DOE/NV/13609-35

Validation, Proof-of-concept, and Postaudit of the Groundwater Flow and Transport Model of the Project Shoal Area

Prepared by

Ahmed Hassan

submitted to

Nevada Site Office
National Nuclear Security Administration
U.S. Department of Energy
Las Vegas, Nevada

SEPTEMBER 2004

Publication No. 45206

DOE/NV/13609-35

Validation, Proof-of-concept, and Postaudit Plans for the Groundwater Flow and Transport Model of the Project Shoal Area

Prepared by

Ahmed Hassan
Division of Hydrologic Sciences
Desert Research Institute
University and Community College System of Nevada

Publication No. 45206

Submitted to

Nevada Site Office
National Nuclear Security Administration
U.S. Department of Energy
Las Vegas, Nevada

September 2004

The work upon which this report is based was supported by the U.S. Department of Energy under Contract #DE-AC52-00NV13609. Approved for public release; further dissemination unlimited.

THIS PAGE LEFT INTENTIONALLY BLANK

EXECUTIVE SUMMARY

The groundwater flow and radionuclide transport model characterizing the Shoal underground nuclear test has been accepted by the State of Nevada Division of Environmental Protection. According to the Federal Facility Agreement and Consent Order (FFACO) between the U.S. Department of Energy and the State of Nevada, the next steps in the closure process for the site are the model validation (or postaudit), the proof-of-concept, and the long-term monitoring stage. This report addresses the development of the validation strategy for the Shoal model, needed for preparing the subsurface Corrective Action Decision Document-Corrective Action Plan (CADD/CAP) and the development of the proof-of-concept tools needed during the five-year monitoring/validation period. The proposed approach builds on the previous model validation work conducted for the Central Nevada Test Area (CNTA). The CNTA validation approach is adapted and modified to the site-specific conditions and challenges of the Shoal area, which are significantly different from those at CNTA (faster transport rates are predicted for Shoal).

Groundwater model validation definitions and literature debates are briefly discussed. Most of the literature debate is on the terminology and not on the process itself. No one disputes the importance and necessity of the model validation or evaluation process and no one disagrees on the concept of using an independent data set to test the model. The disagreement is on what the process should be called and what the implications are for the term used. This has resulted in a lack of focus on development of model validation processes or tools. Although there is rough agreement on the goal of the process, there is no agreement on or development of a uniform methodology for executing model validation. We postulate that there is an urgent need for a validation process that can be used to evaluate model performance and build confidence in its suitability for making decisions. This confidence building is a long-term iterative process and it is this process that should be termed model validation.

The FFACO's 10-step model validation process and the 12-step modeling protocol established in the hydrogeologic literature are discussed and contrasted. It is noted that one of the eliminated steps in the conversion from the hydrogeologic literature process to the FFACO process is the "model redesign" step. This model redesign step is a necessary complement to the model postaudit step, which appears in both processes. It is only meaningful to have a postaudit step together with the "possibility" for a model redesign step. Otherwise, there is no need for the model postaudit step.

Previous model postaudit studies are reviewed and reasons for poor model performance are discussed. Based on these reviews and on the validation process proposed for CNTA, a similar validation approach for Shoal is proposed. The details of the proposed approach are not repeated here, but rather more explanation and discussion of some of the challenging aspects of the process are addressed as they apply to the Shoal model. In particular, more focus is placed on two main aspects: 1) the selection of the validation targets at Shoal, and 2) the selection of the acceptance criteria for the stochastic model realizations.

Individual parametric uncertainty analysis is conducted to identify the contribution of a number of flow and transport parameters to the overall Shoal model's output uncertainty. Validation targets are selected based on this analysis and the practical feasibility of measuring the selected targets in the field. For input parameters to the Shoal model, the analysis indicates that hydraulic conductivity is one of the important validation targets. On the output side, the

hydraulic head as well as the head gradient are viable validation targets. In addition, presence or absence of certain radionuclides at the locations of the validation wells can be used as validation targets.

A few metrics are proposed to be used for the acceptance criteria for the conceptual model as well as for individual model realizations. These metrics are tested with well-known statistical tests and are shown to be reasonable acceptance measures for the model. Hypothetical scenarios are used to show how these metrics perform when field data are consistent with and support the existing model, or indicate model deficiencies. A hierarchical approach to determine whether the number of acceptable model realizations is sufficient is described by a decision tree. This decision tree is proposed for the acceptance of the realizations and for passing the first decision point on the validation approach.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

CADD	Corrective Action Decision Document
CAP	Corrective Action Plan
CAIP	Corrective Action Investigation Plan
CNTA	Central Nevada Test Area
DOE	U.S. Department of Energy
DDA	Data Decision Analysis
FFACO	Federal Facilities Agreement and Consent Order
MCL	maximum contaminant level
NDEP	Nevada Division of Environmental Protection
NTS	Nevada Test Site
PSA	Project Shoal Area
RMSE	root mean square error
RWPT	random-walk particle-tracking
UGTA	Underground Test Area

1. INTRODUCTION

The U.S. Department of Energy (DOE) and State of Nevada are working together through a Federal Facilities Agreement and Consent Order (FFACO) to identify sites in Nevada of potential historic contamination and implement proposed corrective actions based on public health and environmental considerations. This includes completing environmental corrective action activities at facilities where nuclear-related operations were conducted. The closure process involves thorough investigations of the potential impacts of facilities on public health and the environment, providing the information needed to choose appropriate remedies. For underground nuclear tests specifically, Appendix VI of the FFACO defines a Corrective Action Strategy to define boundaries around each test area that contain water that may be unsafe for domestic and municipal use. The strategy is to characterize groundwater flow and contaminant transport through modeling utilizing site-specific hydrologic data.

Though the vast majority of underground nuclear tests occurred at the Nevada Test Site (NTS), a limited number of tests were conducted at other locations. The Project Shoal Area (PSA), about 50 km southeast of Fallon, Nevada, is the location of the Shoal test. Shoal was a 12-kiloton-yield nuclear detonation that occurred on October 26, 1963 (U.S. DOE, 2000). The test was part of a program (Vela Uniform) to enhance seismic detection of underground nuclear tests in active earthquake areas. Figure 1 shows the location of the Shoal site relative to cities in Nevada. The nuclear device was emplaced 367 m below land surface, or about 65 m below the water table, in fractured granite. Details of the geology and hydrogeology at the site are provided elsewhere (Pohll *et al*., 1998, 1999a). Characterization of groundwater contamination resulting from the Shoal test is being conducted by DOE under the FFACO with the State of Nevada Division of Environmental Protection (NDEP) and the U.S. Department of Defense (Pohlmann *et al*., 2004).

The original Corrective Action Investigation Plan (CAIP) for the PSA was approved in September 1996 and described a plan to drill and conduct testing of four characterization wells, followed by flow and transport modeling. The resultant drilling is described in a data report (U.S. DOE, 1998a) and the data analysis and modeling in an interim modeling report (Pohll *et al.*, 1998). After considering the results of the modeling effort, DOE determined that the degree of uncertainty in transport predictions for Shoal remained unacceptably large. As a result, a second CAIP was developed by DOE and approved by the NDEP in December 1998 (U.S. DOE, 1998b). This plan prescribed a rigorous analysis of uncertainty in the Shoal model and quantification of methods of reducing uncertainty through data collection. This analysis is termed a Data Decision Analysis (DDA) (Pohll *et al.*, 1999b) and formed the basis for a second major characterization effort at Shoal. The details for this second field effort are presented in an Addendum to the CAIP, which was approved by NDEP in April 1999 (U.S. DOE, 1999). Four additional characterization wells were drilled at Shoal during summer and fall 1999; details of the drilling and well installation are in IT Corporation (2000), with testing reported in Mihevc *et al.* (2000). A key component of the second field program was a tracer test between two of the new wells (Carroll *et al.*, 2000; Reimus *et al.*, 2003).

The objectives of the characterization effort of this field program include the evaluation of alternative conceptual radionuclide transport models in the saturated, fractured granite and the estimation of transport parameters for use in radionuclide transport models. To achieve these objectives, a cross-hole tracer test involving the simultaneous injection of both nonsorbing and sorbing solute tracers was conducted at the site in 1999 and 2000. A secondary objective of the

tracer test was to determine how well the field-scale transport behavior of a sorbing solute could be predicted based on laboratory derived sorption parameters.

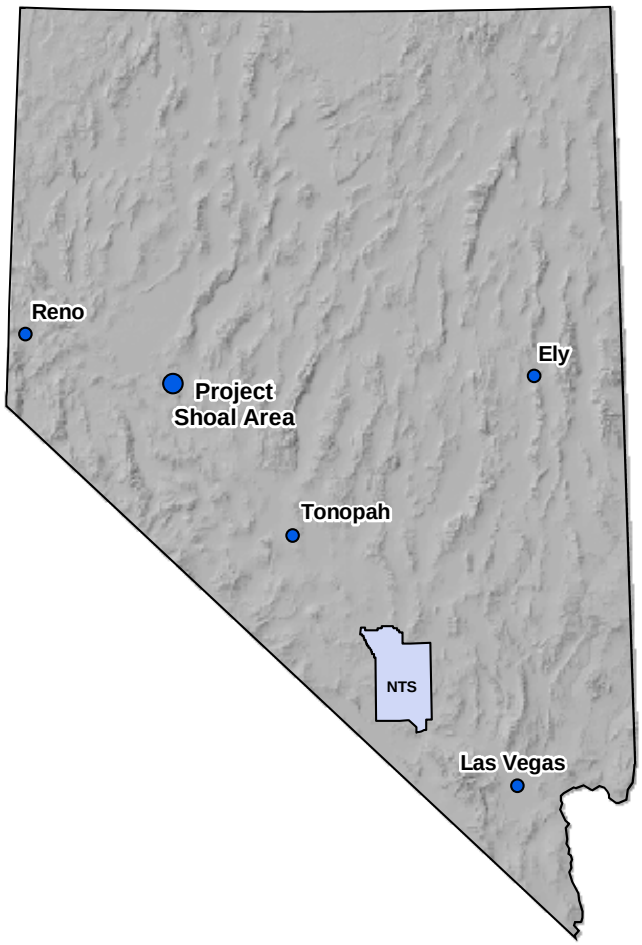

Figure 1. Location of Project Shoal Area.

As a result of the tracer test and the new characterization efforts of 1999 and 2000, a new groundwater flow and radionuclide transport model was developed and submitted to DOE in 2001 (Pohlmann *et al*., 2001). This model report was subsequently revised in response to the comments of the Modeling Subcommittee of the Underground Test Area (UGTA) Technical Working Group and submitted to DOE and NDEP in 2003 (Pohlmann *et al*., 2002). Based on comments received from the NDEP, the model report was again revised and submitted to DOE and NDEP in 2004 (Pohlmann *et al*., 2004).

The NDEP has concurred with the Shoal model of Pohlmann *et al*. (2004) and the natural next step in the closure process of the site is the development of the validation and long-term monitoring approaches for presentation in the Corrective Action Decision Document/Corrective Action Plan (CADD/CAP). This report discusses the validation approach for the Shoal model and builds on the analysis presented in Hassan (2003a, 2004a,b) for the Central Nevada Test Area (CNTA) model. The remainder of this report is organized as follows. In Section 2, a brief discussion of the meaning and the literature controversy regarding the term “model validation” is presented. Section 3 then discusses the link between the model validation process and the

FFACO (2000) model validation and postaudit steps. Section 4 presents a review of reported model postaudit studies and discusses the limitation of these postaudits. Section 5 briefly describes the proposed model validation process (adapted from Hassan [2003a]), then the selection of validation targets at Shoal using the results of parametric uncertainty analysis, and on the selection of validation criteria for the stochastic Shoal model. The report is summarized in Section 6.

2. WHAT IS MODEL VALIDATION?

The term validation has been the subject of literature debate for more than a decade. Many skeptics argue that the term should not be used, as it implies correctness and accuracy for a groundwater model that no one can claim. Regulators and decision makers should understand that there is no way to guarantee that a model-based decision is always correct, or that a model can ever be proven to be valid in the ***strictest sense*** of the term (van der Heijde, 1990). Many assert that it is impossible to validate a groundwater numerical model because such a claim would assert a demonstration of truth that can never be attained for our approximate solutions to subsurface problems (Oreskes *et al*., 1994). Anderson and Bates (2001) recently edited a book covering the issue of model validation in Hydrological Science. Summarized here are some of the quotes they presented from the skeptics and the opponents to the use of the term "validation."

"Absolute validity of a model is never determined" (National Research Council, 1999).

"What is usually done in testing the predictive capability of a model is best characterized as calibration or history matching; it is only a limited demonstration of the reliability of the model. We believe the terms validation and verification have little or no place in groundwater science; these terms lead to a false impression of model capability. More meaningful descriptors of the process include model testing, model evaluation, model calibration, sensitivity testing, benchmarking, history matching, and parameter estimation. Use of these terms will help to shift emphasis towards understanding complex hydrogeological systems and away from building false confidence into model predictions" (Konikow and Bredehoeft, 1992).

"A modeling system can, in principle, never be validated. Instead of a full validation we can think of the degree of validity as the credibility of a given modeling system. The degree of validity of a modeling system is expressed in the first and most immediate place by the sum of all successful validation of all models that have been constructed and operated to date using the modeling system. As the number of such successful models increases, so the credibility of the system itself grows in strength. Behind this, most superficial of views, lies the assumption that the modeling system is in fact being improved on the basis of the operating experience; that it is functioning within its market, tracking the needs of that market and thus learning from this market. From this point of view, the development of a modeling system is not one that leads directly to a finished, rounded and complete product, but it is rather a process of adaptation through evolution. Thus, although the modeling system is indirectly a product, it is one that is constantly evolving, so that its evolution corresponds to a process" (Refsgaard and Storm, 1996)

"The inherent uncertainties of models have been widely recognized, and it is now commonly acknowledged that the term 'validation' is an unfortunate one, because its root -valid - implies a legitimacy that we may not be justified in asserting (Tsang, 1991, 1992; Anderson and Woessner, 1992a; Konikow, 1992; Konikow and Bredehoeft, 1992; Oreskes *et al*., 1994; Oreskes, 1998; Beck *et al*., 1997; Steefel and van Cappellen, 1998). But old habits die hard, and the term persists. In formal documents of major national and international agencies that sponsor modeling efforts and in the work of many modelers, 'validation' is still widely used in ways that

assert or imply assurance that the model accurately reflects the underlying natural processes, and therefore provides a reliable basis for decision-making. This usage is misleading and should be changed. Models cannot be validated" Oreskes and Blitz (2001).

The above views consider the validation from the strictest definition of the word. That is, they refer to the validation as a demonstration of the accuracy of the model in representing the true system and they warn against misconception of the public about the meaning of the term. As discussed in Hassan (2004a), most models, if not all, are not being used to reveal the truth of a system. Of course it would be great if models could do so, but they simply cannot. Models are in many cases decision-making tools. When a model successfully passes a rigorous development, calibration, and testing process, it becomes a reasonable decision-making tool given the limited data used in the development process. Acknowledging the role of uncertainty, the model validation process is one crucial stage in the entire process that should be regarded as an additional filter for independent model evaluation. The fact is that most of the literature debate is on the terminology and not on the process itself. No one argues that the process is unimportant, unneeded, or useless and no one disagrees on the concept of using an independent data set to test the model. The disagreement is on what it is called and what the implications are for the term used.

This lack of focus on development of model validation processes or tools is reflected clearly in Vogel and Sankarasubramanian's (2003) discussion about the validation of watershed models. "When one considers the wide range of watershed models and the heavy emphasis on their calibration (Duan *et al.*, 2003), it is surprising how little attention has been given to the problem of model validation. In three recent reviews of watershed modeling (Singh, 1995; Hornberger and Boyer, 1995; Singh and Woolhiser, 2002) and watershed model calibration (Duan *et al.*, 2003), there was little attention given to developments in the area of model validation. Although there is rough agreement on the goal of model validation, no agreement exists on a uniform methodology for executing model validation."

There is an urgent need for a validation process that can be used to evaluate model performance and build confidence in a model's suitability for making decisions. This confidence building is a long-term iterative process and it is the author's belief that this process is what should be termed model validation. Model validation is a process, not an end result. That is, the process of model validation cannot always assure acceptable prediction or quality of the model. Rather, it provides safeguards against faulty models or inadequately developed and tested models. If model results end up being used as the basis for decision making, then the validation process indicates that the model is valid for making decisions (not necessarily an absolute representation of truth).

Again, such a process should not be viewed as a mechanism for proving that the model is valid, but rather as a mechanism for enhancing the model, reducing its uncertainty, and improving its predictions through an iterative, long-term, confidence-building process. The process should contain trigger mechanisms that will drive the model back to the characterization-conceptualization-calibration-prediction stage (i.e., back to the beginning), but with a better understanding of the modeled system.

As described by a regulator (Shah Alam, 1998), regulators want to be certain that human health and the environment are being protected and general public participation is a key element in a regulatory decision making process for a contaminated site. Affected public should be able to comprehend and concur with the model in their terms. This is difficult to achieve without a

long-term commitment of evaluating and re-evaluating the model results (thus going through an iterative validation process) based on data collected for validation and for the long-term monitoring of the site. Most regulators understand modeling well enough to know that a model cannot be proven to be "correct." Rather, they are seeking evidence that the model is sufficient for decision-making and that model predictions are being thoroughly tested against site-specific data.

3. MODEL VALIDATION AND THE FFACO

Appendix VI of the FFACO between the State of Nevada, DOE, and DoD describes a 10-step model validation procedure. These ten steps are essentially extracted from a 12-step modeling protocol that was devised by Anderson and Woessner (1992b). Thus, the FFACO (2000) defines the model validation process as the performance of ten modeling steps based on Anderson and Woessner's (1992b) description. To further explain this aspect, Figure 2 presents a side-by-side comparison of the 10-step model validation process as outlined in the FFACO (2000) and the 12-step modeling protocol of Anderson and Woessner (1992b), who assert that "each of these twelve steps builds support in demonstrating that a given site-specific model is capable of producing meaningful results, i.e., that the model is valid." It should be mentioned that the how-to steps related to the FFACO (2000) model postaudit step and the link to the proof-of-concept monitoring are not mentioned in any quantitative manner in the FFACO (2000).

Models are essential in performing complex analyses and in making informed predictions. However, no matter how sophisticated the modeling approach is, when applied to a subsurface flow and transport problem, it provides no more than a simplistic representation of very complex field conditions. In this regard, Konikow (1986) states that models should be considered as dynamic representations of nature, subject to further refinements and improvements. As new data become available (e.g., through new wells), model predictions can be evaluated, validated or invalidated, and then modified if necessary. As articulated by Ewing *et al*. (1999), no question exists that mathematical models can be extrapolated over time; however, the question remains as to whether these extrapolations actually capture the physical and chemical behavior of the geologic systems over extended time. Thus, geologic systems are extremely difficult to model over extended periods, and uncertainties are so large that careful attention has to be paid to possibilities of alternative conceptual models, incomplete descriptions of natural systems, and difficulty of describing future stresses and patterns correctly. With all these uncertainties, it is crucial to account for the possibility that the site-specific model being postaudited may not be perfect.

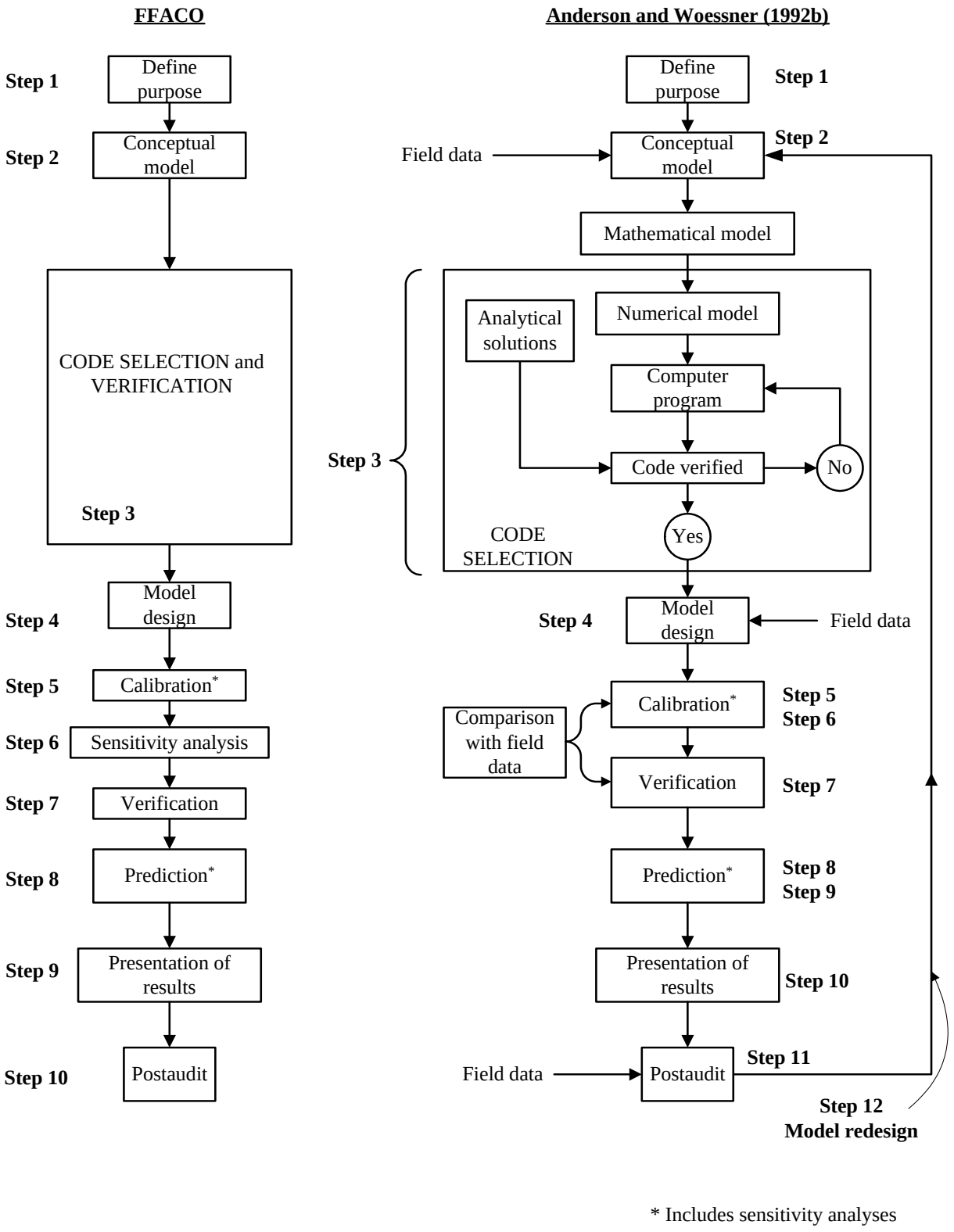

Figure 2. Comparison between the 10-step model validation process as outlined in the FFACO (2000) on the left and the 12-step modeling protocol and model validation process as outlined by Anderson and Woessner (1992b) on the right.

In discussing the strategy of scientific model building, Box (1979) explains how the model builder can go about knowing what aspects of the studied system to include and what to omit so that models that are illuminating and useful result from the model-building process. His argument is as follows. “It is fruitless to attempt to allow for all contingencies in advance so in practice model building must be accomplished by iteration. For example, preliminary graphical analysis of data, and careful thought about what is known about the system under study, may suggest a first model worthy to be tentatively entertained. From this model a corresponding first tentative analysis may be made as if we believed it. The tentative inferences made, like all inferences are conditional on the applicability of the implied model, but the investigator now quickly switches his attitude from that of sponsor to that of critic. In fact the tentative analysis provides a basis for criticism of the model. This criticism phase is accomplished by examining residual quantities using graphical procedures and sometimes more formal tests of fit. Such diagnostic checks may fail to show cause for doubting the model's applicability, otherwise it may point to modification of the model leading to a new tentative model and, in turn, to a further cycle of the iteration.”

Box (1979) further makes the following analogy. “It should be remembered that just as the Declaration of Independence promises the pursuit of happiness rather than happiness itself, so the iterative scientific model building process offers only the pursuit of the perfect model. For even when we feel we have carried the model building process to a conclusion some new initiative may make further improvement possible. Fortunately, to be useful a model does not have to be perfect.”

Bredehoeft (2003) also discusses the issue of modeling process. He indicates that good modeling is an iterative process. As new data are acquired, the model is revisited and adjusted (or recalibrated) so that the model predictions are consistent with all the data including the new data. The model becomes a living tool for analysis. With this paradigm, the modeling strategy changes; it requires continued monitoring and model updating. The iterative process is important in addressing the adequacy of the conceptual model. A mismatch between the model prediction and observed data should raise the issue of conceptualization: Is the mismatch a result of poor parameter adjustments or does it suggest that a rethinking of the conceptual model (Bredehoeft, 2003). Part of the validation plan discussed later exactly addresses this issue.

As modeling tools increase in complexity and ability to handle natural geologic variability, parameterization of the models becomes more difficult. The model complexity and the subsequent high-dimensional parameterization make objective calibration very difficult, if not impossible. The calibration process that is used to determine model parameters can be guided by the principle of parsimony. This principle implies that the best model is the simplest possible model that adequately describes the available data. Using the principle of parsimony, the model is kept as simple as possible while still accounting for the system processes and characteristics evident in the observations and while respecting other information about the system (Hill, 1998). The bottom line is, therefore, that one can strive for model parsimony as long as this does not impair the ability of the model to simulate or approximate observed data values (Perrin *et al*., 2001).

The principle of parsimony can also be used to differentiate between alternative conceptual models. In this context, the principle states that among all plausible models that one can use to explain a given set of experimental data, one should select the model that is conceptually least complex and involves the smallest number of unknown (fitting) parameters.

This principle was implemented in developing the conceptual models and in the calibration procedure for the Shoal model. Whether to simplify the model or add some complexities was always dependent upon the relevance of the simplified aspects to the issue of concern that derived the modeling effort in the first place. When it comes to model validation, the validation process as defined earlier in Section 1 will provide a mechanism for testing whether the parsimony principle was applied correctly or whether oversimplification led to major model deficiencies.

Hassan (2003a, 2004a) presented a thorough review of groundwater model validation including definitions, literature debates, validation attempts, and proposed strategies. The literature review and the discussions presented in these studies make it clear that even the simplest deterministic subsurface model is very difficult to evaluate. This discussion and the debate in the literature on model validation make it essential that models be refined and updated whenever new information becomes available, especially if the new information represents site-specific data. In the context of performance assessment (PA) of proposed high-level waste repositories, Ewing *et al*. (1999) postulate that although PA is required as a condition for licensing certain waste-disposal sites, there is no *a priori* reason that the public or scientific community should uncritically accept a PA analysis; on the contrary, any PA analysis should be carefully scrutinized and improvements will inevitably result. These improvements cannot be made if the model postaudit stage outlined in the FFACO (2000) does not allow for revisiting and refining the model conceptualization and prediction based on newly collected field data.

The validation process proposed for the CNTA groundwater flow and transport model (Hassan, 2003a, 2004a, b) is consistent with the 10-step FFACO (2000) strategy provided that the tenth step of model postaudit in that strategy is perceived as presented by Anderson and Woessner (1992b). Figure 3 shows the linkage between the FFACO closure process for Nevada offsite test areas and the 10-step modeling protocol with detailed conceptualization for the tenth step (the postaudit step). As can be seen from the figure, steps 1 to 9 of the FFACO modeling protocol or model validation strategy belong to the development stage that is represented in the figure by the top large box. Step 10 in the FFACO strategy (the postaudit) is expanded in Figure 3 and it represents the model validation process as conceptualized in this study. It is important to notice that the five-year proof-of-concept monitoring network development that is required in the FFACO (2000) can start once the development stage is completed and can be performed simultaneously with the model validation analyses, monitoring network development and the postaudit. As stated in the FFACO (2000), measurement of field parameters through this proof-of-concept monitoring network will be used to demonstrate that the model is capable of making reasonable predictions that fall within an acceptable level of confidence. This is what the lower loop in Figure 3 is designed to provide, which corresponds to the CAP, Closure Report (CR), and long-term stewardship stages of the closure process.

When the monitoring network is installed, data collected, and tests performed for evaluating and validating the model, the question will arise as to how the model predictions compare to the collected field measurements. If the results indicate major model deficiencies, the process will be driven back to step 2 for model reconceptualization. This implies that steps 2 through 9 of the modeling protocol (Anderson and Woessner, 1992b) and the FFACO model validation strategy will be repeated if needed and this repetition is considered as part of the model postaudit or model validation stage. Once the model postaudit step, the proof-of-concept

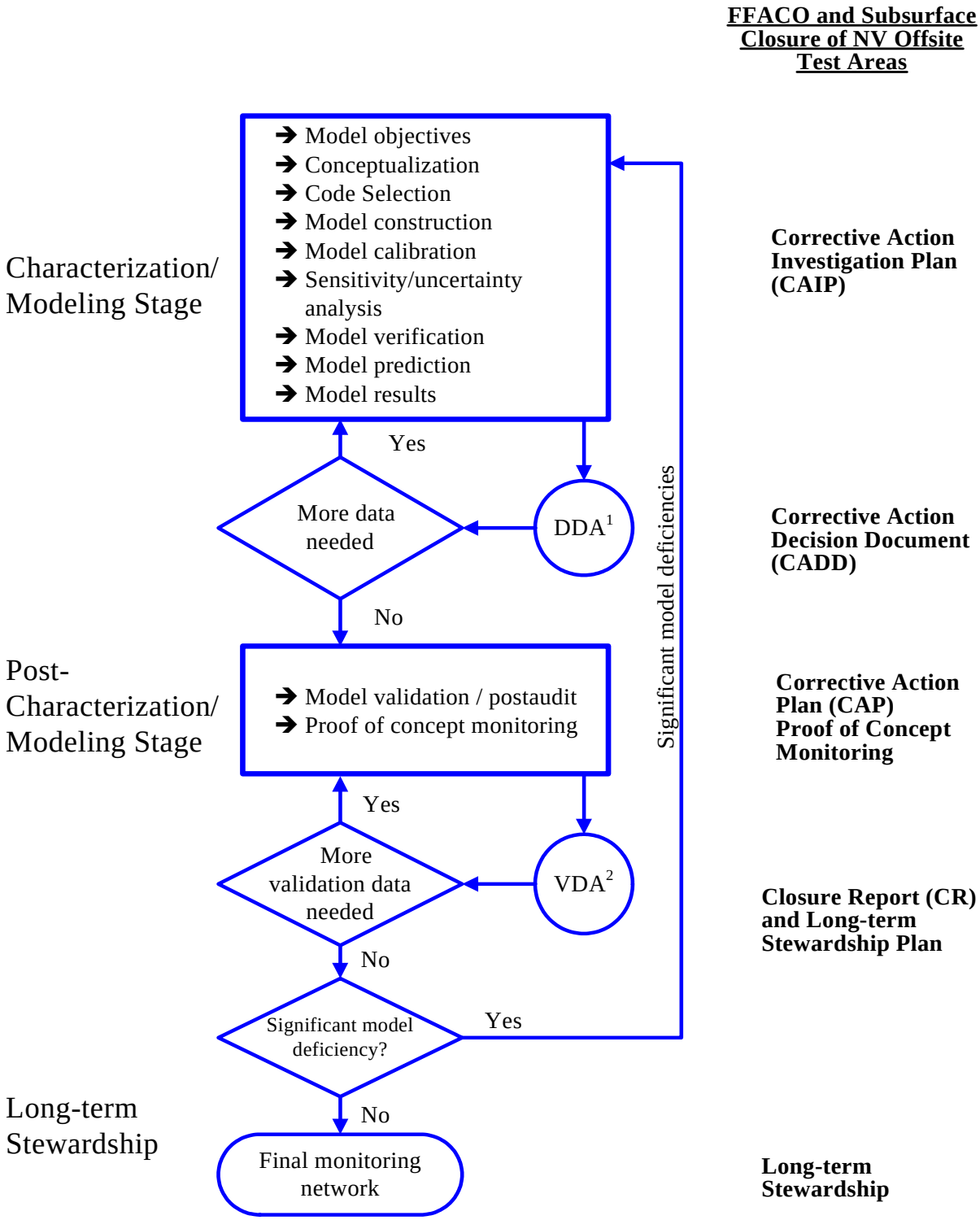

Figure 3. The linkage between the FFACO closure process and the 10-step modeling strategy and how the postaudit step (step number 10) leads to an iterative validation process.

stage, and confidence-building loop (model validation process) have been completed successfully and the model is deemed validated, the long-term monitoring network will be amended if necessary to provide sufficient surveillance for the site. Using the long-term monitoring data to re-evaluate the model over time is considered as a continuous model validation and postaudit process and is necessary for the long-time period of concern at these nuclear testing sites. Again, geologic systems are extremely difficult to model over extended periods of time, and uncertainties are so large that careful attention has to be paid to the possibilities of imperfect models.

