DOE/NV/13609-51

Validation Analysis of the Groundwater Flow and Transport Model of the Central Nevada Test Area

Prepared by

Ahmed Hassan, Jenny Chapman, Hesham Bekhit, Brad Lyles, and Karl Pohlmann

submitted to

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

September 2006

Publication No. 45221

DOE/NV/13609-51

Validation Analysis of the Groundwater Flow and Transport Model of the Central Nevada Test Area

Prepared by

Ahmed Hassan, Jenny Chapman, Hesham Bekhit, Brad Lyles, and Karl Pohlmann

Division of Hydrologic Sciences
Desert Research Institute
Nevada System of Higher Education

Publication No. 45221

Submitted to

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

September 2006

The work upon which this report is based was supported by the U.S. Department of Energy under Contracts #DE-AC52-00NV13609 and #DE-AC52-06NA26383. Approved for public release; further dissemination unlimited.

THIS PAGE LEFT INTENTIONALLY BLANK

EXECUTIVE SUMMARY

A proof-of-concept groundwater model postaudit period for the Central Nevada Test Area (CNTA), site of the Faultless underground nuclear test, was begun with the drilling of three monitoring/validation wells in 2005. This five-year period is prescribed by a Federal Facility Agreement and Consent Order to establish if the groundwater model developed for the site is capable of producing meaningful results with an acceptable degree of uncertainty. The wells are denoted as MV-1, MV-2, and MV-3, with locations selected to meet monitoring objectives and provide data for model validation. Completion depths for the wells varied from 1115 to 1286 m below land surface. A well and two piezometers were constructed in each borehole.

The wells provide lithologic data, water level measurements, hydraulic conductivity data, and water chemistry data for comparison to the groundwater flow model. After analyzing and interpreting the data, 19 real-number validation targets and 60 categorical values (related to lithology) were identified. The real number targets include nine head measurements, four hydraulic conductivity measurements, and six inferred (or computed) vertical head gradients in the three wells. The categorical data include 21 model layers in MV-1, where field data indicate the type of geologic unit (i.e., alluvium, tuffaceous sediments, or densely welded tuff) for each layer, 18 layers in MV-2, and 21 layers in MV-3, with known categories identified from the resistivity logs of these wells.

These data sets are used to test different model components separately and to also test the model as a whole. The model validation approach detailed in the Corrective Action Decision Document/Corrective Action Plan (CADD/CAP) for CNTA is implemented and step-by-step analysis is performed using a variety of statistical tests. Goodness-of-fit measures, linear regression analysis, hypothesis testing, a stochastic perturbation approach to validation, model structure evaluations, and lithologic comparisons are all performed for evaluating the model. The acceptance criteria presented in the CADD/CAP are also applied to the model.

Most of the tests and evaluations indicate the model has a major deficiency. In particular, measured heads are much higher than the model predicted at the elevation of the nuclear test, and are also higher in the alluvium. The field data indicate an upward vertical gradient in the upper portion of the model domain, whereas the model predicted vertical downward gradients throughout the domain at the MV-1 and MV-3 locations.

The model performed well in a number of respects. Hydraulic conductivity data fit within the distributions used in the model with the field values close to the modal values (highest frequency values) of the model distributions. However, the field data indicate much lower hydraulic conductivities than the values used in the model for the densely welded tuff. Lithology data are similar to the model assigned categories for most of the logged sections in the three wells. Water chemistry data indicate no tritium above background levels, thereby supporting the transport model finding that no far-field transport is expected to occur in the 1,000 year regulatory time frame.

A rigorous quantitative analysis relies on a number of statistical tests but lacks the value of hydrogeologic expertise and a broader view of model attributes and their performance. Therefore, a holistic model evaluation is conducted where model assumptions are reviewed in light of the MV data. The overall assessment of this holistic evaluation is that

the main conceptual components of the model are valid, but near-field conditions that were deliberately neglected in the original model are responsible for the discrepancies between the model and the validation data. The effects of the nuclear test and surrounding faults likely account for the observations in the MV wells. A preliminary revised model is developed and used to test the hypothesis that near-field conditions and faults can cause persistence of the elevated pressure pulse (the high heads observed in the MV wells). The revised model indicates that if the faults are barriers to flow, they could lead to the persistence of the elevated heads.

The revised model is a simplified representation of the system. However, it shows the potential for reproducing the head and flow patterns observed in the MV wells. Adding the necessary details (e.g., refining the model discretization, including heterogeneity, conditioning on all available data, etc.) may result in a model capable of reasonably representing the near-field conditions. Such a model would necessarily introduce significant uncertainties by virtue of the absence of data regarding the nuclear test effects and fault geometry and properties, unless near-field data are collected. Near-field data collection and model revision to incorporate nuclear-test effects on the flow system are major efforts that may or may not be needed to meet the regulatory objectives of the site. This is a decision for the U.S. Department of Energy and Nevada Division of Environmental Protection.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

CADD	Corrective Action Decision Document
CAP	Corrective Action Plan
CAIP	Corrective Action Investigation Plan
CAU	Corrective Action Unit
CNTA	Central Nevada Test Area
DOE	U.S. Department of Energy
DDA	Data Decision Analysis
FFACO	Federal Facility Agreement and Consent Order
GLUE	generalized likelihood uncertainty estimates
MCL	maximum contaminant level
NDEP	Nevada Division of Environmental Protection
NTS	Nevada Test Site
RMSE	root mean squared error
RWPT	random-walk particle-tracking
UGTA	Underground Test Area

