Groundwater Model Validation

Ahmed E. Hassan

Irrigation and Hydraulics Department, Faculty of Engineering, Cairo University, Giza, Egypt
Also at: Division of Hydrologic Sciences, Desert Research Institute, Las Vegas, Nevada, USA

Table of Content

1. Introduction

The term ‘validation’ has been the subject of literature debate for more than a decade. Some skeptics argue the term should not be used, as it implies correctness and accuracy for a groundwater model, which no one can claim. Similarly, others believe 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 the approximate solutions to subsurface problems (Oreskes et al., 1994). Anderson and Bates (2001) edited a book covering the issue of model validation in hydrological sciences. Some of the quotes they presented from the skeptics and opponents to the use of the term “validation” include:

“Absolute validity of a model is never determined” (National Research Council, 1990).

“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” (Konikow and Bredehoeft, 1992).

The above views consider validation in the strictest definition of the word. That is, they refer to 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 (an objective they simply cannot achieve). 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, often the best available for subsurface problems that demand answers now. The model validation process can be regarded as an additional filter for independent model evaluation to assess the suitability of a model for its given purpose (likely to be decision-making). Most of the literature debate focuses on validation terminology and not on the process. No one argues the process is unimportant, unneeded, or useless and no one disagrees with the concept of using an independent data set to test the model. The disagreement is with what to call it and what the implications are for the term.

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 model 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 model predictions 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 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.

Following this introduction, a brief review of practical definitions of the term ‘validation’ is presented in Section 2. A description of a proposed model validation process for stochastic groundwater models (Hassan, 2004b) is presented in Section 3. The details of this process will not be repeated here; rather, more explanation and discussion of some of the challenging aspects of the process will be addressed as they apply to two actual field sites: the Project Shoal Area (PSA) in Nevada and the Amchitka site in Alaska. In particular, two main aspects are highlighted: 1) the selection of the acceptance criteria for the stochastic model realizations for the PSA, presented in Section 4, and 2) the evaluation of model input parameters and the reduction of their uncertainty at Amchitka, presented in Section 5. A wrap-up and concluding remarks are presented in Section 6.

2. Practical Views of Validation

Refsgaard (2001) defines model validation as the process of demonstrating a given site-specific model is capable of making accurate predictions for periods outside a calibration period. He also states that a model is said to be validated if its accuracy and predictive capability in the validation period have been proven to lie within acceptable limits or errors. De Marsily et al.

(1992) argue that using a model in a predictive mode and comparing it with new data does not (and does not have to) prove the model is correct for all circumstances; it only increases the confidence in the model's value. They added, "We do not want certainty; we will be satisfied with engineering confidence."

Neuman (1992) defines the validation of safety assessment models as the process of building scientific confidence in the methods used to perform such assessment. However, he recognizes that this confidence-building approach to validation is possibly open-ended, as many iterations between modelers and regulators may be needed. Eisenberg et al. (1994) support the idea of confidence building and indicate this term recognizes that full scientific validation of models of performance assessment may be impossible and the acceptance of mathematical models for regulatory purposes should be based on appropriate testing, which will lead to a reasonable assurance that the results are acceptable. Hassanizadeh (1990) differentiates between research (or analysis) model validation and safety assessment modeling (or predictive modeling) validation. The former is a tool to help understand processes, uncertainties, etc., whereas the latter is a goal to help the decision-making process. Sargent (1990) regards validation as a process that consists of performing tests and evaluations during model development until sufficient confidence is obtained that a model can be considered valid for its intended application.

There is a general consensus that the main concern is whether or not a model is adequate for its intended use and whether or not there is sufficient evidence that the model development followed logical and scientific approaches and did not fail to account for important features and processes. It should be noted that determining the adequacy of a model or building confidence in its prediction is not ideally a one-time exercise. It can be considered an iterative process that should be viewed as part of an integral loop with trigger mechanisms or decision points that force the model back to the data collection-conceptualization-calibration-prediction phase loop if the model validation process indicates the need to do so. Key to this process is the use of a diverse set of tests that should be designed to evaluate a diverse set of aspects related to the model.

It is clear that often-quoted statements such as “groundwater models cannot be validated” and “groundwater models can only be invalidated” (Konikow and Bredehoeft, 1992, 1993; Bredehoeft and Konikow, 1992, 1993) refer to validation in only the strictest sense, responding to a concern regarding layman’s possible misconceptions. Unfortunately, such statements may lead to a relaxed attitude on the part of researchers, consultants, and even regulatory agencies when it comes to evaluating model predictions. With the perception that the groundwater model will never be validated, there may be a temptation to believe good model development, building, and calibration are sufficient and nothing more can be done. All groundwater modelers agree that their models cannot be validated in the strictest sense (at least with present-day technology) but similarly agree on the importance of post-prediction testing and evaluation. By expanding the definition of validation to encompass a long-term process of confidence building, modelers and model users can develop rigorous validation processes that will ultimately improve the models and the quality of decisions based on those models.

3. The Proposed Model Validation Process

Hassan (2004b) proposed an approach for the validation process of stochastic numerical models. The details of the process will not be repeated here; rather, more explanation and discussion of some of the challenging aspects of the process will be addressed as they apply to two actual field sites. However, for completeness, the overall validation approach is briefly discussed in the following.

The validation process is designed to account for the stochastic nature of some groundwater models. The focus of the process is centered around the following three main themes: (1) testing if predictions of numerical groundwater models and the underlying conceptual models are robust and consistent with regulatory purposes, (2) improving model predictions and reducing uncertainty using data collected through field activities designed for model validation, and (3) linking validation efforts to long-term monitoring and management of the site.

Figure 1, adapted from Hassan (2004b), displays the step-by-step approach for performing the validation and postaudit processes for a typical stochastic groundwater model. The proposed steps of the model validation process 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.

Step 2: Install the wells and collect the largest amount of data possible from the wells. The data should include geophysical logging, head measurements, conductivity measurements, contaminant concentrations, and any other information that could be used to test the model structure, input, or output.

Step 3: Perform different validation tests to evaluate various components of the model. The stochastic validation approach proposed by Luis and McLaughlin (1992) is an example of an approach that can be used to test the flow model output (heads) under saturated conditions. Other goodness-of-fit tests can be used for the heads to complement this stochastic approach. 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 4.

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 in many of the tested aspects and focus on those that "passed" the majority of the tests. 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 as presented in Section 5 for the Amchitka site.

Step 5: Step 4 results will determine the subsequent step and guide the decision of whether there is a sufficient number of realizations with high scores or not.

5a. If the number of realizations with high scores is very small compared to the total number of model realizations (the P_1 measure discussed in Section 4), 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, the model may be conceptually good, but the input parameter distributions are skewed one way or another. Generating more realizations and keeping those fitting 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 will be obvious. The importance of the distinction between conceptualization problems and inappropriate input distributions has been discussed by Bredehoeft (2003). He states 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 we rethink the conceptual model. The use of different metrics such as P_1 and P_2 discussed in Section 4 provides a tool for making this distinction.

5b**.** If the number of realizations with high scores is found sufficient, this indicates the model does not have any major deficiencies or conceptual problems. Based on the realizations retained in the analysis and deemed acceptable, the regulatory quantity of interest can be calculated and compared to the original model-predicted quantity. 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 1. 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 is no, then the left-hand side path in Figure 1 begins, and a new iteration of model development will begin using the original data plus data collected for validation. Steps 1 to 6 will eventually be repeated.

6b. If the answer to the question is yes, validation is deemed sufficient and the model is considered adequate or robust and the process proceeds to step 7, which starts the development of the long-term monitoring program for the site.

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 1) the criteria for determining the sufficiency of the number of acceptable realizations, which are demonstrated for the Project Shoal Area (PSA) model, and 2) the reduction of uncertainty of model input parameters, which is demonstrated for the Amchitka model. Other aspects of this validation approach need further analysis and development. The work on developing a model validation process is still in its infancy and the hope is that more attention will be devoted to the development of rigorous approaches for conducting groundwater model validation processes.

4. Acceptance Criteria for Stochastic Model Realizations and Application to PSA Model

The PSA, about 50 km southeast of Fallon, Nevada, is the location of the Shoal underground nuclear test. The nuclear device was emplaced 367 m below ground surface, about 65 m below the water table, in fractured granite. Details of the geology and hydrogeology of the site are provided in Pohll *et al*. (1999).

Ongoing environmental remediation efforts at the PSA have successfully progressed through numerous technical challenges related to the substantial uncertainties present when characterizing a heterogeneous subsurface environment. The original Corrective Action Investigation Plan (CAIP) for the PSA described the drilling and testing of four characterization wells, followed by flow and transport modeling. The data analysis and the first modeling effort are described in Pohll et al. (1999). After evaluating the modeling effort, the model sponsor determined the degree of uncertainty in transport predictions for PSA remained unacceptably large. Thus, a second CAIP was developed prescribing a rigorous analysis of uncertainty in the PSA model and quantification of methods for reducing uncertainty by collecting additional data. This analysis formed the basis for a second major characterization effort, where four additional characterization wells were drilled during summer and fall 1999. A key component of the second field program was a tracer test conducted between two of the new four wells, which is described in Reimus et al. (2003).

Results of the tracer test and new characterization efforts produced a new groundwater flow and radionuclide transport model that was approved by the model sponsor and regulators (Pohlmann et al. 2004). The next step in the environmental management of the PSA is to start the validation process and develop the long-term monitoring network. This section focuses on the selection of acceptance criteria and how they can be applied in the PSA model. An attempt is made to develop and test some acceptance criteria to be used through the model validation process. It is important to note that this analysis is still in the development stage and the application here is made for hypothetical cases (i.e., hypothetical validation data). Field data will be collected for the validation process of the PSA model at the time this book will be in press.

4.1. Proposed Acceptance Criteria for PSA

According to the validation approach shown in Figure 1, 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 validation (new). The determination of whether a 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% of the probability density function (pdf) of these targets as used in the model is the second criterion, P_2.

3. The results of hypothesis testing to be conducted using the stochastic perturbation approach of Luis and McLaughlin (1992) as described in detail in Hassan (2004b), P_3 , represent the third criterion.

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 data set obtained in the field), P_4, are considered the fourth criterion.

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 is 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 2. First, S_j is evaluated to determine whether the percentage of realizations with scores above the reference value, P_1, is more than 40%, between 30% and 40%, or less than 30%. If the number is more than 40%, it is deemed sufficient. If it is between 30% and 40% or less than 30%, the second criterion, P_2, is used as shown in Figure 2. The second criterion represents the number of validation targets where the field data lie within the inner 95% of the pdf for that target as used in (input) or produced by (output) the model. Then if P_1 is between 30% and 40% and P_2 is between 40% and 50% or if P_1 is less than 30% but P_2 is greater than 50%, the number of realizations is deemed sufficient. If P_1 is less than 30% and P_2 is less than 40%, 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, the model may be conceptually good but the input parameter distribution is skewed or inappropriate and by generating more realizations and keeping the ones fitting 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% to 40% for P_1 and 40% to 50% for P_2) is described through an example and when these metrics are evaluated with statistical hypothesis testing later in this Section.

4.2 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 a well). Figure 3 shows the head distribution as produced by the stochastic PSA 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 can be computed as

$$\text{Reference Value }(RV) = \begin{cases} \exp\left[-\dfrac{(h_o - h_{2.5})^2}{(h_{97.5} - h_{2.5})^2}\right] & \text{for } h_o < h_{50} \\ \\ \exp\left[-\dfrac{(h_o - h_{97.5})^2}{(h_{97.5} - h_{2.5})^2}\right] & \text{for } h_o > h_{50} \end{cases} \tag{1}$$

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

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

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 PSA model. This leads to all realizations with absolute errors smaller than $\min\{(|h_o - h_{2.5}|), (|h_o - h_{97.5}|)\}$, attaining a score higher than the reference value. Figure 4 below shows the resulting scores and how they compare to the reference value, RV as obtained from the above equations.

It can be seen from Equations (1) through (3) that the maximum value *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 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 50th percentile of the target output, h_{50}, then P_1 will be approximately 95% indicating that 95% of the realizations attained scores higher than *RV*.

4.3. 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 for the model output is assumed. 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 5. The *RV* metric decreases rapidly as the observation value approaches the median, h_{50}. When the measured value lies outside the middle 95% of the output distribution (i.e., outside the range [-1.96, 1.96]), *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.

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 for the head. 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 6.

Figure 6 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 evaluating the statistical difference between the mean of the model output (assumed standard normal distribution) and each observed value (assuming 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 $n - 1 = 0$.

To avoid this limitation, the *Z* test, which is commonly used for the same purpose, but it assumes the variances of the distributions are known, is employed. It is assumed 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 tested:

$$\begin{aligned} &\mathrm{H}_0 : h_j = h_o \qquad && for\ \ j = 1, \ldots, NMC \\ &\mathrm{H}_1 : h_j \neq h_o \qquad && for\ \ j = 1, \ldots, NMC \end{aligned} \tag{4}$$

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 5.