An attempt is made in Hassan (2003a, 2004a, b) to quantify the model validation process and this report adapts this validation process to the Shoal model. What is provided in Figure 4 and discussed later in Section 5.1 is a detailed approach for performing this process in the case of stochastic numerical models that rely on Monte Carlo simulations, which is the case for the Shoal model.

4. REVIEW OF POSTAUDIT STUDIES

A postaudit of a numerical model compares the model's predictions with field observations. Model postaudits test the predictive capabilities of groundwater models and shed light on their practical limitations. Postaudits can provide valuable perspectives on the use of numerical groundwater models to predict the behavior of groundwater systems. The predictive ability of numerical groundwater models is often, incorrectly, associated with a model's ability to reproduce a calibration data set (Freyberg, 1988). Without assessing the model performance over the prediction period through a postaudit, it is not possible to make a general assessment of the predictive ability of numerical models, or to discover common reasons for poor performance (Stewart and Langevin, 1999).

A number of studies have explored the predictive reliability of reasonably calibrated models by a posterior comparison between model predictions and observed data (e.g., Person and Konikow, 1986; Konikow, 1986; Freyberg, 1988). However, the few documented examples of postaudits of groundwater models (e.g., Konikow, 1986; Konikow and Person, 1985; Konikow and Swain, 1990; Alley and Emery, 1986; Stewart and Langevin, 1999) have not provided high confidence in the predictive accuracy of these models. These studies showed that prediction accuracy was moderate. For example, Stewart and Langevin (1999) suggest that the actual drawdown around a major well field in Florida exceeded model predictions by a factor of two. In their study, they postaudited a numerical groundwater flow model that was created in 1978 and revised in 1981 to predict the drawdown effects of a proposed municipal well field permitted to withdraw 30 million gallons per day from the Floridan Aquifer.

In their postaudit, observed drawdowns were derived by taking the difference between pre-pumping and post-pumping potentiometric surface levels. Comparison of predicted and observed drawdowns suggested that actual drawdown over a 12-year period exceeded predicted drawdown by a factor of two or more. Analysis of the source of error in the 1981 predictions suggested that the values used for transmissivity, storativity, specific yield, and leakance were reasonable at the well field scale. Simulation using actual 1980 to 1992 pumping rates improved the agreement between predicted and observed drawdowns. The principle source of error was the assumption that water levels in a semiconfined aquifer would achieve a steady-state condition after a few days or weeks of pumping (Stewart and Langevin, 1999). Simulations using a version of the 1981 model modified to include recharge and evapotranspiration suggested that it could

take hundreds of days or several years for water levels in the linked Surficial and Floridan aquifers to reach an apparent steady-state condition, and that slow declines in levels continued for years after the initiation of pumping.

Person and Konikow (1986) and Konikow (1986) recalibrated a groundwater flow and solute transport model of an irrigated stream-aquifer system because of the discrepancies between prior predictions of groundwater salinity trends during 1971 to 1982 and the observed outcome in February 1982. The original model was calibrated with a 1-year record of data collected during 1971 to 1972 in an 18-km reach of the Arkansas River Valley in southeastern Colorado. The model was improved by incorporating additional hydrologic processes (salt transport in the unsaturated zone) and through reexamination of the reliability of some input data. Extended simulations using the recalibrated model were made to investigate the usefulness of the model for predicting long-term trends of salinity and water levels within the study area. Person and Konikow (1986) found that predicted groundwater levels during 1971 to 1982 were in good agreement with the observed, indicating that the original 1971 to 1972 study period was sufficient to calibrate the flow model. However, the recalibrated solute transport model underestimated the observed 1982 groundwater salinity by about 10 percent when the 1971 to 1972 surface water salinities were assumed to recur in succeeding years. Person and Konikow (1986) concluded that the calibration period (covering some seasonal variations in the river-aquifer interaction and irrigation cycles) needed for accurate transport prediction is longer than that required for the flow model predictions. The metric used to judge the prediction accuracy of the model was the spatially averaged groundwater level for the flow model and the spatially averaged groundwater salinity for the solute transport model.

Anderson (1968) constructed and calibrated a two-dimensional electric-analog model of an alluvial aquifer system in the Salt River Valley and the lower Santa Cruz River basin located near Phoenix, Arizona. As discussed in Konikow (1995), this model study represents one of the first well-documented, ***deterministic***, distributed-parameter model analyses of a groundwater system. The model was constructed to determine the probable future effects of continued groundwater withdrawals in central Arizona. The model was based on the assumption that prior to 1923 an equilibrium existed in the aquifer in which recharge balanced discharge. The model was calibrated by adjusting aquifer properties and boundary conditions to match observed historical changes in water levels during 1923 to 1964. The model was then used to predict future water level changes as groundwater withdrawals from the aquifer continued. Water level measurements from 77 wells were used to evaluate the model predictions where the 1974 predicted and observed water levels were compared. For these wells, the water table decline was predicted to average about 25 m after 10 years and ranged from 4.5 to 65 m. However, measurements of the actual change in water level in the same wells showed an average decline of only 2.7 m, and the observed change ranged from a decline of 28 m to a rise of 45 m. The relation between observed and predicted water level changes indicated poor predictive accuracy and the presence of a bias in the model predictions. Also, the data showed a relatively wide scatter, indicating that the model predictions were imprecise.

This postaudit example indicates the weakness of basing a prediction of aquifer responses on a single set of assumed parameters or future stresses (Konikow, 1995). Because the uncertainty of the 1965 to 1974 stresses was not assessed, it could not be known whether the actual 1965 to 1974 responses fall within some associated confidence interval. Hence, one cannot make a judgment based solely on these predictive errors as to whether the model was

"good" or "bad" (Konikow, 1995). In cases like this, it would be preferable to assess the uncertainty in estimated (or assumed) future stresses and then present the forecasts as a range of responses with associated probabilities of occurrence or confidence intervals. Because Anderson (1968) indicated correctly that the assumed stresses were probably greater than would occur, the predictions can be viewed as a "worst-case" scenario. From that perspective, the model predictions were reasonably accurate (Konikow, 1995). Interestingly, Anderson's (1968) prediction of additional significant water table declines was one factor contributing to water management decisions and actions leading to reduced groundwater withdrawals after 1964, which in turn was a major source of predictive errors.

Alley and Emery (1986) examined predictions of 1982 water-level declines and stream flow depletions for the Blue River Basin, Nebraska, made in 1965 using an electric analog model. The postaudit showed that the model underestimated the depletion of the stream flow and overestimated the decline of the groundwater levels. Analysis performed by Alley and Emery (1986) suggested that net groundwater withdrawals used in the analog simulation were too low. The model overestimated groundwater level declines because it assumed that all of the new groundwater withdrawals would come from storage in the aquifer, when in fact some water comes from induced recharge from the stream. Furthermore, they speculated that storage coefficients used in the model were too low. Alley and Emery (1986) concluded that the error in the prediction was a result of uncertainty in the conceptual model of the Blue River Basin.

Lewis and Goldstein (1982) postaudited a two-dimensional groundwater flow and solute transport model that was developed and calibrated by Robertson (1974) and then used to predict flow and transport in a basalt aquifer beneath the Idaho National Engineering Laboratory. The flow model was calibrated to an assumed steady-state flow field and the transport model was calibrated to observed concentrations of chloride in groundwater in 1958 and 1969. Robertson (1974) then used the calibrated model to predict chloride and tritium concentrations in 1980. Through the postaudit study, Lewis and Goldstein (1982) found that the contaminant plumes predicted by the model extended farther downgradient than the actual plumes and attributed this deviation to the conservative worst-case assumptions in the model input, the simplicity of the conceptual model, and the inaccurate estimate of subsequent waste discharge and aquifer recharge conditions. The original model of Robertson (1974) viewed the aquifer as a continuous porous medium, and it is likely that the flow in this aquifer would be better approximated using a dual-porosity model that includes fracture flow as well as matrix diffusion (Anderson and Woessner, 1992a).

Flavelle *et al*. (1990) presented another postaudit of a model that simulates the release of hydrogen ions from a tailings pile situated in glaciofluvial deposits in Ontario, Canada. The flow model of that study was calibrated to measured heads in 1989 within the inner part of the plume where pH was less than 4.8. The solute transport model was calibrated by matching plume position to observed position in 1983 and 1984 through varying the distribution coefficient. The calibrated model was then used to predict the plume distribution in 1989. Data collected in 1989 showed that the model accurately predicted the pH values in the inner core of the plume but not at the outer edges. Flavelle *et al*. (1990) concluded that even though their site is one of the most thoroughly studied uranium tailings sites in Canada, the data were not complete enough for a successful model postaudit.

In the study of Anderson and Lu (2003), they postaudited groundwater model predictions that were used to design an extraction-treatment-injection system at a military ammunition

facility located in western Tennessee, halfway between Memphis and Nashville. These predictions were evaluated using site-specific water-level data collected approximately one year after system startup. The water-level data indicated that performance specifications for the design, i.e., containment, had been achieved over the required area, but that predicted water-level changes were greater than observed, particularly in the deeper zones of the aquifer. Probable model error was investigated by determining the changes that were required to obtain an improved match to observed water-level changes. Anderson and Lu's (2003) analysis suggested that the originally estimated hydraulic properties were in error by a factor of two to five. To determine the importance of these errors to the predictions of interest, the original and updated models were used to simulate the capture zones resulting from the originally estimated and updated parameter values. Anderson and Lu (2003) found that these capture zones were not significantly different in the shallow part of the aquifer where the contaminant plume was primarily located. Unlike the other postaudits, Anderson and Lu's (2003) postaudit suggests that while the base model had errors that were revealed through the postaudit, the groundwater model contributed in a very positive way to designing a remedial system that achieved the project goals.

Weaver *et al*. (1996) performed a postaudit on two groundwater flow models that were used to design a well array for a groundwater capture and containment system installed along the boundary of a manufacturing facility. The first model was an analytical model for which the postaudit indicated that the performance of the initial system design provided by this model did not meet expectations. This led to using a numerical model to design an enhanced system, for which detailed postaudit could not be performed, as the system was in place for a short period of time. However, a cursory review of the numerical model results versus observed conditions was performed. The results of the postaudit indicated that the deviations of the models' predictions from actual water levels could be mainly attributed to changes in system conditions (pumping rates, variations in well efficiencies, and limitations on total available drawdown) and aquifer heterogeneity.

An interesting discussion related to the postaudit concept is presented by Brown (1996). The previous studies all focused on evaluating the model and conducting the postaudit long after the model had been accepted and used for decision making. So, although the postaudit may enable the modeler to improve the model and benefit from the knowledge gained by the new field data, the improvement can only take place after actions have been taken that were based upon the prediction. Therefore, the postaudit is not something that helps a model withstand attempts at invalidation prior to decision making (Brown, 1996). An alternative to this type of model postaudit is the field postaudit that can be performed after the prediction but before the final decision is made based on the prediction. If some test of a modeling prediction is required prior to decision making, a field audit will provide information of a direct and relevant nature to evaluate the adequacy of the model's prediction. This type of evaluation is what model sponsors and regulators usually call model validation.

It can be seen that in general, postaudit studies found that errors in model predictions were caused by errors in the conceptual model and the failure to use appropriate values for assumed future stresses (Anderson and Woessner, 1992a). The lack of successful demonstrations that groundwater models can provide accurate predictions has led to a certain amount of skepticism regarding the utility of groundwater models as predictive tools. However, the common situation in these studies was that the calibrated model was used to predict system behavior under modified conditions (future predicted system stresses, modified boundary

conditions, or different parameter values). In particular, Freyberg (1988) showed that the ability of a calibrated parameter set of a groundwater flow model to reproduce observed data was not an indicator of the ability of that parameter set to predict system response under modified conditions. He reports that good calibration does not necessarily guarantee equally good prediction.

Another common feature of the models previously postaudited is that they all, except Flavelle *et al*. (1990), dealt with transient flow conditions under external stresses (e.g., pumping, extraction-injection systems, interactions with changing surface water bodies). The steady-state flow conditions at both CNTA and Shoal and the long time frame of the predictions make the simple postaudit application inherent in previous studies inappropriate. This is further exacerbated by the fact that the CNTA and Shoal models are cast in a stochastic framework that provides the predictions as ranges of values as opposed to single deterministic values. The challenge is thus how to conduct a model postaudit (or a model validation process) for such stochastic predictions.

In discussing the role of postaudits in model validation and after reviewing a number of posaudits that indicated poor model predictions, Anderson and Woessner (1992a) commented "All of the postaudits indicate that errors in the predictions could be attributed at least partly to errors in the conceptual model. Model validation, therefore, requires a good conceptual model. Herein lies a major difficulty because a good conceptual model requires accurate and complete field characterization. Field characterization is always incomplete, thereby introducing uncertainty into the conceptual model. Continual improvement of the conceptual model requires periodic collection of field data and a trial and error process of model improvement over many years. It is rare to find such a large commitment of time and money to a modeling effort." Once again, this highlights the importance of the iterative process of model building, calibration, monitoring, validation, and refinement.

Similarly, Konikow (1995) summarized the results of a number of model postaudits and commented "The postaudits show that predictions should be accompanied by an assessment of their reliability that is based on the uncertainty in all model parameters, including stresses and boundary conditions. A major value of the postaudit is that the evaluation of the nature and magnitude of predictive errors may itself lead to a large increase in the understanding of the system and in the value of a subsequently revised model. As new information become available, previous forecasts could and should be modified. Feedback from preliminary models not only helps an investigator to set improved priorities for the collection of additional data, but also helps test hypotheses concerning governing processes in order to develop an improved conceptual model of the system and problem of concern."

5. PROPOSED VALIDATION APPROACH FOR SHOAL

Hassan (2003a, 2004b) proposed an approach for the validation process of stochastic numerical models. As mentioned earlier, the details of the process will not be repeated here, but rather more explanation and discussion of some of the challenging aspects of the process will be addressed as they apply to the Shoal model. In particular, the focus is on two main aspects: 1) the selection of the validation targets at Shoal, and 2) the selection of the acceptance criteria for the stochastic model realizations. However, for completeness, we briefly discuss the overall validation approach in Section 5.1, then follow it with the detailed analyses of validation targets and acceptance criteria.

5.1 Proposed Validation Approach

The proposed plan for the validation process of the Shoal model accounts for the stochastic nature of the model and aims at building enough confidence in the model predictions before using these predictions to design the long-term monitoring network necessary for site surveillance. The focus of the proposed validation methodology is centered around the following three main themes: (1) testing how predictions of numerical groundwater flow and transport models of Shoal and the underlying conceptual models and assumptions are robust and consistent with regulatory purposes, (2) reevaluating and refining model predictions and reducing the uncertainty level based on data collected in the proposed field activities for model validation, and (3) linking validation efforts to long-term monitoring that benefits from and builds on the validation-phase field activities.

Figure 4, adapted from Hassan (2003a), displays the step-by-step approach for performing the validation and postaudit processes for the Shoal model. There is one clearly defined point in the validation process where a significant revision of the model can be triggered. This trigger point occurs at Step 6, where the results may be determined to not meet regulatory objectives. All of the validation steps are described below.

Step 1: Identify the data needed for validation, the number and location of the wells, and the type of laboratory or field experiments needed. Well locations can be determined based on the existing model and should favor locations likely to encounter fast migration pathways. The monitoring design process (Hassan, 2003b) will be implemented to help select these locations. Other factors such as the location of contaminant and compliance boundaries, and the cost of drilling and collecting data have to be considered. Sequencing of data collection is also important, as is the ability to adjust the plan as information is gathered.

Step 2: Carry out the fieldwork to install the wells and obtain the largest amount of data possible from the wells. The data should include geophysical logging, head measurements, conductivity measurements, concentrations (e.g., checking for tritium or ^{14}C), and any other information (e.g., temperature logs, fracture information) that could be used to test the model structure, input, or output.

Step 3: Perform the different validation tests that will help evaluate the different submodels and components of the model. The stochastic validation approach proposed by Luis and McLaughlin (1992) 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 that is based on hypothesis testing. Some data will be used to check the occurrence or lack thereof of failure scenarios (e.g., whether tritium or ^{14}C exists much farther from the cavity than is predicted by any realization of the stochastic Shoal model). The philosophy here is to test each individual realization with as many diverse tests (in terms of the statistical nature of the test and the tested aspect of the model) as possible and have a quantitative measure of the adequacy of each realization in capturing the main features of the modeled system. It is important to note that goodness-of-fit results and other statistical results for the current realization will be used after analyzing all realizations to obtain some of the acceptance criteria measures, P_1 through P_5, which are discussed in detail in Section 5.3.

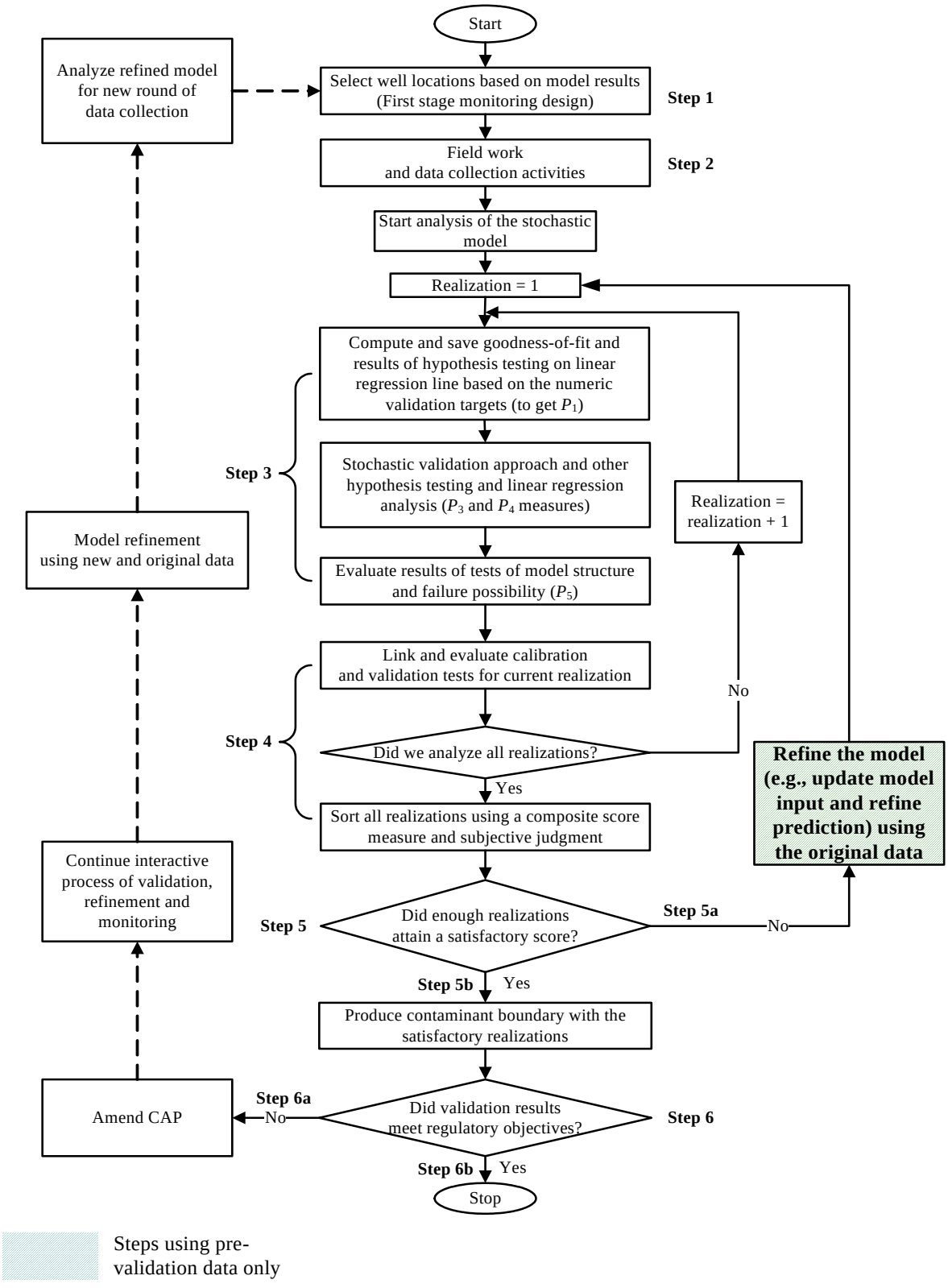

Figure 4. Details of the proposed model validation and postaudit processes for the Shoal model with the selection criteria measures (P_1 through P_5) explained in Section 5.3.

Step 4: Link the results of the calibration accuracy evaluations performed during the model-building stage and the validation tests (step 3) for all realizations and sort the realizations in terms of their adequacy and closeness to the field data. A subjective element may be invoked in this sorting based on expert judgment and hydrogeologic understanding. The objective here is to filter out the realizations that show a major deviation or inadequacy in many of the tested aspects and focus on those that "passed" the majority of the tests and evaluations. By doing so, the range of output uncertainty is reduced and the subsequent effort can be focused on the most representative realizations/scenarios. To continue reducing the uncertainty level, a refinement of the conductivity or other input distribution can be made based on information collected from the validation wells. More details about the acceptance criteria for individual realizations and for the conceptual model are presented in Section 5.3.

Step 5: Step 4 results will determine the forward path and guide the decision as to whether there is a sufficient number of realizations that attained a satisfactory high score (thus building confidence in the original model) and are considered sufficient for further analysis or whether this number of realizations is not sufficient in comparison to the realizations with low scores indicating that the original flow and transport models or their input parameter distributions need adjustment.

5a. If the number of realizations with low scores is very large compared to the total number of model realizations (The P_1 measure discussed in Section 5.3), it could be an indication that the model has a major deficiency or conceptual problem or it could be that the input is not correct. In the latter case, it may be that the model is conceptually good, but the input parameter distributions may be skewed one way or another. Generating more realizations and keeping those that fit the validation criteria can shift the distribution to the proper position. This can be done using the existing model without conditioning or using any of the new validation data. If the model has a major deficiency or conceptual problem, generating additional realizations will not correct it and continued failure per the validation criteria (see Section 5.3) will be obvious. The importance of the distinction between conceptualization problems and inappropriate input distributions has been discussed by Bredehoeft (2003). He states that one should carefully ask the question of whether the mismatch between the model and the field observations is a result of poor parameter adjustments or does it suggest that the conceptual model be revisited. The use of different metrics such as P_1 and P_2 discussed in Section 5.3 provides a tool for making this distinction.

5b**.** If the number of realizations with high scores is found sufficient, this indicates that the model does not have any major deficiencies or conceptual problems. This determination will be made according to a number of metrics as detailed in the validation criteria section. These metrics are tested and supported by statistical hypothesis testing and provide good evaluation criteria for the model realizations. Based on the realizations retained in the analysis and deemed acceptable, a contaminant boundary will be calculated and compared to the original contaminant boundary. This comparison will be presented for reference by decision makers in step 6.

Step 6: Once the model performance has been evaluated per the acceptance criteria, the model sponsors and regulators have to answer the last question in Figure 4. This question will determine whether the validation results meet the regulatory objectives or not. This is the trigger point that could lead to significant revision of the original model.

6a. If the answer to the question posed is no, then the left-hand side path in Figure 4 begins with an evaluation of the investigation strategy, consistent with the process flow diagram in Appendix VI of the FFACO. If the original strategy is deemed sound, a new iteration of model development will begin, using the data originally collected for validation, and steps 1 to 6 will be eventually repeated. Whatever strategy is selected, the CADD/CAP will be amended before execution.

6b. If the answer to the question posed is yes, validation is deemed sufficient and the model is considered adequate or robust.

Numerical groundwater models, and in particular stochastic models, are very complex and modifying or changing any aspect of the model may produce unanticipated consequences in a different aspect of the model. To get the best outcome of the validation process, one needs to both consider the different details separately and take the broader view of the entire model while working step-by-step through the different decisions and trade-offs.

It can be seen and expected that the process of validating a site-specific groundwater model is not an easy one. Throughout the structured process described above, there may be a desire to confirm that the work is on the right track. The way to this confirmation is the cumulative knowledge gained from the different stages of the validation process. That is, a set of independent tests and evaluations will provide knowledge about the model performance, and the test results will provide some incremental, but additive, pieces of information that will be of importance. While there are no guarantees of success (attaining a conclusive outcome about model performance), the combined presence of these different results and evaluations sharply improves the odds that one can make a good decision about the model performance.

Two aspects of the validation process are explained in detail in the following subsections. These are the validation targets for the Shoal model, and the acceptance criteria for determining the sufficiency of the number of acceptable realizations. It should be emphasized here that the proposed validation targets and the acceptance criteria may need to be changed depending on the type of data that can be obtained and the practicality of collecting certain types of information. In other words, from a statistical point of view, one would desire to obtain a large number of head measurements for example, but the practical limitations may render this desire unrealistic. Therefore, the tools and tests to be used and the targets to be analyzed must consider the role they play in the model as well as practical considerations.

5.2 Validation Targets

To select the validation targets at Shoal, a multi-parameter uncertainty analysis is performed to identify the sensitivity of contaminant boundaries to various parameters. One should distinguish between sensitivity analysis and uncertainty analysis. Sensitivity analysis is the systematic investigation of the model responses to extreme values of the model input or to drastic changes in the model structure (Kleijnen, 1999). Sensitivity analysis can support validation: such analysis shows whether factors have effects that agree with the modeler’s prior

qualitative knowledge (for example, faster transport rates result from lower fracture porosity). Unfortunately, not all subsurface processes have effects with known signs; yet, many do have factors with known signs. Sensitivity analysis further shows which factors are important. If possible, information on these factors should be collected for validation purposes (Kleijnen, 1999).

Related to sensitivity analysis is uncertainty analysis. Uncertainty analysis also runs a simulation model for various combinations of input parameter values. Uncertainty analysis is performed when the input parameter values of the simulation model are not accurately known, and thus uncertainty analysis samples from a pre-specified probability distribution for these parameters. The simulation results for uncertainty analysis are commonly presented as ranges of values with associated probabilities or confidence intervals. Uncertainty analysis can be performed for individual parameters and for combinations of parameters. When it is desired to determine the impact of uncertainty in an individual parameter on the model output, the input values for that parameter are generated from its respective probability distribution while keeping all other model parameters at fixed values. In a combined parametric uncertainty analysis, all uncertain model parameters are simultaneously sampled from their respective probability distributions.

The numerical model of groundwater flow and transport at Shoal requires quantitative descriptions of numerous aspects of the conceptual model including fracture geometry and hydraulic properties, groundwater recharge, matrix diffusion, and rates of radionuclide release from glass puddles in the cavity. All of these components contribute to the transport predictions, but the most critical are those that determine the pattern and magnitude of groundwater velocities and, as a consequence, influence the travel times of radionuclides away from the cavity. Large-scale flow and transport models have shown that the results of radionuclide transport calculations are most profoundly impacted by parameters that affect travel time (Pohll *et al*., 1999a; Pohlmann *et al*., 1999; Hassan *et al*., 2002). Naturally, all of the flow and transport parameters are subject to the uncertainties that are always present when representing subsurface conditions. These parametric uncertainties are incorporated and carried through the Shoal numerical modeling process, and are therefore ultimately included in predictions of the contaminant boundaries.

Though the Shoal model presented by Pohlmann *et al*. (2004) analyzed combined parametric uncertainty, sensitivity of the transport predictions to individual parameter uncertainties was not assessed. Thus, the following analysis presents the results of individual parametric uncertainty analysis and the impact on an estimated contaminant boundary for ^{14}C. Carbon-14 is selected because of the long half-life (thus minimum decay in the 1,000-year regulatory time frame) and because it does not adsorb onto the solid matrix, thereby representing one of the main elements determining the shape and extent of a contaminant boundary (Pohll *et al*., 2003). The use of ^{14}C in this analysis will yield contaminant boundaries that are similar, but not identical, to the contaminant boundary calculated using the entire test inventory by Pohll and Pohlmann (2004). Note that the ^{14}C boundary analysis presented here is based on the regulatory Maximum Contaminant Level (MCL) of 2,000 pCi/L.

The components of the Shoal groundwater flow and transport model that incorporate uncertainty are listed below. It should be noted that the shear zone and hydraulic divide are considered to be no-flow boundaries with known geometry and are therefore treated as deterministic aspects of the model (Pohlmann *et al*., 2004).

1. Orientation of fracture zones
2. Spatial continuity of fracture zones
3. Hydraulic conductivity of fracture zones
4. Hydraulic conductivity of intervening zones of small random fractures
5. Hydraulic conductivity of the cavity and damaged zone
6. Recharge from precipitation entering the top surface of the model
7. Flux originating from upland recharge that enters the up-gradient vertical face of the model
8. Porosity of fracture zones
9. Porosity of the cavity
10. Porosity of the damaged zone around the cavity
11. Rate of nuclear glass dissolution in the cavity
12. Diffusion of radionuclides between fractures and matrix

All these parameters were explicitly considered uncertain in the Shoal model analysis except the hydraulic conductivity of intervening zones of small random fractures that was considered as a calibration parameter and the matrix diffusion parameter that was spatially and temporally varying (Pohlmann *et al*., 2004). Although a statistical distribution for the hydraulic conductivity of the intervening small fractures is not directly specified, it is indirectly handled as an uncertain parameter via the automated calibration process. Similarly, the matrix diffusion parameter was varying in space and in time based on the spatial variability of the velocity field and the residence time of each particle in the fractures (implicitly uncertain). However, it was based on deterministic values of the fracture spacing, the fracture retardation, the matrix retardation, fracture porosity, matrix porosity, and molecular diffusion coefficient. These values were obtained by calibrating the transport model to the Shoal tracer test data (see Appendix D in Pohlmann *et al*. [2004] for more details.)