1.0 INTRODUCTION

The Central Nevada Test Area (CNTA) is a U.S. Department of Energy (DOE) site undergoing environmental restoration. The CNTA is located about 95 km northeast of Tonopah, Nevada, and 175 km southwest of Ely, Nevada (Figure 1.1). It was the site of the Faultless underground nuclear test conducted by the U.S. Atomic Energy Commission (DOE's predecessor agency) in January 1968. The purposes of this test were to gauge the seismic effects of a relatively large, high-yield detonation completed in Hot Creek Valley (outside the Nevada Test Site [NTS]) and to determine the suitability of the site for future large detonations. The yield of the Faultless underground nuclear test was between 200 kilotons and 1 megaton (DOE, 2000). A three-dimensional flow and transport model was created for the CNTA site (Pohlmann *et al*., 1999) and determined acceptable by DOE and the Nevada Division of Environmental Protection (NDEP) for predicting contaminant boundaries for the site.

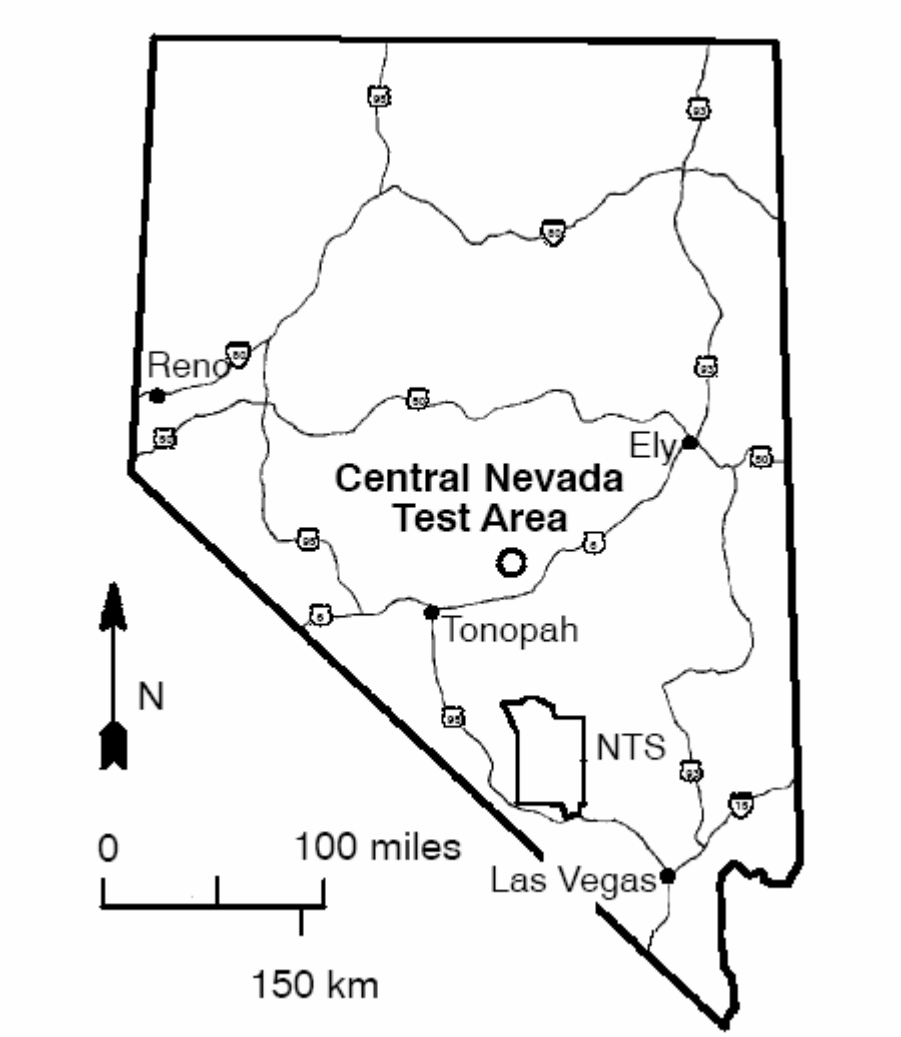


Figure 1.1. Location map of the Central Nevada Test Area in the state of Nevada.

Hassan (2003a, 2004a,b) proposed a validation approach for stochastic groundwater models in general and for the CNTA model in particular. The approach addresses some of the important issues recognized in previous validation studies, conferences, and symposia as crucial to the process. The integrated approach uses a number of tools for evaluating the predictive CNTA model. This is particularly critical in radionuclide transport models such as the CNTA model, since only a few aspects of the transport modeling results can be tested because the predictions of the model extend a thousand years into the future and no data can be used at this time scale. The key strategy is to focus on evaluating other model elements (e.g., geologic model, model structure, and flow model) using validation data to evaluate transport predictions and reduce their uncertainty.

Model validation will not eliminate uncertainty from the model predictions; some uncertainty is inherent and irreducible in models of subsurface processes. Monitoring is the final step in addressing uncertainty in environmental problems. Groundwater monitoring not only serves to build confidence that the system is performing as predicted, it acknowledges the uncertainties inherent in the modeling process and the possibility, however remote, of unexpected outcomes.

The details of the CNTA numerical model (Pohlmann *et al*., 1999, 2000), the validation approach (Hassan, 2003a), and the monitoring network design approach (Hassan, 2003b) were included in the CNTA Corrective Action Decision Document / Corrective Action Plan (CADD/CAP) together with other relevant details and submitted to the state regulator (DOE, 2004). The CADD/CAP for CNTA was subsequently approved by the State of Nevada in December 2004. Drilling and data collection at the three wells occurred during 2005 and 2006. Details of drilling the three monitoring/validation wells (MV-1, MV-2, and MV-3; Figure 1.2) can be found in DOE (2006) and hydrologic data and analyses are presented by Lyles *et al*. (2006). The wells were drilled and completed to collect geologic, geophysical, hydrologic, and geochemical data in support of the validation and monitoring efforts for the CNTA site. Actual completion depths were 1,250.3 m for MV-1, 1,115.6 m for MV-2, and 1,286.3 m for MV-3.