The plots in Figure 6 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% 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% significance level and if the observed value coincides with the median of the model output, only 95% 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% 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 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 reduce the model user's risk (both model sponsor and regulators) at the expense of increasing model builder's risk (supposedly the research team constructing the model).

4.4 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

$$RV = \exp\Big(-\sum_{i=1}^{N}\min\big[\,(h_{\text{o}_i} - h_{2.5_i})^2\,,\,(h_{\text{o}_i} - h_{97.5_i})^2\,\big]/\sum_{i=1}^{N}[h_{97.5_i} - h_{2.5_i}]^2\Big) \tag{5}$$

$$S_j = \exp\Big(-\sum_{i=1}^{N}[h_{\text{o}_i} - h_j]^2 / \sum_{i=1}^{N}[h_{97.5_i} - h_{2.5_i}]^2\Big) \quad \text{for} \quad j = 1, \ldots, NMC \tag{6}$$

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 PSA model provides a distribution of values. It is then assumed that the values of the field data are known (one realization is chosen at random to provide an example observation for all targets.) Figures 7 and 8 show the results of this example (Example 1) where P_1 is found to be about 76.7%. In this case, there is no check for P_2 and the sufficiency of the number of realizations having acceptable scores is accepted. Note, however, that if P_2 were checked, it would be about 94% (=17/18) because 17 data points lie between the 2.5th and the 97.5th percentiles for the corresponding targets.

Using another set of random values to hypothesize the field data, a different result is obtained as shown in Figures 9 and 10 for Example 2. In this case, both P_1 and P_2 are less than 40% (since the number of validation targets where the red circle is between the 2.5th and the 97.5th percentiles is only 2 ~ 11%). 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.

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 exactly matching one of the model realizations, the P_1 metric was found to be about 76.7%. 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 (similar to a jackknife method). The P_1 metric is obtained for these 1,000 experiments and its mean value was found to be about 43%. Given the actual field data to be collected for the validation analysis are very unlikely to exactly match any of the PSA model realizations, the 30% to 40% threshold for P_1 is considered realistic. In other words, if one, on average, obtains 43% 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% and 40% when using the actual validation data.

4.5. 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 the observation set being constant. A range of observation set 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].

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, with mean values far away from zero. This experiment allows us to compare the rejection region by 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 11. 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.

For the Z test, we assume each output realization represent a mean of a normal distribution. For each observation mean value, we then test the following hypothesis:

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

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 11. According to the figure, the t test would suggest accepting all model realizations if the mean value of the observations is inside the range [–2.2, 2.2] at 95%. The P_1 criterion has a narrower acceptance region ([–1.6, 1.6] according to the black or magenta line) again suggesting 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.

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% significance level and if the observed mean value coincides with the median of the model output, 95% of the realizations are deemed acceptable using the Z test, whereas only 60% of the model realizations are deemed acceptable using the P_1 measure.

Therefore a rejection region of less than 30% for the P_1 criteria is very stringent and should not be confused with the 95% confidence interval used for presenting the output uncertainty.

4.6 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% 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, we repeated the above experiment about 100 times and the results are shown in Figure 12. The figure shows that if 50% is chosen as the rejection threshold for the P_2 metric, then the model would be accepted for μ = [–1.96, 1.96]. This is a very interesting result as one might initially think 95% should be the acceptance threshold, but 50% yields the same acceptance region as a standard *t* test at a 95% confidence level. This warrants further analysis, and as stated earlier, the different aspects of the validation process need more attention for ultimate goal of arriving at a rigorous set of steps for conducting a model validation process.

5. Uncertainty Reduction Using Validation Data and Application to the Amchitka Model

The use of validation data to evaluate input parameter distributions and reduce their uncertainty level is demonstrated for the Amchitka site as a case study. Amchitka is the southernmost island of the Rat Island Group of the Aleutian Island chain extending southwestward from mainland Alaska. The focus here is on one of three underground nuclear tests conducted on the island, a test known as Milrow.

5.1 Model Background

The groundwater flow and radionuclide transport at the Milrow test was modeled using two-dimensional numerical simulations as described in Hassan et al. (2002). The Milrow model involved solving a density-driven flow problem (seawater intrusion problem) to obtain the salinity distribution and the groundwater velocities followed by solving the transport problem to predict radionuclide mass fluxes leaving the groundwater system and discharging into the sea. A multi-parameter uncertainty analysis was adapted and used to address the effects of the uncertainties associated with the definition of the modeled processes and the values of the parameters governing these processes. Details of the modeling efforts at Amchitka can be found in Hassan et al. (2001, 2002) and Pohlmann et al. (2002).

With uncertain parameters, the output of the seawater intrusion problem (i.e., the location of the transition zone between freshwater and saltwater in the groundwater system) varies between model realizations with significant impacts on the solution of the transport equation. This location is mainly determined by the recharge-conductivity ratio. This ratio varies dramatically in the model due to the large parametric uncertainty built into the conductivity and recharge distributions. Figure 13 shows the model domain, boundary conditions of the density-driven flow problem, the radionuclide source, transport processes considered and an example transition zone location and velocity vectors as produced by one realization.

Although the analysis conducted in Hassan et al. (2002) conservatively accounted for parametric uncertainty, the need remained validation of the numerical groundwater model with an independent data set. The Consortium for Risk Evaluation with Stakeholder Participation II (CRESP II), which represents an Organization of the Institute for Responsible Management, initiated a field expedition in the summer of 2004 for the purpose of collecting data as part of the Amchitka Independent Assessment Science Plan.[1] One of the objectives of this data collection campaign was to provide data to reduce uncertainty about the risk of radionuclide release through the groundwater system to the marine environment.

Of great importance to the groundwater model are the magnetotelluric (MT) measurements for determining the subsurface salinity and porosity structure of the Island, which were recently released in a report by Unsworth et al. (2005). Magnetotelluric data were collected on profiles that passed through the contaminant source at Milrow. The data presented in Unsworth et al. (2005) showed the presence of a pattern of increasing, decreasing and increasing resistivity with increasing depth at the site. The depth at which there is an inflection point of the resistivity

[1] http://www.cresp.org/

profile where resistivity begins to decrease is interpreted by Unsworth et al. (2005) as corresponding to the top of the transition zone (TZ) as the salinity increases. The deeper inflection point (increase in resistivity after its decrease with depth) is interpreted as corresponding to the base of the transition zone, as salinity remains constant and the decreasing porosity causes a rise in resistivity (Unsworth et al., 2005).

This independent set of data provides an opportunity for applying the validation process to the groundwater flow model at the Milrow site and reducing the uncertainty in the model parameters and subsequently the model output. To accomplish this, data pertaining to the location of the freshwater lens (from the interpretation of the MT data) could be compared to the groundwater model input distributions as part of the validation process, and then the distributions tightened around the new data for uncertainty reduction.

5.2 Bayesian Framework for Uncertainty Reduction

In hydrologic modeling, the uncertainty of parameter estimation needs to be accounted for and the impact on the model output uncertainty needs to be quantified. Bayesian inference provides a framework within which these issues can be addressed with the end product of models being a probability distribution known as the posterior distribution of the model parameters (input) and predictions (output) quantifying uncertainty after data have been collected and incorporated. Although a number of recent studies used this framework for rainfall-runoff modeling and parameter estimation (e.g., Kuczera, 1983; Freer et al., 1996; Kuczera and Parent, 1998; Kuczera and Mroczkowski, 1998; Bates and Campbell, 2001; Marshall et al., 2004), its application to other areas has been limited due to computational difficulties. The advent of Markov Chain

Monte Carlo (MCMC) methods has helped address some of these computational difficulties (Marshall et al., 2004).

For certain simple analyses, the posterior distributions (on which inferences are usually made) can be derived analytically, which means they can be written down in standard statistical notation. When this is not possible, a Bayesian solution can still be obtained through the use of simulation methods, such as the MCMC methods. Some details about the Bayesian framework and the MCMC method are presented next in relation to the Milrow model.

For the stochastic groundwater flow model at Milrow, it is assumed that the newly collected data are expressed by the data vector **D**. One of the elements in this data vector would be the groundwater salinity profile beneath the island, *C*, which also represents the steady-state output of the model. One can express the model as

$$C(\mathbf{x}) = M(\mathbf{G}, \mathbf{x}; \boldsymbol{\Theta}) + \varepsilon(\mathbf{x}) \qquad (8)$$

where $C(\mathbf{x})$ is the observed data at location $\mathbf{x}$, $M(\mathbf{G}; \boldsymbol{\Theta})$ is the model output for location $\mathbf{x}$, **G** is the set of model input describing domain geometry, boundary conditions, and discretizations, $\boldsymbol{\Theta}$ is the vector of unknown model parameters to be estimated from the data (e.g., hydraulic conductivity, recharge, porosity), and $\varepsilon(\mathbf{x})$ is an error term which is assumed to be a normally distributed random variable having zero mean and variance σ_{ε}^{2}. At this point it is also assumed that the error terms are all mutually independent.

The set of model input **G** includes those aspects that do not change from one realization to another. The set of model parameters $\boldsymbol{\Theta}$ includes all the uncertain parameters needed to run the model and that change from one model realization to another. The vector $\boldsymbol{\Theta}$ is treated as a random variable distributed according to a probability density function. This density function expresses our uncertainty about **Θ**. Before considering the newly collected data, the knowledge about the model parameter set can be summarized in a distribution *P*(**Θ**) called the prior distribution.

The posterior distribution of the parameter set, *P*(**Θ** |*C*) can be obtained through the application of Bayes' theorem

$$P(\mathbf{\Theta} \mid C) = \frac{P(\mathbf{\Theta})\, P(C \mid \mathbf{\Theta})}{P(C)} \tag{9}$$

where *P*(*C*| **Θ**) is the likelihood function that summarizes the model relation to the collected data given the parameters used in the model and *P*(*C*) is a proportionality constant required so that *P*(**Θ**|*C*) is a proper density function. It is important to note that the posterior distribution assumes a shape similar to the prior when available data are limited. But when there are large data sets, the posterior distribution will be influenced more by the data than by the assumed prior distribution. Also, the information in the new sample data will dominate the posterior if the prior distribution is selected to represent vague prior knowledge. *P* (**Θ**|*C*) thus contains all of the available information about **Θ** coming from both prior knowledge and collected data. Bayesian inference, therefore, reduces to summarizing the posterior density *P* (**Θ**|*C*).

MCMC sampling explores the posterior distribution by generating a random process (a Markov process) that eventually converges to the stationary, posterior distribution of the parameters. While there exist many different MCMC sampling algorithms, the Metropolis-Hastings algorithm is the most commonly used. Details of this algorithms are beyond the scope of this chapter, but can be found in Bates and Campbell (2001), Campbell et al. (1999), Marshall et al. (2004), Kuczera and Parent (1998), to name a few sources.

5.3 Application to the Milrow Site in Amchitka

The validation data at Milrow are obtained from the MT data and the interpretation of the measured MT signals. Specifically, this provided identification of the top and bottom bounds of the transition zone from freshwater to seawater beneath the island. This set of data is used with MCMC to develop the posterior distributions for conductivity, recharge, and conductivity-recharge ratio. The development of the posterior distribution serves to evaluate the original (prior) distribution selection and to reduce the uncertainty in the parameter distributions, both of which are objectives of the validation process.

In the original Milrow model (Hassan et al., 2002), lognormal distributions were assumed for both recharge and conductivity and they were assumed to be uncorrelated. These prior distributions were based on calibrating the model to a set of salinity data that was available from one borehole at the site. This data set clearly defines the increase in salinity with depth to near-seawater concentrations. Therefore, the location of the transition zone separating the shallow freshwater from the deep saltwater provided by this data set guided the selection of the distributions used in the original model. These distributions are used here as the prior distribution for the MCMC analysis and with the help of the MT data, posterior distributions are developed.

Figure 14 shows the potential of the MCMC approach to help in the validation process and uncertainty reduction by conditioning on field data. The figure compares the prior distributions for the recharge, conductivity and the recharge-conductivity ratio to the posterior distributions for these parameters obtained utilizing the prior knowledge and the new MT-related data. It is clear that the posterior distributions are very different from the prior ones, especially for the recharge-conductivity ratio. The interpretation of the MT data indicated a deeper transition zone than what was used in the model. This translated into the posterior distribution of conductivity being skewed to lower values relative to the prior and that of recharge being skewed to the higher values. The end result is a higher recharge-conductivity ratio for which the posterior distribution has a sharp peak and a much smaller range compared to the prior, indicating a dramatic reduction in uncertainty of this ratio.