Hereafter, Pohlmann *et al*.'s (2004) model is referred to as the base-case model. Not all of the model parameters listed above are included in the individual parametric uncertainty analysis performed here. The flux from the upgradient face, and nuclear melt glass dissolution rate is the calibration parameter of hydraulic conductivity of small fracture zones. These parameters have limited usefulness as model validation targets because they would be difficult to quantify even with further field work. Other parameters included in the individual uncertainty analysis that are similarly ill-suited for validation are the conductivity of the cavity and damaged zone, porosity of the cavity and damaged zone, and length of fracture zones. For example, the lengths of fracture zones have been extensively studied through geologic mapping of the land surface at Shoal. It is therefore unlikely that further data collection would provide information that differs from the model because virtually all of the visible fractures have been mapped. Therefore, five flow parameters and four transport parameters are selected to perform the individual parametric uncertainty analysis. Table 1 shows the nine cases and the notation to be used in the results discussion presented later in this section.

Table 1. Uncertainty cases for five flow parameters and four transport parameters. Two flow realizations are selected to run the transport parameter uncertainty analyses: a) one of the fastest realizations in the base-case model, and b) the realization with the least root mean squared error (RMSE) in the base-case model.

	Parameters									
Case	**Fracture Orientation (both orientations)**	**Fracture Length (both orientations)**	**Cavity *K* (all 3 zones)**	**Fracture *K***	**Recharge**	**Fracture Porosity**	**Cavity Porosity**	**Damage Zone Porposity**	**Matrix Diffusion**	**Comments**
1	**Uncertain**	Fixed	Fixed	Fixed	Fixed	Fixed	Fixed	Fixed	Fixed	Uncertain flow parameters that require generating 500 flow realizations each
2	Fixed	**Uncertain**	Fixed	Fixed	Fixed	Fixed	Fixed	Fixed	Fixed	
3	Fixed	Fixed	**Uncertain**	Fixed	Fixed	Fixed	Fixed	Fixed	Fixed	
4	Fixed	Fixed	Fixed	**Uncertain**	Fixed	Fixed	Fixed	Fixed	Fixed	
5	Fixed	Fixed	Fixed	Fixed	**Uncertain**	Fixed	Fixed	Fixed	Fixed	
6a	Fixed	Fixed	Fixed	Fixed	Fixed	**Uncertain**	Fixed	Fixed	Fixed	Uncertain transport parameters using one of the fastest flow realizations in the base case model
7a	Fixed	Fixed	Fixed	Fixed	Fixed	Fixed	**Uncertain**	Fixed	Fixed	
8a	Fixed	Fixed	Fixed	Fixed	Fixed	Fixed	Fixed	**Uncertain**	Fixed	
9a	Fixed	Fixed	Fixed	Fixed	Fixed	Fixed	Fixed	Fixed	**Uncertain**	
6b	Fixed	Fixed	Fixed	Fixed	Fixed	**Uncertain**	Fixed	Fixed	Fixed	Uncertain transport parameters using the flow realization with the smallest RMSE in the base case model
7b	Fixed	Fixed	Fixed	Fixed	Fixed	Fixed	**Uncertain**	Fixed	Fixed	
8b	Fixed	Fixed	Fixed	Fixed	Fixed	Fixed	Fixed	**Uncertain**	Fixed	
9b	Fixed	Fixed	Fixed	Fixed	Fixed	Fixed	Fixed	Fixed	**Uncertain**	
Base Case	**Uncertain**								**Spatially and temporally varying**	

The analysis of parametric uncertainty independently evaluates the selected uncertain components of the model, producing a comparison of the relative contribution of each to the overall output uncertainty. The output uncertainty for the base case and for all the uncertainty cases is obtained in terms of an estimated ^{14}C contaminant boundary. However, virtually any model result can be used as the output metric. The analysis proceeds by running the calibrated model in Monte Carlo mode for each of the individual uncertain parameters that are under consideration; all of the other uncertain parameters are held at constant values that are representative of their distribution. For each uncertain parameter, the Monte Carlo values are chosen from the input parameter distributions used in the contaminant-boundary model (Pohlmann *et al*., 2004). However, only 500 realizations are used for this uncertainty analysis (to reduce computational time) as opposed to the 1,000 realizations used in Pohlmann *et al*. (2004) due to the large number of cases studied here. This does not impact the results, the interest is in the relative contribution of each uncertain parameter to the overall output uncertainty of the model.

For each of the uncertain flow parameters (Case 1 through Case 5), the MODFLOW-2000 flow solver is run to generate a set of flow realizations that incorporate the uncertainty in each targeted flow parameter. These flow realizations are then used to model the radionuclide transport with the transport parameters held fixed at their mean values for all flow cases. When

the effects of uncertain transport parameters are studied, a single realization of the flow field is used and the transport problem is solved 500 times with the uncertain transport parameter value being drawn from the respective parameter distribution. The selection of the flow realization to be used for these analyses is based on criteria discussed in the results section. The results of the uncertainty analysis for the five flow cases and the four transport cases will be analyzed for their relative impact on the overall uncertainty in the size of the contaminant boundary. The ^{14}C contaminant boundary is calculated using an unclassified estimate of ^{14}C mass for the Shoal test (0.24 Curies), based on data from Smith (2001). In that report, Smith presents the average source term for the UGTA tests on Pahute Mesa (about 85 tests). The average Pahute Mesa source term for ^{14}C is about 7.3 Curies. However, when this value is scaled by the ratio of tritium (^{3}H) source terms (the unclassified ^{3}H estimate for Shoal relative to the average ^{3}H source term of Pahute Mesa), the resulting ^{14}C source term for Shoal is about 0.24 Curies.

5.2.1 Uncertain Flow Parameters

The five flow parameters considered here are the fracture orientation, the fracture length that indicates the continuity of fractures, the hydraulic conductivity, K, of the cavity and surrounding zones (three conductivity values), the hydraulic conductivity of fractures, and the recharge at the domain top. The fracture orientation plays an important role in determining how far from the cavity radionuclides can migrate. When fractures are oriented parallel to the general flow direction from the cavity outward, the migration distance increases. However, if the fractures are oriented normal to that flow direction, migration distances decrease. Therefore, the uncertainty associated with the fracture orientation may impact the contaminant boundary and lead to uncertainty in its extent. In the base-case model (Pohlmann *et al*., 2004), four sets of fractures were chosen to represent the distribution of orientations for one of the flow categories (Flow Category 2 according to Pohlmann *et al*. [2004]), which represents zones of strongly oriented, large fractures. These fracture sets have the most well-defined orientations and contain the highest proportion of the total fractures. Flow Category 1 (Pohlmann *et al*., 2004) is assumed to have no preferred spatial orientation. For those cases in which the fracture orientation was kept fixed (i.e., deterministic), conditioned fracture orientations were set to the mean orientations for each class, and unconditioned fractures were first randomly sampled to determine the class, and then the mean orientations were applied.

Case # 1

In uncertainty case # 1, the fracture orientation for the two flow categories are considered uncertain while fixing all other flow parameters (and subsequently transport parameters) at their best estimate or mean values. The empirical orientation distribution used in Pohlmann *et al*. (2004) is used here to draw the values of the fracture orientation for each realization. Figures 5 and 6 display the impact of the fracture orientation uncertainty on the ^{14}C contaminant boundary and the associated uncertainty. Transport simulations and contaminant boundary computation are performed in the same manner described in Pohlmann *et al*. (2004).

Figure 5 shows the uncertainty case # 1 compared to the base-case model of Pohlmann *et al*. (2004). The uncertainty case has few realizations with larger RMSE than the base-case model, but the range of RMSE for the uncertainty case is smaller than the base case. For the plume extent, it is clear that the base-case model has a much larger uncertainty range compared to the fracture orientation uncertainty case. If one removes the fastest two realizations in the uncertainty case, it can be concluded that the uncertainty range of the plume extent for the

base-case model is at least four times larger than that for the uncertainty case. It is important to note that the plume extent is obtained by determining the *y* coordinate of the farthest cell where the maximum ^{14}C concentration in the *x*-*y* projection over 1,000 years is above zero and then subtracting the *y* coordinate of the working point.

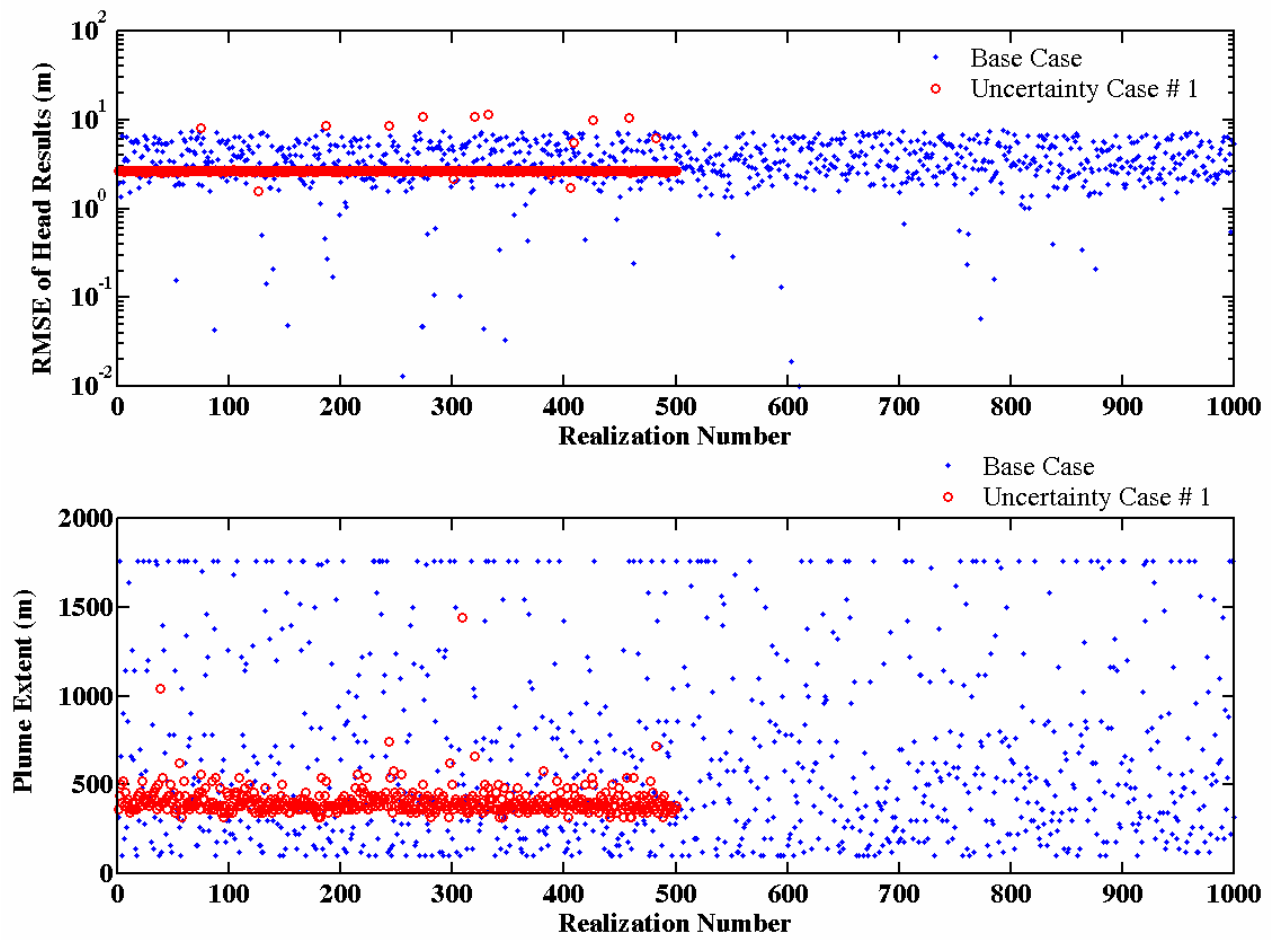

Figure 5. Comparison between the base-case model and the fracture orientation uncertainty case # 1 in terms of the root mean square error (RMSE) of the head results for each flow realization and the plume extent expressed as the distance between the working point and the farthest point traveled by any particle in the transport simulations.

Figure 6 shows the impact on the contaminant boundary and its uncertainty. The subplots a), b), and c) show the different two-dimensional projections of the three-dimensional contaminant boundary volume. Again, these are computed in the same manner described in Pohlmann *et al*. (2004). These subplots compare the contaminant boundaries of the base-case model, which are plotted to scale and in the correct cavity location indicated by the red square, to those of the uncertainty case # 1, which are shifted from their right location to clarify the comparison. The contaminant boundaries of the uncertainty case are, however, drawn to scale in terms of their size so that they could be directly compared to the base-case model.

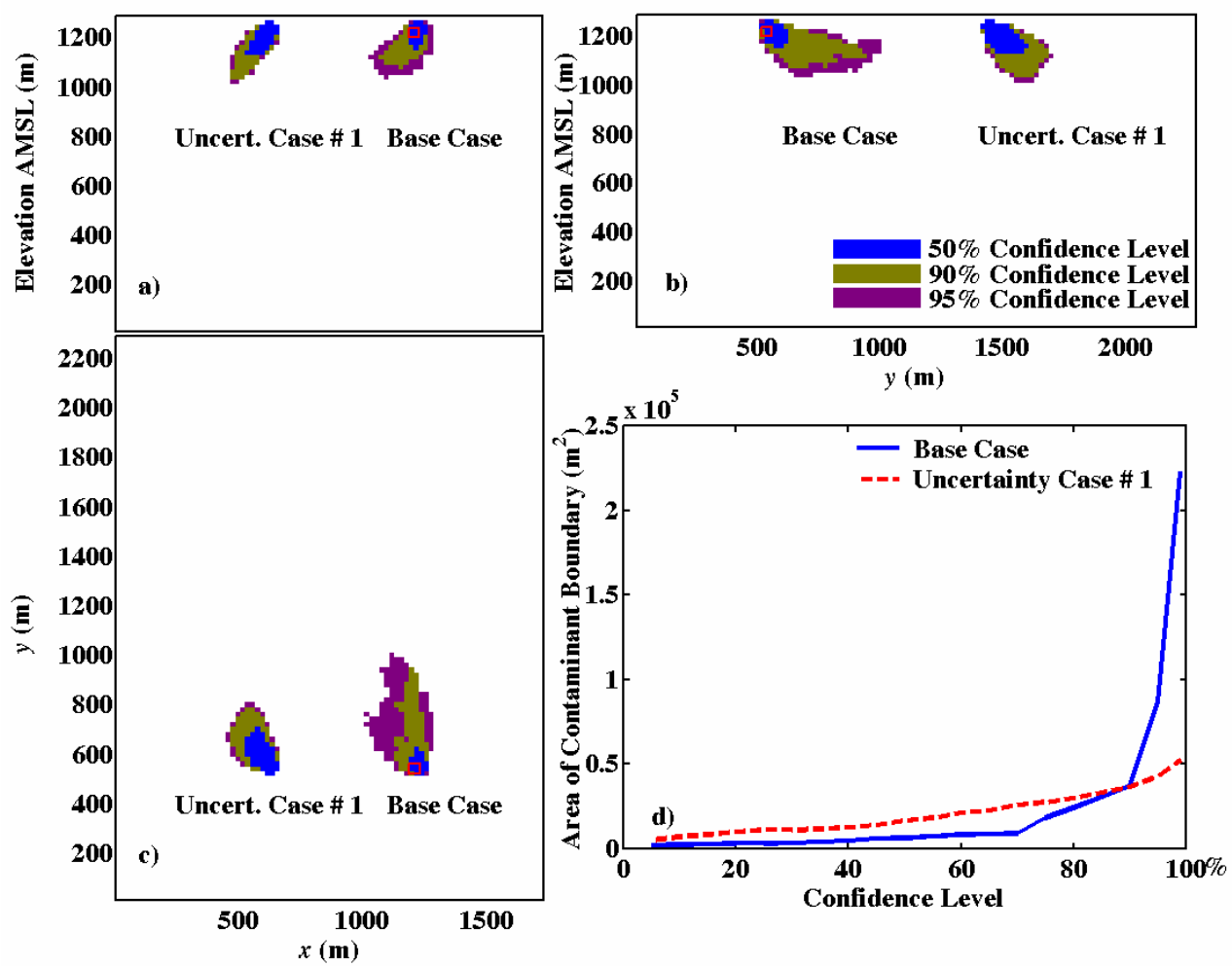

Figure 6. Contaminant boundary maps for ^{14}C delineating the areas exceeding 2,000 pCi/L at different confidence levels. Subplot a) shows the *x*-*z* (or elevation) map, subplot b) shows the *y*-*z* map, subplot c) shows the *x*-*y* map, and subplot d) shows the *x*-*y* contaminant boundary area at different confidence levels. The uncertainty case # 1 is drawn to scale, but is spatially shifted on the plots for comparison to the base-case model (Pohlmann *et al*., 2004).

The subplot d) shows the area of the *x*-*y* projection of the contaminant boundary obtained at the 5th, 10th, …, 99th confidence levels. This subplot provides an easy way to compare the impact of individual uncertain parameters to the impact of having all parameters as uncertain on the contaminant boundary uncertainty. The steeper the curve, the more uncertain the resulting contaminant boundary is. If one assumes no uncertainty, the 500 (or the 1,000) realizations will give the same exact result and the contaminant boundary area will be the same at all confidence levels (i.e., horizontal line in subplot d).

It is clear that the contaminant boundaries are smaller in area for the uncertainty case at the 90th and 95th confidence levels. It is also clear from subplot d) that the base-case model has a much larger degree of uncertainty than the uncertainty stemming from the fracture orientation

alone. Therefore, fracture orientation contributes to the overall uncertainty but to a small extent. New field data on fracture orientation would not thus yield significant reduction in the overall model uncertainty. This aspect, added to the difficulty of better characterizing fracture orientations at Shoal, means that the fracture orientation is not a useful validation target.

Case # 2

The fracture continuity expressed by fracture length is the second uncertain parameter that requires new flow realizations. The fractures and fracture networks are the principle pathways of water and contaminants through an otherwise impermeable or low-permeability rock such as the granite at Shoal. Field and laboratory experiments in natural fractures have demonstrated strong evidence of highly preferential flowpaths in individual fractures and fracture networks (Neretnieks *et al*., 1982; Neretnieks, 1993). Field data, for example from large-scale investigation of fracture flow in a granite uranium mine at Fanay-Augeres, France, show four orders of magnitude difference between the largest and smallest injection flow rates despite very good fracture connectivity (Berkowitz, 2002). Cacas *et al*. (1990a, b) concluded that the high degree of heterogeneity is due to a broad distribution of fracture conductivities, and that it overwhelmingly governs flow and transport behavior. Therefore, the nature of fractures, their connectivity and conductivity distribution, the surrounding porous rock and its characteristics, and the combined effect of fractures and porous matrix on groundwater flow and contaminant transport make the analysis significantly different than for classical porous media.

As discussed by Berkowitz (2002), studies of well tests often report that of a large number of fractures intersecting a well, only one or two actually transmit fluid. For example, at the Fanay-Augeres site in France, it was found that only 0.1 percent of the fractures contributed to flow on a large scale (Long and Billaux, 1987). Thus, the question of fracture connectivity is of prime importance to flow and contaminant transport in fractured systems and to interpretation of data from single and multiple well tests. Even domains that appear to be heavily fractured may not in fact be well connected (Berkowitz, 2002). In percolation theory terminology, the salient question one has to ask here is whether the fracture network is above the percolation threshold (i.e., connectivity of fractures is sufficient to permit flow through the network from point A to point B) or conversely near the percolation threshold and thus poorly connected. Again, this question is of paramount importance when it comes to interpreting and using well test data in fractured geologic units.

At Shoal, the spatial persistence of fracture features was quantified using a digitized map of fractures observed at land surface (Pohll *et al*., 1998). Although fractures are not truly linear features, they were treated as linear so that an average strike could be calculated for each fracture. A bimodal distribution of strike orientations resulted and the data were grouped as follows (Pohlmann *et al*., 2004):

- Group1: Strike of 0 to 70 and 130 to 180 degrees East of North
- Group2: Strike of 70 to 130 degrees East of North

Fracture lengths in Group 1 are described by a lognormal distribution with mean length of 572 m and a natural log standard deviation of 0.86 m. Group 2 contained very few fractures, so a uniform distribution having a range of 100 to 750 m was used. The distribution of fracture lengths along the dip direction was assumed to be identical to the distribution of lengths along the strike direction. In the uncertainty case # 2, the fracture lengths along the dip and the strike

are kept uncertain and are generated from the respective distributions, while fixing all other flow and transport parameters as seen in Table 1. Figures 7 and 8 illustrate the results of this uncertainty case.

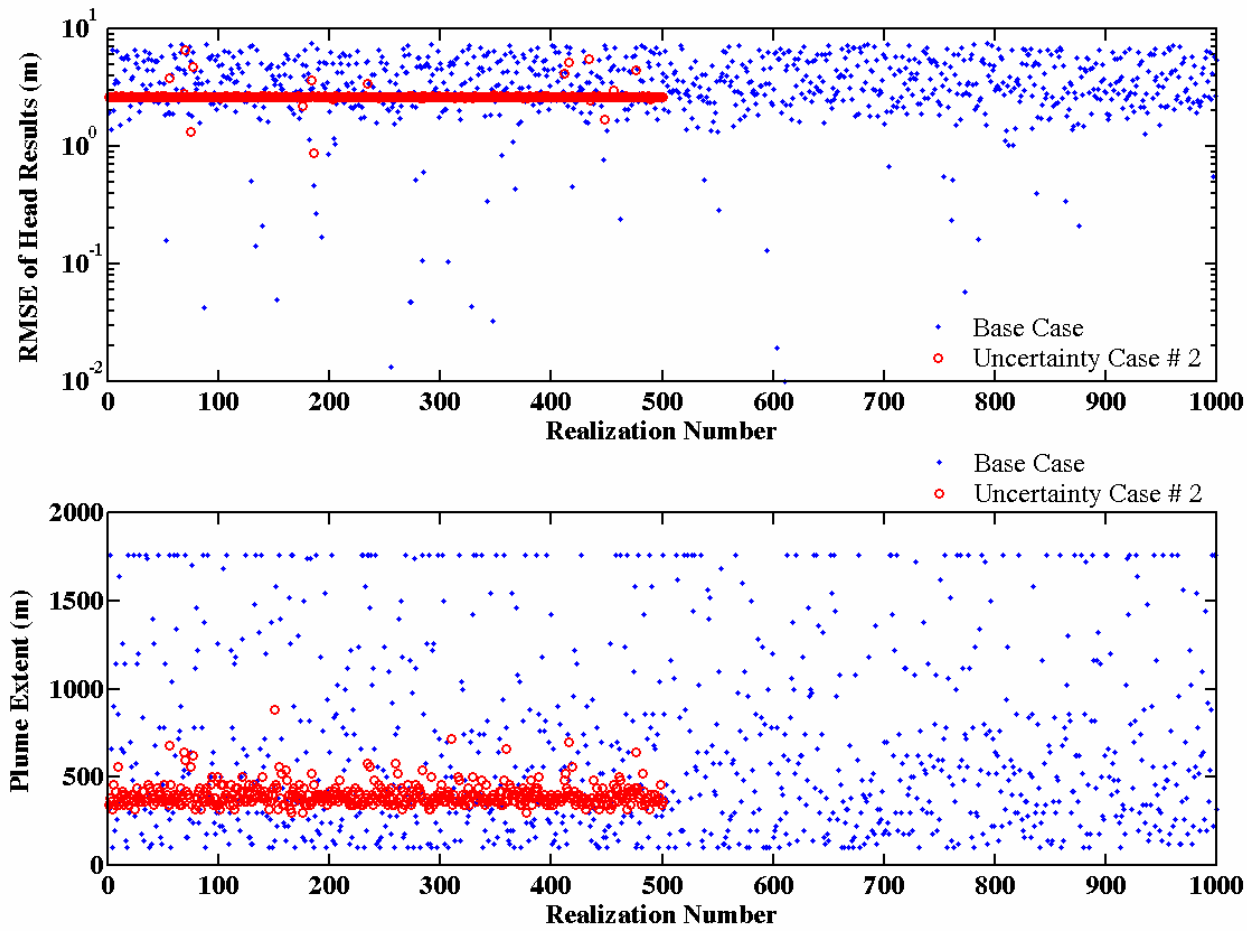

Figure 7. Comparison between the base-case model and the fracture length uncertainty case # 2 in terms of the root mean square error (RMSE) of the head results for each flow realization and the plume extent expressed as the distance between the working point and the farthest point traveled by any particle in the transport simulations.

As can be seen from Figure 7, the base-case model has a wider range of uncertainty compared to the fracture length uncertainty case. This is reflected in both the head RMSE and the plume extent. In terms of the RMSE, the base-case model has both higher and lower RMSE than the uncertainty case. This is because the base-case model has more parameter combinations that would lead to both better and worse calibration results than the uncertainty case where all parameters are fixed except the fracture lengths.

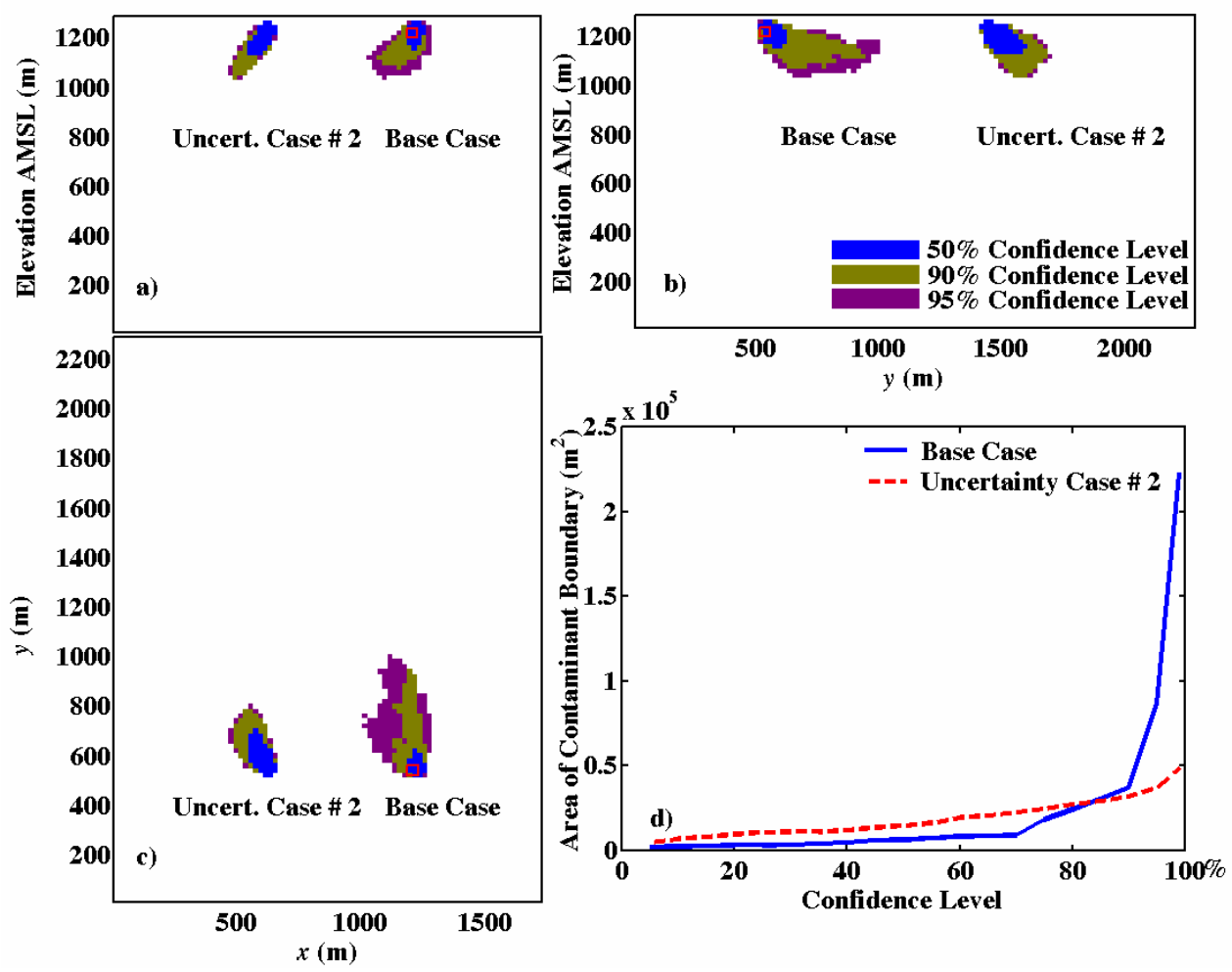

Figure 8. Contaminant boundary maps for ^{14}C delineating the areas exceeding 2,000 pCi/L at different confidence levels. Subplot a) shows the *x*-*z* (or elevation) map, subplot b) shows the *y*-*z* map, subplot c) shows the *x*-*y* map, and subplot d) shows the *x*-*y* contaminant boundary area at different confidence levels. The uncertainty case # 2 is drawn to scale, but is spatially shifted on the plots for comparison to the base-case model (Pohlmann *et al*., 2004).

Figure 8 displays the contaminant boundary maps at the 50 percent, 90 percent, and 95 percent confidence levels (subplots a, b, and c) as well as the change in the contaminant boundary area as a function of the confidence level (subplot d). The results are very similar to uncertainty case # 1, indicating that the uncertainty in fracture lengths minimally contributes to the overall model uncertainty. This can probably be attributed to the fact that all generated values for fracture lengths provide very well connected networks of fractures in the simulation domain (i.e., way above percolation threshold even for the smallest fracture lengths). Therefore, changing the fracture lengths above the connectivity threshold does not significantly change the migration distances away from the cavity. Given this result and the relative abundance of fracture data available for the Shoal site, the fracture length is deemed well characterized, and

any new information on fracture lengths will not significantly reduce model uncertainty or help the model validation/postaudit process.

Case # 3

Uncertainty case # 3 has the hydraulic conductivity of the cavity zone and the two zones surrounding it as uncertain. Three spherical zones are identified around the Shoal working point and are approximated in the base-case model by a rectangular grid (Pohlmann *et al*., 2004, Figure 3.1). The first zone represents the cavity and chimney. Values of *K* for the cavity and chimney are chosen from a random distribution and are assigned randomly to the cells corresponding to the cavity and chimney; no spatial correlation within this region of the model is assumed. The mean of this distribution is taken to be 43.0 m/d, which is about two and one half orders of magnitude higher than the mean of the *K* distribution of the undisturbed granite (the fracture conductivity studied in case # 4). A $\log_{10}$-normal distribution having a mean and standard deviation of 1.6 and 0.18 m/d, respectively, is used for the simulations. This zone is also given a porosity distribution with a higher mean as will be discussed in case # 7 below. The second zone extends from approximately 1 to 2 R_c from the working point with R_c being the cavity radius that is about 26.0 m at Shoal. This zone is also assigned *K* values that are substantially higher than that of the undisturbed granite. These *K* values are chosen from a distribution having a mean of 8.6 m/d. A $\log_{10}$-normal distribution having a mean and standard deviation of 0.94 and 0.26 m/d, respectively, is used for the simulations. Zone 3 begins the transition from the highly disturbed, near-cavity region, to the unaffected rock. This zone extends from 2 to 4 R_c from the test and is assigned *K* values at the upper end of the distribution from which the *K* values of the undisturbed granite are selected. The distribution of *K* values assigned to this zone has a mean value of 0.86 m/d. A $\log_{10}$-normal distribution having a mean and standard deviation of -0.063 and 0.26 m/d, respectively, is used for the simulations.