This report addresses the use of the data collected from MV-1, MV-2, and MV-3 in conducting the model validation process for CNTA as detailed in Hassan (2003a). Following this introduction, the report is organized as follows. Section 2 presents a brief review of the validation process and the relevant acceptance criteria. The detailed validation analysis is then presented in Section 3 along with a more qualitative, broader-view evaluation of the model. Section 4 discusses the implications of the validation results and the vision for the forward steps in the corrective action process for the site. The report is summarized and the main conclusions are discussed in Section 5.

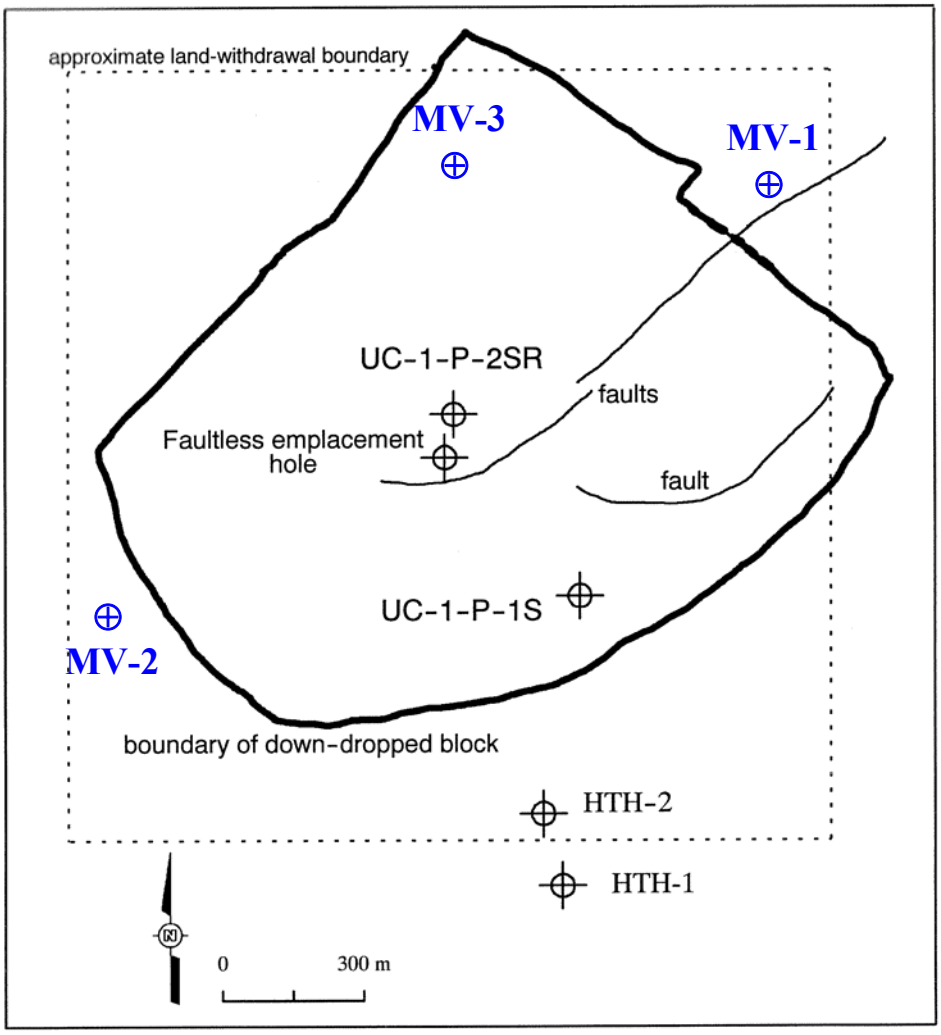


Figure 1.2. Faultless land withdrawal (dashed line) showing surface ground zero and nearby wells. The MV wells are the monitoring-validation wells constructed in 2005.

2.0 VALIDATION PROCESS AND ACCEPTANCE CRITERIA

Even the simplest deterministic subsurface model is challenging to validate (Hassan, 2004b). The validation approach for CNTA accounts for the stochastic nature of the Faultless model and evaluates the large number of realizations present in the Monte Carlo analysis (Hassan, 2004a). A brief review of the proposed validation procedure is presented below.

2.1 Proposed Step-by-step Procedure for CNTA Model Validation

Figure 2.1 describes the steps of the process to validate the model predictions. The validation steps are described below.

Step 1: Identify the data needed for validation, and the number of wells and their location. The first stage of the monitoring strategy was implemented to help select the well locations (Hassan, 2003b) shown in Figure 1.2.