Although the posterior distributions of model input parameters are different from the prior, no parameter value in the posterior range is outside the ranges implemented in the orginal model. The MCMC tool has the potential to be a very useful tool in the validation process and in particular for uncertainty reduction of model input parameters and consequently model output.

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 model validation is postulated as a confidence building and long-term iterative process (Hassan, 2004a). Model validation should be viewed as a process not an end result.

Following Hassan (2004b), an approach is proposed for the validation process of stochastic groundwater models. The approach is briefly summarized herein and detailed analyses of acceptance criteria for stochastic realizations and of using validation data to reduce input parameter uncertainty are presented and applied to two case studies.

During the validation process for stochastic models, a question arises as to the sufficiency of the number of acceptable model realizations (in terms of conformity with validation data). Using a hierarchical approach to make this determination is proposed. This approach is based on computing five measures or metrics and following a decision tree to determine if a sufficient number of realizations attain satisfactory scores regarding 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 using the first case study and assuming field data consistent with the model or significantly different from 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 they are appropriate measures for evaluating model realizations.

The use of validation data to constrain model input parameters is shown for the second case study using a Bayesian approach known as Markov Chain Monte Carlo. The approach shows a great potential to be helpful in the validation process and in incorporating prior knowledge with new field data to derive posterior distributions for both model input and output.

Acknowledgements

The author would like to express his special thanks to Greg Pohll, Jenny Chapman, and Karl Pohlmann for very helpful comments and stimulating discussions and also for the help on the statistical evaluation of the acceptance criteria. Appreciation is extended to Hesham Bekhit for his help on the Metropolis Algorithm. This work was performed under the auspices of the U.S. Department of Energy by the Desert Research Institute under contract number DE-AC52-00NV13609.

For Further Information

This chapter attempts to introduce a newly proposed validation methodology for numerical groundwater models cast in a stochastic framework. Research on model validation is extensively reported in the literature, but mostly focuses on process models and definitions of the term. In the area of toxic waste management, a number of authors (e.g., Moran and Mezgar, 1982; Huyakorn et al., 1984; van der Heijde et al., 1985; van der Heijde, 1987; Beljin, 1988) have considered the question of whether a model used in a safety assessment program is valid in making appropriate long-term predictions. During the late 1980s, an effort was made to establish a groundwater research data center for the validation of subsurface flow and transport models (Miller and van der Heijde, 1988; van der Heijde et al., 1989).

In the area of nuclear waste management, the need to validate groundwater models has received increased emphasis. This has led to institutionalized and publicized programs for validation of hydrogeological models. A number of international cooperative projects such as INTRACOIN (1984, 1986), HYDROCOIN (Grundfelt, 1987; Grundfelt et al., 1990), INTRAVAL (Andersson et al., 1989; Nicholson, 1990), STRIPA (Herbert et al., 1990), CHEMVAL (Broyd et al., 1990), BIOMOVS (SSI, 1990) were devoted to the validation of models. Model validation was also extensively discussed in symposia including GEOVAL87 (1987), GEOVAL90 (1990) and GEOVAL94 (1994). The journal *Advances in Water Resources* dedicated two special issues to the topic of model validation (AWR, 1992a, b). Additionally, a wealth of literature has been published on validation in the field of systems engineering and operations research (Tsang, 1991), some of which may be useful for subsurface model validation. Examples cited by Tsang (1991) include Balci (1988, 1989), Balci and Sargent (1981, 1982, 1984), Gass (1983), Gass and Thompson (1980), Oren (1981), Sargent (1984, 1988), Schruben (1980), and Zeigler (1976).

The Swedish Nuclear Power Inspectorate, SKI, initiated and completed three international cooperation projects to increase the understanding and credibility of models describing groundwater flow and radionuclide transport. The INTRACOIN project is the first of these, and it focused on verification and validation of transport models. The HYDROCOIN study was the second study and represented an international cooperative project for testing groundwater-modeling strategies for performance assessment of nuclear waste disposal. The SKI initiated the study in 1984, and the technical work was finalized in 1987 (Swedish Nuclear Power Inspectorate, 1987). The participating organizations were regulatory authorities as well as implementing organizations in 10 countries. The study was devoted to testing of groundwater flow models and was performed at three levels: computer code verification, model validation, and sensitivity/uncertainty analysis.

Based upon lessons learned from INTRACOIN and HYDROCOIN, international consensus grew prior to and during the GEOVAL Symposium in Stockholm in April 1987 to begin a new project dealing with validation of geosphere transport models. This new international cooperative project, named INTRAVAL, began in October 1987. As with the preceding projects, INTRAVAL was organized and managed by the SKI.

The INTRAVAL project was established to evaluate the validity of mathematical models for predicting the potential transport of radioactive substances in the geosphere (Swedish Nuclear Power Inspectorate, 1990). The unique aspect of INTRAVAL was the interaction between the experimentalists and modelers simulating the selected test cases for examining model validation issues. The test cases selected consisted of laboratory and field transport experiments and natural analogue studies that incorporate hydrogeologic and geochemical processes relevant to safety assessments of radioactive waste repositories.

References

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

Andersson, K., B. Grundfelt, A. Larsson, and T. Nicolson. 1989. INTRAVAL as an integrated international effort for geosphere model validation – A status report. Proceedings of Symposium on Safety Assessment of Radioactive Waste Repositories, Paris, October 9-13, OECD Nuclear Energy Agency.

AWR. 1992a. Special issue: Validation of Geo-Hydrological Models - Part 1. *Advances in Water Resources* 15, no. 1.

AWR. 1992b. Special issue: Validation of Geo-Hydrological Models - Part 2. *Advances in Water Resources* 15, no. 3.

Balci, O. 1988. Credibility assessment of simulation results: The state of the art, methodology and validation. *Simulation Series*, *The Society for Computer Simulation* 19, no. 1: 19-25.

Balci, O. 1989. How to assess the acceptability and credibility of simulation results. In *Proceedings of the 1989 Winter Simulation Conference*, Washington, DC. Edited by E. MacNair, K. Musselman, and P. Heidelberger.

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

Balci, O., and R.G. Sargent. 1982. Validation of multivariate response simulation models by using Hotelling's two-sample T^2 test. *Simulation* 39, no. 6: 185-192.

Balci, O., and R.G. Sargent. 1984. A bibliography on the credibility, assessment and validation of simulation and mathematical models. *Simuletter* 15, no. 3: 15-27.

Bates, B. C., and E. P. Campbell. 2001. A Markov chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall runoff modeling. *Water Resour. Res*. 37(4), 937– 947.

Beljin, M.S. 1988. Testing and validation of models for solute transport in groundwater: Code intercomparison and evaluation of validation methodology. International Ground Water Modeling Center, Holcomb Institute, Butler Univ., Indianapolis, Indiana. Report GWMI 88-11.

Bredehoeft, J.D., and L.F. Konikow. 1992. Reply to comment. *Advances in Water Resources* 15, 371-372.

Bredehoeft, J.D., and L.F. Konikow. 1993. Ground water models: Validate or invalidate. *Ground Water* 31(2), 178-179.

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

Broyd, T., D. Read, and B. Come. 1990. The CHEMVAL Project: An international study aimed at the verification and validation of equilibrium speciation and chemical transport computer programs. In *Proceedings of GEOVAL90 Symposium*, Stockholm, May 14-17, 1990, Swedish Nuclear Power Inspectorate (SKI), Stockholm.

Campbell, E. P., D. R. Fox, and B. C. Bates. 1999. A Bayesian approach to parameter estimation and pooling in nonlinear flood event models. *Water Resour. Res*. 35(1), 211 – 220.

de Marsily, G., P. Combes, and P. Goblet. 1992. Comment on 'Groundwater models cannot be validated,' by L.F. Konikow and J. D. Bredehoeft. *Advances in Water Resources* 15, 367-369.

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.

Eisenberg, N., M. Federline, B. Sagar, G. Wittmeyer, J. Andersson, and S. Wingefors. 1994. Model validation from a regulatory perspective: A summary. In *GEOVAL 94, Validation Through Model Testing, Proceedings of an NEA/SKI Symposium*, Paris, France, 11-14 October, 421-434.

Freer, J., K. Beven, and B. Ambroise. 1996. Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach. *Water Resour. Res.* 32(7), 2161– 2173.

Gass, S.I. 1983. Decision-aiding models: Validation, assessment, and related issues for policy analysis. *Operations Research* 31, no. 4: 601-663.

Gass, S.I., and B.W. Thompson. 1980. Guidelines for model evaluation: An abridged version of the U.S. General Accounting Office exposure draft. *Operations Research* 28, no. 2: 431-479.

GEOVAL87. 1987. *Proceedings of Symposium on Verification and Validation of Geosphere Performance Assessment Models*, organized by Swedish Nuclear Power Inspectorate, Stockholm, Sweden, April 7-9.

GEOVAL90. 1990. *Proceedings of Symposium on Validation of Geosphere Flow and Transport Models,* organized by Swedish Nuclear Power Inspectorate, Stockholm, Sweden, May 14-17.

GEOVAL94. 1994. *Proceedings of Symposium on Verification Through Model Testing*, organized by OECD Nuclear Energy Agency and the Swedish Nuclear Power Inspectorate, Paris, France, October 11-14.

Grundfelt, B., B. Lindbom, A. Larsson, and K. Andersson. 1990. HYDROCOIN level 3 - Testing methods for sensitivity/uncertainty analysis. *Proceedings of GEOVAL9O Symposium*, Stockholm, May 14-17, 1990. Swedish Nuclear Power Inspectorate, (SKI), Stockholm.

Grundfelt, B. 1987. The HYDROCOIN Project - Overview and results from level one. *Proceedings of International GEOVAL-87 Symposium*, Swedish Nuclear Power Inspectorate, (SKI), Stockholm. April 7-9, 1987.

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.

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

Hassanizadeh, S.M. 1990. Panel discussion. In *GEOVAL-90, Symposium on Validation of Geosphere Performance Assessment Models*, Stockholm, Sweden, 14-17 May, 631-658.

Herbert, A., W. Dershowitz, J. Long, and D. Hodgkinson. 1990. Validation of fracture flow models in the Stripa project. *Proceedings of GEOVAL90 Symposium*, Stockholm, May 14-17, Swedish Nuclear Power Inspectorate (SKI).

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.

Huyakorn, P.S., A.G. Kretschek, R.W. Broome, J.W. Mercer, and B.H. Lester. 1984. Testing and validation of models for simulating solute transport in groundwater. International Ground Water Modeling Center, Holcomb Research Institute, Butler University, Indianapolis, IN. Report GWMI 84-13.

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

INTRACOIN. 1986. Final report levels 2 and 3, Model validation and uncertainty analysis. Report SKI 86:2, Swedish Nuclear Power Inspectorate, Stockholm, Sweden.

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

Konikow, L.F., and Bredehoeft, J.D. 1993. The myth of validation in groundwater modeling. In *Proceedings 1993 Groundwater Modeling Conference*, IGWMC, Golden, CO, pp. A-4.

Kuczera, G. 1983. Improved parameter inference in catchment models: 1. Evaluating parameter uncertainty. *Water Resour. Res*. 19(5), 1151–1162.

Kuczera, G., and M. Mroczkowski (1998), Assessment of hydrologic parameter uncertainty and the worth of multiresponse data, *Water Resour. Res*., 34(6), 1481– 1489.

Kuczera, G., and E. Parent. 1998. Monte Carlo assessment of parameter uncertainty in conceptual catchment models: The Metropolis algorithm. *J. Hydrol.* 211, 69– 85.

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

Marshall, L., D. Nott, and A. Sharma. 2004. A comparative study of Markov chain Monte Carlo methods for conceptual rainfall-runoff modeling. *Water Resour. Res.* 40, W02501, doi:10.1029/2003WR002378.

Miller, R.E., and P.K.M. van der Heijde. 1988. A groundwater research data center for model validation. *Proceedings of the Indiana Water Resources Assoc. Symposium*, June 8-10, Greencastle, IN.

Moran, M.S., and L.J. Mezgar. 1982. Evaluation factors for verification and validation of low-level waste disposal site models. Oak Ridge National Laboratory, Oak Ridge, TN. Report DOE/OR/21400-T119.

National Research Council, 1990. *Groundwater Models; Scientific and Regulatory Applications*. National Academy Press, Washington, D.C., 303 pp.

Neuman, S.P. 1992. Validation of safety assessment models as a process of scientific and public confidence building. In *Proceedings of HLWM Conference*, Vol. 2, Las Vegas.

Nicholson, T.J. 1990. Recent accomplishments in the INTRAVAL Project - A status report on validation efforts. *Proceedings of GEOVAL90 Symposium*, Stockholm, Sweden, May 14-17, Swedish Nuclear Power Inspectorate (SKI).