In this uncertainty case, the conductivity values for these three zones are kept uncertain while fixing all other flow and transport parameters. For each realization, the *K* value for each of the three zones is drawn at random from their respective distributions as described above. The flow problem is solved for each realization followed by the transport simulations for ^{14}C. Figures 9 and 10 show the results of this uncertainty case. As can be seen from Figure 9, the uncertainty in the three *K* values yields a small range of calibration accuracy (in terms of RMSE) and a small uncertainty in the plume extent.

Figure 10 indicates a smaller contaminant boundary for this case compared to the base-case model at the 90^{th} and the 95^{th} confidence levels. The uncertainty in the contaminant boundary produced by the uncertainty in *K* values of the three zones around the cavity is again much smaller than the overall uncertainty built into the base-case model as shown in subplot d) of Figure 10. Since these three zones are small in scale relative to the migration distances of radionuclides from the cavity, their impact on the contaminant movement is small and thus they contribute very little to the overall output uncertainty of the model. With the difficulty (or impossibility due to the high risk involved) of characterizing the conductivity close to the cavity and the minor impact on the model output, these *K* values are not suitable for being validation targets.

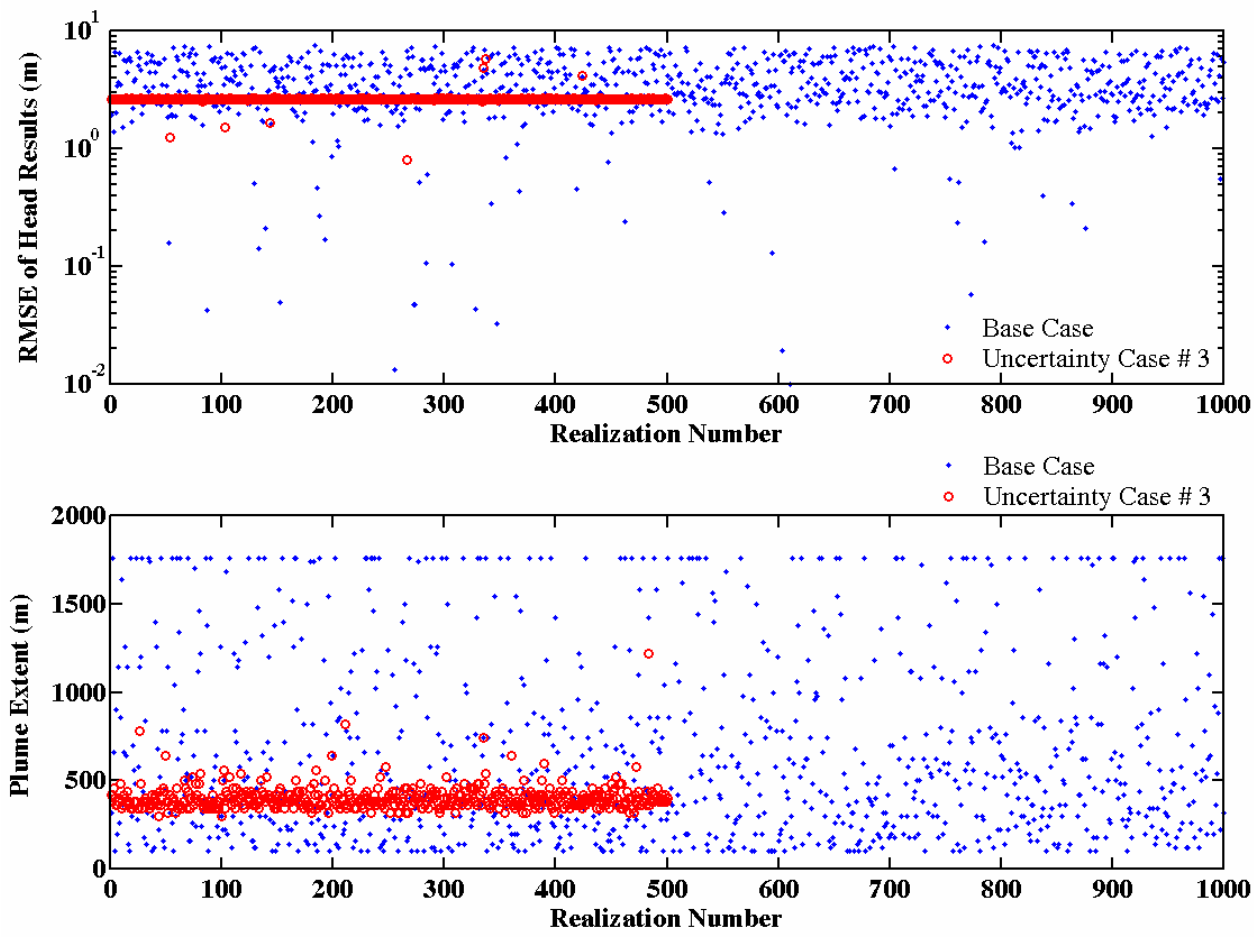

Figure 9. Comparison between the base-case model and uncertainty case # 3 (conductivity of cavity and the surrounding zones) in terms of the root mean square error (RMSE) of the head results for each flow realization and the plume extent expressed as the distance between the working point and the farthest point traveled by any particle in the transport simulations.

Case # 4

Here, the conductivity of the fractured, undisturbed granite is the only uncertain parameter. In the original model, two fracture flow categories are used to characterize the fracture flow system at Shoal. Flow Category 1 represents the small and less conductive fractures, whereas Flow Category 2 is designed to represent the more conductive portions of the granite. This latter flow category is represented in the model by a $\log_{10}$-transformed triangular distribution with the minimum at -5.0 (i.e., $K_{min} = 1.0\times10^{-5}$ m/d), the maximum at 0.0 (i.e., $K_{max} = 1.0$ m/d), and the mean at -2.5 (i.e., $K_{mean} = 3.2\times10^{-3}$ m/d). The mean of this distribution is consistent with the calibrated K values obtained from the transient numerical analysis of cross-hole pumping during the 320-day tracer test conducted at HC-6 and HC-7 (3.4×10^{-3} m/d) and the regional flow model for Shoal (6.9×10^{-3} m/d). In addition, the range of the K distribution for Flow Category 2 captures the full range of the field data. The K values for Flow Category 1 were obtained during model calibration using the methodology described in Pohlmann *et al*. (2004).

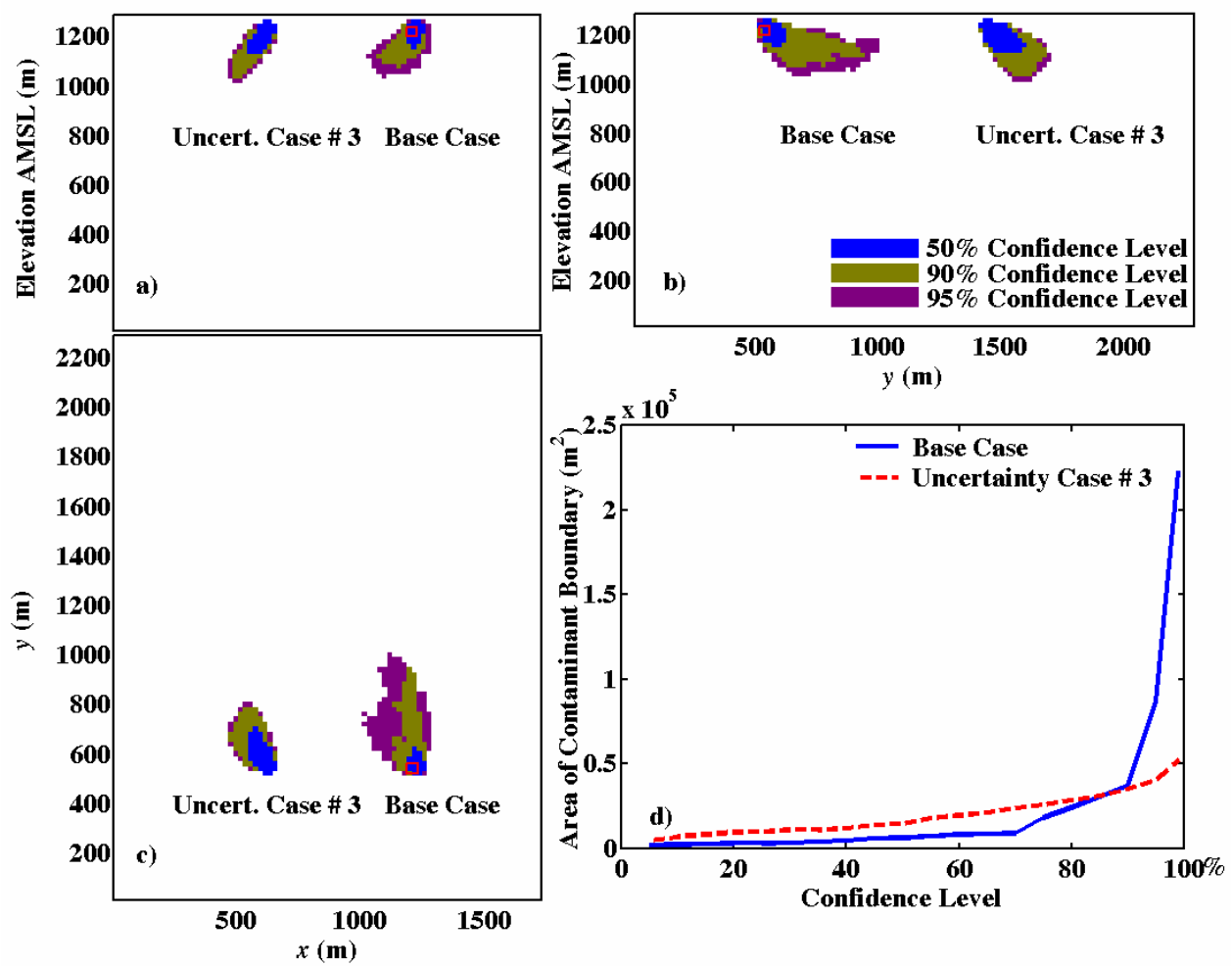

Figure 10. Contaminant boundary maps for ^{14}C delineating the areas exceeding 2,000 pCi/L at different confidence levels. Subplot a) shows the *x*-*z* (or elevation) map, subplot b) shows the *y*-*z* map, subplot c) shows the *x*-*y* map, and subplot d) shows the *x*-*y* contaminant boundary area at different confidence levels. The uncertainty case # 3 is drawn to scale, but is spatially shifted on the plots for comparison to the base-case model (Pohlmann *et al*., 2004).

Hydraulic conductivity values are assumed constant for each fracture in the simulated domain and no spatial correlation is assumed within either flow category. Every cell within a given individual fracture zone simulated in Flow Category 2 is assigned the same value of *K* that is chosen from the distribution described above. Likewise, every cell in Flow Category 1 of a given realization is assigned the same value of *K* as determined during model calibration that relies on the generalized likelihood uncertainty estimate methodology.

The conductivity value for the Flow Category 2 is drawn from the log-transformed triangular distribution, and during the flow simulations, the *K* value of Flow Category 1 is adjusted as the calibration parameter. The results of this uncertainty case are exhibited in Figures 11 and 12.

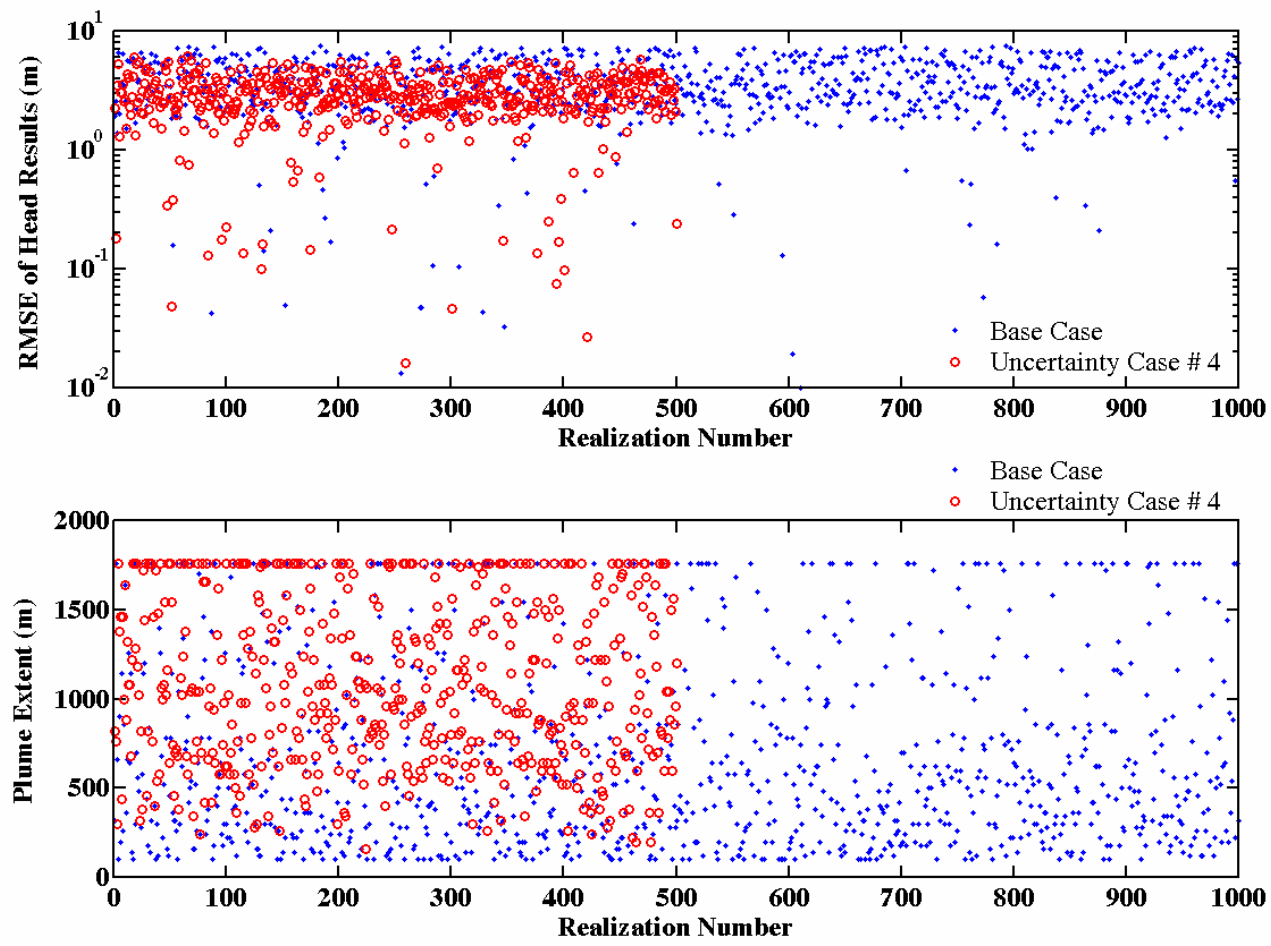

Figure 11. Comparison between the base-case model and uncertainty case # 4 (fracture conductivity) in terms of the root mean square error (RMSE) of the head results for each flow realization and the plume extent expressed as the distance between the working point and the farthest point traveled by any particle in the transport simulations.

Unlike the previous three cases, this uncertainty case shows that fracture conductivity is a major contributor to the overall model uncertainty. By far, it is the most important parameter driving the output uncertainty. As shown in Figure 11, the fracture conductivity alone produces a range of RMSE and a range of plume extent that are only slightly narrower than the range produced by the base-case model having all 11 flow and transport parameters as uncertain. Since the flow velocity in fracture Flow Category 2 is directly proportional to the hydraulic conductivity assigned to these fractures, the range of output uncertainty is derived by the uncertain velocity stemming from the conductivity uncertainty.

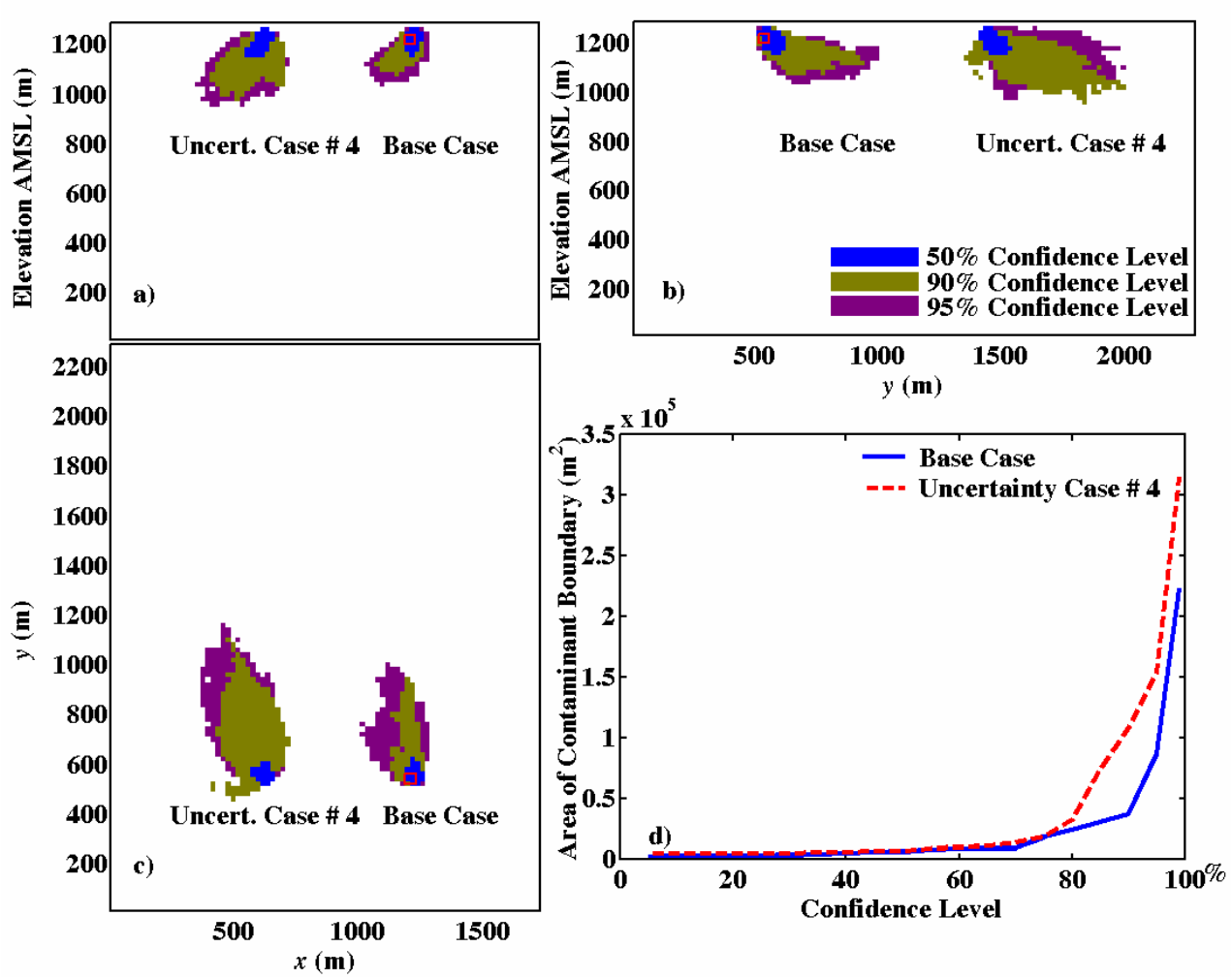

Figure 12. Contaminant boundary maps for ^{14}C delineating the areas exceeding 2,000 pCi/L at different confidence levels. Subplot a) shows the *x*-*z* (or elevation) map, subplot b) shows the *y*-*z* map, subplot c) shows the *x*-*y* map, and subplot d) shows the *x*-*y* contaminant boundary area at different confidence levels. The fracture conductivity uncertainty case # 4 is drawn to scale, but is spatially shifted on the plots for comparison to the base-case model (Pohlmann *et al.*, 2004).

The unclassified ^{14}C contaminant boundaries shown in Figure 12 indicate that the uncertainty in the fracture conductivity yields larger boundaries compared to the base-case model. It also yields slightly more uncertainty than the base-case model as shown by the comparison in subplot d) of Figure 12. In the base-case model, the uncertain fracture conductivity is normally combined with other uncertain parameters, which yields different results than when all parameters are fixed at their means and only the fracture conductivity is changing. That is, when other parameters are uncertain and in the absence of any restriction on parameter correlations, high fracture conductivity values in the base case may be accompanied by high fracture velocity values and vice versa which yield less uncertainty and smaller travel distances than when the high fracture conductivity values in this uncertainty case are always associated with the mean fracture porosity.

It is apparent that the fracture conductivity is an important parameter impacting the results of the flow and transport model at Shoal. It can thus be considered as one of the validation targets for the purpose of narrowing down the range of uncertainty for that parameter and the resulting contaminant boundary uncertainty. Thus, fracture conductivity measurements should be considered as validation targets. Multiple measurements in each new well are desirable and should be considered when feasible.

Case # 5

The recharge entering the top surface of the model domain is the last flow parameter studied in this uncertainty analysis. Recharge to the groundwater system by infiltration of precipitation through the land surface is a critical parameter controlling the velocity and direction of groundwater flow. Precipitation is thus expected to be the driving factor determining how much recharge can occur. However, recharge to the groundwater system is constrained by the observed hydraulic properties of hydraulic conductivity and head distribution. For example, a recharge rate too high relative to the hydraulic conductivity of the aquifer results in modeled water levels far higher than those observed in the field. The Data Decision Analysis of Pohll *et al*. (1999b) identified uncertainty in recharge as a significant contributor to overall model uncertainty. As a result, recharge was one of the parameters treated stochastically in the base-case flow model (Pohlmann *et al*., 2004). Recharge is applied evenly over the top of the model domain with the value for each realization selected from a triangular distribution ranging between 0.05 and 0.70 cm/yr.

The recharge model employed in Pohlmann *et al*. (2004) uses the most recent and accurate recharge data in combination with a robust statistical model and it is in general agreement with the vadose zone model of Pohll (1999), which suggests that these two methods are more favorable as compared to the Maxey-Eakin recharge model and the thermal profile methods. Given the uncertainties in all models, a parsimonious model seems appropriate. Therefore, a triangular distribution is chosen to represent the potential range of recharge over the model domain as 0.05 to 0.70 cm/yr.

Using this distribution, the flow model is run with all other parameters fixed at their mean values. However, the hydraulic conductivity of fracture Flow Category 1 (small fractures) is again used as a calibration parameter and is adjusted automatically during the flow simulation. The hydraulic conductivity of Flow Category 2 is kept fixed at its mean value. The results of this uncertainty case are shown in Figures 13 and 14.

Figure 13 shows that the recharge uncertainty case produces both smaller and larger RMSE than the base-case model in only a few realizations. In most of the realizations, the range of the head RMSE in the recharge uncertainty case is very similar to that for the base-case model. The plume extent, however, does exhibit a much smaller range in the recharge uncertainty case than the base-case model. Although recharge impacts the groundwater flow velocity, fixing the *K* value of Flow Category 2 reduces the range of velocity variability induced by the variability in recharge. It is important also to note that recharge variability alone produces a maximum plume extent of about 1,000 m across all realizations, whereas the base-case model produces a maximum plume extent of about 1,800 m. It is also of interest to note that this uncertainty case includes one realization that has a very small RMSE that is smaller than the best realization of the base-case model. This realization is further analyzed in the discussion of the transport parameters uncertainty results.

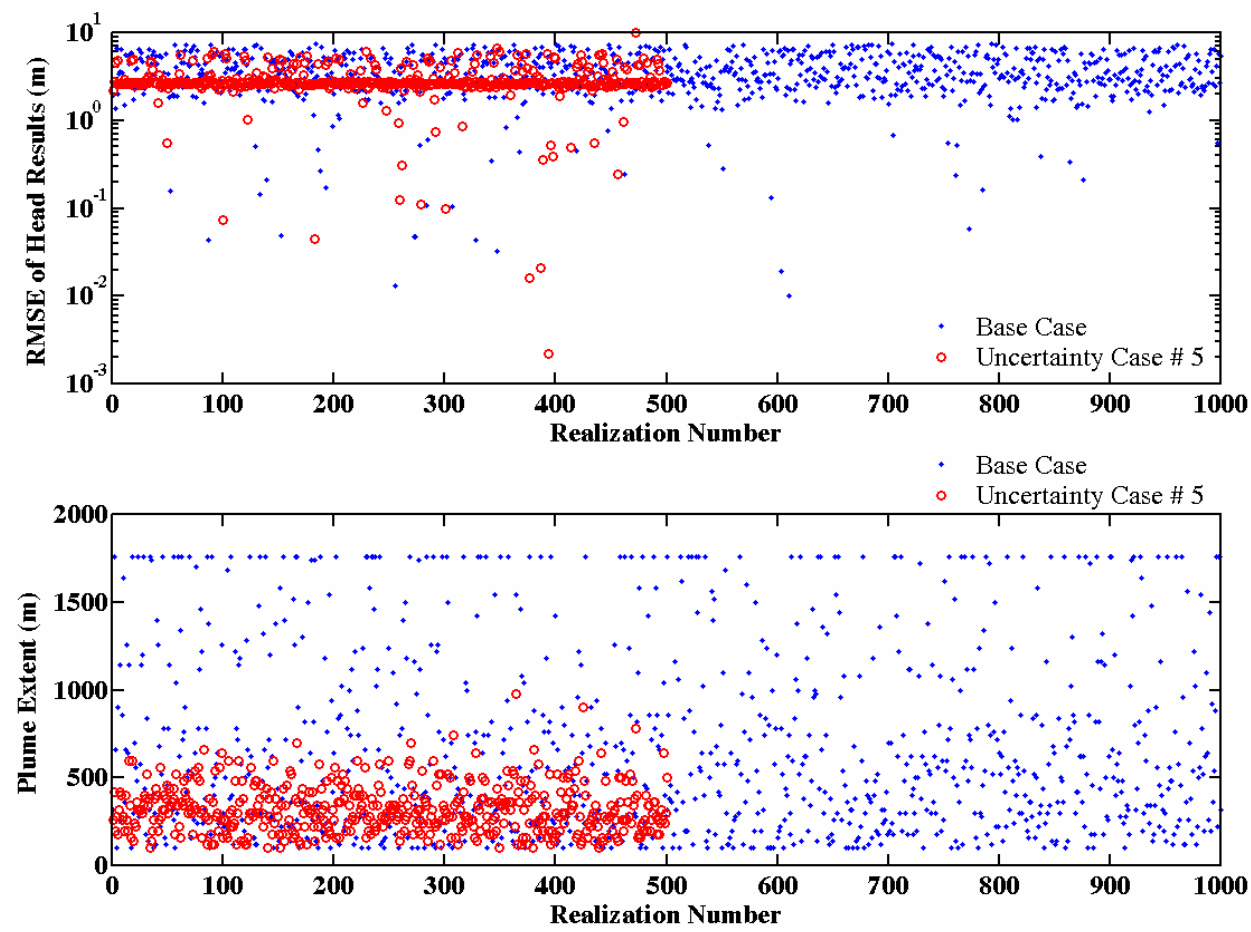

Figure 13. Comparison between the base-case model and uncertainty case # 5 (recharge uncertainty) in terms of the root mean square error (RMSE) of the head results for each flow realization and the plume extent expressed as the distance between the working point and the farthest point traveled by any particle in the transport simulations.

Figure 14 shows the resulting contaminant boundary maps and how they compare to the base-case model. Interestingly, the uncertainty in the recharge value yields minor uncertainty in the contaminant boundaries. Only between the 90 percent and the 99 percent confidence levels does uncertainty appear in the results. So despite what one might expect, the recharge variability contributes insignificantly to the contaminant boundary uncertainty. In terms of the contaminant boundary size, the recharge uncertainty case yields very small sizes compared to the base-case model. Therefore, recharge may not be considered as a validation target. However, if new temperature logs or other data provide independent recharge values, then these values could be used to confirm or invalidate the range of recharge used in the model.

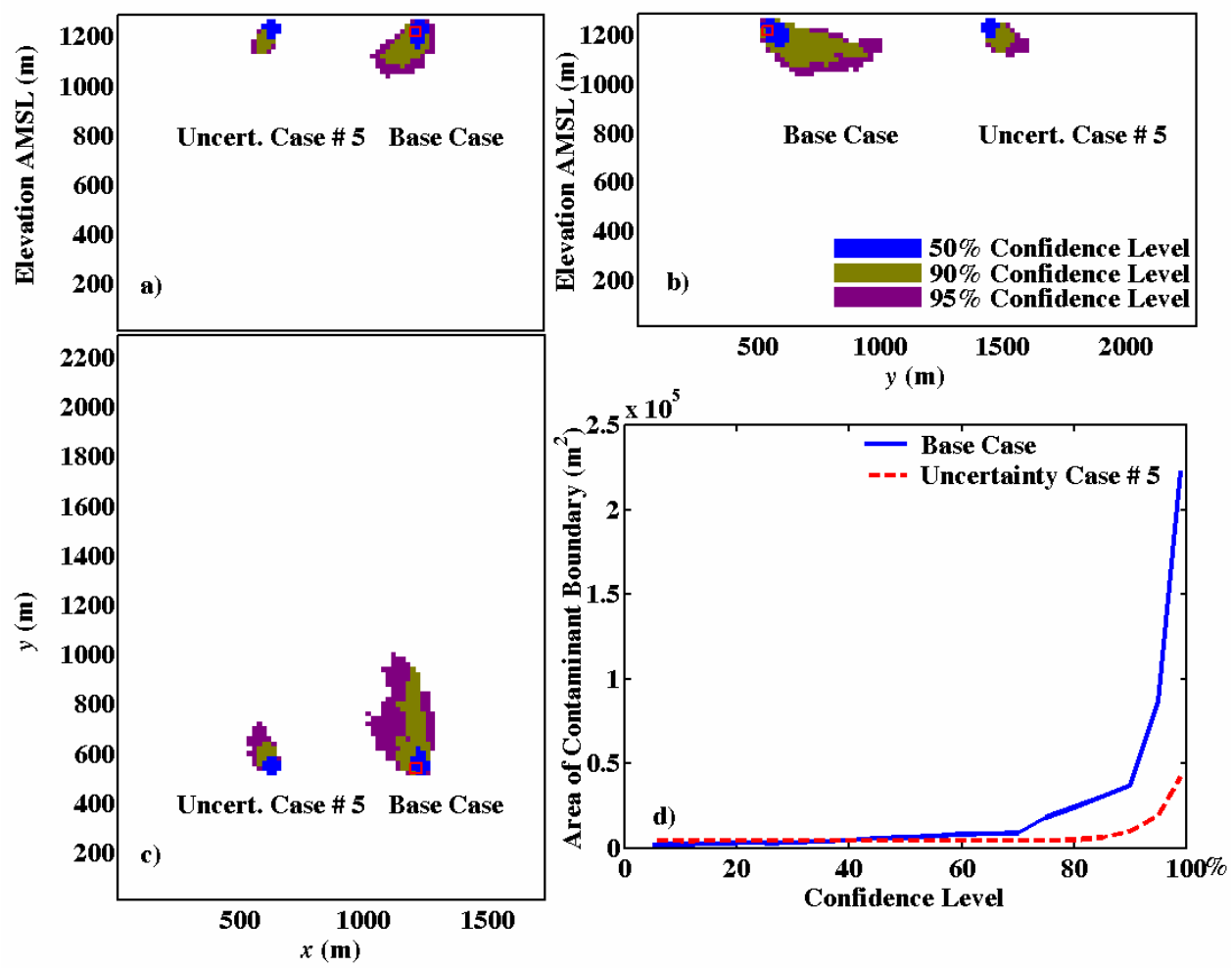

Figure 14. Contaminant boundary maps for ^{14}C delineating the areas exceeding 2,000 pCi/L at different confidence levels. Subplot a) shows the *x*-*z* (or elevation) map, subplot b) shows the *y*-*z* map, subplot c) shows the *x*-*y* map, and subplot d) shows the *x*-*y* contaminant boundary area at different confidence levels. The recharge uncertainty case # 5 is drawn to scale, but is spatially shifted on the plots for comparison to the base-case model (Pohlmann *et al.*, 2004).