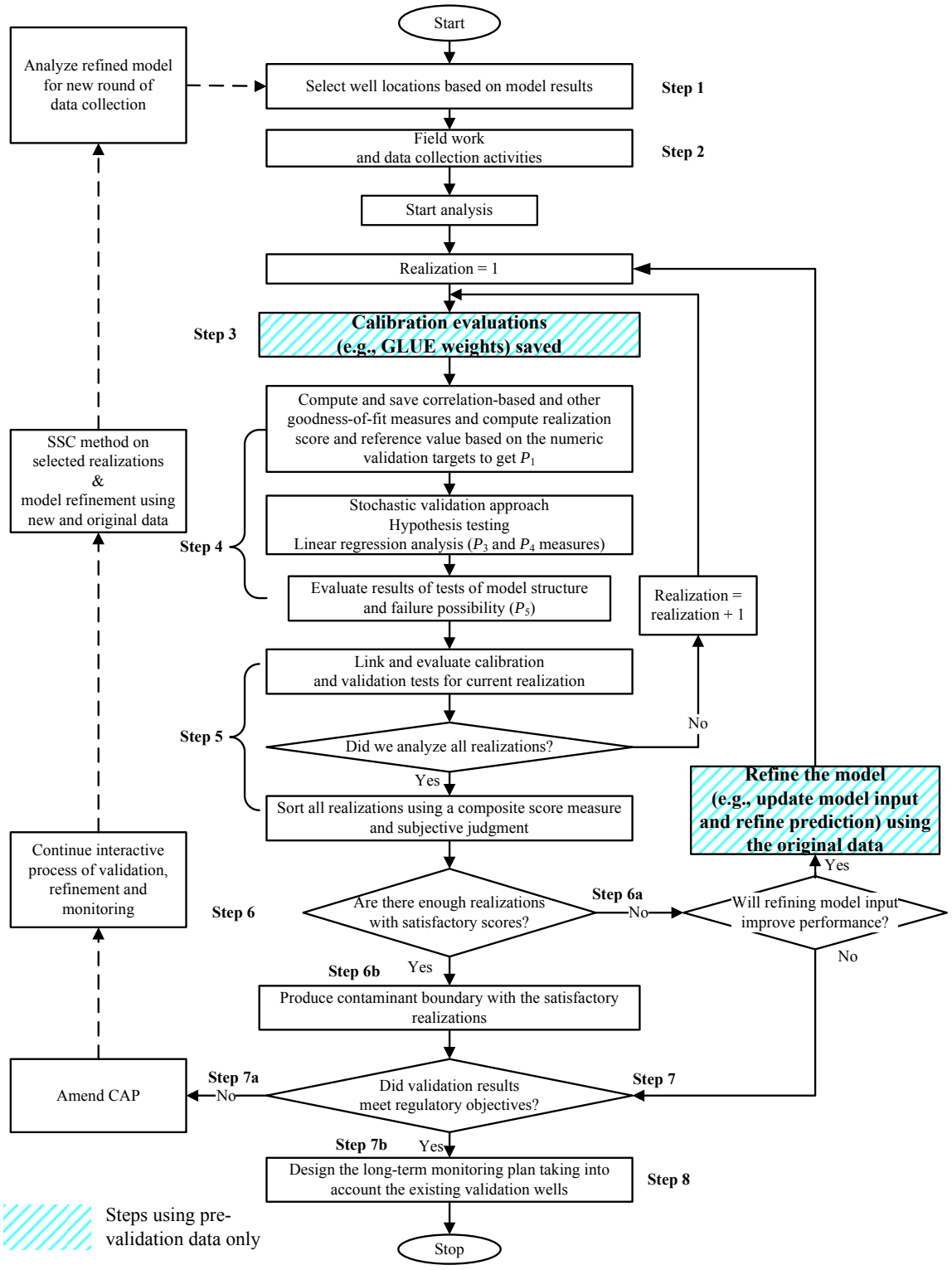


Figure 2.1. Details of the proposed model validation process for the CNTA model with the acceptance criteria measures (P_1 through P_5) explained in Section 2.2. This plan has been slightly modified from the one in the CADD/CAP (DOE, 2004).

Step 2: Install the wells and obtain the largest amount of data possible from the wells. The data should be diverse to be able to test the model structure, input, and output.

Step 3: Evaluate the model calibration accuracy for each individual realization using goodness-of-fit measures and using the calibration data only (prevalidation data; the data used to construct the original model).

Step 4: Perform the different validation tests to evaluate the different submodels and components of the model. Goodness-of-fit tests using the validation data (previously, it was calibration data) can be used for the heads as well as hypothesis testing. Data will also be used to check the occurrence of failure scenarios (e.g., whether tritium exists farther from the cavity than is predicted by any realization of the stochastic CNTA model [Pohlmann *et al*., 1999; Pohll *et al*., 2003]).

Step 5: Link the different results of the calibration accuracy evaluation (Step 3) and the validation tests (Step 4) for all realizations and sort the realizations in terms of their adequacy and closeness to the field data. The objective is to filter out 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, with the passing score determined using hydrogeologic expertise and subjective assessment. As a result of this filtering, the range of output uncertainty is reduced and the subsequent effort can be focused on the most representative realizations/scenarios.

Step 6: Results of Step 5 guide the decision as to whether there is a sufficient number of acceptable realizations (thus building confidence in the original model) or insufficient number of acceptable realizations, indicating that the original model needs adjustment.

Step 6a: If the number of unacceptable realizations is very large compared to the total number of model realizations, it indicates that either the model has a major deficiency or the input is not correct. In the latter case, the model may be conceptually good, but the input parameter distributions may be skewed. Generating more realizations and keeping those that fit the validation data 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 conceptual problem, generating additional realizations will not correct it and continued failure per the validation criteria will be obvious. In this case, the answer to the question of whether refining model input distributions may improve model performance is no, and Step 6a leads to Step 7.

Step 6b: If the number of acceptable realizations is sufficient, it indicates the model does not have conceptual problems. This determination will be made according to a number of metrics described in Section 2.2. Based on the acceptable realizations, a contaminant boundary is calculated and compared to the original contaminant boundary. This comparison will be presented to decision makers for evaluation in Step 7.