Oren, T. 1981. Concepts and criteria to assess acceptability of simulation studies. A Frame of Reference. *Communication of the ACM* 24, no. 4: 180-189.

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

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

Pohlmann, K., A.E Hassan, and J. Chapman. 2002. Modeling density-driven flow and radionuclide transport at an underground nuclear test: Uncertainty analysis and effect of parameter correlation. *Water Resour. Res.* 38(5), 10.1029/2001WR001047.

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. Division of Hydrologic Sciences, Desert Research Institute, Publication No. 45184-revised, pp. 197.

Refsgaard, J.C. 2001. Discussion of model validation in relation to the regional and global scale. In *Model Validation: Perspectives in Hydrological Sciences*, M. G. Anderson and P. D. Bates (Eds.). John Wiley and Sons, New York.

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 Resour. Res.* 39(12),1356-1370.

Sargent, R.G. 1984. Simulation Model Validation. In *Simulation and Model-based Methodologies: An Integrative View*. Edited by Oren *et al.*, Springer-Verlag.

Sargent, R.G. 1988. A tutorial on validation and verification of simulation models. *Proceedings of 1988 Winter Simulation Conference*. Edited by M. Abrams, P. Haigh, and J. Comfort, 33-39.

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.

Schruben, L.W. 1980. *Establishing the Credibility of Simulations - The Art and the Science*. Prentice-Hall.

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

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

SSI (Swedish National Institute of Radiation Protection). 1990. *Proceedings of Symposium and Workshop on the Validity of Environmental Transfer Models (BIOMOVS)*, October 8-12, 1990. Swedish National Institute of Radiation Protection (SSI), Stockholm.

Swedish Nuclear Power Inspectorate. 1987. The International HYDROCOIN Project—Background and Results. OECD, Paris.

Swedish Nuclear Power Inspectorate. 1990. The International INTRAVAL Project—Background and Results. OECD, Paris.

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

van der Heijde, P.K.M., P.S. Huyakorn, and J.W. Mercer. 1985. Testing and validation of groundwater models. *Proceedings of the NWWA/IGWMC Conference on "Practical Applications of Ground Water Models*." August 19-20, Columbus, OH.

van der Heijde, P.K.M. 1987. Quality assurance in computer simulations of groundwater contamination. *Environmental Software* 2, no. 1.

van der Heijde, P.K.M., W.I.M. Elderhorst, R.A. Miller, and M.F. Trehan. 1989. The establishment of a groundwater research data center for validation of subsurface flow and transport models. International Ground Water Modeling Center, Holcomb Research Institute, Butler Univ., Indianapolis, In Report GWMI 89-01.Vogel, R. M., Sankarasubramanian. 2003. Validation of a watershed model without calibration, *Water Resour. Res.* 39(10), 1292, doi:10.1029/2002WR001940.

Unsworth, M., W. Soyer, and V. Tuncer, 2005. Magnetotelluric measurements for determining the subsurface salinity and porosity structure of Amchitka Island, Alaska. In *Biological and Geophysical Aspects of Potential Radionuclide Exposure in the Amchitka Marine Environment*, edited by Powers et al. Consortium for Risk Evaluation with Stakeholder Participation, Institute for Responsible Management, Piscataway, New Jersey.

Zeigler, B.P. 1976. *Theory of Modeling and Simulation*. John Wiley and Sons, Inc., New York.

Glossary

Model: An abstraction or a simple representation of a real system or process

Conceptual Model: A hypothesis for how a system or a process operates

Mathematical Model: A quantitative expression of the system processes, forces, and events

Computer Code: An algorithm to implement the mathematical model and perform the model computations

Generic Models: The computer codes that are used to solve the mathematical flow and transport equations

Site-specific Models: The computer codes combined with the conceptual models (model structure), input data and boundary conditions for a particular geographical area

Research or Analysis Models: Models used for studying and understanding different phenomena in the subsurface and they usually rely on hypothetical domains or very well characterized field sites

Predictive or Decision-making Models: Models that are mainly used to support and aid a regulatory decision regarding a subsurface issue

Model Calibration: The process of tuning the model to identify the independent input parameters by fitting the model results to some field data or experimental data, which usually represent the dependent system parameters

History Matching: Using historical field data during the model building and calibration process and trying to match them before using the model to predict future system response

Computer Code Verification: Verification of a mathematical model or its computer code; it is obtained when it is shown that the model behaves as intended and that the equations are correctly encoded and solved

Model Verification: A process aimed at establishing a greater confidence in the model by using a set of calibrated parameter values and stresses to reproduce a second set of field data that is available during the model building process

Model Validation: A process, not an end result, aimed at building confidence in the model predictions through a structured set of tests and evaluations with trigger mechanisms that force the process back to the model building stage if model is not consistent with field data collected for validation

Research Model Validation: Validation of a research model that helps one understand processes, uncertainties, etc.

Predictive Validation: Performing a validation process for a site-specific, predictive model for the goal of helping the decision-making process

Validation Target: The model output for which filed data will be collected for validation purposes

Bayesian Inference: An alternative to the classical approach to statistical analysis that benefits from prior knowledge as well as newly collected data to quantify parameter uncertainty

Markov Chain Monte Carlo: A form of Bayesian inference that can be used to simulate the posterior distribution of model parameters

Prior Distribution: The parameter distribution that summarizes all the knowledge about the parameter before collecting new data

Posterior Distribution: The parameter distribution that is obtained through Bayesian inference and it relies on both prior knowledge and newly collected data

Likelihood Function: A probability function describing the probability of seeing the observed data (newly collected data) given that a certain model/assumption is true

Density-driven Flow: A groundwater flow problem in which the flow depends on (or is driven by) the variations in groundwater density (e.g., due to thermal effects or salinity effects)

Transition Zone: The varying salinity zone separating shallow freshwater (mainly from recharge) from deep saltwater (from the ocean or the sea)

List of Figures

Figure 14. Comparison between the prior (solid lines) and posterior (dotted lines) distributions obtained using validation data and the MCMC approach for conductivity, *K*, recharge, *R*, and recharge-conductivity ratio.

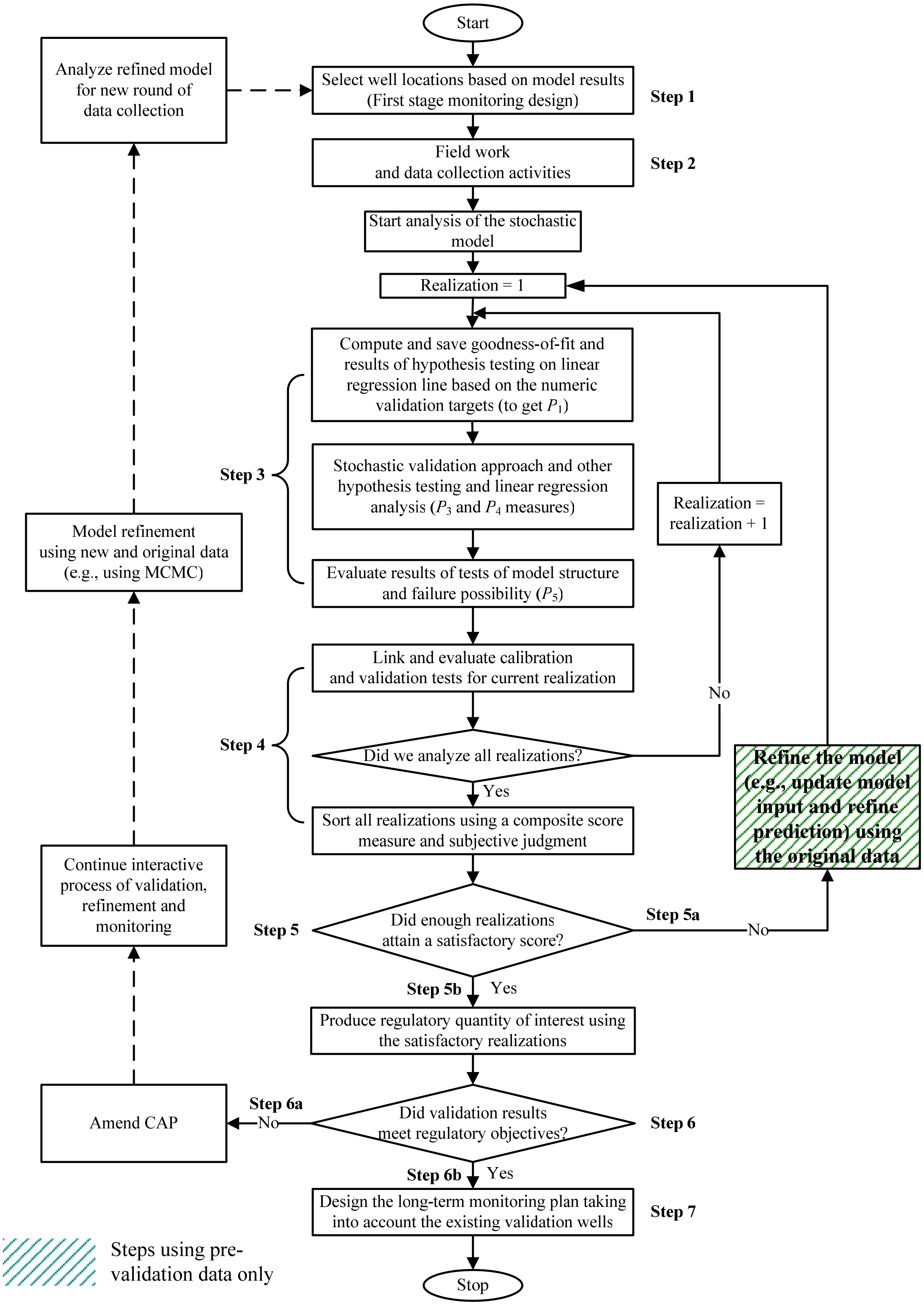

Start
Select well locations based on model results (First stage monitoring design)
Step 1
Analyze refined model for new round of data collection
Field work and data collection activities
Step 2
Start analysis of the stochastic model
Realization = 1
Compute and save goodness-of-fit and results of hypothesis testing on linear regression line based on the numeric validation targets (to get P_1)
Step 3
Stochastic validation approach and other hypothesis testing and linear regression analysis (P_3 and P_4 measures)
Realization = realization + 1
Model refinement using new and original data (e.g., using MCMC)
Evaluate results of tests of model structure and failure possibility (P_5)
Link and evaluate calibration and validation tests for current realization
No
Step 4
Did we analyze all realizations?
Yes
Refine the model (e.g., update model input and refine prediction) using the original data
Sort all realizations using a composite score measure and subjective judgment
Continue interactive process of validation, refinement and monitoring
Step 5
Did enough realizations attain a satisfactory score?
Step 5a
No
Step 5b
Yes
Produce regulatory quantity of interest using the satisfactory realizations
Step 6a
Amend CAP
No
Did validation results meet regulatory objectives?
Step 6
Step 6b
Yes
Design the long-term monitoring plan taking into account the existing validation wells
Step 7
Stop
Steps using pre-validation data only