5.2.2 Uncertain Transport Parameters

Four transport parameters are considered in this analysis. These are the fracture porosity, the cavity porosity, and the damaged zone porosity. These are shown in Table 1 presented earlier. To run the uncertainty analysis for any of these parameters, a single flow realization needs to be selected and the transport problem solved for different values of the parameter being studied. The flow realization selected for the analysis has a major influence on the results. Therefore, two criteria were used and two realizations selected to run the analysis for transport parameters. First, one of the fastest flow realizations was used in the base-case model. About 88 realizations were identified in the base-case model where particles reached the northern domain boundary. Note that individual particles reaching the end of the domain do not necessarily equate to a concentration exceeding the MCL, and thus do not necessarily lead to inclusion in the

contaminant boundary. Among these realizations, the one with the least RMSE was selected. This realization is used for cases 6a, 7a, 8a, and 9a as shown in Table 1.

The second criterion used to select a flow realization is the goodness of fit in the calibration process. Based on the RMSE of the modeled heads, the base-case flow realization that has the least RMSE value was selected. This realization is the basis for the transport parameters uncertainty analyses presented in cases 6b, 7b, 8b, and 9b as shown in Table 1 and discussed below. As mentioned earlier, one of the flow realizations in uncertainty case # 5 attained an RMSE value smaller than this selected realization. Although not shown here, we ran the transport uncertainty analyses was run using that realization. It was found that the contaminant boundary did not exceed the cavity area for all cases and with no uncertainty. The reason for this result is the fact that this realization produces very flat head gradients around the cavity, which lead to very low velocities in the cavity vicinity. Therefore, regardless of the value of any transport parameter, these low velocities inhibit the particle migration and lead to small contaminant boundaries.

Cases 6a and 6b

In case 6, the fracture porosity is considered uncertain where all other transport parameters are fixed at their mean values. The base-case flow realization # 284 that produces fast flow velocities is used for case 6a, whereas realization # 610 with the least RMSE is used for case 6b. Similar to the base-case model, the effective fracture porosity value is selected from a lognormal distribution that is used to describe the empirical distribution determined from the numerical analysis conducted using the tracer test results (Reimus *et al*., 2003). The lognormal distribution has a mean of 0.025 and a standard deviation of 0.023. The 90 percent confidence interval ranges between 0.005 and 0.07 (Pohlmann *et al*., 2004).

Figure 15 shows the contaminant boundary results for case 6a and Figure 16 shows the results for case 6b. These figures compare the contribution of the fracture porosity to the model output uncertainty to the contributions of all uncertain flow and transport parameters as incorporated into the base-case model. It can be seen that the fracture porosity contributes significantly to the uncertainty in the contaminant boundary, especially in case 6a with a fast flow realization. In case 6b, however, the fracture porosity does not produce as much uncertainty as it does in case 6a. In both cases, the flow is mainly controlled by one set of fractures that do not change between realizations, as only one flow realization is used and only the fracture porosity is changed.

To better grasp the contribution of fracture porosity (or any other transport parameter) to the overall model uncertainty, one would ideally repeat the simulations conducted here using other “representative” realizations of the base-case model that span the range of all possible flow scenarios. Then one would compute the average uncertainty by, for instance, averaging the curves in subplot d) of Figures 15 and 16 together with all other results based on a sufficiently large number of representative realizations. This, however, is computationally prohibitive.

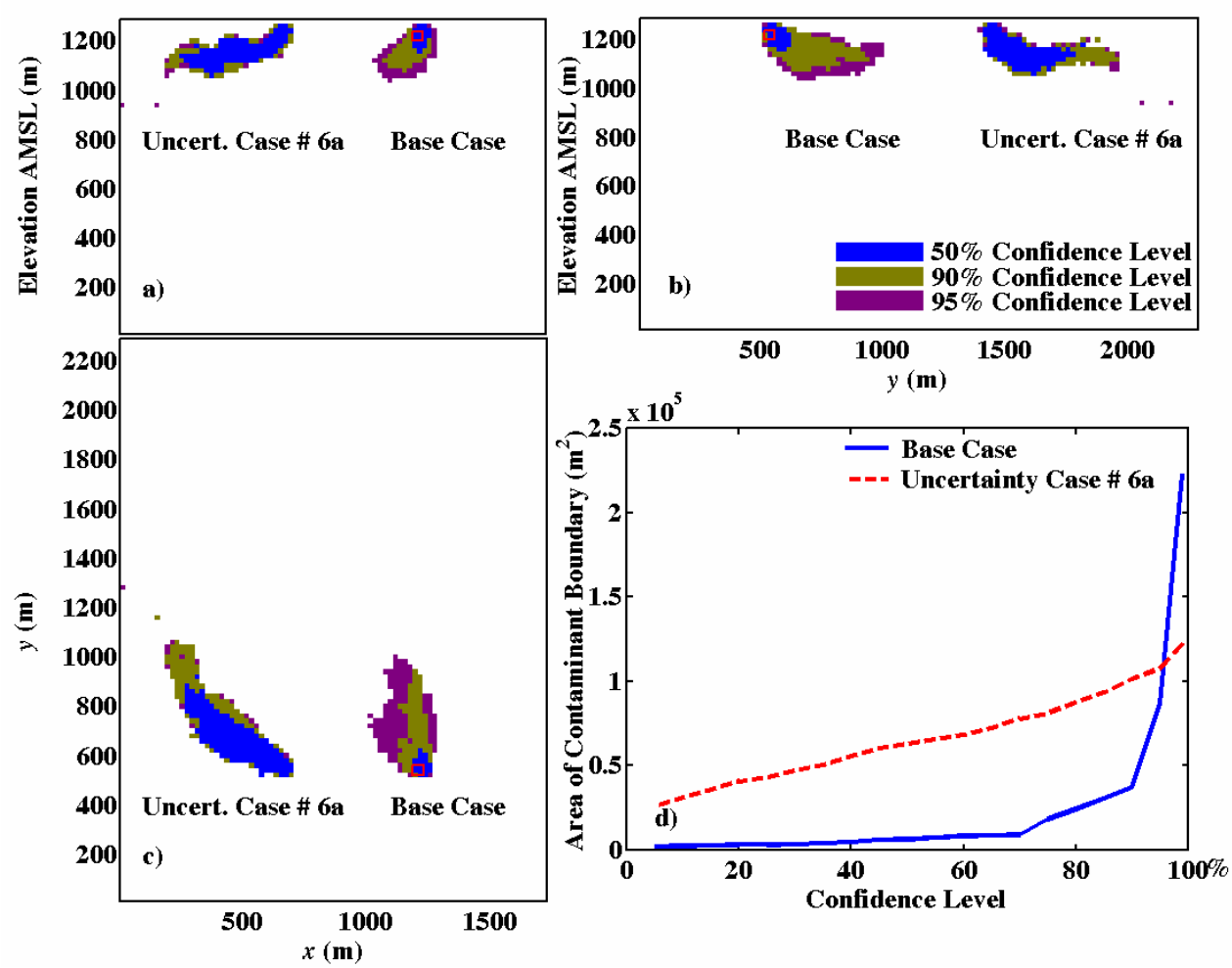

Figure 15. Contaminant boundary maps for ^{14}C delineating the areas exceeding 2,000 pCi/L at different confidence levels. Subplot a) shows the *x*-*z* (or elevation) map, subplot b) shows the *y*-*z* map, subplot c) shows the *x*-*y* map, and subplot d) shows the *x*-*y* contaminant boundary area at different confidence levels. The fracture porosity uncertainty case # 6a is drawn to scale, but is spatially shifted on the plots for comparison to the base-case model (Pohlmann *et al*., 2004).

Field characterization efforts in 1999 and 2000 provide additional information on fracture porosity as a validation target. A long-term, sustained and substantial effort was expanded through a two-well tracer test to provide information on fracture porosity andf transport behavior at Shoal (Reimus *et al*., 2003). This effort substantially reduced prior uncertainties in porosity, but similar reduction is unlikely in future efforts due to uncertainties in field configuration (orientation and contribution of fractures to the test) and approximations required in the analysis. Also, given that any tracer test in a fractured medium will likely span a short travel distance compared to the migration distances simulated for the 1,000-year regulatory time frame, representative fracture porosity on a large scale is very difficult to impossible to obtain. Thus, given the uncertainty of the measurement, length of time to obtain, and large cost, fracture porosity is not an optimum validation target. However, other field measurements (e.g., ^{3}H or ^{14}C concentration measurements) may indirectly help reduce the range of plausible values of fracture

porosity. These measurements may confirm the predicted presence or absence of these elements in certain locations, thereby indicating what velocity range may be plausible for such findings.

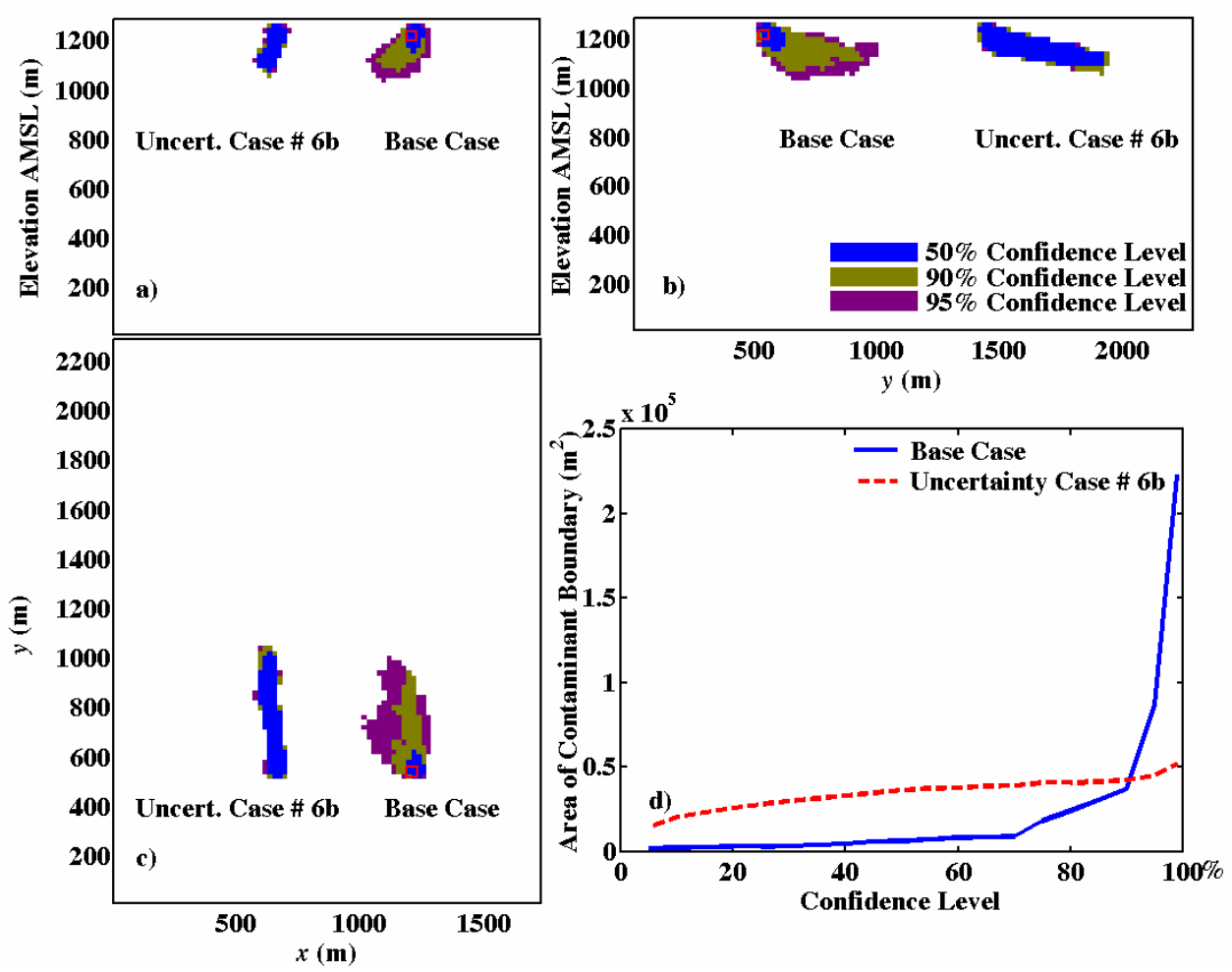

Figure 16. Contaminant boundary maps for ^{14}C delineating the areas exceeding 2,000 pCi/L at different confidence levels. Subplot a) shows the *x*-*z* (or elevation) map, subplot b) shows the *y*-*z* map, subplot c) shows the *x*-*y* map, and subplot d) shows the *x*-*y* contaminant boundary area at different confidence levels. The fracture porosity uncertainty case # 6b is compared to the base-case model (Pohlmann *et al*., 2004).

Cases 7a and 7b

The cavity porosity is the uncertain parameter for this case. The Shoal nuclear test created a cavity that collapsed and formed a rubble chimney. The chimney did not propagate to the land surface (i.e., there is no collapse crater). The cavity radius is reported to be 26 m (Hazelton-Nuclear Science, 1965). The top of the chimney is located 108.5 m above the test location. Borg *et al*. (1976) reported a range for cavity and chimney porosity of 18 to 35 percent

for competent rocks such as granite, basalt, and indurated tuffs. In the base case transport model, uncertainty is incorporated using Borg's values as endpoints for a uniform distribution of porosity in the cavity and chimney. Thus, the porosity value is randomly selected from this uniform distribution and assigned to the eight model cells representing the cavity, and the overlying 16 cells that represent the chimney through the water table (Pohlmann *et al*., 2004).

Using the same flow realizations as the previous case, the results of the cavity porosity uncertainty was obtained, as shown in Figures 17 and 18. The two figures indicate that minor differences exist between the contaminant boundaries at the 50 percent, 90 percent, and 95 percent confidence level. This leads to the conclusion that the cavity porosity uncertainty does not add to the model's overall uncertainty. Again, because of the locality of the cavity and chimney, the uncertainty in their porosity value does not impact far field transport as does other parameters such as conductivity and fracture porosity. This parameter is thus not a major contributor to the output uncertainty and thus cannot be considered as a validation target.

Cases 8a and 8b

The next spherical zone around the cavity represents the damaged zone of highly disturbed rock that is assigned a range of hydraulic conductivity roughly an order of magnitude higher than the highest end of the range assigned to the fractured granite. Consistent with this conceptualization of a damaged zone, the porosity of these cells are assigned random values from a uniform distribution having endpoints of 0.07 to 0.18, substantially higher than the undisturbed rock, but lower than that of the cavity (Pohlmann *et al*., 2004).

The uncertainty analysis for the damaged zone porosity is performed on the same two flow realizations as before. The results of the 500 realizations are summarized in Figures 19 and 20. Similar to the cavity porosity case, the damaged zone porosity contributes very little to the overall uncertainty of the model. However, the damaged zone porosity impacts the contaminant boundary uncertainty more than does the cavity porosity. These results added to the fracture porosity uncertainty (Case # 6) are consistent with the conceptual model of Shoal. The gradual transition from the cavity outward is reflected in a gradual increase in the porosity contribution to the model uncertainty as one moves from the cavity/chimney area to the damaged zone to the undisturbed rock. Once again, none of these three porosity values could be better characterized in the field and thus are not considered as targets for the validation process.

Cases 9a and 9b

The transport simulations for the base-case Shoal model included the process of matrix diffusion. The process was incorporated into the random-walk particle-tracking (RWPT) method using a particle transfer approach. Although similar transport models for the CNTA site utilized the particle transfer approach developed by Liu *et al*. (2000), more accurate methods have been developed recently and were used in the Shoal transport model. These methods were based on the studies of Hassan (2002), Liu *et al*. (2002), Hassan and Mohamed (2003), Pan and Bodvarsson (2002), and Pan *et al.* (2001).

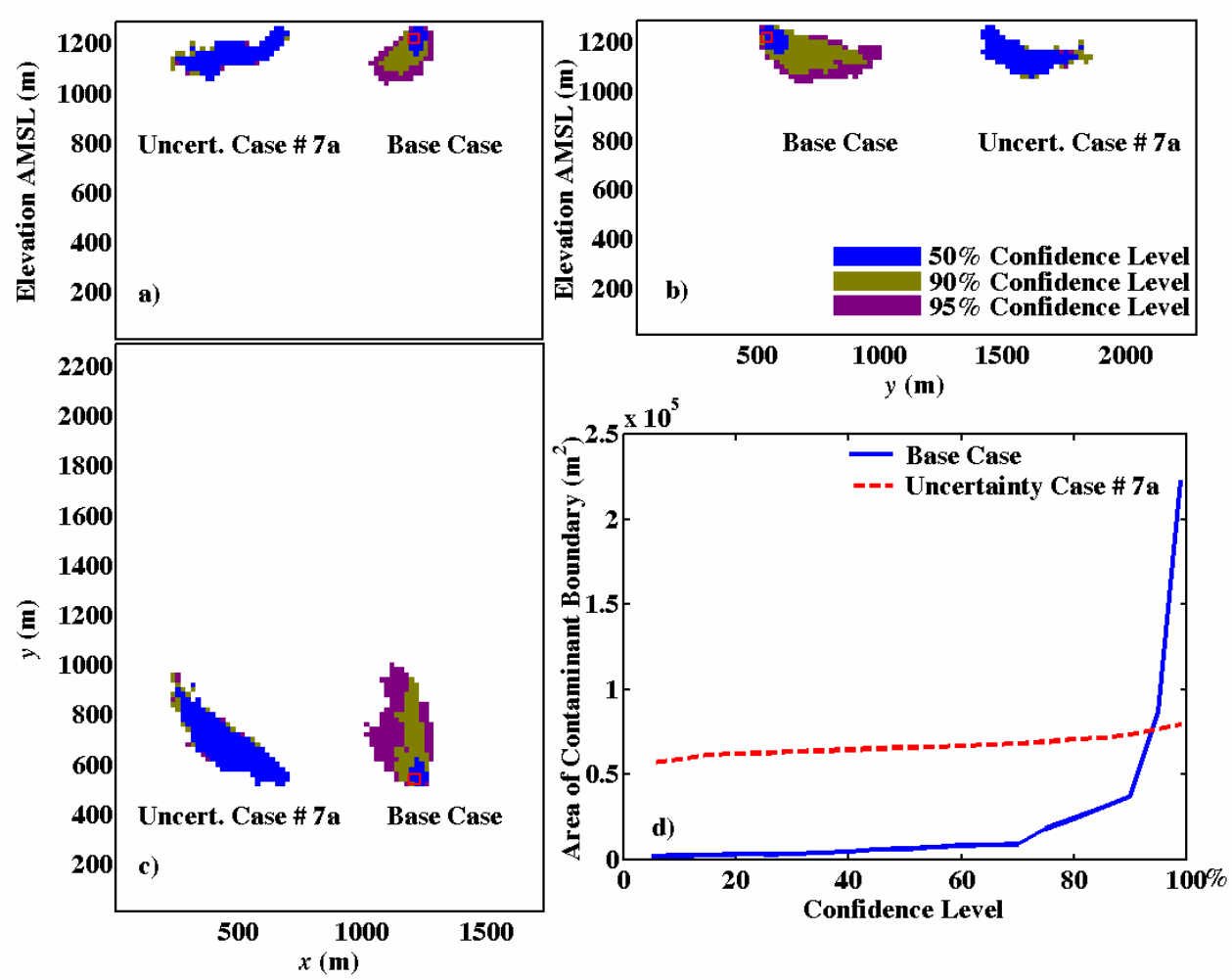

Figure 17. Contaminant boundary maps for ^{14}C delineating the areas exceeding 2,000 pCi/L at different confidence levels. Subplot a) shows the *x*-*z* (or elevation) map, subplot b) shows the *y*-*z* map, subplot c) shows the *x*-*y* map, and subplot d) shows the *x*-*y* contaminant boundary area at different confidence levels. The cavity porosity uncertainty case # 7a is compared to the base-case model (Pohlmann *et al*., 2004).

In the base-case model, Pohlmann *et al*. (2004) adopted a Markov chain model to simulate the particle transfer between the fracture and matrix waters. This Markov chain model is extended to simulate the particle transfer across the fracture/matrix interface by coupling the "active diffusion range" as developed by Pan and Bodvarsson (2002) into the model. The advantage of this approach is that it allows a relatively large time interval (Δt), which dramatically reduces the computation time. The model also allows for multiple fractures within a single grid cell, which effectively increases the effective fracture-matrix interface area. As mentioned earlier, the different parameters affecting the transfer probability were obtained by calibration to the Shoal tracer test data and then kept constant in all realizations.

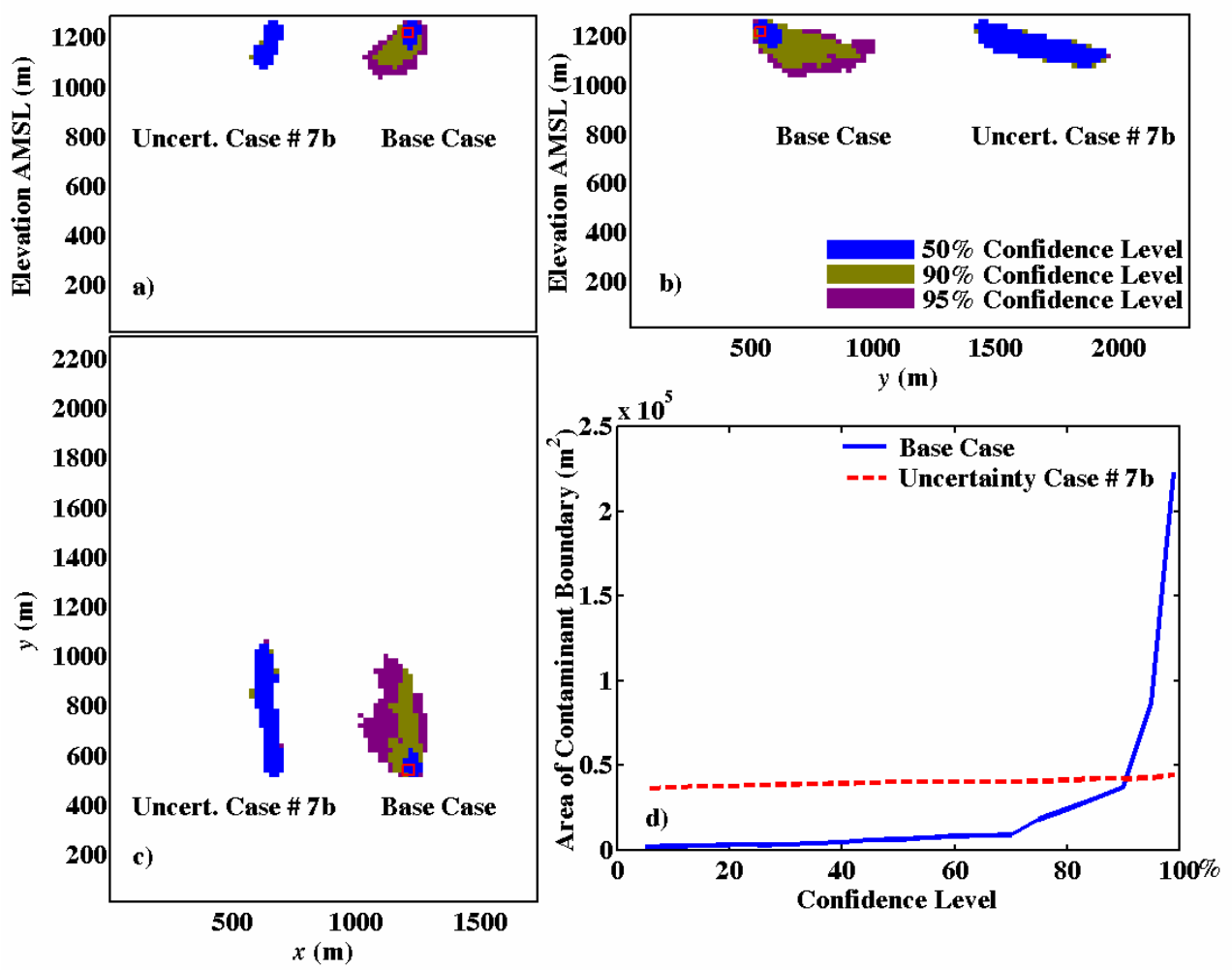

Figure 18. Contaminant boundary maps for ^{14}C delineating the areas exceeding 2,000 pCi/L at different confidence levels. Subplot a) shows the *x*-*z* (or elevation) map, subplot b) shows the *y*-*z* map, subplot c) shows the *x*-*y* map, and subplot d) shows the *x*-*y* contaminant boundary area at different confidence levels. The cavity porosity uncertainty case # 7b is drawn to scale, but is spatially shifted on the plots for comparison to the base-case model (Pohlmann *et al*., 2004).

Here, the impact of uncertainty stemming from matrix diffusion parameters is evaluated. The fracture spacing parameter is selected as the uncertain parameter but any other parameter could also be used for this purpose. A lognormal distribution with mean -0.938 and a standard deviation of 0.7 is used for the fracture spacing, which yields values for this parameter with a mean of 0.5 m (equivalent to the deterministic values used in the base case) and a range of about 0.05 to 2.5 m. This range leads to uncertainty in the transfer probability and, in turn, it changes the strength of matrix diffusion from one realization to another. Figures 21 and 22 show the results of this uncertainty case.

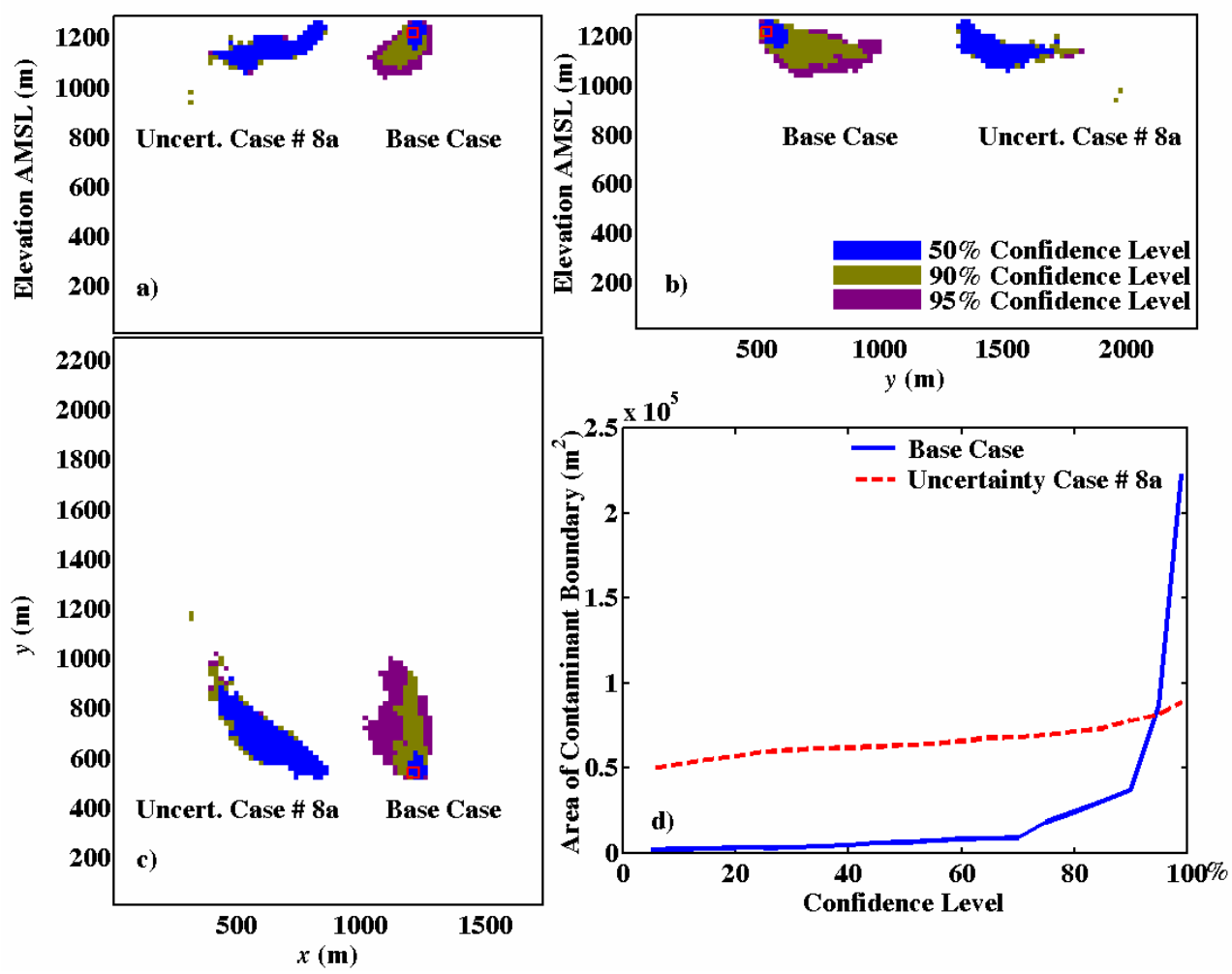

Figure 19. Contaminant boundary maps for ^{14}C delineating the areas exceeding 2,000 pCi/L at different confidence levels. Subplot a) shows the *x*-*z* (or elevation) map, subplot b) shows the *y*-*z* map, subplot c) shows the *x*-*y* map, and subplot d) shows the *x*-*y* contaminant boundary area at different confidence levels. The damaged zone porosity uncertainty case # 8a is drawn to scale, but is spatially shifted on the plots for comparison to the base-case model (Pohlmann *et al.*, 2004).

Case 9a shows that the effect of uncertainty in matrix diffusion is enhanced by the set of strong fractures having high velocities, as was the case for previous parameters (i.e., cases 6a, 7a, and 8a). The results, however, exhibit less sensitivity to matrix diffusion compared to other transport parameters such as damaged zone porosity and fracture porosity. Case 9b, which used the flow realizations having the least RMSE, shows very little sensitivity to the matrix diffusion parameter. So, given these results and the fact that matrix diffusion parameters were obtained by calibration to the tracer test data, the matrix diffusion parameter would not be considered as a useful validation target.