Step 7: Once the model performance has been evaluated per the acceptance criteria, the model sponsors and regulators have to answer the last question in Figure 2.1. 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.

Step 7a: If the results do not meet regulatory requirements, the left-hand-side path in Figure 2.1 begins with an evaluation of the investigation strategy, consistent with the process flow diagram in Appendix VI of the Federal Facility Agreement and Consent Order (FFACO). If

the original strategy is deemed sound, a new iteration of model development begins, using the data originally collected for validation, and steps 1 to 6 are eventually repeated. If the original strategy is deemed unsound, a new strategy will be developed. In either case, the CADD/CAP will be amended before execution.

Step 7b: If the results meet regulatory requirements, validation is deemed sufficient, the model is considered adequate for its intended use, and the process proceeds to the long-term monitoring plan development for the site closure.

While there are no guarantees of success (attaining a conclusive outcome about model performance), the combined presence of the different results and evaluations improves the odds that one can make a good decision about the model performance.

2.2 Acceptance Criteria

According to the validation plan (Figure 2.1), the first set of analyses using the validation data will yield results that are evaluated to determine the path forward. The first "if" statement in the validation process pertains to whether there is a sufficient number of acceptable realizations that are consistent with the field data used for calibration (old) and validation (new). This determination will be based on five criteria, with the decision made in a hierarchical manner. The five criteria are:

1. Individual realization scores (S_j, j = 1, …, number of realizations) are computed based on how well each realization fits the validation data, and the first criterion, P_1, is the percentage of these scores that exceeds a certain reference value.
2. The second criterion, P_2, represents the number of validation targets where field data fit within the inner 95 percent of the target probability distribution as used in the model.
3. The third criterion, P_3, relies on hypothesis testing based on the stochastic perturbation approach of Luis and McLaughlin (1992) as described in detail in Hassan (2003a).
4. The results of linear regression analysis and hypothesis testing represent the fourth criterion, 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 is compared against the field validation data give the fifth criterion, P_5.

The hierarchical approach to making the above determination is described by a decision tree (Figure 2.2). The process starts with evaluating S_j and determining the percentage of realizations with scores above the reference value, P_1. If P_1 is more than 40 percent, the number of acceptable realizations is deemed sufficient. If it is less than 40 percent, then the second criterion, P_2, is used (Figure 2.2). If P_1 is between 30 and 40 percent and P_2 is between 40 and 50 percent or if P_1 is less than 30 percent but P_2 is greater than 50 percent, the number of acceptable 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 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 a detailed example in Hassan (2003b).

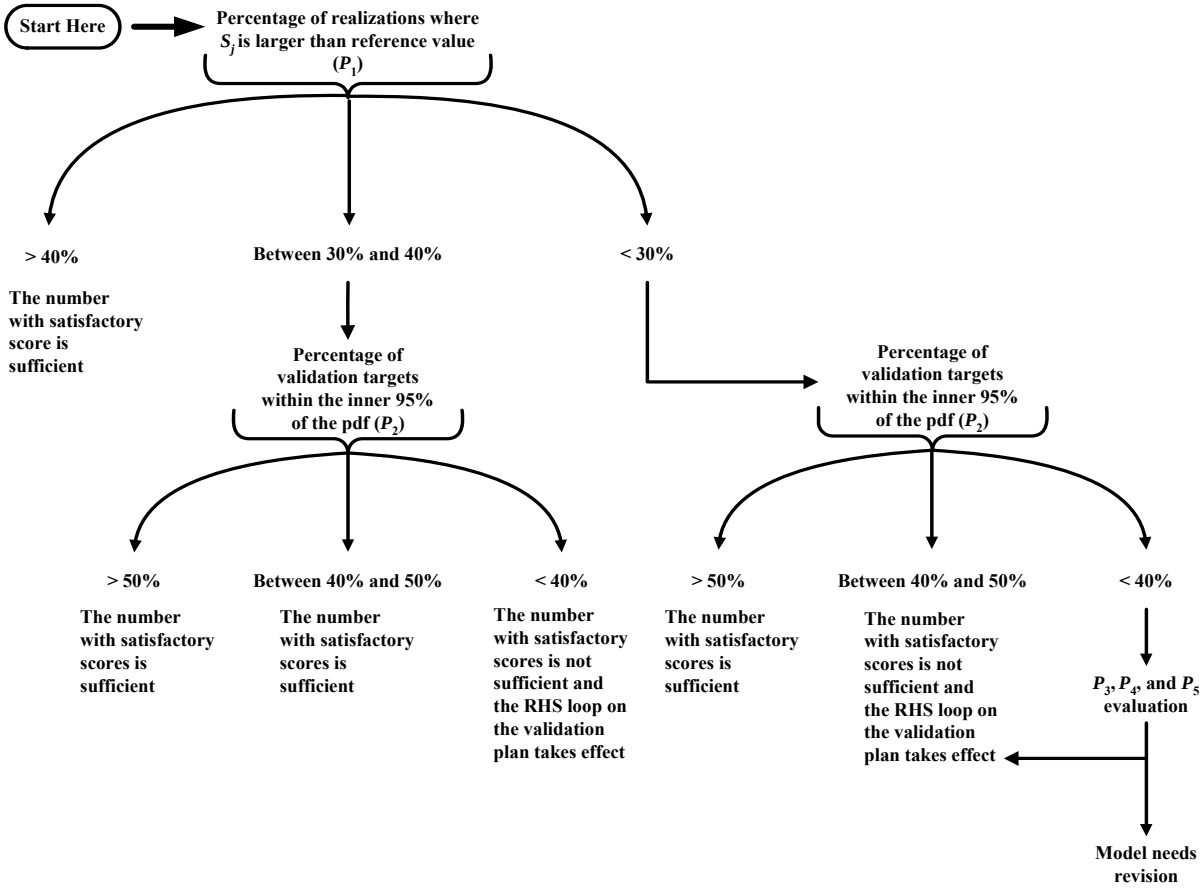


Figure 2.2. A decision tree chart showing how the first decision (Step 6) in the validation process is made and the criteria for determining the sufficiency of the number of acceptable realizations.