Figure 1

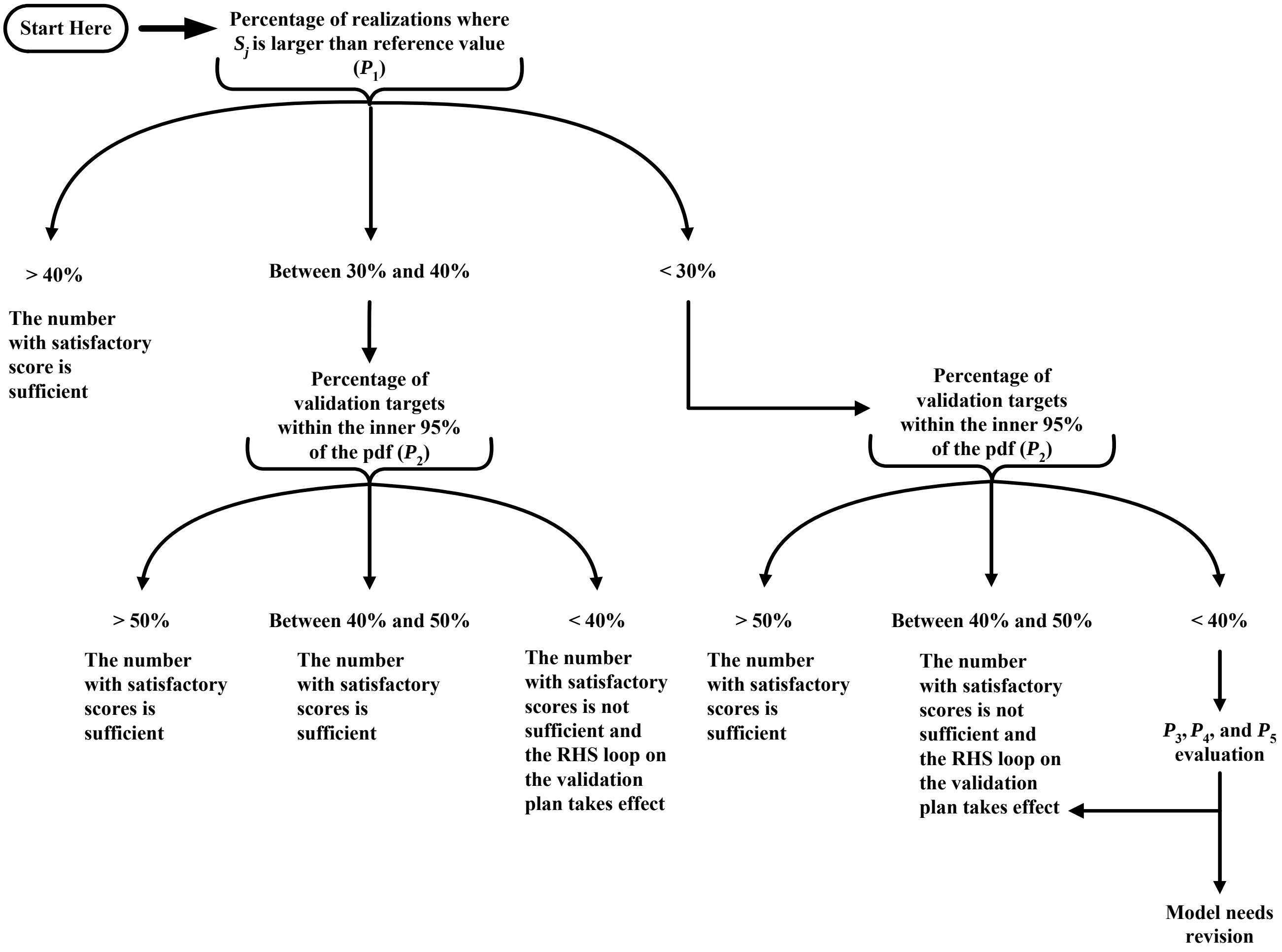

Start Here
Percentage of realizations where S_j is larger than reference value (P_1)
> 40%
The number with satisfactory score is sufficient
Between 30% and 40%
< 30%
Percentage of validation targets within the inner 95% of the pdf (P_2)
Percentage of validation targets within the inner 95% of the pdf (P_2)
> 50%
The number with satisfactory scores is sufficient
Between 40% and 50%
The number with satisfactory scores is sufficient
< 40%
The number with satisfactory scores is not sufficient and the RHS loop on the validation plan takes effect
> 50%
The number with satisfactory scores is sufficient
Between 40% and 50%
The number with satisfactory scores is not sufficient and the RHS loop on the validation plan takes effect
< 40%
P_3, P_4, and P_5 evaluation
Model needs revision

Figure 2

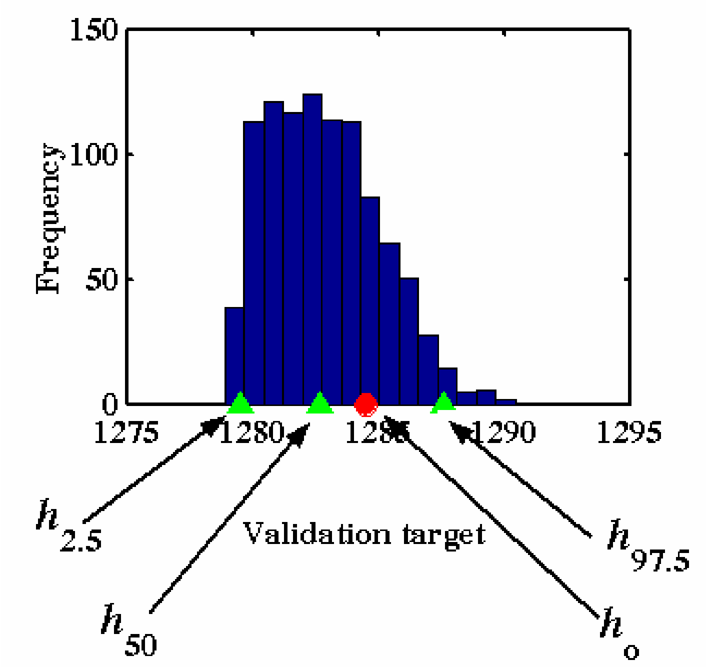

150
100
50
0
Frequency
1275
1280
1285
1290
1295
Validation target
$h_{2.5}$
h_{50}
$h_{97.5}$
h_o

Figure 3

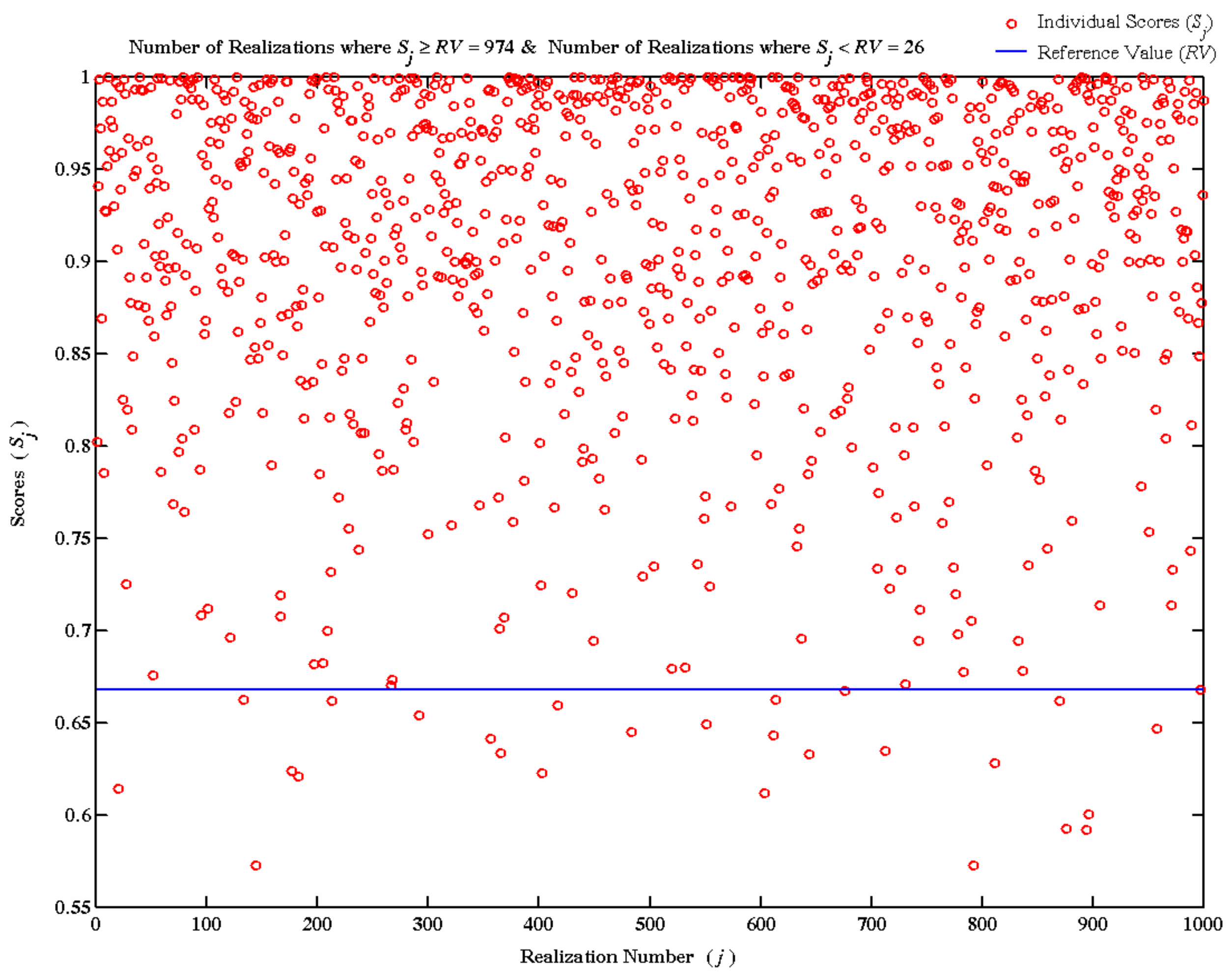

Number of Realizations where $S_j \geq RV = 974$ & Number of Realizations where $S_j < RV = 26$
Individual Scores (S_j)
Reference Value (RV)
Scores (S_j)
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0
100
200
300
400
500
600
700
800
900
1000
Realization Number (j)