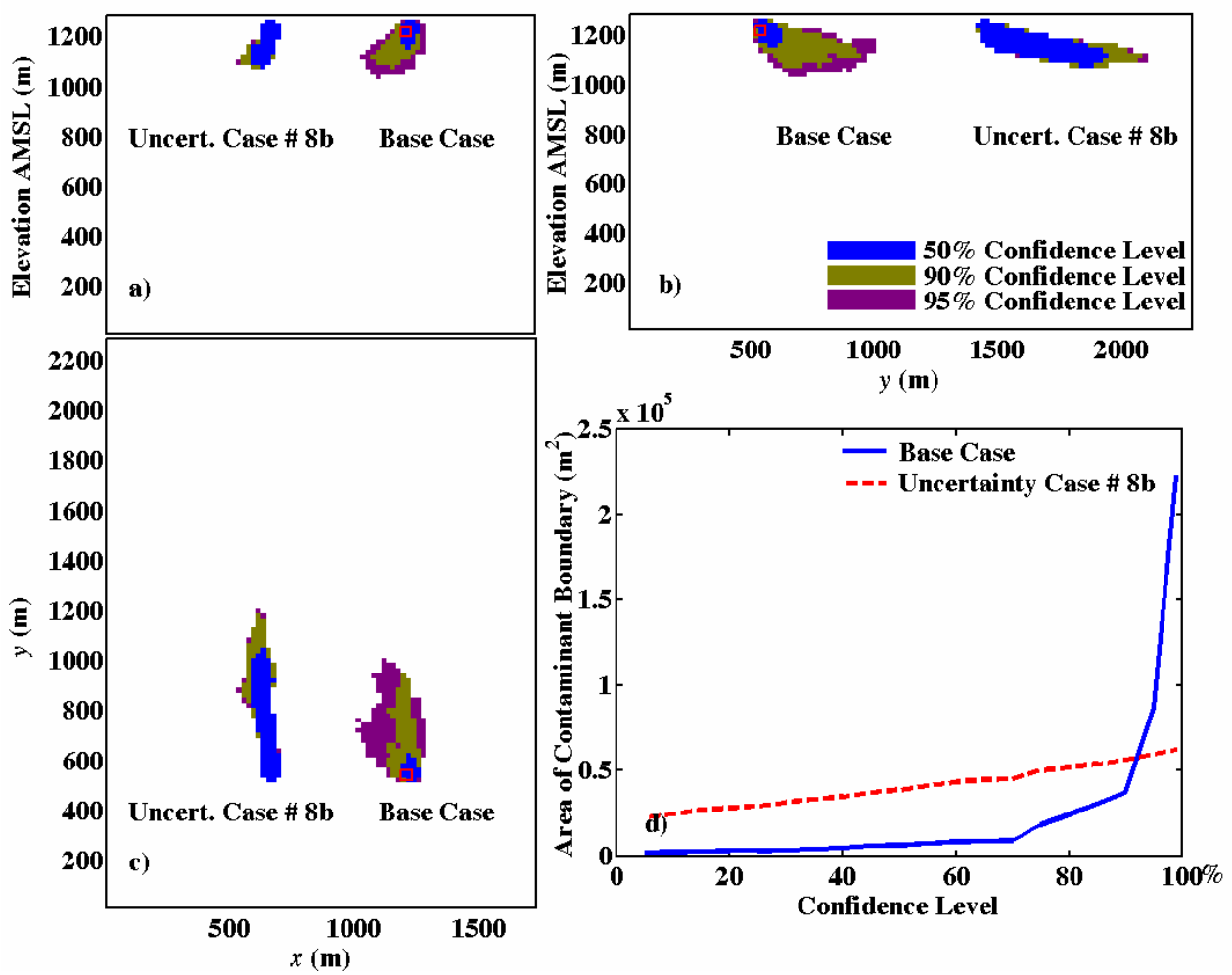

Figure 20. Contaminant boundary maps for ^{14}C delineating the areas exceeding 2,000 pCi/L at different confidence levels. Subplot a) shows the *x*-*z* (or elevation) map, subplot b) shows the *y*-*z* map, subplot c) shows the *x*-*y* map, and subplot d) shows the *x*-*y* contaminant boundary area at different confidence levels. The damaged zone porosity uncertainty case # 8b is drawn to scale, but is spatially shifted on the plots for comparison to the base-case model (Pohlmann *et al*., 2004).

5.2.3 Selection of Validation Targets

Tsang (1987) highlights the importance of the choice of the measurable quantities that are to be used for validation purposes, as there are measurable quantities that are almost impossible to use for model validation (e.g., point and instantaneous concentration data). The averaged solute concentration over a large region and over a period of time is a more relevant quantity for certain purposes such as determining the effectiveness of geological isolation of nuclear or toxic waste (Tsang, 1987).

Most of the model components (conceptual model, mathematical model, computer code, and input data) contain some degree of uncertainty due to lack of perfect knowledge about the subsurface conditions no matter how well the system is characterized. Furthermore, experimental results (e.g., field measurements) that are designed for model validation studies contain some

errors or uncertainty. The validation tests should consider these sources of uncertainty, they make it difficult to ascertain whether or not the model results agree with the experimental data; the more uncertain the data are, the more difficult it is to conclude that the model is acceptable (Davis *et al*., 1991). However, these uncertainty effects should be viewed in terms of whether or not they affect the quantity of regulatory interest. In some cases, input uncertainty may have minor impact on the resulting regulatory quantity such as the size and location of the water volume having contaminant concentration exceeding a certain threshold (Pohll and Mihevc, 2000). These effects should therefore be carefully studied prior to designing the validation study, which was the subject of the analysis presented in this section.

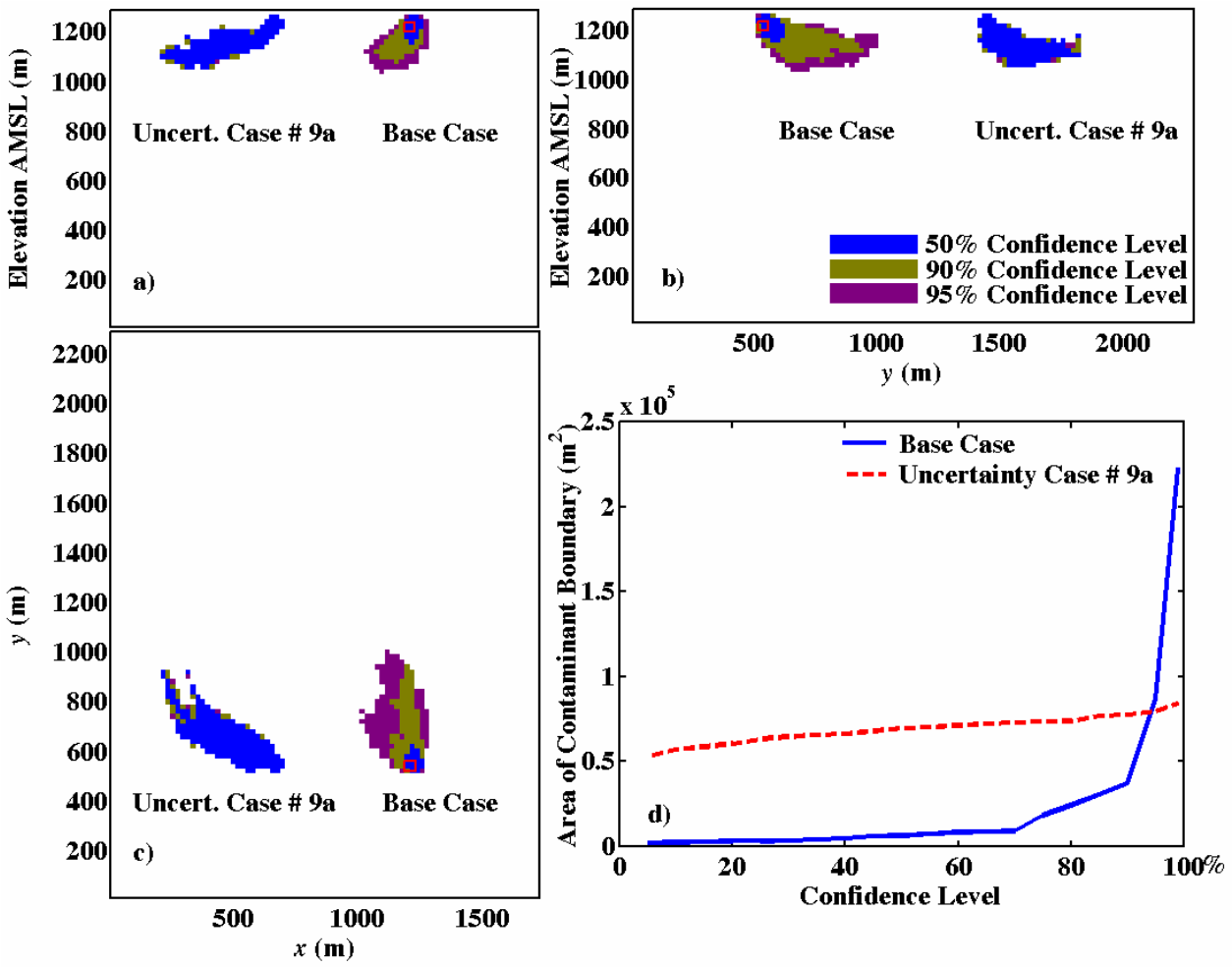

Figure 21. Contaminant boundary maps for ^{14}C delineating the areas exceeding 2,000 pCi/L at different confidence levels. Subplot a) shows the *x*-*z* (or elevation) map, subplot b) shows the *y*-*z* map, subplot c) shows the *x*-*y* map, and subplot d) shows the *x*-*y* contaminant boundary area at different confidence levels. The matrix diffusion uncertainty case # 9a is drawn to scale, but is spatially shifted on the plots for comparison to the base-case model (Pohlmann *et al*., 2004).

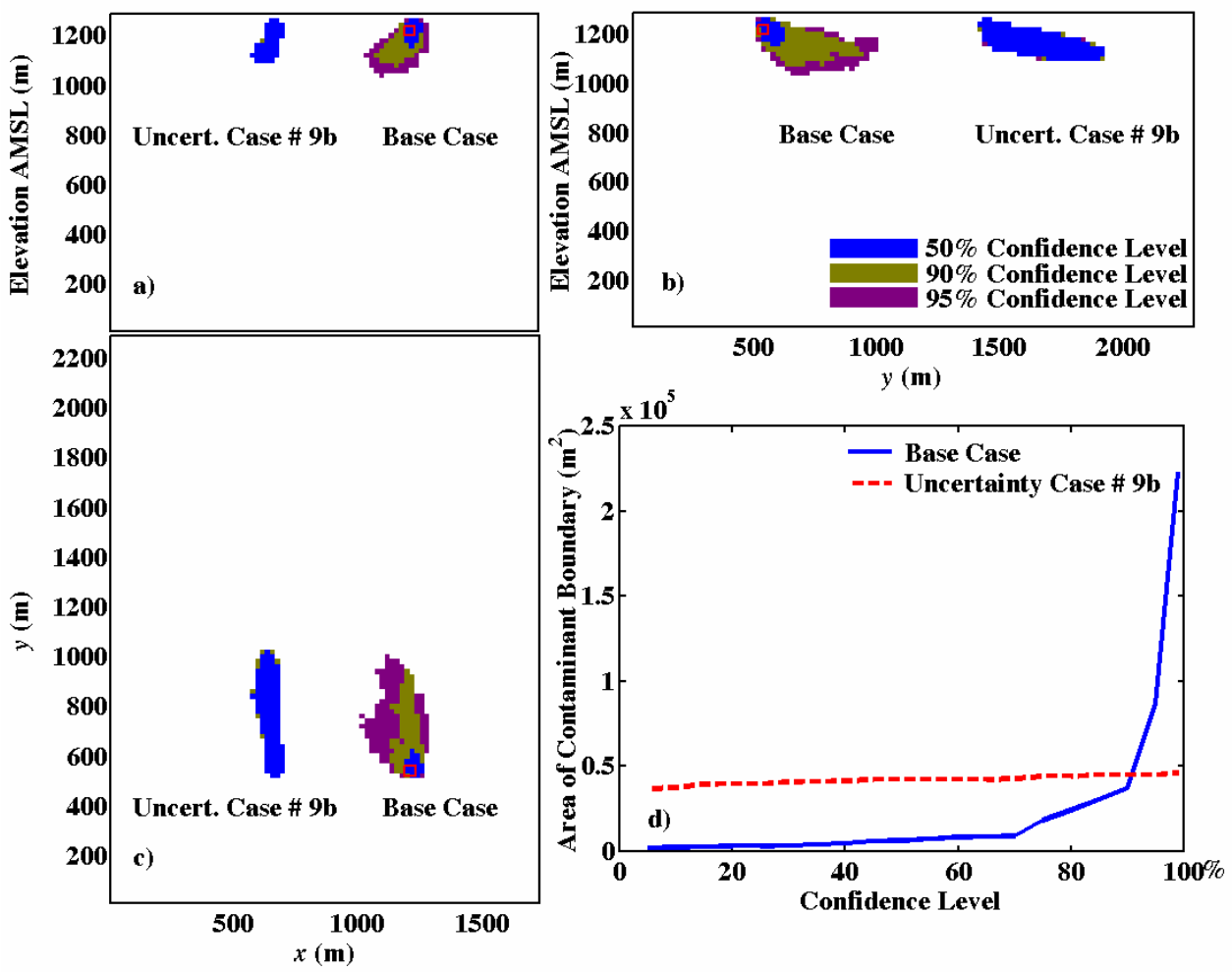

Figure 22. Contaminant boundary maps for ^{14}C delineating the areas exceeding 2,000 pCi/L at different confidence levels. Subplot a) shows the *x*-*z* (or elevation) map, subplot b) shows the *y*-*z* map, subplot c) shows the *x*-*y* map, and subplot d) shows the *x*-*y* contaminant boundary area at different confidence levels. The matrix diffusion uncertainty case # 9b is drawn to scale, but is spatially shifted on the plots for comparison to the base-case model (Pohlmann *et al*., 2004).

Based on the results of the individual parametric uncertainty analysis presented above, hydraulic conductivity is the most viable validation target on the input side of the model. Fracture porosity is a similarly sensitive target, but obtaining porosity data is quite difficult and the data remain subject to large uncertainties due to both test constraints and complexity of interpretation. On the output side, the hydraulic head as well as the head gradient are viable validation targets. In addition, presence or absence of certain radionuclides at the locations of the validation wells will be used as validation targets. However, point concentration measurements may not be helpful for the validation as much as the determination of the presence or absence of radionuclides (i.e., the binary aspect of radionuclide presence as opposed to the value of their concentration). Also, if geophysical logging or other information could be gathered about the fracture sizes and intensity as a function of depth in each of the new validation wells, they can be used as validation targets for the purpose of conditioning the model and reducing the uncertainty built into the fracture characteristics in the model.

One should not also eliminate the possibility to obtain new data that would help in narrowing down the range of recharge values used in the model. For example, temperature logs in the new wells to be drilled for validation can be utilized to derive independent values of recharge and thus can help validate and subsequently constrain the distribution of recharge used in the model. Although recharge did not impact the model output dramatically, the constraints on its impact imposed by changing the conductivity of Flow Category 1 (intervening zones of small random fractures) during the calibration process contribute to this result. With this type of correlation, an independent recharge value can be used to validate the range of selected recharge and K_1 values as produced by the calibrated flow model.

5.3 Acceptance Criteria

Davis *et al.* (1991) discuss some acceptance criteria when validating performance assessment models. To declare that a model is acceptable or adequate for a specific regulatory requirement, the model structure, as well as the model input data, has to be acceptable. Model structure should reflect how the real system behaves. All assumptions inherent in the conceptual model should be justified using site-specific information and field data collected for validation purposes. Accepting the model structure implies that the model results will exhibit a system behavior that is independent of the input data used. That is to say that changing the input data for a structurally accepted model only changes the output results in a quantitative sense but not in a qualitative sense.

To declare that the model input is adequate, one has to build confidence in the model over a wide range of experimental conditions. That is, by changing the conditions under which the laboratory or field validation experiment is performed (e.g., different flow or pumping rates), the model predictions can be compared to a wide range of input conditions that will help build confidence in model input. When changing experimental conditions and thus some portions of the input data for the model, the adequate model should predict the experimental results with a reasonable accuracy without changing other input data. If other input data are correlated to those changing conditions, then the model input should reflect this type of correlation to accept the model input and declare the model not invalid.

5.3.1 Proposed Acceptance Criteria

According to the validation plan shown in Figure 4, the first set of analyses using the field data collected for validation purposes will yield results that will be evaluated to determine the path forward. The first “if” statement in the validation approach pertains to whether a sufficient number of realizations attained satisfactory scores on how they represent the field data used for calibration (old) and for validation (new). The determination of whether sufficient number exists will be based on five criteria with the decision made in a hierarchical manner as will be discussed later. The five criteria are summarized below.

1. Individual realization scores (S_j, j = 1, …, number of realizations), obtained based on how well each realization fits the validation data, will be evaluated. The first criterion then becomes the percentage of these scores, P_1, that exceeds a certain reference value.
2. The number of validation targets where field data fit within the inner 95 percent of the probability density function (pdf) of these targets as used in the model (P_2) is the second criterion.

3. The results of hypothesis testing to be conducted using the stochastic perturbation approach of Luis and McLaughlin (1992) as described in detail in the validation report of Hassan (2003a) (P_3).
4. The results of linear regression analysis and other hypothesis testing (e.g., testing error variance based on calibration data and based on validation data) that could be feasible (depending on the size of the data set obtained in the field), P_4.
5. The results of the correlation analysis where the log-conductivity variance is plotted against the head variance for the targeted locations and the resulting plot for the model are compared against the field validation data (P_5).

The hierarchical approach to make the above determination is described by a decision tree. This decision tree for the acceptance of the realizations and for passing the first decision point on the validation approach is shown in Figure 23 below. It begins with evaluating S_j and determining whether the percentage of realizations with scores above the reference value, P_1, is more than 40 percent, between 30 percent and 40 percent, or less than 30 percent. If the number is more than 40 percent, it is deemed sufficient. If it is between 30 percent and 40 percent or less than 30 percent, then the second criterion, P_2, is used as shown in Figure 23.

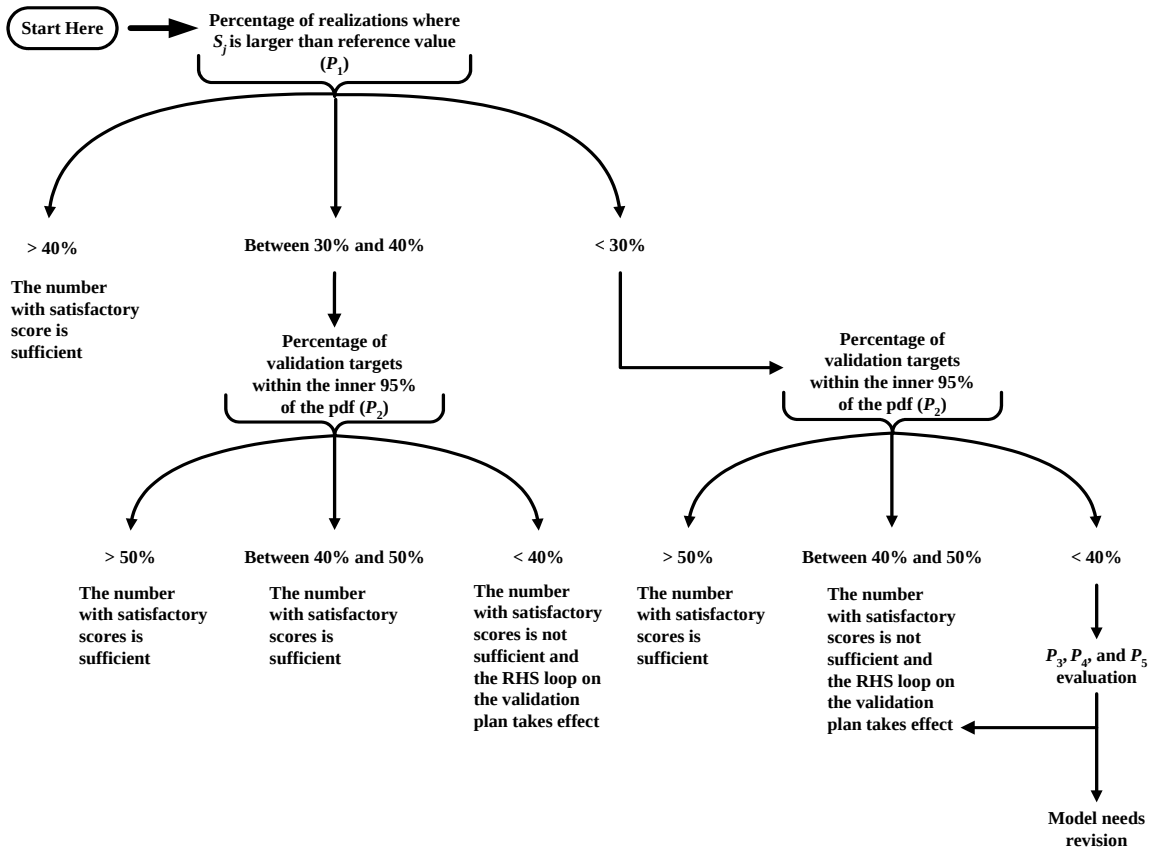

Figure 23. A decision tree chart showing how the first decision (step 6) in the validation plan will be made and the criteria for determining the sufficiency of the number of acceptable realizations.

The second criterion represents the number of validation targets where the field data lie within the inner 95 percent of the pdf for that target as used in (input) or produced by (output) the Shoal model. Then if P_1 is between 30 percent and 40 percent and P_2 is between 40 percent and 50 percent or if P_1 is less than 30 percent but P_2 is greater than 50 percent, the number of realizations is deemed sufficient. If P_1 is less than 30 percent and P_2 is less than 40 percent, then the remaining three measures, P_3, P_4, and P_5, are used to determine whether the model needs revision or whether more realizations can be generated to replace some of the current realizations. In this latter case, it may be that the model is conceptually good but the input parameter distribution is skewed one way or another and by generating more realizations and keeping the ones that fit the above criteria, the distribution attains the proper position. This can be done using the existing model without conditioning or using any of the new validation data (i.e., no additional calibration). The rationale for selecting the above thresholds (30 percent to 40 percent for P_1 and 40 percent to 50 percent for P_2) is described through an example and when these metrics are evaluated with statistical hypothesis testing later in this section.

5.3.2 Single Validation Target Illustration

The first criterion is to compute the number of realizations with scores S_j above a reference value. To demonstrate how this reference value is computed, assume there is only one validation target (e.g., the head measurement in one interval in one well). Figure 24 shows the pdf for this head value as produced by the stochastic Shoal model where the triangles represent the 2.5th, 50th, and 97.5th percentiles and the circle indicates a hypothesized field measurement, h_o. The reference value and the score for any individual realization for this simple case are computed as

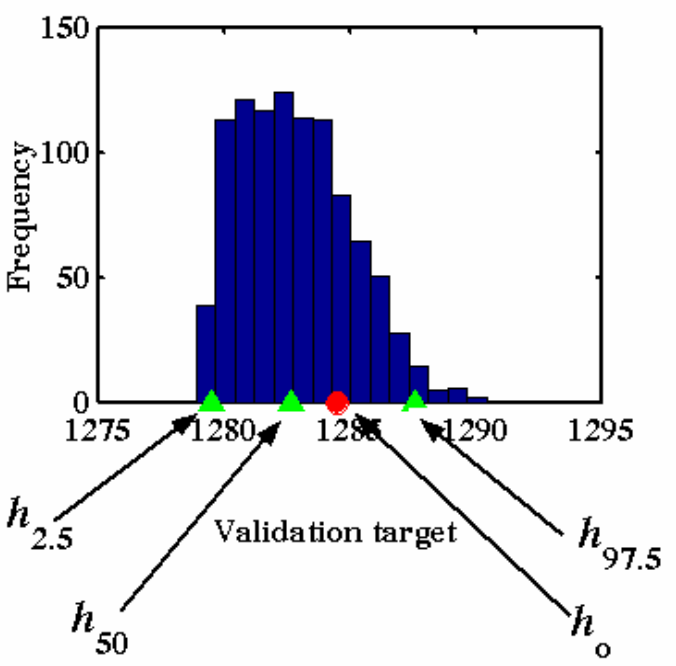

Figure 24. The head distribution (or pdf) as obtained from the Shoal model with the 2.5th, 50th, and 97.5th percentiles shown with the green triangles and the hypothesized field data shown by the red circle.

$$RV = \exp\left[-\frac{(h_o - h_{2.5})^2}{(h_{97.5} - h_{2.5})^2}\right] \quad \text{for } h_o < h_{50}$$
$$RV = \exp\left[-\frac{(h_o - h_{97.5})^2}{(h_{97.5} - h_{2.5})^2}\right] \quad \text{for } h_o > h_{50} \tag{1}$$

$$\text{Realization Score } (S_j) = \exp\left[-\frac{(h_o - h_j)^2}{(h_{97.5} - h_{2.5})^2}\right] \quad \text{for } j = 1, \ldots, NMC \tag{2}$$

$$\text{First Criterion } (P_1) = \frac{\#\text{ of Realizations where } S_j > RV}{NMC} \tag{3}$$

where j is the realization index and it varies from 1 to NMC (number of Monte Carlo realizations) with NMC being 1,000 realizations for the Shoal model. This leads to all realizations with absolute errors smaller than ($|h_o - h_{2.5}|$) or ($|h_o - h_{97.5}|$), whichever is smaller, attaining a score higher than the reference value. Figure 25 below shows the resulting scores and how they compare to the reference value, RV, as obtained from the above equations.

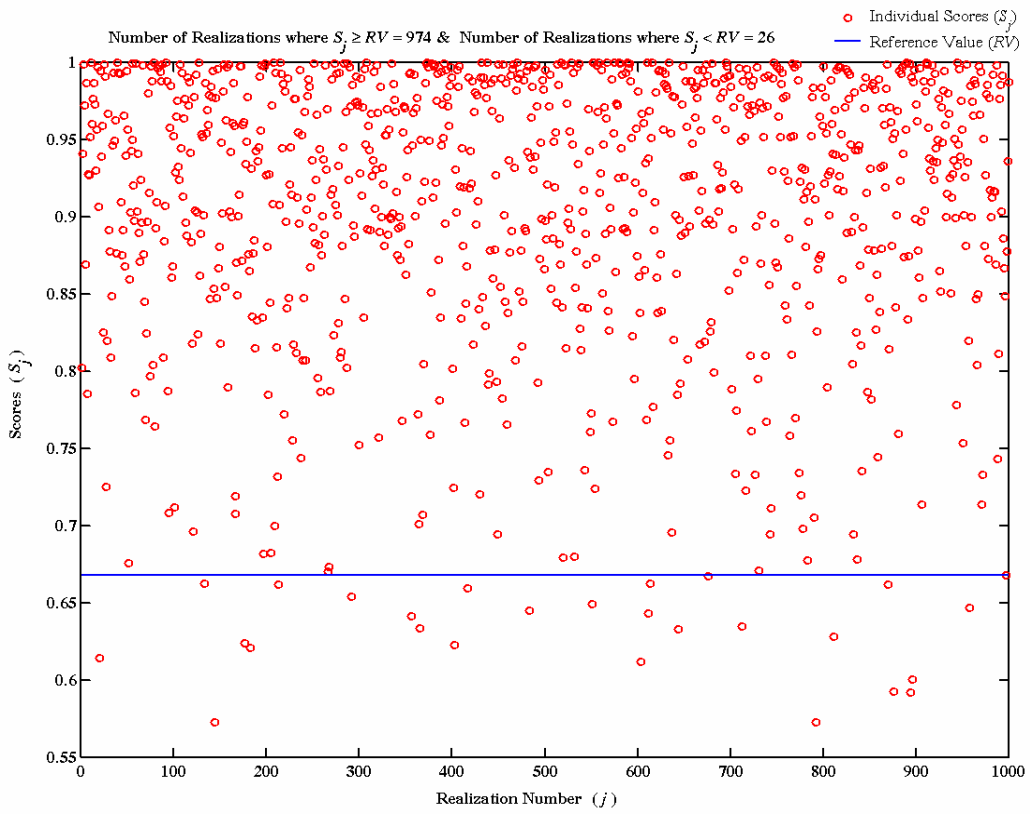

Figure 25. Realizations scores, S_j, relative to the Reference Value, RV, for the single validation target case presented in Figure 24. The P_1 value here is 97.4 percent (= 974/1,000).

It can be seen from Equations (1) through (3) that the maximum value that *RV* or S_j can attain is 1.0. Thus, if the observed value, h_o, is equivalent to the 2.5th or the 97.5th value, P_1 becomes zero because *RV* becomes 1.0 and all S_j values will be less than 1.0. Also, if the observed value is found to be less than $h_{2.5}$ or greater than $h_{97.5}$, P_1 will be automatically set to zero. In such cases, one may conclude that the model output is skewed toward higher or lower values than indicated by field data. However, this does not necessarily indicate conceptual problems and it may be an indication of incorrect input parameter distributions. The other tests and evaluations can help identify the reasons for this output skewness. When the measured value coincides with the mean value (or 50th percentile) of the target output, h_{50}, then P_1 will approximately be 95 percent indicating that 95 percent of the realizations attained scores higher than *RV*.

5.3.3 Testing the Efficacy of P_1 for a Single Validation Target

To investigate the P_1 metric for the case of a single validation target, a distribution form is assumed for the model output. For simplicity, it is assumed that the model predictions follow a standard normal distribution with zero mean and unit variance, so h_{50} = 0.0, $h_{2.5}$ = -1.96, and $h_{97.5}$ = 1.96. The performance of this metric is tested for a range of measurement values (hypothesized values for the single field data point) between -10.0 and +10.0. For each one of these hypothesized values, the *RV* can be obtained according to Equation (1) and the results are shown in Figure 26. The *RV* metric decreases rapidly as the observation value approaches the median, h_{50}. When the measured value lies outside the middle 95 percent of the output distribution (i.e., outside the range [-1.96, 1.96]), the *RV* is not computed, since P_1 becomes zero. Also, as shown in the figure, when h_o equals -1.96 ($h_{2.5}$) or 1.96 ($h_{97.5}$), *RV* equals 1.0. Due to the exponential form in Equation (2), all S_j values will be less than 1.0, resulting in a zero value for P_1 when h_o is at the 2.5th or 97.5th percentile.

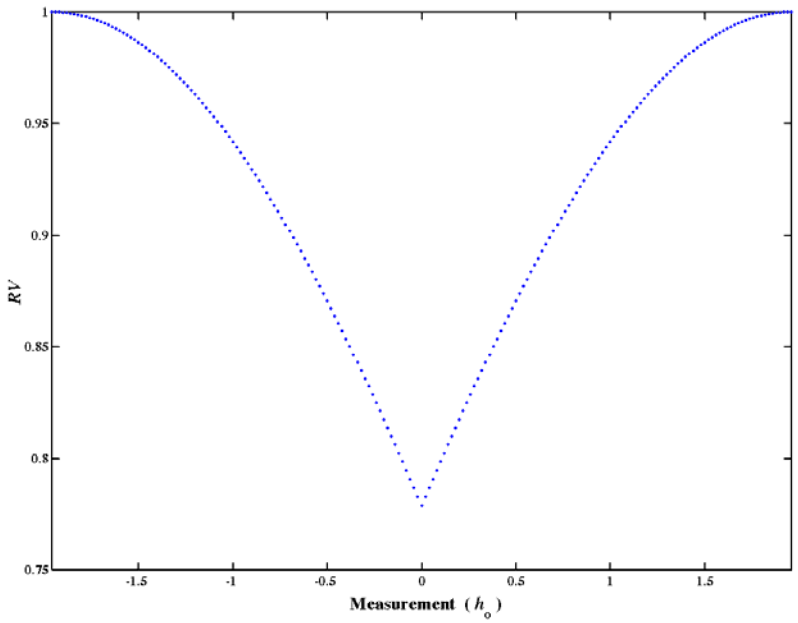

Figure 26. The Reference Value, *RV*, for the single validation target case as a function of the measured field value.