3.0 VALIDATION ANALYSIS FOR CNTA

The first step in the validation process, identification of data needs, was documented in the CADD/CAP. The aspects of the Faultless flow and transport model selected as validation targets are those considered key for effective monitoring and for estimating the contaminant boundary. The validation targets selected for CNTA, as presented in the CADD/CAP, are as follows:

1. Hydraulic head
2. Presence or absence of densely welded tuff near emplacement location
3. Contaminant transport predictions (confirming absence of transport above maximum contaminant levels [MCLs])
4. Hydraulic conductivity range

Each of these is important for different reasons. Comparing hydraulic head values confirms flow directions. Determining whether or not the densely welded tuff exists near the emplacement horizon is important because only those simulations with densely welded tuff predict any significant transport. Confirming the transport predictions (essentially ruling out fast pathways) is desirable, despite the low probability of detectable transport predicted by the model. Comparing the range of hydraulic conductivity in new wells with that used in the model will support the slow predicted velocities.

The second step in the validation process (data collection) was accomplished with a drilling and testing program designed to meet the data objectives (DOE, 2006; Lyles *et al*., 2006). For the first target, hydraulic head was measured in units distributed both laterally and vertically around the test. These measurements were performed in the well bore and in piezometers installed in the annular space. For the second target (presence or absence of welded tuff), the lithologic section in the three wells was logged (by evaluation of cuttings and geophysical tools). For the contaminant transport target, samples of groundwater were collected and analyzed for Faultless-related contaminants, principally tritium. General groundwater characteristics were determined to confirm conditions used in the transport model. These included major ions, silica, pH, EC, temperature, and stable isotopes of oxygen and hydrogen. For the final target (the hydraulic conductivity range), aquifer tests were performed in the wells, and in one of the piezometer tubes in MV-2.

Following the collection of the validation data from the three MV wells, steps 3 through 7 of the validation process were performed and are documented here. To organize the analysis and the discussion of the results, a summary is first presented of the data relevant to the validation process, along with discussion of data interpretation issues and conversion to model input or output parameters. The data are linked to the model domain and its discretized cells so that comparisons between field data and model simulation can be made. Steps 3 through 7 of the validation process are then implemented and the results are discussed.

3.1 Validation Data and Interpretation

The MV well locations are as follows (all in NAD27 Nevada State Plane coordinates system): MV-1 is at 192369.3 m easting and 431789.9 m northing, MV-2 is at 190962.3 m easting and 430579.0 m northing, and MV-3 is at 191652.3 m easting and 431745.7 m northing. Based on the locations, the corresponding column, i, (easting) and row, j, (northing) for the three wells in the model coordinate system are ($i = 50$, $j = 50$) for MV-1, ($i = 22$, $j = 26$) for MV-2, and ($i = 36$, $j = 49$) for MV-3 (Figure 3.1). This association allows the model input or output parameters at the locations of the validation wells to be extracted from the model realizations.

The validation data can be categorized into two sets. One set pertains to the model input parameters and the other pertains to the model-produced output. Lithology-related data (i.e., resistivity logs and identification of the densely welded tuff units) and hydraulic conductivity data belong to the first set, whereas measured heads and “inferred” gradients belong to the second set.

Resistivity was measured using downhole logging tools during the advancement of the boreholes. A resistivity log was not collected from the upper portion of well MV-3 (from land surface to a depth of 321 m). This section was in alluvium, as confirmed by geologic evaluation of drill cuttings.