Figure 4

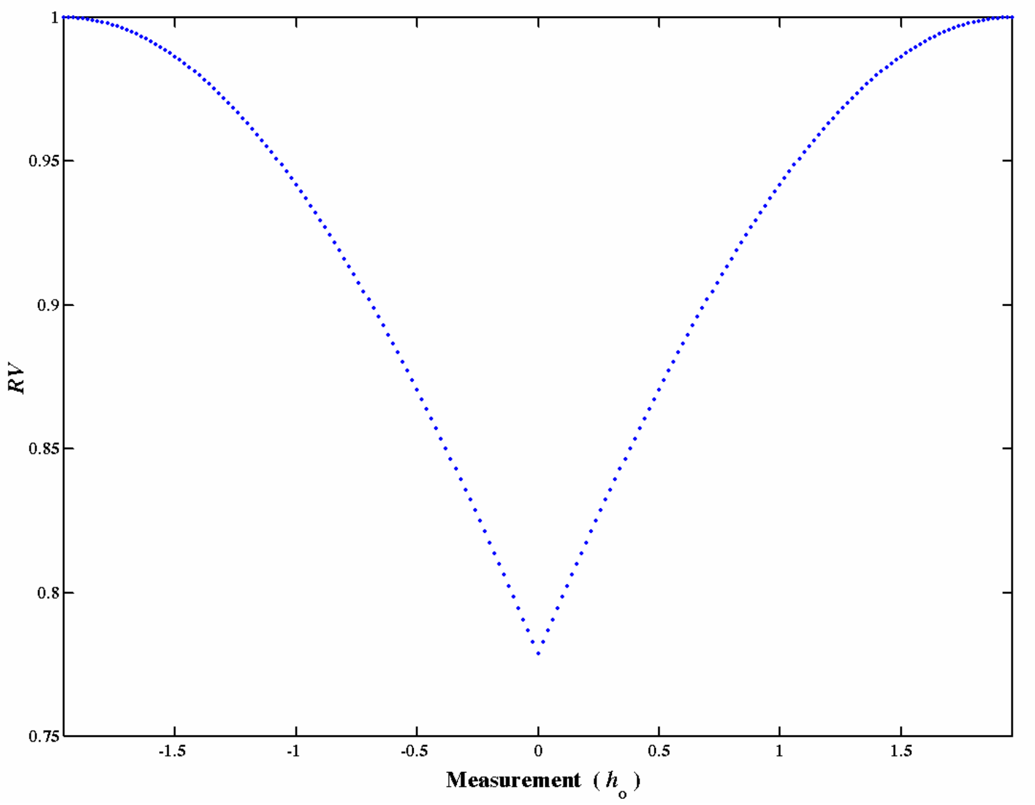


Figure 5

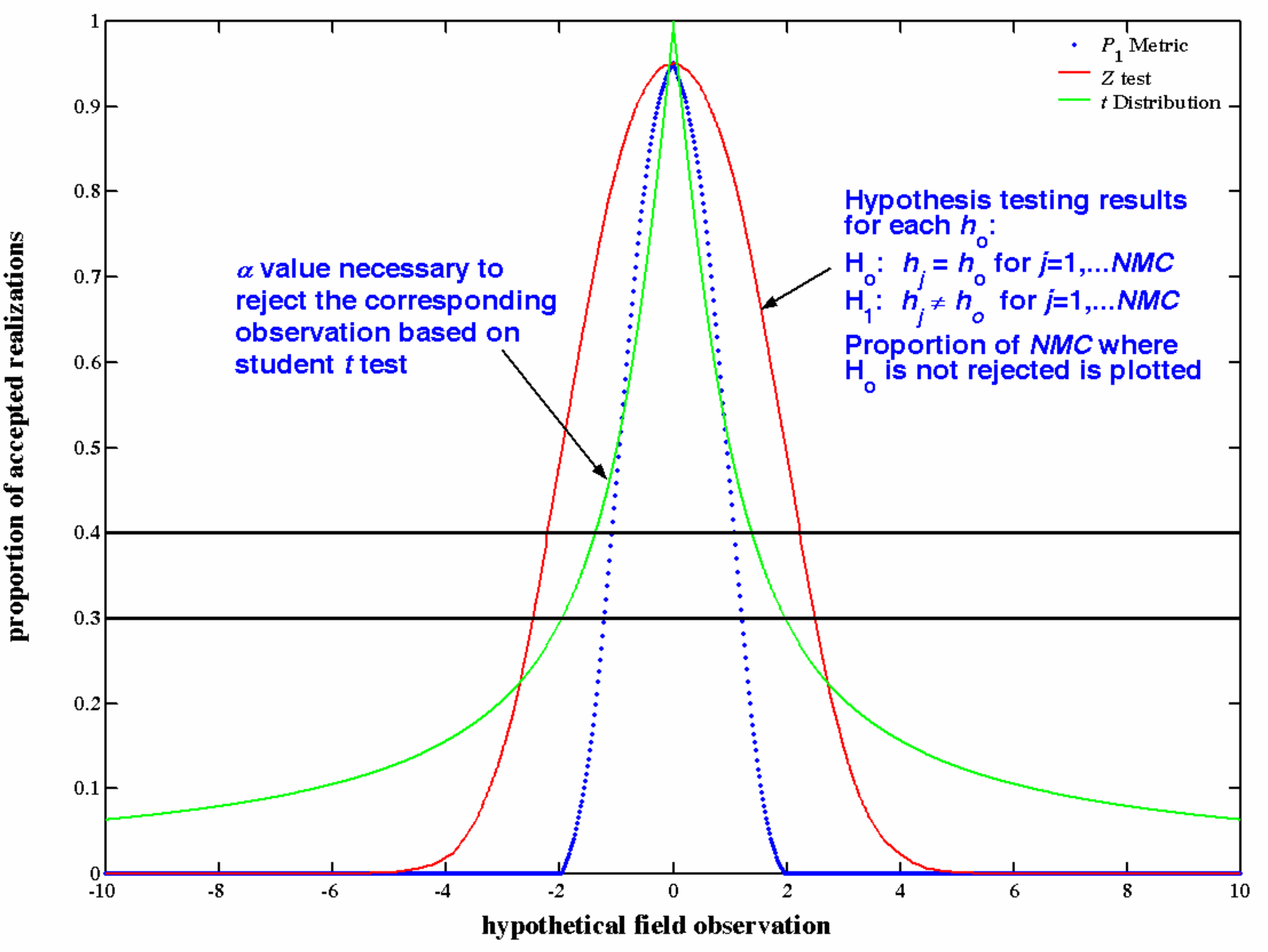
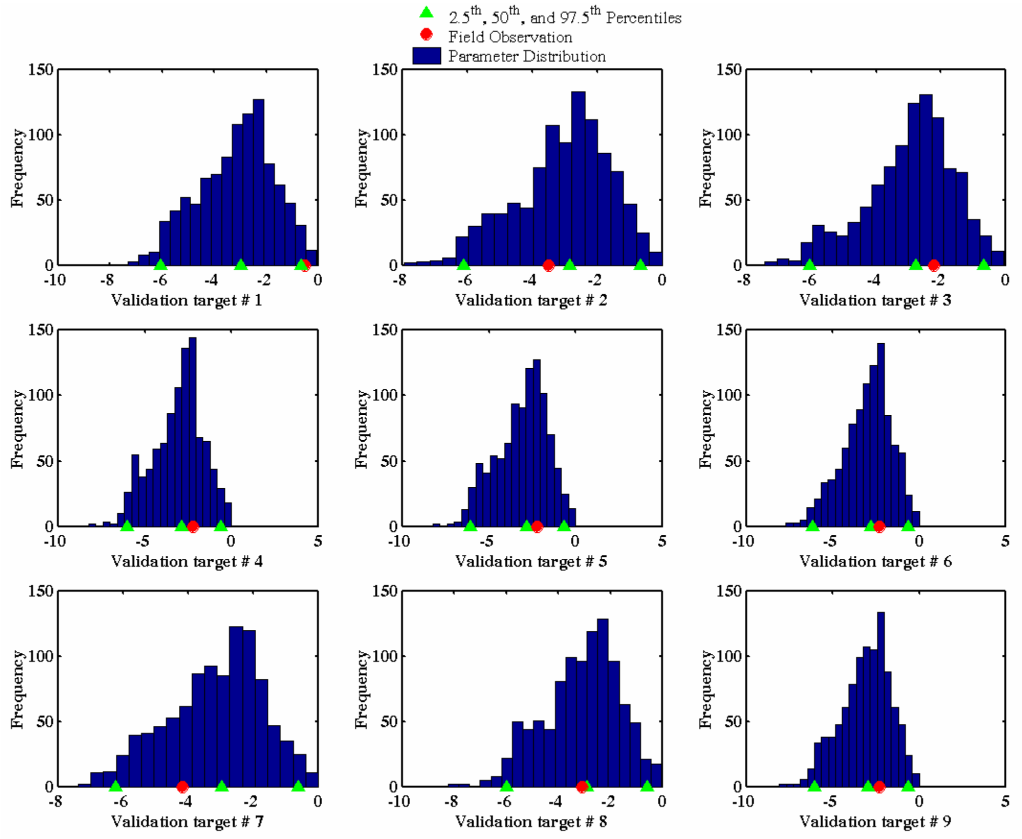
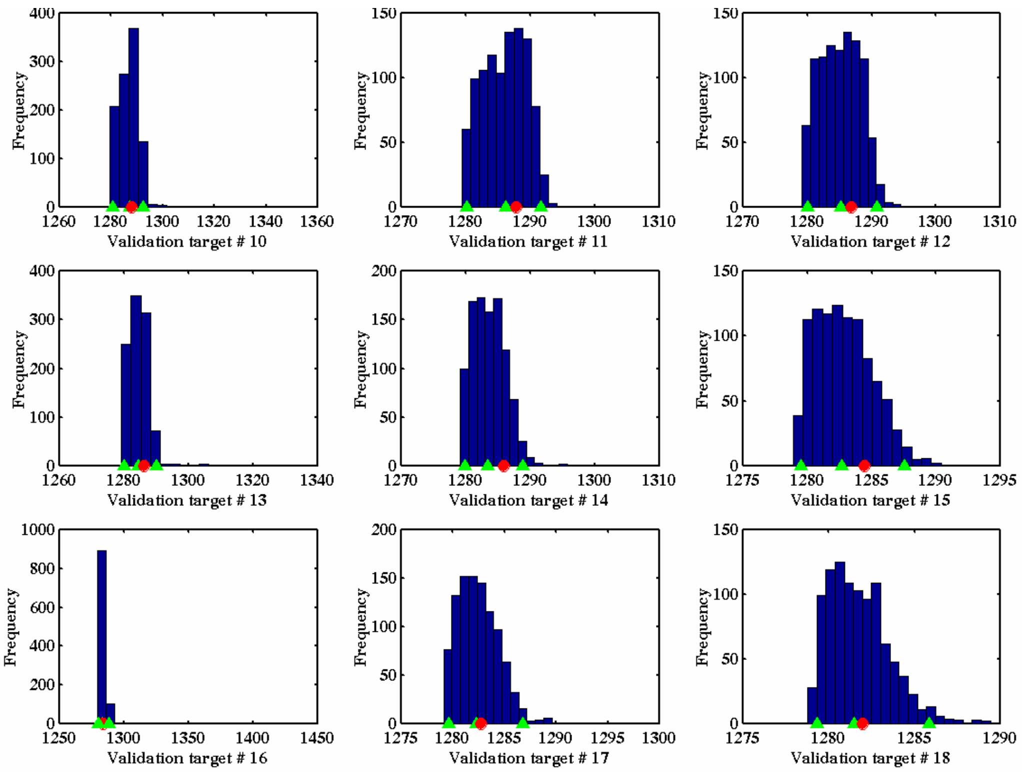

1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-10
-8
-6
-4
-2
0
2
4
6
8
10
proportion of accepted realizations
hypothetical field observation
P_1 Metric
Z test
t Distribution
α value necessary to reject the corresponding observation based on student t test
Hypothesis testing results for each h_o:
H_o: $h_j = h_o$ for j=1,...NMC
H_1: $h_j \neq h_o$ for j=1,...NMC
Proportion of NMC where H_o is not rejected is plotted

Figure 6



2.5th, 50th, and 97.5th Percentiles
Field Observation
Parameter Distribution
Frequency
Validation target # 1
Validation target # 2
Validation target # 3
Validation target # 4
Validation target # 5
Validation target # 6
Validation target # 7
Validation target # 8
Validation target # 9
0
50
100
150
-10
-8
-6
-4
-2
-5
5



Frequency
Validation target # 10
Validation target # 11
Validation target # 12
Validation target # 13
Validation target # 14
Validation target # 15
Validation target # 16
Validation target # 17
Validation target # 18
0
50
100
150
200
300
400
600
800
1000
1250
1260
1270
1275
1280
1285
1290
1295
1300
1310
1320
1340
1350
1360
1400
1450

Figure 7

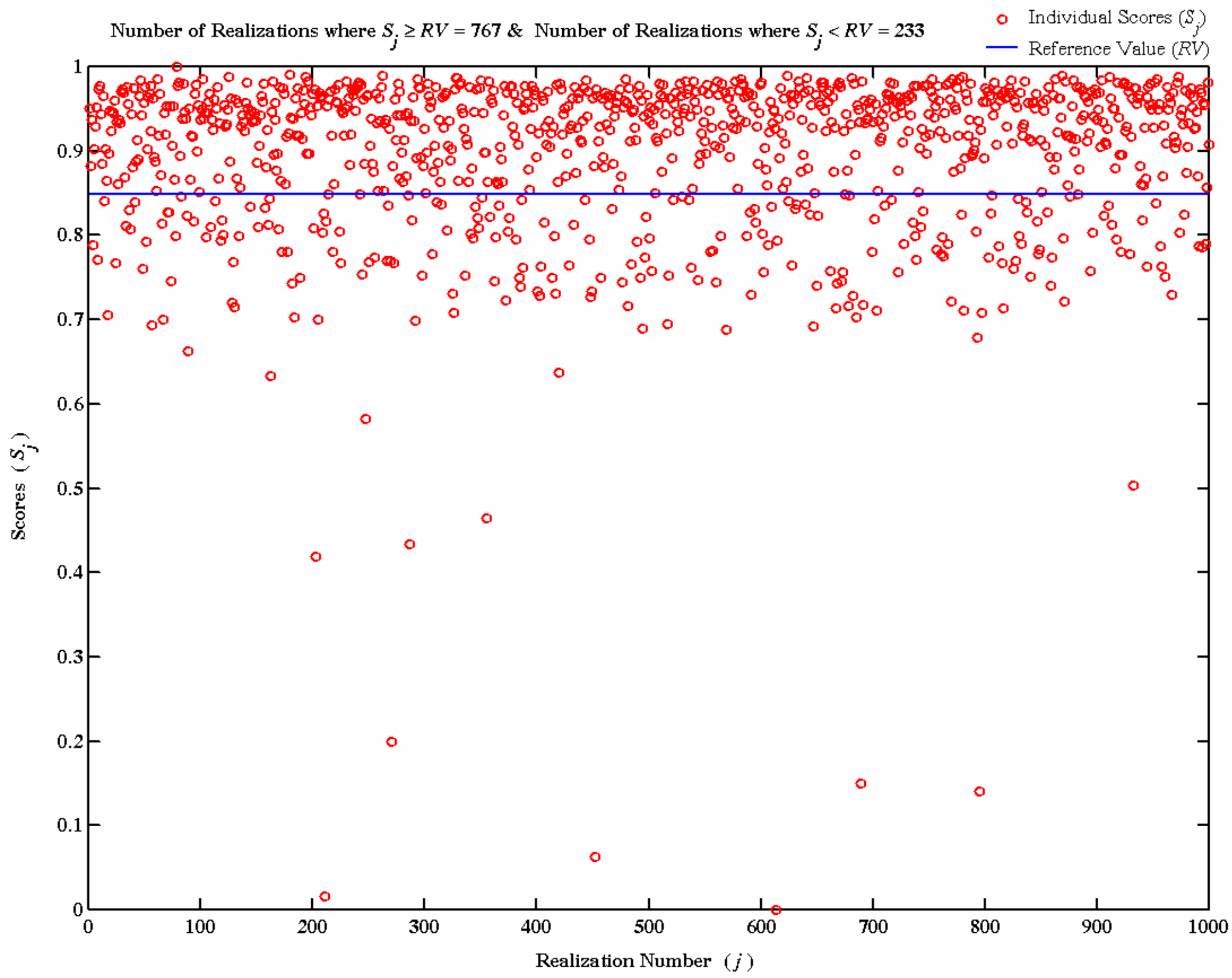
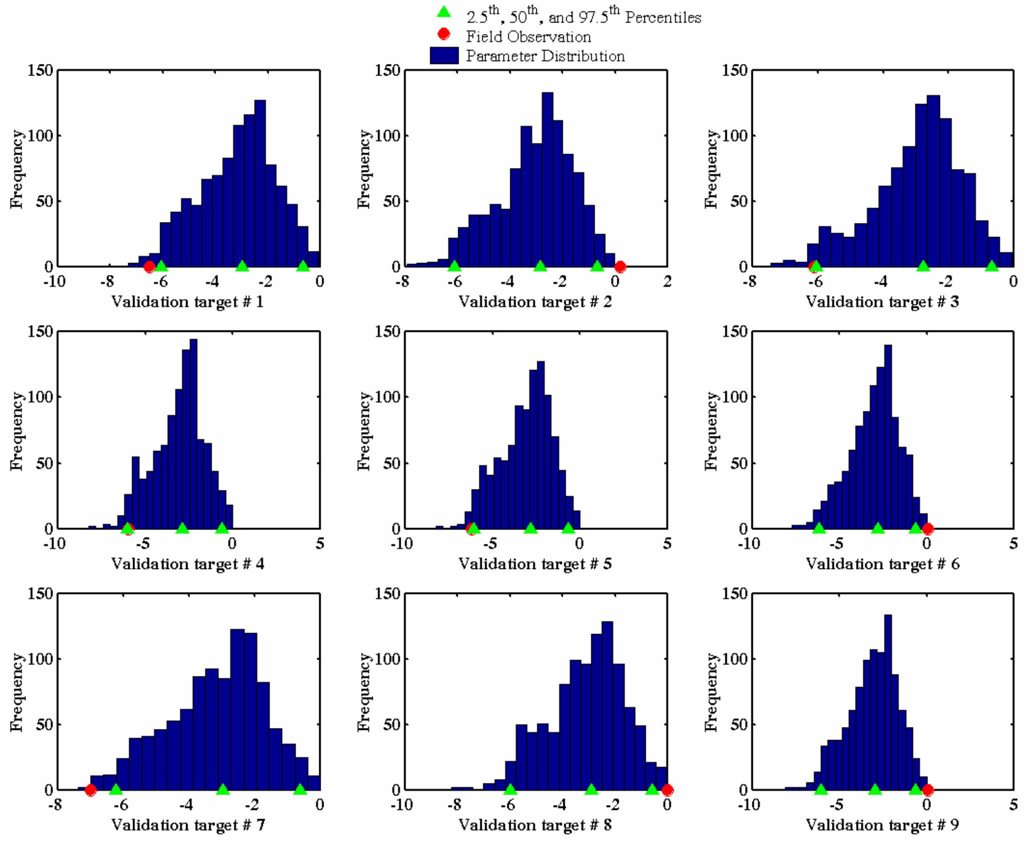