The next step is to calculate the S_j score for each Monte Carlo realization, with S_j being a similar measure to the *RV*, but using individual realization predictions. The S_j score is compared to the *RV* score and the relative number of S_j values that exceed the *RV* are tallied to obtain P_1. The S_j values and the corresponding P_1 value were tallied for a range of single observation values in the range [-10, 10] as shown in Figure 27.

Figure 27 also compares the P_1 metric to the *t*-distribution with one degree of freedom. The *t*-distribution is commonly used to test the statistical differences among means when the variance of the distribution is not known. The distribution plotted with green in the figure simply shows the value of the significance level, α, at which each observation on the range [-10, 10] would be rejected in a hypothesis testing that evaluates the statistical difference between the mean of the model output (assumed standard normal distribution) and each observed value (assuming that each observed value represents a distribution with only one [n = 1] sample). The one-degree of freedom used in this plot is not exactly correct, as the degrees of freedom are actually n - 1 = 0.

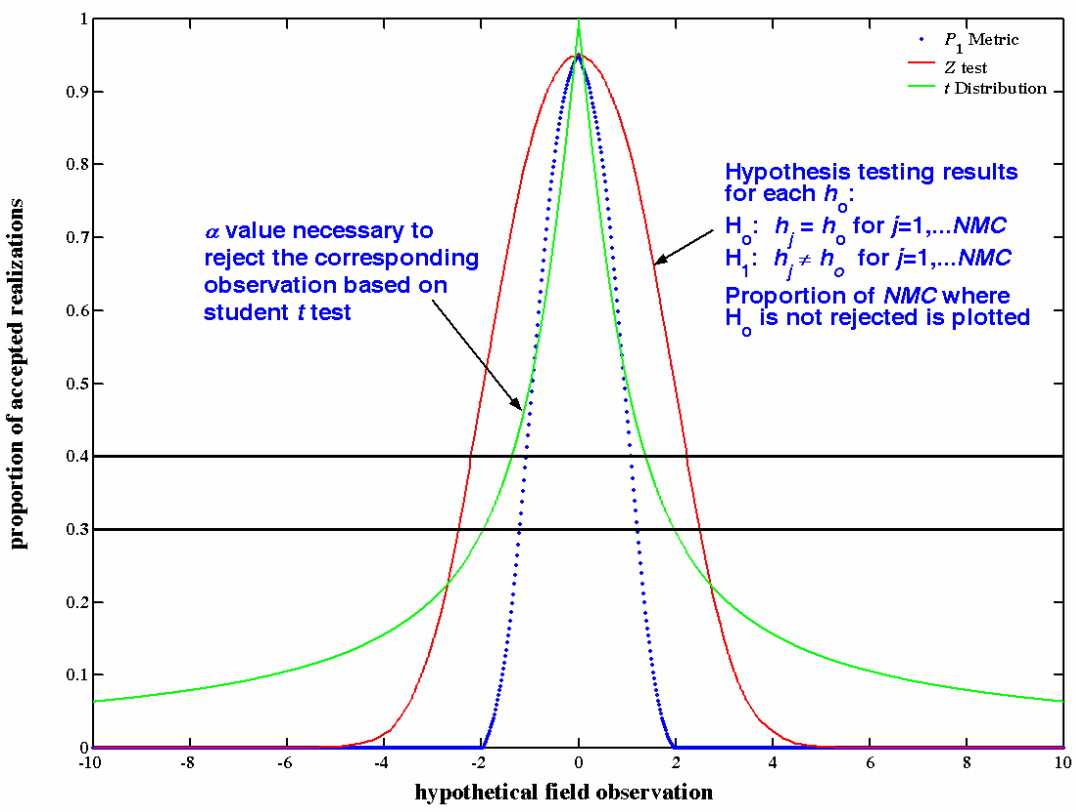

Figure 27. The P_1 metric, student *t*-distribution, and the results of hypothesis testing using the *Z* test.

To avoid this limitation, the *Z* test that is commonly used for the same purpose is employed, but it assumes that the variances of the distributions are known. An assumption is made that each observation is a mean of a normal distribution and each output realization represents a mean of a normal distribution. For each observation value, the following hypothesis is then tested:

$$\begin{aligned} H_0 &: h_j = h_o \qquad for\ \ j = 1,..., NMC \\ H_1 &: h_j \neq h_o \qquad for\ \ j = 1,..., NMC \end{aligned} \tag{4}$$

Then the proportion of Monte Carlo realizations where the null hypothesis, H_0, above is not rejected is plotted against each observation value as shown with the red line in Figure 27.

The plots in Figure 27 provide an indication of how the P_1 test compares against standard statistical tests. According to the figure, one would accept all model realizations for any of the observed values [-10, 10] based on the student *t*-test. In other words, if the *t*-test is used, one would not reject any of the model realizations until approximately the absolute value of the observation is well above 10 (at the 95 percent confidence level). On the other hand, the P_1 measure and the *Z*-test both indicate decreasing proportions of acceptable realizations as one deviates from the median of the model output distribution, which is zero in this test case. At the 5 percent significance level and if the observed value coincides with the median of the model output, only 95 percent of the realizations are deemed acceptable using the P_1 measure and the *Z*-test. When the observed value deviates from the median, the proportion of acceptable realizations drops faster using the P_1 measure compared to the *Z*-test. For example, 40 percent or more of the model realizations would be accepted using the *Z*-test for any observation value in the range [-2.22, 2.22], whereas the P_1 measure gives this level of acceptance for a narrower range of observation values [-1.07, 1.07].

At first glance it appears that the two methods (the P_1 measure versus the *Z*-test or the *t*-test) are in large disagreement. But Type I error (rejecting a model realization when in fact it is a good one) versus Type II error (accepting a poor model realization) must be considered. The P_1 metric is essentially reducing the Type II error at the expense of Type I error. As discussed by Sargent (1990), the probability of Type I error is called model builder's risk, whereas the probability of Type II error is called model user's risk, and in model validation, model user's risk is extremely important and must be kept small. As a result, it is believed that the restrictiveness of the P_1 measure helps minimize Type II error and thus reduces the model user's risk (both DOE and/or NDEP) at the expense of increasing the model builder's risk (supposedly the research team evaluating the model).

5.3.4 Multiple Validation Targets Illustration

For the general case of having *N* validation targets, the above equations should be modified to account for these different validation targets. In this case, the *RV* and the individual scores, S_j, will depend on the sum of squared deviations between each observation, h_o, and the corresponding $h_{2.5}$ or $h_{97.5}$. The equations thus become

$$\text{Reference Value } (RV) = \exp\left(-\sum_{i=1}^{N} \min\left[(h_{o_i} - h_{2.5_i})^2 , (h_{o_i} - h_{97.5_i})^2 \right] / \sum_{i=1}^{N} [h_{97.5_i} - h_{2.5_i}]^2 \right) \tag{5}$$

$$\text{Realization Score } (S_j) = \exp\left(-\sum_{i=1}^{N} [h_{o_i} - h_j]^2 / \sum_{i=1}^{N} [h_{97.5_i} - h_{2.5_i}]^2 \right) \quad \text{for } j = 1,..., NMC \tag{6}$$

For demonstration purposes and as an example, assume the hypothetical case that data are collected on 18 validation targets. These, for example, could be conductivity data in three wells, three measurements each (i.e., 9 intervals) and head data for the same intervals. For each one of these targets, the current stochastic Shoal model provides a distribution of values, as each

realization of the model has different values for these targets than other realizations. It is then assumed that the values of the field data are known (one realization is picked at random to provide an example observation for all targets.) Figures 28 through 30 show the results of this example (Example 1) where P_1 is found to be about 76.7 percent. In this case, P_2 is not checked for and the sufficiency of the number of realizations having acceptable scores is accepted. Note, however, if P_2were checked, it would be about 94 percent (=17/18).

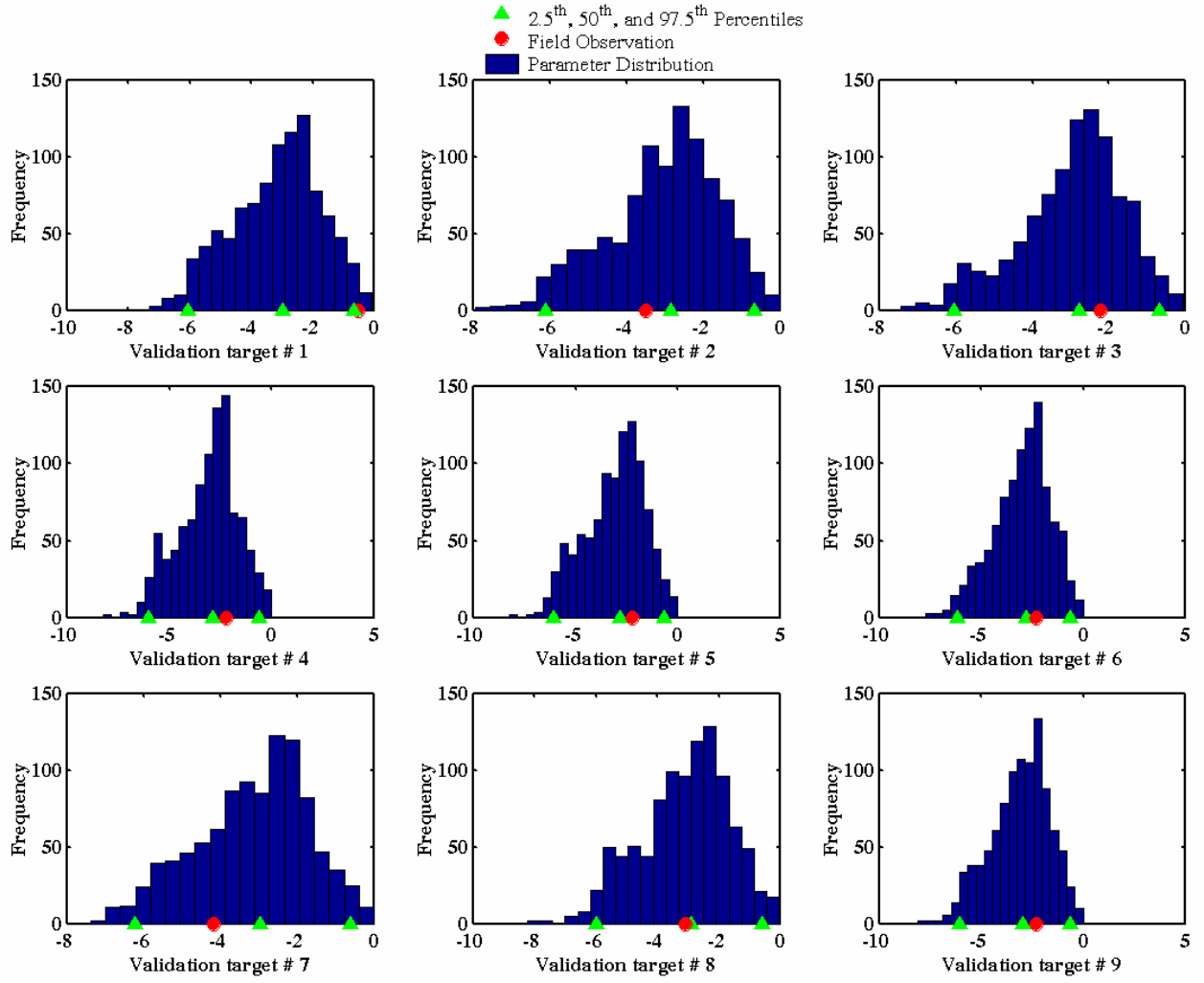

Figure 28. Example 1 showing the pdf distributions for validation targets 1 through 9 with the 2.5th, 50th, and 97.5th percentiles shown with the green triangles and the hypothesized field data shown by the red circles.

Using another set of random values to hypothesize the field data, a different result is obtained as shown in Figures 31 through 33 for Example 2. In this case, both P_1 and P_2 are less than 40 percent (since the number of validation targets where the red circles are between the 2.5th and the 97.5th percentiles is only 2, or approximately 11 percent). In this case, the additional hypothesis tests and linear regression evaluations will be performed to assert whether the model needs to be revised or if the parameter distributions need to be modified.

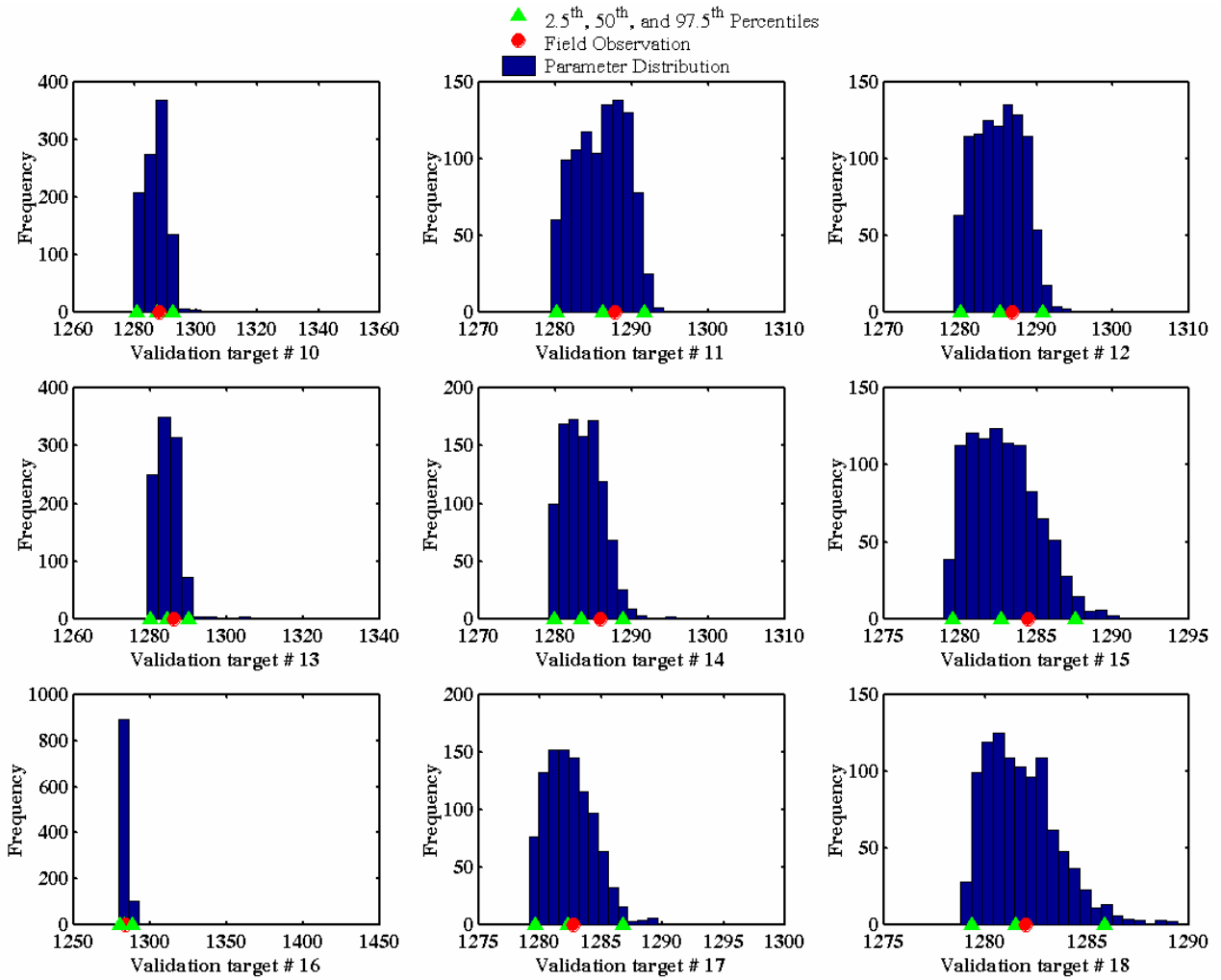

Figure 29. Example 1 showing the pdf distributions for validation targets 10 through 18 with the 2.5th, 50th, and 97.5th percentiles shown with the green triangles and the hypothesized field data shown by the red circles.

In example 1 above, the field data values are hypothesized to be equivalent to one of the model realizations. That is, the values of the 18 validation targets are obtained from one single realization and assumed to represent field data collected for the validation analysis. In spite of assuming field values that exactly match one of the model realizations, the P_1 metric was found to be about 76.7 percent. This value is obviously dependent on which realization is selected. Therefore, the above example was repeated 1,000 times with each of the model realizations assumed to represent the field data in one of those times. The P_1 metric is obtained for these 1,000 experiments and its mean value was found to be about 43 percent. Given that the actual field data to be collected for the validation analysis are very unlikely to exactly match any of the Shoal model realizations, the 30 percent -40 percent threshold for P_1 is considered realistic. In other words, if one, on average, obtains 43 percent for P_1 when one of the model realizations is assumed to match real field conditions, one can safely assume the model conceptually valid if P_1 is between 30 percent and 40 percent when using the actual validation data.

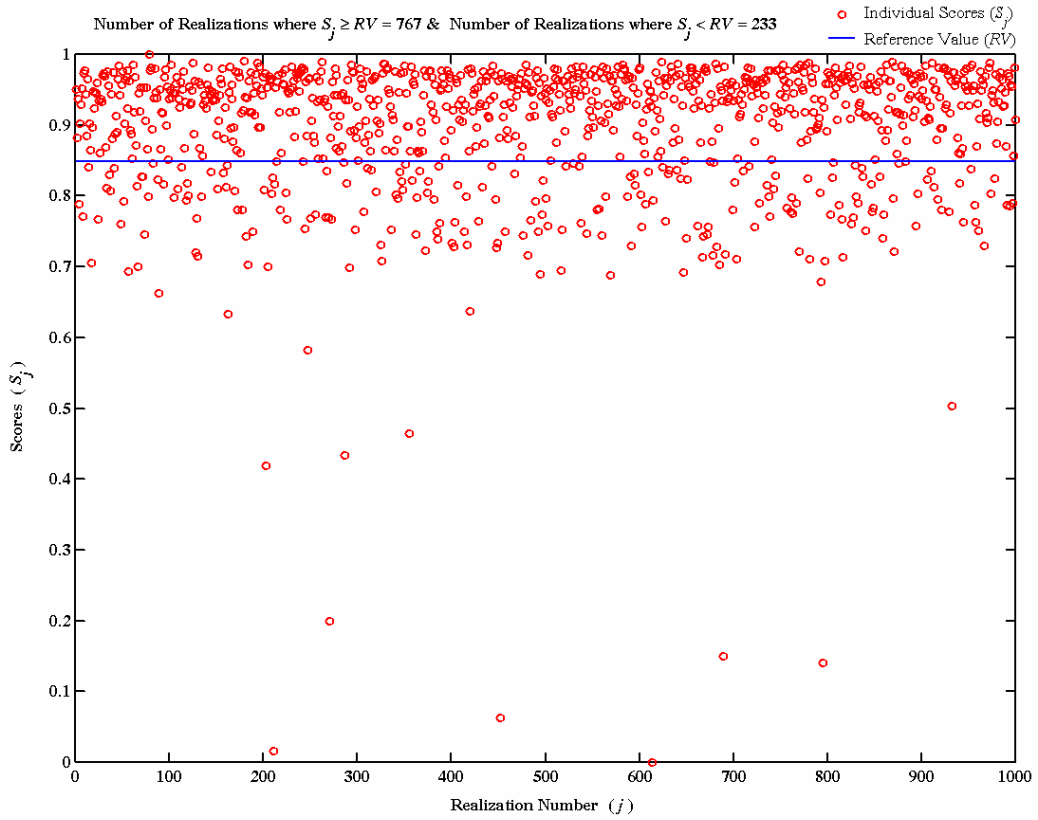

Figure 30. Example 1 showing individual realization scores, S_j, relative to the Reference Value, *RV;* the P_1 value here is about 76.7 percent (= 767/1,000).

5.3.5 Testing the Efficacy of P_1 for Multiple Validation Targets

A numerical experiment is performed to evaluate the P_1 metric for the case of multiple validation targets. The experiment is run as follows:

1. A model is assumed to produce multiple outputs, each following a standard normal distribution with zero mean and unit variance.
2. To test the sensitivity of the P_1 metric, 30 observations are randomly selected, with the mean value of each observation being constant. A range of observation means is used to determine at what point the model will be rejected. The mean of each observation set is tested over the range -4.0 to 4.0 (i.e., -4.0, -3.9,...,4.0).
3. For each mean value, 30 observations are randomly drawn from a normal distribution with the mean equal to the current mean value (i.e., -4.0, -3.9, …, 4.0) and a standard deviation = 1.0.
4. The *RV* value for the 30 validation targets is computed using Equation (5).
5. For each observation mean, the scores S_j for 10,000 realizations of a model (model is assumed to be standard normal) are computed and the metric P_1 is obtained according to Equation (3).
6. Steps 3 through 5 are then repeated for each observation mean in the range [-4.0, 4.0].

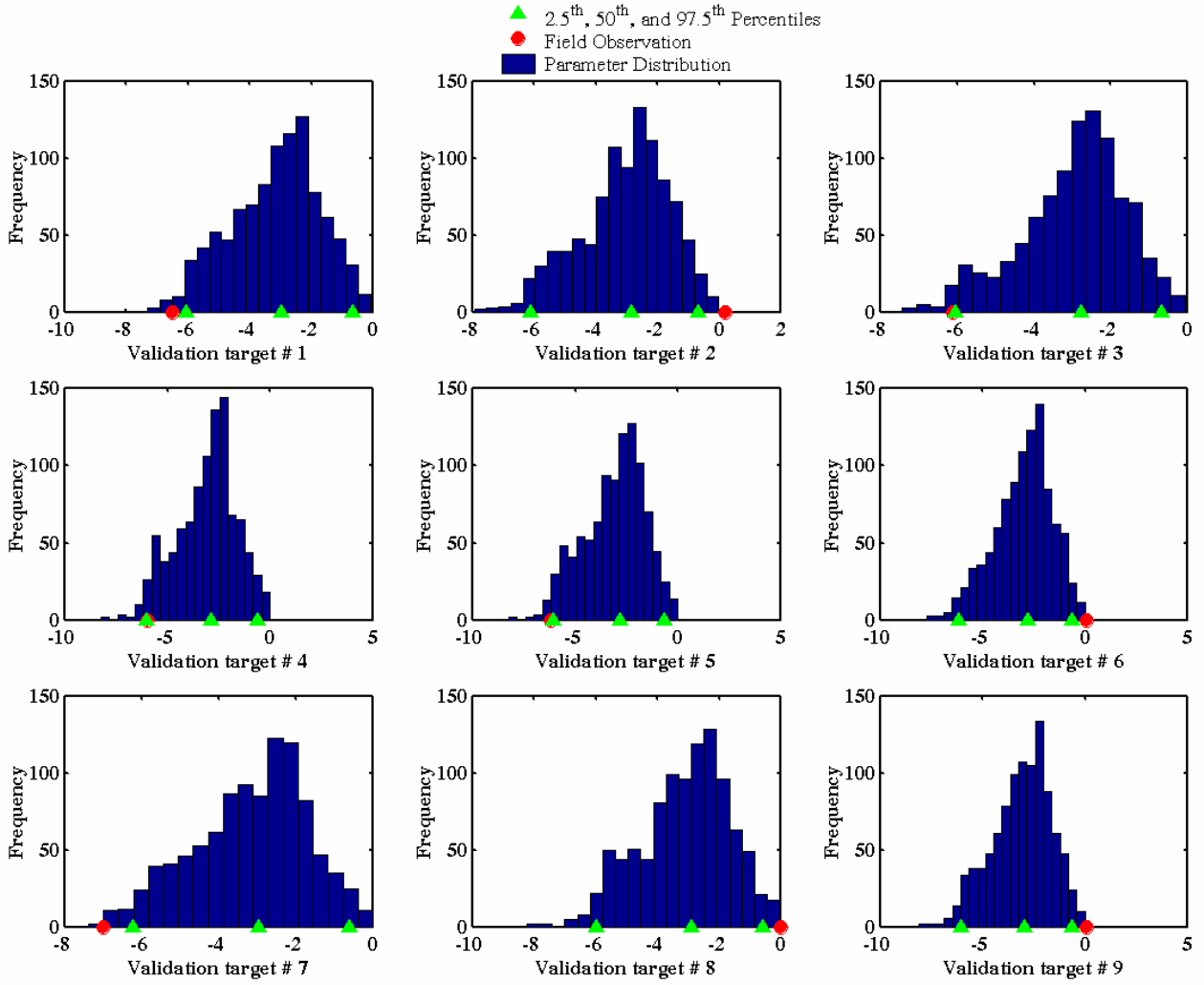

Figure 31. Example 2 showing the pdf distributions for validation targets 1 through 9 with the 2.5th, 50th, and 97.5th percentiles shown with the green triangles and the hypothesized field data shown by the red circles.

The purpose of this experiment is to determine the point at which a model will be considered invalid. Each observation set represents data that are either close to the model predictions (i.e., mean values close to zero), or poor fitting data with mean values far away from zero. This experiment allows us to compare the rejection region for using a simple hypothesis test (i.e., *Z*-test) versus the P_1 measure.

Due to the random nature of the distributions generated in the above procedure, the above experiment was repeated 100 times and the average results are shown in Figure 34. The blue dots in the figure represent the results for the P_1 metric, the red line shows the results of the *Z*-test that is similar to the test conducted for the single validation target case, the magenta line represents the mean value (of 100 values) of the P_1 metric at each observation mean, and the black line represent a normal distribution that best fits the P_1 results.

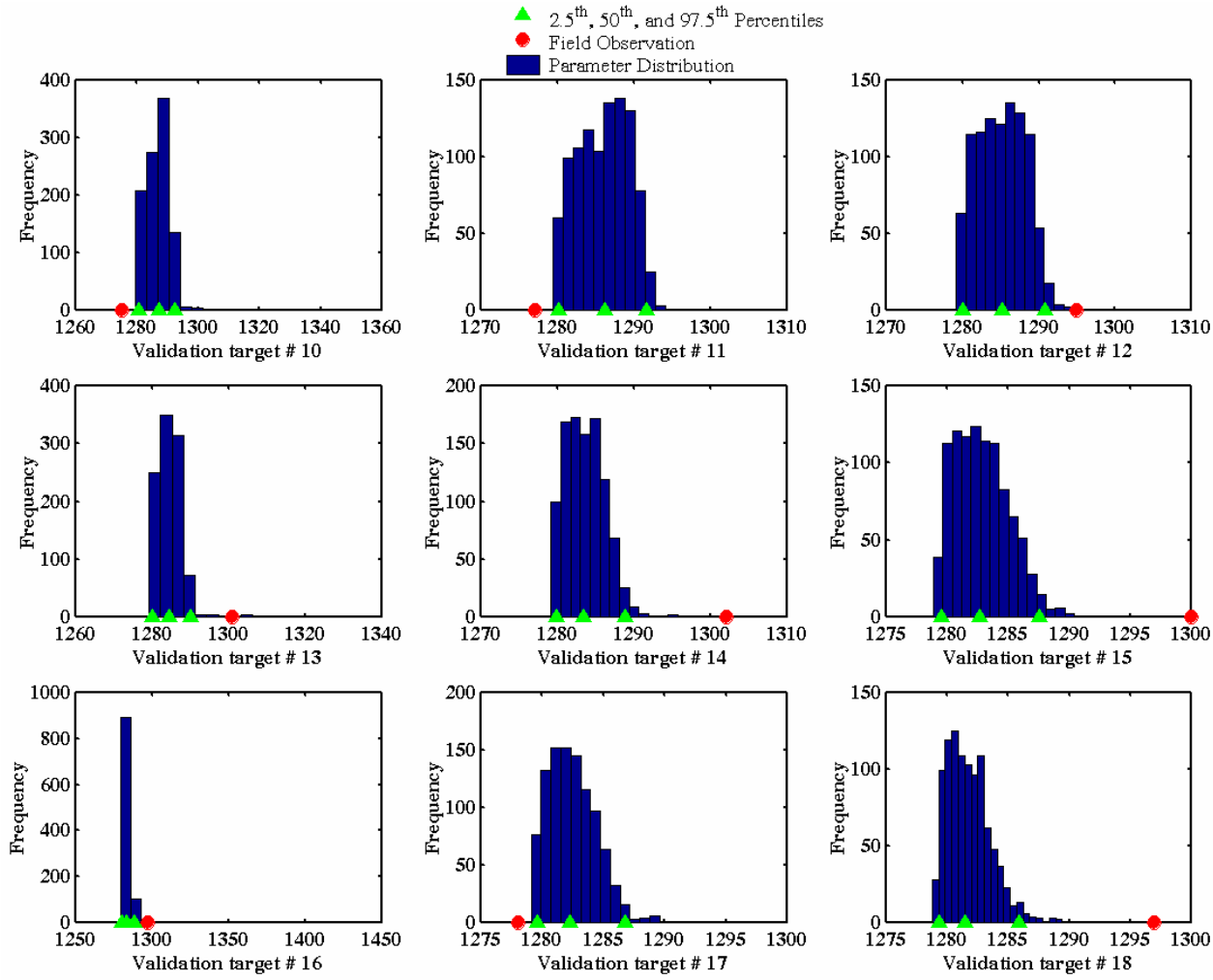

Figure 32. Example 2 showing the pdf distributions for validation targets 10 through 18 with the 2.5th, 50th, and 97.5th percentiles shown with the green triangles and the hypothesized field data shown by the red circles.

For the *Z*-test, each output realization is assumed to represent a mean of a normal distribution. For each observation mean value, the following hypothesis is tested:

$$\begin{aligned} \mathrm{H}_0 &: h_j = h_o \qquad for\ \ j = 1,\ldots,NMC \\ \mathrm{H}_1 &: h_j \neq h_o \qquad for\ \ j = 1,\ldots,NMC \end{aligned} \tag{7}$$

Then, the proportion of Monte Carlo realizations (assumed 10,000 in this experiment) where the null hypothesis, H_0, above is not rejected is plotted against each observation mean as shown with the red line in Figure 34. According to the figure, the *t*-test would suggest that all model realizations are accepted if the mean value of the observations was inside the range [-2.2, 2.2] at 95 percent. The P_1 criterion has a narrower acceptance region ([-1.6, 1.6] according to the

black or magenta line), again suggesting that the P_1 metric is overemphasizing (i.e., trying to reduce) Type II error. Therefore, the P_1 criterion is more stringent than typical hypothesis tests and provides a useful method to test multiple validation targets, which is a more difficult task with standard hypothesis test procedures.