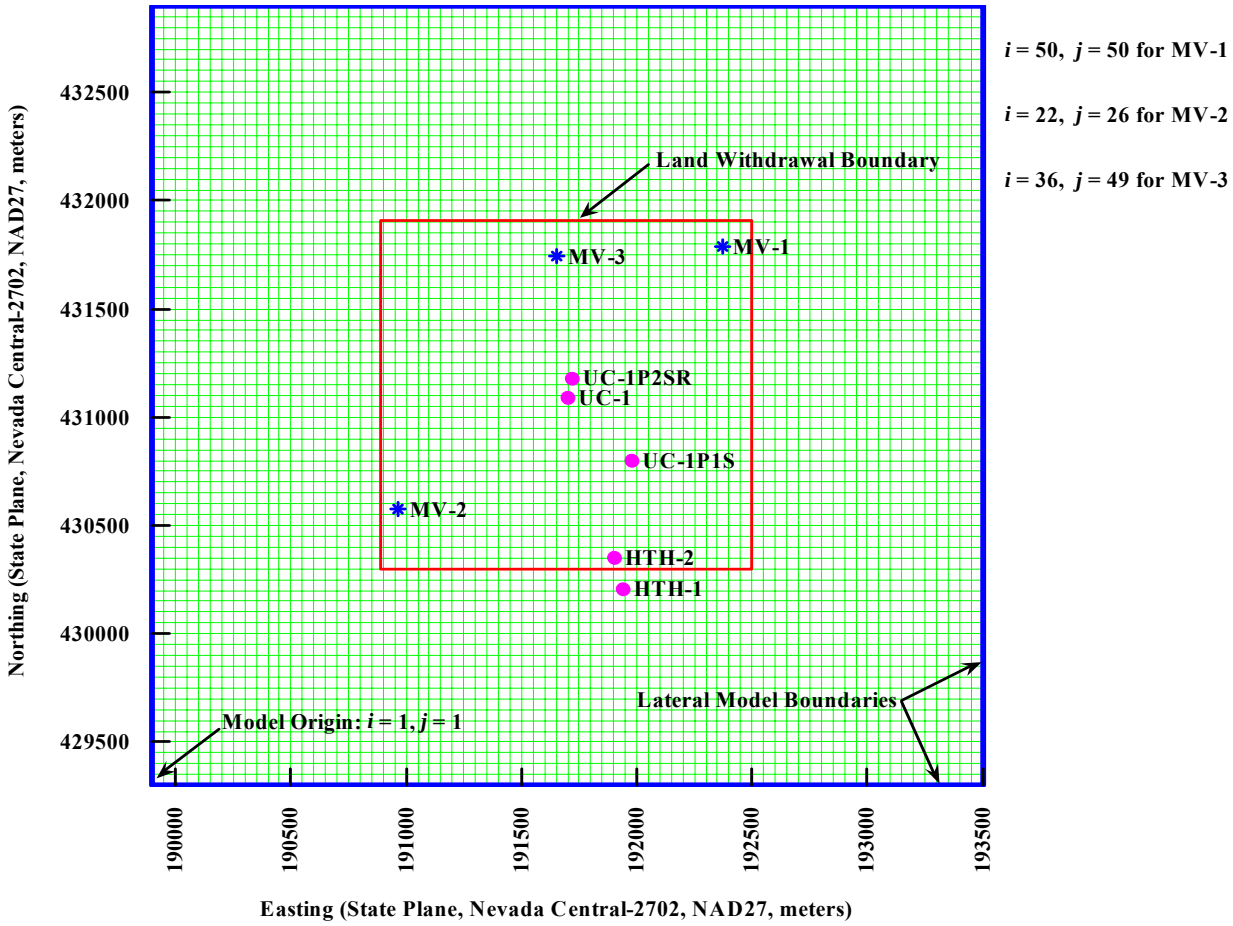


Figure 3.1. Map view of the model used for the calculation of Faultless contaminant boundaries. The model grid cells and the old as well as new wells are shown on the model domain.

Two aspects determine the lithology profile relevant to validation at each borehole; the location of the contact between the alluvium (flow category 1 in the 1999 model) and tuffaceous sediments (flow category 2) and the existence of densely welded tuff (flow Category 3). The base of the alluvium was determined by lithologic inspection (DOE, 2006). This determines the interface between flow category 1 and the other two flow categories associated with volcanic rock. Within the volcanic interval, the determination between flow categories 2 and 3 is based on resistivity data. A threshold of 30 ohm-m is used to identify densely welded tuff sections (i.e., sections exhibiting resistivity values higher than 30 ohm-m). To associate the identified densely welded tuff sections with model cells (or layers, *k*) in the original model, the resistivity data for each borehole are upscaled to the 50-m scale of the model using a simple 50 percent rule (Pohlmann *et al*., 1999). A cell was considered to belong to flow Category 3 (densely welded tuff) if 50 percent or more of the vertical 50 m cell interval had densely welded tuff present (identified from resistivity data). However, an exception to this rule was made for HTH-1 in the original model where the densely welded tuff zone identified on the HTH-1 litholog was manually assigned to flow

Category 3 even though resistivity exceeded 30 ohm-m for less than 2 m (4 percent of the model cells). This was the only exception made to the rule in the 1999 model.

The assignment of flow Category 3 to model cells based on the MV resistivity logs follows a slightly different rule compared to the 1999 model. Due to the importance of the densely welded tuff and the critical role it plays in the transport simulations, a 25 percent rule is applied where model cells are assigned to Category 3 if resistivity exceeded 30 ohm-m in 25 percent of the cell thickness. Spikes in the resistivity logs (very high resistivity in a thin section of the log) are ignored.

Figure 3.2 displays the resistivity logs for the three wells and the identification of the densely welded tuff intervals. Applying the 25 percent rule to MV-1 results in one model layer, layer 8, at the well location being assigned as densely welded tuff. Validation data indicate that layers 13 and 14 from MV-2 and layers 7 and 9 from MV-3 are densely welded tuff. Therefore, in model coordinates cells, (i = 50, j = 50, k = 8), (i = 22, j = 26, k = 13 and 14) and (i = 36, j = 49, k = 7 and 9) belong to flow category 3, the densely welded tuff.