Figure 8


▲ 2.5th, 50th, and 97.5th Percentiles
● Field Observation
■ Parameter Distribution
Frequency
0
50
100
150
-10
-8
-6
-4
-2
0
2
5
Validation target # 1
Validation target # 2
Validation target # 3
Validation target # 4
Validation target # 5
Validation target # 6
Validation target # 7
Validation target # 8
Validation target # 9

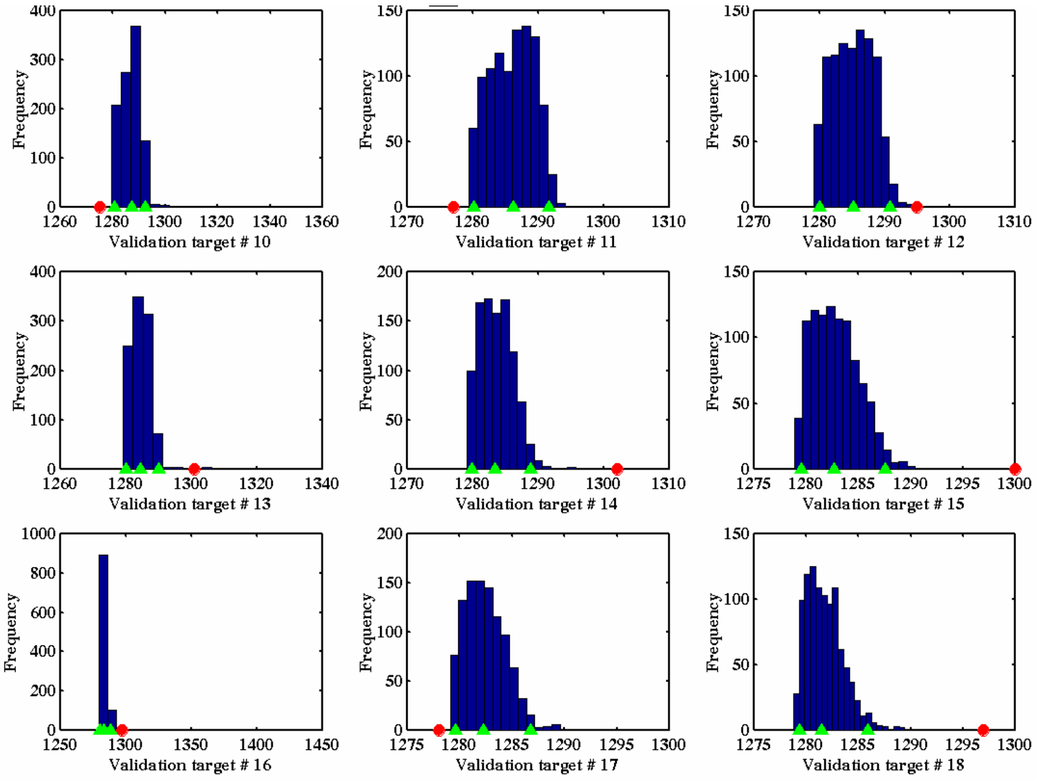

Frequency
0
50
100
150
200
300
400
600
800
1000
1250
1260
1270
1275
1280
1285
1290
1295
1300
1310
1320
1340
1350
1360
1400
1450
Validation target # 10
Validation target # 11
Validation target # 12
Validation target # 13
Validation target # 14
Validation target # 15
Validation target # 16
Validation target # 17
Validation target # 18

Figure 9

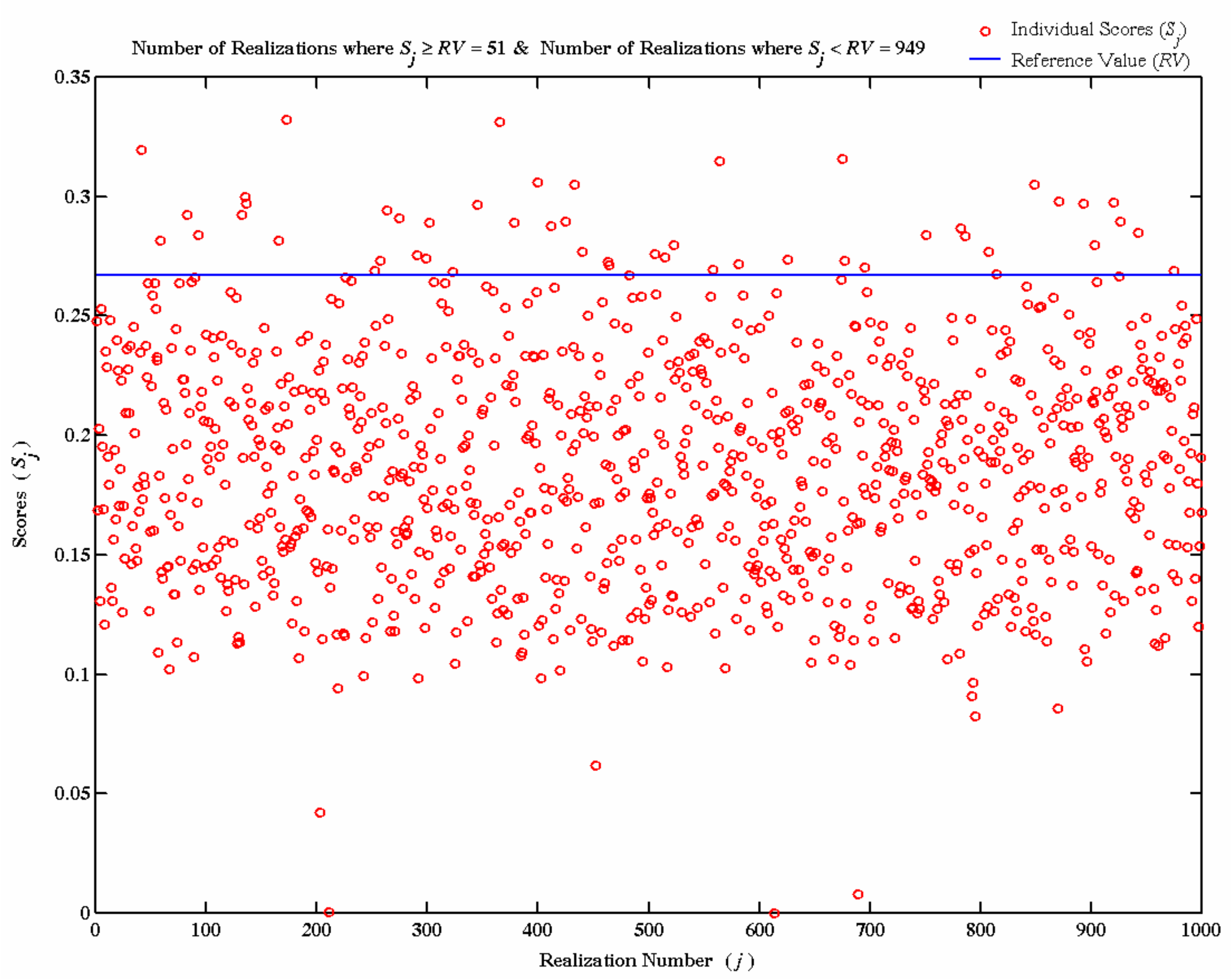


Figure 10

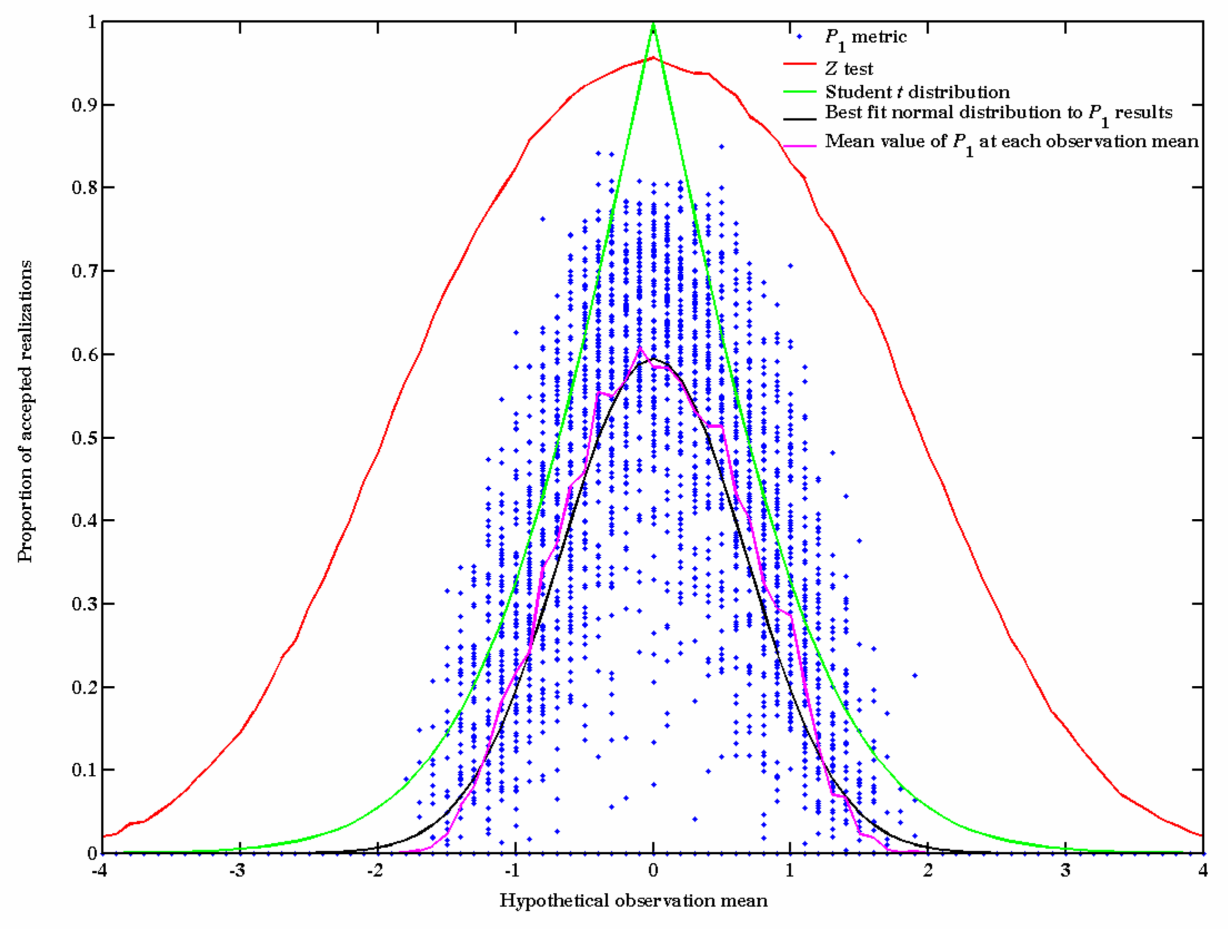

P_1 metric
Z test
Student t distribution
Best fit normal distribution to P_1 results
Mean value of P_1 at each observation mean
Proportion of accepted realizations
Hypothetical observation mean
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
-4
-3
-2
-1
0
1
2
3
4