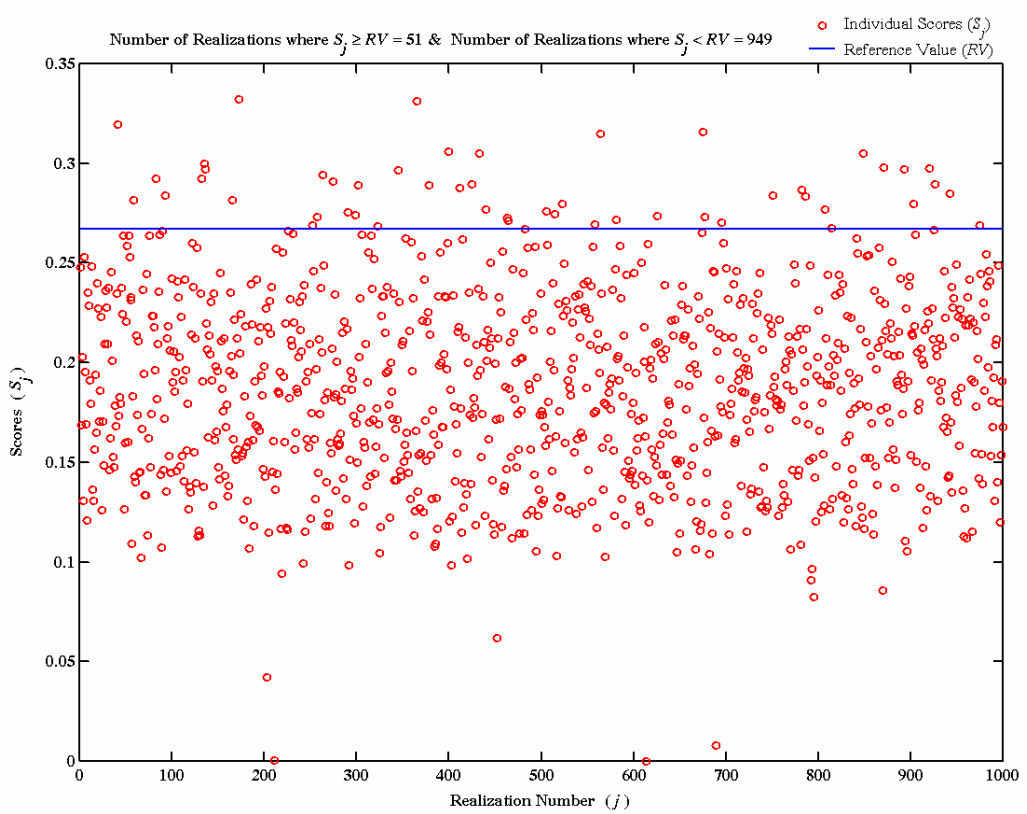

Figure 33. Example 2 showing individual realization scores, S_j, relative to the Reference Value, RV. The P_1 value here is about 5.1 percent (= 51/1,000).

It is important to note that according to P_1 and the Z-test, decreasing proportions of acceptable realizations are obtained as one deviates from the median of the model output distribution (zero in this test case.) At a 5 percent significance level and if the observed mean value coincides with the median of the model output, 95 percent of the realizations are deemed acceptable using the Z-test, whereas only 60 percent of the model realizations are deemed acceptable using the P_1 measure. Therefore, a rejection region of less than 30 percent for the P_1 criteria is very stringent and should not be confused with the 95 percent confidence interval used for presenting the output uncertainty.

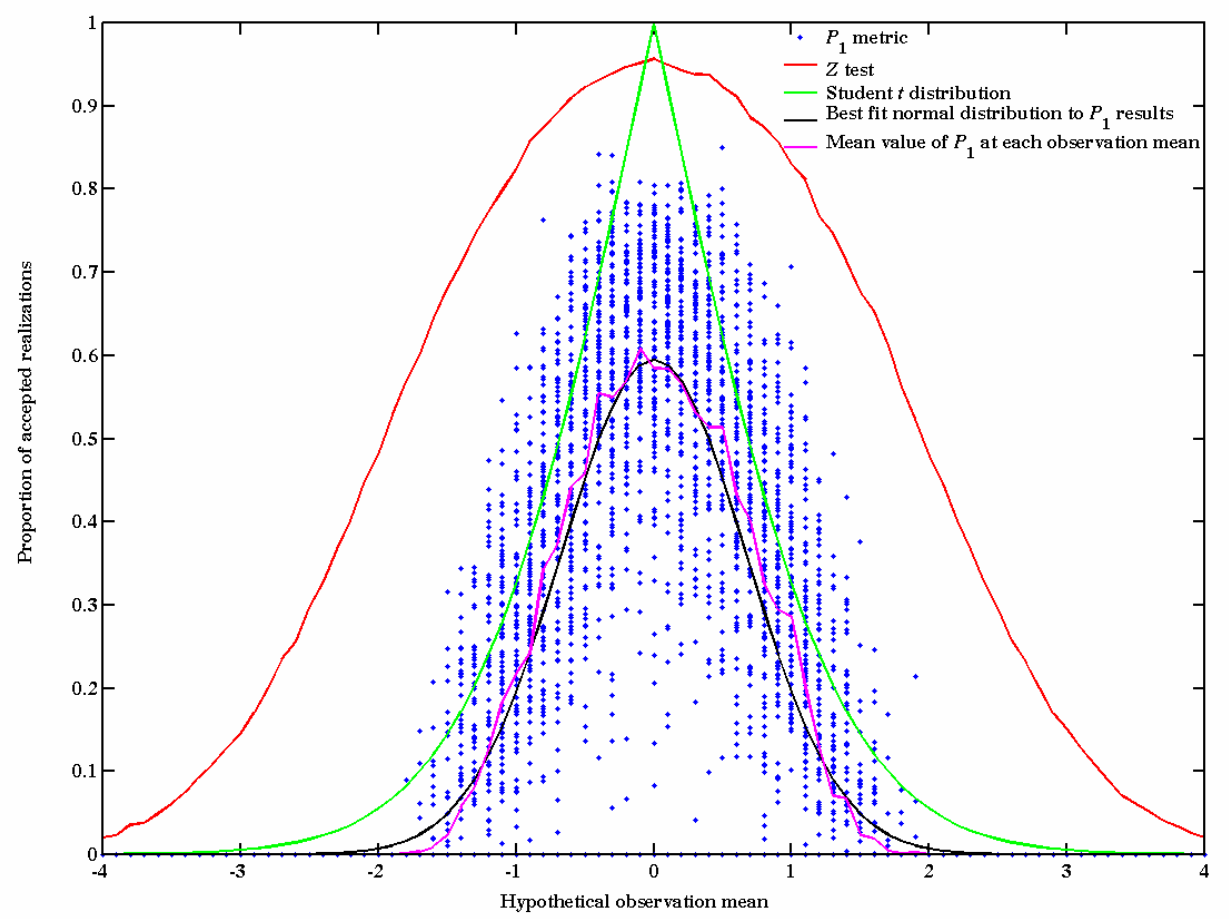

Figure 34. The P_1 metric (blue), its mean (magenta), its best-fit normal distribution (black) student t distribution (green), and the results of hypothesis testing using the Z-test (red) for the multiple validation targets case.

5.3.6 Testing the Efficacy of P_2 for Multiple Validation Targets

A numerical experiment is constructed to test the efficacy of the P_2 metric as follows:

1. A model is assumed to produce output according to a standard normal distribution.
2. Observations are assumed to follow a normal distribution with mean μ and unit variance. The numerical experiment chooses mean values μ from an observation distribution range -4.0 to 4.0 (i.e., -4.0, -3.9, …, 4.0).
3. For each mean value, a random sample of 30 observations is drawn from a normal distribution with the mean equal to the current mean value (i.e., -4.0, -3.9, …, 4.0) and a standard deviation equal to 1.0.
4. Each of the 30 observations is then compared to the model's distribution N (0,1) to determine what percentage falls outside of the 95 percent confidence interval (i.e., -1.96 to 1.96).
5. The process is repeated for all observation means [-4.0, 4.0].

Due to the random nature of the distributions generated in the above procedure, the above experiment was repeated about 100 times and the results are shown in Figure 35. The figure shows that if 50 percent is chosen as the rejection threshold for the P_2 metric, then the model would be accepted for $\mu = [-1.96, 1.96]$. This is a very interesting result, as one might initially think that 95 percent should be the acceptance threshold, but 50 percent yields the same acceptance region as a standard *t*-test at a 95 percent confidence level.

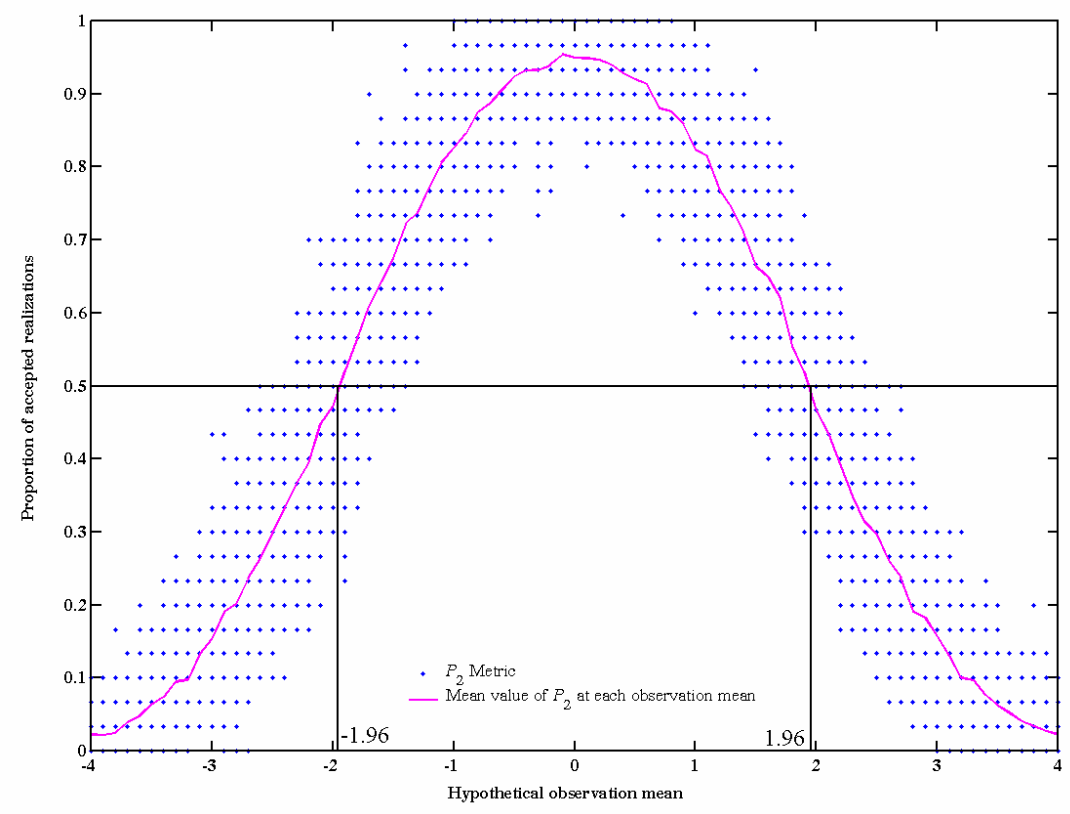

Figure 35. The P_2 metric (blue) and its mean (magenta) for the multiple validation targets case. The black lines show that at the 50 percent threshold, the acceptance region is [-1.96, 1.96], which is the same acceptance region for a standard *t*-test at the 95 percent confidence level.

6. SUMMARY AND CONCLUSIONS

Models have an inherent uncertainty. The difficulty in fully characterizing the subsurface environment makes uncertainty an integral component of groundwater flow and transport models, which dictates the need for continuous monitoring and improvement. Building and sustaining confidence in closure decisions and monitoring networks based on models of subsurface conditions require developing confidence in the models through an iterative process.

The definition of the model validation-model postaudit process is reaffirmed here in a way similar to Hassan (2003a). Namely, it is postulated that the model validation process is a confidence building and long-term iterative process. Model validation should be viewed as a process, not an end result. That is, model validation cannot always assure acceptable prediction or quality of the model. Rather, it provides a safeguard against faulty models or inadequately

developed and tested models. If model results end up being used as the basis for decision making, then the validation process indicates that the model is valid for making decisions (not necessarily a true representation of reality).

This report attempts to reconcile the differences between the FFACO's (2000) 10-step model validation process and the 12-step modeling protocol established in Anderson and Woessner's (1992b) book and adopted in many other hydrogeologic references. It is argued here that if the modeling process is to contain a model postaudit stage, it implies iterative model building, calibration, simulation, validation, monitoring, and re-characterization. Without a model reconfiguration and refinement possibility, the model postaudit step is meaningless. This is simply because the postaudit step requires comparisons be made between model predictions and newly collected field data. This comparison is unnecessary if one does not consider the possibility of revising the model, which gives rise to an iterative process of model building, evaluation, and refinement.

Following Hassan (2003a, 2004b), an approach is proposed for the validation process of the stochastic Shoal model. The proposed approach aims at building confidence in the model predictions.The focus of the proposed validation methodology is on testing how the predictions of the groundwater flow and transport models of Shoal and the underlying conceptual models and assumptions are robust and consistent with regulatory purposes, and on linking validation efforts to long-term monitoring that benefits from and builds on the validation-phase field activities. As the approach is similar to that proposed for CNTA in Hassan (2003a), the details of the approach are not repeated here, but rather more explanation and discussion of some of the challenging aspects of the process are addressed as they apply to the Shoal model. In particular, two main aspects are analyzed here. These are the selection of the validation targets at Shoal, and the selection of the acceptance criteria for the stochastic model realizations.

To select the validation targets at Shoal, a multi-parameter uncertainty analysis is performed to identify which flow and transport parameters the contaminant boundaries are most sensitive to. Five flow parameters and four transport parameters are selected for this analysis. These are the fracture orientation, the fracture lengths, the cavity and disturbed zone conductivity, the fracture conductivity, the recharge, the cavity porosity, the damaged zone porosity, fracture porosity, and matrix diffusion. The analysis proceeds by running the calibrated model in Monte Carlo mode for each of the individual uncertain parameters that are under consideration; all of the other uncertain parameters are held at constant values that are representative of their distribution. The output uncertainty for the base case mode (Pohlmann *et al*., 2004) and for all the uncertainty cases is obtained in terms of the ^{14}C contaminant boundary.

The results of the uncertainty analysis indicate that hydraulic conductivity contributes significantly to the overall output uncertainty, far more than any other flow or transport parameter. This implies the need to consider the hydraulic conductivity of the fractured granite at Shoal as one of the validation targets at the input side of the model. The contribution of all other input parameters to overall uncertainty is less and with the added difficulty to measure these parameters in the field, they are not considered as viable validation targets. At the output side of the model, hydraulic head and head gradients can be used as validation targets. Also, the presence or absence of radionuclides at locations in the downgradient direction can be used as another validation target (i.e., presence of contaminants where model predicts none or vice versa). Also, fracture mapping in wells and temperature logs may provide independent

information about certain aspects that can be tested for the model such as fracture distribution and recharge distribution.

During the validation process, a question arises as to the sufficiency of the number of acceptable (in terms of conformity with validation data) model realizations. A hierarchical approach is proposedto make this determination. This approach is based on computing five measures or metrics, following a decision tree, to determine if a sufficient number of realizations attain satisfactory scores on how they represent the field data used for calibration (old) and used for validation (new).

The first two of these measures are applied to hypothetical scenarios assuming field data consistent with the model or significantly different than the model results. In both cases, it is shown how the two measures would lead to the appropriate decision about the model performance. Standard statistical tests are used to evaluate these measures with the results indicating that they are appropriate measures for evaluating model realizations.

REFERENCES

Alley, W.M. and P.A. Emery, 1986. Groundwater model of the Blue River Basin, Nebraska - twenty years later. *Journal of Hydrology* 85, 225-250.

Anderson, M.G. and P.D. Bates. 2001. *Model Validation: Perspectives in Hydrological Science*. New York, NY. John Wiley & Sons, Ltd.

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

Anderson, M.P. and W.W. Woessner, 1992b. *Applied Groundwater Modeling: Simulation of Flow and Advective Transport*. New York, NY. Academic Press.

Anderson, P.F. and S.Lu, 2003. A post audit of a model-designed groundwater extraction system. *Groundwater* 41(2), 212-218.

Anderson, T.W., 1968. Electric analog analysis of ground-water depletion in central Arizona. U.S. Geological Survey, Water Supply Paper 1860, 21 pp.

Beck, M.B., J.R. Ravetz, L.A. Mulkey and T.O. Barnwell, 1997. On the problem of model validation for predictive exposure assessments. *Stochastic Hydrology and Hydraulics* 11, 229-254

Berkowitz, B., 2002. Characterizing flow and transport in fractured geological media: A review, *Advanced Water Resources* 25, 861-884.

Borg, I.Y., R. Stone, H.B. Levy and L.D. Ramspott, 1976. Information Pertinent to the Migration of Radionuclides in Ground Water at the Nevada Test Site, Part 1: Review and Analysis of Existing Information. Lawrence Livermore National Laboratory, UCRL-52078 Pt. 1, 216 p.

Box, G.E.P., 1979. Robustness in the Strategy of Scientific Model Building. In *Robustness in Statistics*, R.L. Launer and G. N. Wilkinson (eds.), Academic Press, New York, p.201-236.

Bredehoeft, J.D., 2003. From models to performance assessment: The conceptualization problem. *Groundwater* 41(5), 571-577.

Brown, D.M., 1996. Reducing modeling uncertainty using ASTM ground-water modeling standards. In *Subsurface Fluid-Flow* (*Ground-Water and Vadose Zone*) *Modeling, ASTM STP 1288,* J.D. Ritchey and J.D. Rumbaugh (eds.). American Society for Testing and Materials, 24-41.

Cacas, M.C., E. Ledoux, G. de Marsily, B. Tillie, A. Barbreau, E. Durand, B. Feuga, P, Peaudecerf, 1990a. Modeling fracture flow with a stochastic discrete fracture network: Calibration and validation, 1. The flow model. *Water Resources Research* 26(3), 479-489.

Cacas, M.C., E. Ledoux, G. de Marsily, A. Barbreau, P. Calmels, B. Gaillard and R. Margrita, 1990b. Modeling fracture flow with a stochastic discrete fracture network: Calibration and validation, 2. The transport model. *Water Resources Research* 26(3): 491-500.

Carroll, R., T. Mihevc, G. Pohll, B. Lyles, S. Kosinski and R. Niswonger, 2000. Project Shoal Area Tracer Test Experiment. Desert Research Institute, Publication No. 45177, DOE/NV/ 13609--05, 35p.

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.

Duan, Q., H.V. Gupta, S. Sorooshian, A.N.Rousseau and R. Turcotte (eds.), 2003. Calibration of watershed models. *Water Sci. Appl. Ser.*, vol 6, AGU, Washington, D.C.

Ewing, R.C., M.S. Tierney, L.F. Konikow and R.P. Rechard, 1999. Performance assessment of nuclear waste repositories: A dialogue on their value and limitations. *Risk Analysis* 19(5), 933-958.

FFACO, 2000. Federal Facilities Agreement and Consent Order, Appendix VI: Corrective Action Strategy.

Flavelle, P., S. Nguyen and W. Napier, 1990. Lessons learned from model validation: A regulatory perspective. *In* GEOVAL 1990: Symposium on Validation of Geosphere Flow and Transport Models, Organization for Economic Cooperation and Development, Nuclear Energy Agency, Paris, France, 441-448.

Freyberg, D.L., 1988. An exercise in groundwater model calibration and prediction. *Ground Water* 26(3), 350-360.

Hassan, A.E., 2002. Comment on "Determination of particle transfer for random walk particle methods in fractured porous media " by H.H. Liu *et al*., *Water Resources Research,* 38(11), 1221, doi:10.1029/2000WR000132.

Hassan, A.E., 2003a. A Validation Process for the Groundwater Flow and Transport Model of the Faultless Nuclear Test at Central Nevada Test Area. Desert Research Institute, Division of Hydrologic Sciences, Publication No. 45197, U.S. Department of Energy, Nevada Site Office report DOE/NV/13609-24, 70p. Las Vegas, NV.

Hassan, A.E., 2003b. Long-term Monitoring Plan for the Central Nevada Test Area. Desert Research Institute, Division of Hydrologic Sciences Publication No. 45201, U.S. Department of Energy, Nevada Site Office report DOE/NV/13609-30, 56p. Las Vegas, NV.

Hassan, A.E., 2004a. Validation of numerical groundwater models used to guide decision making. *Ground Water* 42(2), 277-290.

Hassan, A.E., 2004b. A methodology for validating numerical groundwater models. *Ground Water* 42(3), 347-362.

Hassan, A., K. Pohlmann and J. Chapman, 2002. Modeling Groundwater Flow and Transport of Radionuclides at Amchitka Island's Underground Nuclear Tests: Milrow, Long Shot, and Cannikin. Desert Research Institute, Division of Hydrologic Sciences, Publication No.45172, DOE/NV/11508--51.

Hazelton-Nuclear Science Corporation, 1965. Post-Shot Hydrologic Safety, Project Shoal Final Report. U.S. Atomic Energy Commission, Vela Uniform Project Shoal, VUF-1014, 50 p.

Hornberger, G.M. and E.W. Boyer, 1995. Recent advances in watershed modeling, *U.S. Natl. Rep. Int. Union Geod. Geophys. 1991 – 1994*. *Rev. Geophys.* 33, 949-957.

IT Corporation, 2000. 1999 Well Installation Report, Project Shoal Area, Churchill County, Nevada. Prepared for U.S. Department of Energy, Nevada Operations Office. Las Vegas, Nevada, ITLV/13052-097, variable paging.

Kleijnen, J.P.C., 1999. Statistical validation of simulation, including case studies. In *Validation of simulation models*, C. van Dijkum, D. de Tombe, and E. van Kuijk (eds.). SISWO, Amsterdam.

Konikow, L.F., 1986. Prediction accuracy of a groundwater model: Lessons from a postaudit. *Groundwater* 24, no. 2: 173-184.

Konikow, L.F., 1992. Discussion of "The modeling process and model validation" by Chin-Fu Tsang, *Ground Water*, 30, 622-623.

Konikow, L.F., 1995. The value of postaudits in groundwater model applications. In Groundwater Models for Resources Analysis and Management, A.I. El-Kadi (ed.), Lewis Publishers, Boca Raton, p. 59- 78.

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

Konikow, L.F. and L.A. Swain, 1990. Assessment of predictive accuracy of a model of artificial recharge effects in the Upper Coachella Valley , California. In *Selected Papers on Hydrogeology*, v. 1, E.S. Simpson and J.M. Sharp, Jr. (eds.), International Assoc. of Hydrogeologists, Proc. 1989 IGC Meeting, Washington, D.C., p. 433-449.

Konikow, L.F. and M.A, Person, 1985. Assessment of long-term salinity changes in an irrigated stream-aquifer system. *Water Resources Resources* 21, 1611-1624.

Lewis, B.D. and F.S. Goldstein, 1982. Evaluation of a predictive groundwater solute transport model at the Idaho National Engineering Laboratory, Idaho. U.S. Geological Survey Water Resources Investigations Report 82-25, 71 pp.

Liu, H.H., G.S. Bodvarsson and L. Pan, 2000. Determination of particle transfer in random walk particle methods for fractured porous media. *Water Resources Research* 36, 707-713.

Liu, H.H., G.S. Bodvarsson and L. Pan, 2002. Reply to ''Comment on 'Determination of particle transfer in random walk particle methods for fractured porous media' by H.H. Liu *et al*.'' *Water Resources Research* 38(11), 1222, doi:10.1029/2002WR001568.

Long, J.C. and D.M. Billaux, 1987. From field data to fracture network modeling: an example incorporating spatial structure. *Water Resources Research* 23, 1201-1216.

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

Mihevc, T., G. Pohll and B. Lyles, 2000. Project Shoal Area Field Data Summary Report. Desert Research Institute, Division of Hydrologic Sciences, Publication No. 45175, DOE/NV/11508--54, 258p.

National Research Council, 1999. *Groundwater Models; Scientific and Regulatory Applications*. National Academy Press, Washington, D.C., 303 pp.

National Research Council, 2000. *Research Needs in Subsurface Science*. National Academy Press, Washington, D.C., 159 pp.

Neretnieks, I., 1993. Solute transport in fractured rock - Applications to radionuclide waste repositories. *In Flow and Transport in Fractured Rock*, Bear *et al*. (eds.), Academic Press, San Diego, 39-127.

Neretnieks, I., T. Eriksen and P. Tahtinen, 1982. Tracer movement in a single fracture in granitic rock: some experimental results and their interpretation. *Water Resources Research* 18(4), 849-858.

Oreskes, N., 1998. Evaluation (not validation) of quantitative models. *Environmental Health Perspectives* 106 (suppl. 6), 1453-1460.

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

Oreskes, N. and K. Belitz, 2001. Philosophical issues in model assessment. In *Model Validation: Perspectives in Hydrological Science*, M.G. Anderson and P.D. Bates (eds.). London: John Wiley and Sons, Ltd., pp. 23-41.

Pan, L. and G.S. Bodvarsson, 2002. Modeling transport in fractured porous media with the random-walk particle method: The transient activity range and the particle transfer probability. *Water Resources Research* 38(6), 16-1 - 16-7.

Pan, L., H.H. Liu, M. Cushey and G.S. Bodvarsson, 2001. DCPT V1.0 - New particle tracker for modeling transport in dual continua media, users' manual, Rep. LBNL-42958, Lawrence Berkeley National Laboratory, Berkeley, CA.

Person, M. and L.F. Konikow, 1986. Recalibration and predictive reliability of a solute-transport model of an irrigated stream-aquifer system. *Journal of Hydrology* 87, 145-165.

Pohll, G., 1999. Evaluation of surface recharge flux for the Project Shoal Area. Desert Research Institute, Division of Hydrologic Sciences, Publication No. 45171, DOE/NV/11508--49, 18p.

Pohll, G. and K. Pohlmann, 2004. Letter Report: Contaminant Boundary at the Shoal Underground Nuclear Test. Desert Research Institute, Division of Hydrologic Sciences, DOE/NV--993, 19 p.

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

Pohll, G., J. Chapman, A. Hassan, L. Papelis, R. Andricevic and C. Shirley, 1998. Evaluation of Groundwater Flow and Transport at the Shoal Underground Nuclear Test: An interim report. Desert Research Institute, Division of Hydrologic Sciences, Publication No. 45162, DOE/NV/11508-35.

Pohll, G., A.E. Hassan, J.B. Chapman, C. Papelis and R. Andricevic, 1999a. Modeling groundwater flow and radioactive transport in a fractured aquifer. *Ground Water* 37(5), 770-784.

Pohll, G., J. Tracy and F. Forsgren, 1999b. Data Decision Analysis: Project Shoal. Desert Research Institute, Publication No. 45166, DOE/NV/11508--42, 27p.

Pohll, G., K. Pohlmann, J. Daniels, A. Hassan and J. Chapman, 2003. Contaminant Boundary at the Faultless Underground Nuclear Test. Desert Research Institute, Division of Hydrologic Sciences, Publication No. 45196, DOE/NV/13609-24, 52 p.

Pohlmann, K.F., A.E. Hassan and J.B. Chapman, 1999. Evaluation of Groundwater Flow and Transport at the Faultless Underground Nuclear Test, Central Nevada Testing Area. Desert Research Institute, Division of Hydrologic Sciences, Publication No. 45165.

Pohlmann, K.F., G. Pohll, J. Chapman, A. Hassan, R. Carroll and C. Shirley, 2001. Modeling of Groundwater Contaminant Boundaries for the Shoal Underground Nuclear Test. Desert Research Institute, Division of Hydrologic Sciences, Publication No. 45184.

Pohlmann, K., G. Pohll, J. Chapman, A.E. Hassan, R. Carroll and C. Shirley, 2002. Modeling of Groundwater Contaminant Boundaries for the Shoal Underground Nuclear Test. Desert Research Institute, Division of Hydrologic Sciences, Publication No. 45184-revised, pp. 132.

Pohlmann, K., G. Pohll, J. Chapman, A. E. Hassan, R. Carroll and C. Shirley, 2004. Modeling to Support Groundwater Contaminant Boundaries for the Shoal Underground Nuclear Test. Desert Research Institute, Division of Hydrologic Sciences, Publication No. 45184 revised, pp. 197.

Refsgaard, J.C. and B. Storm, 1996. Construction, calibration, and validation of hydrological models. *In Distributed Hydrological Modeling*, M.B. Abbott and J.C. Refsgaard (eds.), Kluwer, Dordrecht, 41-45.

Reimus, P., G. Pohll, T. Mihevc, J. Chapman, M. Haga, B. Lyles, S. Kosinski, R. Niswonger and P. Sanders, 2003. Testing and parameterizing a conceptual model for solute transport in a fractured granite using multiple tracers in a forced-gradient test. *Water Resources Research* 39(12):1356-1370.

Robertson, J.B., 1974. Digital modeling of radioactive and chemical waste transport in the Snake River Plain aquifer at the National Reactor Testing Station, Idaho. U.S. Geological Survey Open File Report IDO-22054.

Sargent, R.G., 1990. Validation of mathematical models. In *GEOVAL-90, Symposium on Validation of Geosphere Performance Assessment Models*, Stockholm, Sweden, 14-17 May, 571-579.

Shah Alam, A.H.M., 1998. Regulatory guidance for accepting contaminant fate and transport models. In *Proceedings Modflow 98*, E. Poeter *et al*. (eds.) 387-393. Golden, CO.

Singh, V.P., 1995. *Computer Models of Watershed Hydrology*, 1129 pp., Water Resour. Publ., Highlands Ranch, CO.

Singh, V.P. and D.A. Woolhiser, 2002. Mathematical modeling of watershed hydrology. *J. Hydrol. Eng*. 7(4), 270-292.

Smith, D.K., 2001. Unclassified radiologic source term for Nevada Test Site areas 19 and 20. Lawrence Livermore National Laboratory report UCRL-ID-141706, 4p.

Steefel, C.I. and P. van Cappellen, 1998. Reactive transport modeling of natural systems. *Journal of Hydrology* 209, 1-7.

Stewart, M. and C. Langevin, 1999. Post audit of a numerical prediction of well field drawdown in a semiconfined aquifer system. *Ground Water* 37(2), 245-252.

Tsang, C.F., 1987. Comments on model validation. *Transport in Porous Media* 2(6), 623-630.

Tsang, C.F., 1991. The modeling process and model validation. *Ground Water* 29(6), 825-831.

Tsang, C.F., 1992. Reply to the preceding discussion of "The modeling process and model validation" by L.F. Konikow. *Ground Water* 30, 622-624.

U.S. Department of Energy (U.S. DOE), 1998a. Data Report Project Shoal Area Churchill County, Nevada. Nevada Operations Office, Environmental Restoration Division, DOE/NV--505, variable paging.

U.S. Department of Energy (U.S. DOE),, 1998b. Corrective Action Investigation Plan for Corrective Action Unit 447: Project Shoal Area, Nevada Subsurface Site. Nevada Operations Office, Environmental Restoration Division, DOE/NV--513, 71p.

U.S. Department of Energy (U.S. DOE),, 1999. Addendum to the Corrective Action Investigation Plan for Corrective Action Unit 447: Project Shoal Area, Nevada Subsurface Site. Nevada Operations Office, Environmental Restoration Division, DOE/NV--513-ADD, 10p.

U.S. Department of Energy (U.S. DOE), 2000. United States Nuclear Tests, July 1945 through September 1992. DOE/NV-209, rev. 15, Nevada Operations Office.

van der Heijde, P.K.M., 1990. Quality assurance in the development and application of groundwater models. In *ModelCARE 90: Calibration and Reliability in Groundwater Modeling* IAHS Publ. no. 195, 271-278.

Vogel, R.M., Sankarasubramanian, A., Validation of a watershed model without calibration, Water Resour. Res., Vol. 39, No. 10, 2003.

Weaver, J.D., R.K. Digel and P.V. Rosasco, 1996. A postaudit of groundwater flow models used in design of a groundwater capture/containment system. In *Subsurface Fluid-Flow* (*Ground-Water and Vadose Zone*) *Modeling, ASTM STP 1288,* J.D. Ritchey and J.D. Rumbaugb (eds.), American Society for Testing and Materials, 377-390.