Figure 11

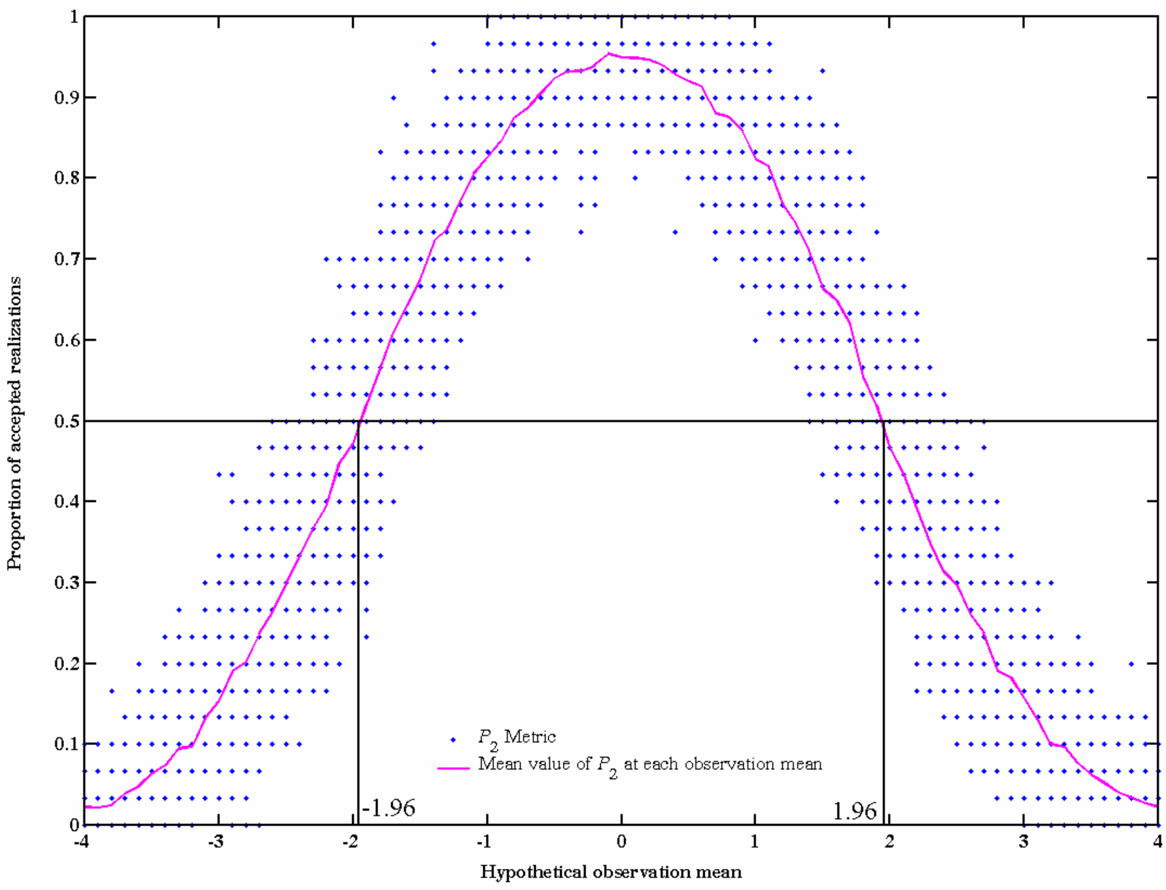

Proportion of accepted realizations
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-4
-3
-2
-1
0
1
2
3
4
Hypothetical observation mean
-1.96
1.96
P_2 Metric
Mean value of P_2 at each observation mean

Figure 12

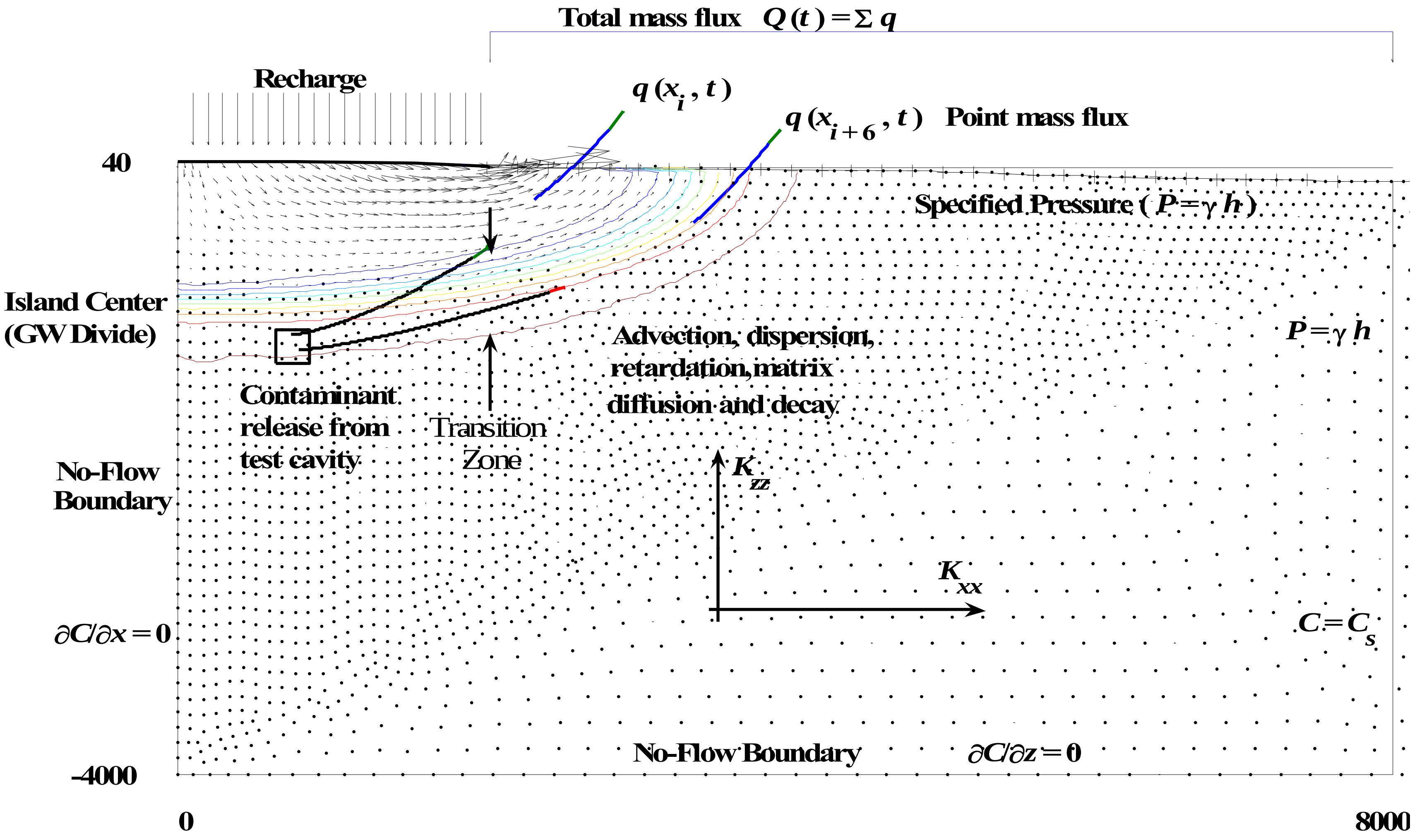

Total mass flux $Q(t) = \Sigma q$
Recharge
$q(x_i, t)$
$q(x_{i+6}, t)$ Point mass flux
40
Specified Pressure ($P = \gamma h$)
Island Center (GW Divide)
$P = \gamma h$
Advection, dispersion, retardation, matrix diffusion and decay
Contaminant release from test cavity
Transition Zone
No-Flow Boundary
K_{zz}
K_{xx}
$\partial C / \partial x = 0$
$C = C_s$
-4000
No-Flow Boundary
$\partial C / \partial z = 0$
0
8000

Figure 13

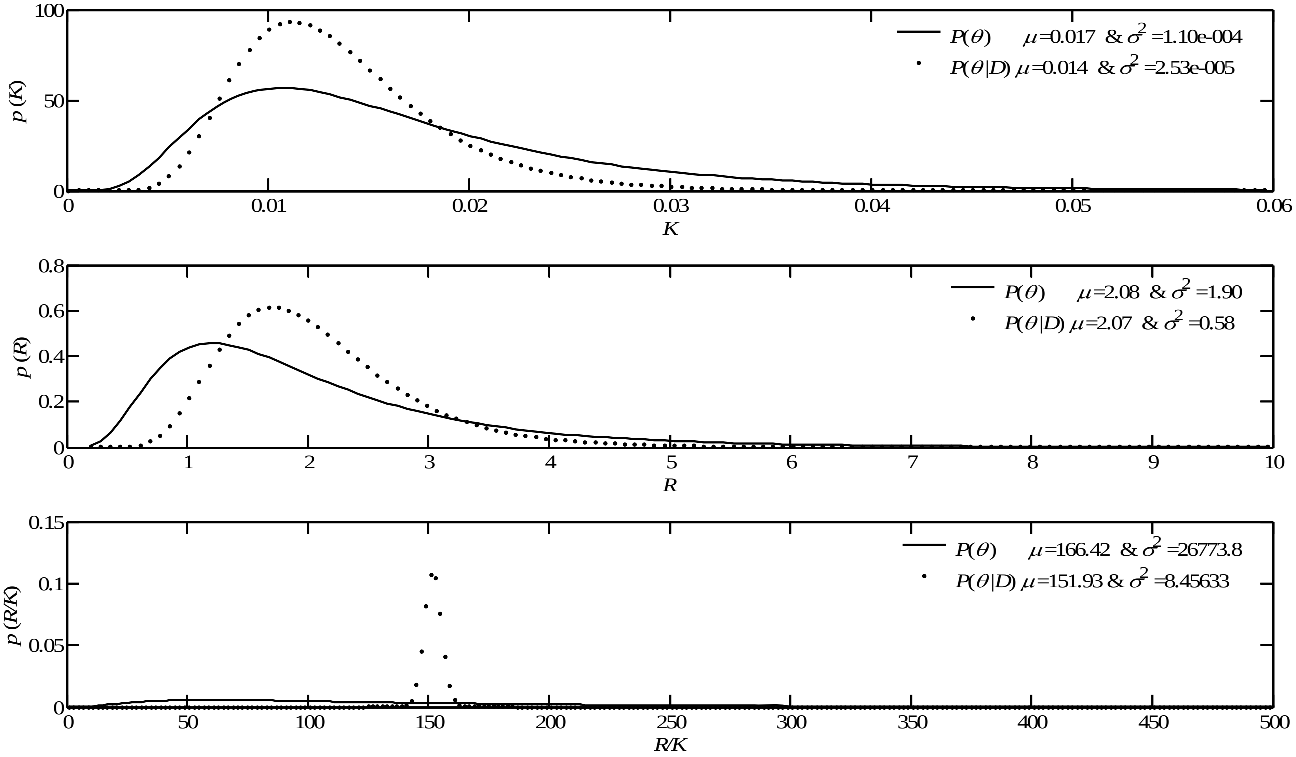

100
50
0
p (K)
0
0.01
0.02
0.03
0.04
0.05
0.06
K
P(θ) μ=0.017 & σ² =1.10e-004
P(θ|D) μ=0.014 & σ² =2.53e-005
0.8
0.6
0.4
0.2
0
p (R)
0
1
2
3
4
5
6
7
8
9
10
R
P(θ) μ=2.08 & σ² =1.90
P(θ|D) μ=2.07 & σ² =0.58
0.15
0.1
0.05
0
p (R/K)
0
50
100
150
200
250
300
350
400
450
500
R/K
P(θ) μ=166.42 & σ² =26773.8
P(θ|D) μ=151.93 & σ² =8.45633

Figure 14