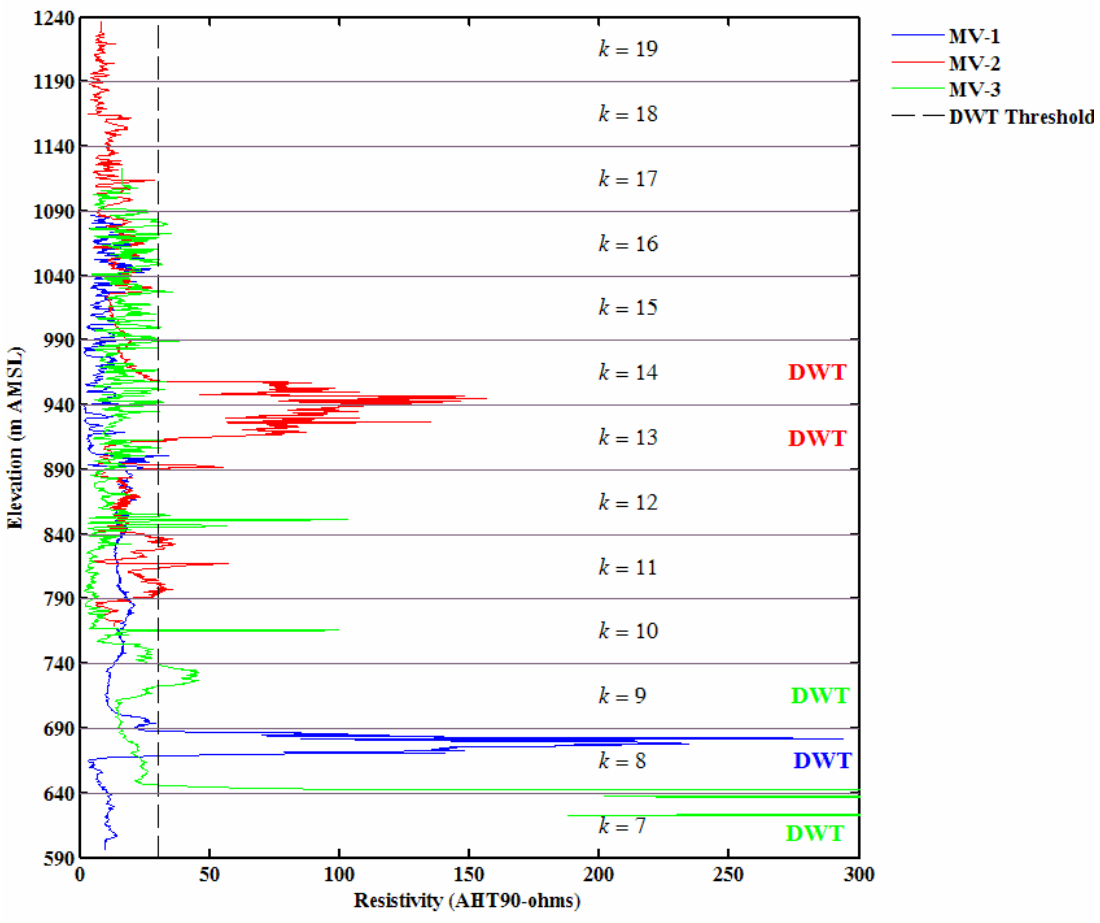


Figure 3.2. Resistivity logs from the three MV wells showing the determination of the densely welded tuff (DWT) sections and the association with model layers (k indicates model layer).

Well purging, water level monitoring, and aquifer testing are detailed by Lyles *et al*. (2006). Lithium bromide was added to drilling fluids during the drilling of the MV wells. The low hydraulic conductivity of the aquifers required a lengthy bromide purge period. MV-1 and MV-3 produced less than 1 gallon per minute (GPM); pump limitations only allowed the wells to be pumped for a few hours before the pump controller would shut off the pump. Therefore, the wells were pumped once weekly for several months until the bromide was less then 1 milligram per liter (mg/L). MV-2 produces about 3 GPM and could sustain pumping for about 6 hours; it was also pumped weekly until bromide levels were less then 1 mg/L.

Once well development at each MV well was completed, transducers were installed in the lower piezometer and main well. Campbell Scientific, Inc., CR-10X dataloggers were used to measure Geokon vibrating wire transducers; 10 PSI transducers were installed in the piezometers and 500 PSI transducers were used in the main wells during aquifer testing. Flowmeters measured the discharge during the bromide purging and during the aquifer tests. Periodic fluid level measurements were made in the upper piezometers.

Aquifer tests were performed in each MV well once the bromide purging was complete. Water level data from the aquifer tests and from the bromide purging were used to compute aquifer hydraulic conductivity and transmissivity. Each of the MV wells tested a densely welded tuff intercepted by the main well screen. The hydraulic conductivity of the densely welded tuff intervals in the MV wells is substantially lower than that reported for other densely welded tuff units in the region.

Given that the screened interval and the surrounding filter pack extend through more than one model cell at each well or piezometer, assigning head, h, and hydraulic conductivity, K, measurements to model cells is not straightforward. Often the filter pack interval is selected for head measurements since under ambient groundwater flow conditions heads will tend to be a composite of the entire section. But when vertical gradients are present as is the case for Faultless, assigning the head measurement to the entire section poses a problem. By choosing an interval covering multiple cells, the vertical gradient is forced to be zero in this zone (and artificially high above and below). Given this and the fact that vertical gradients observed from the validation data are in fact very large, it seems appropriate to assign the head to the single cell that most represents the measurement interval. These are validation data, and so they are not being "assigned" in the model in the traditional sense. They are compared to the simulation results at these locations. This is another reason to choose a single cell in which to make the comparison, because there is only one measured value at each location covering many cells, but the model has different values for adjacent cells.

A K value estimated from an aquifer test is generally considered to represent the screen interval because when the zone is stressed, flow is horizontal, and this is what the analytical methods used to derive the K values from the test results assume. But under low-K conditions, this assumption is generally not met and the entire filter pack may be involved in the hydraulic response. Thus it is tempting in this case to compare the K values over as many cells as covered by the filter pack, though this has the effect of artificially increasing the number of validation targets. Therefore, only one cell is assigned each measured hydraulic conductivity value. The cell selected is the one that is covered the most by the well screen and the filter pack.

The assignment of validation data measured from MV-1, MV-2, and MV-3 is shown in Figures 3.3, 3.4, and 3.5, respectively. The figures display the association between the validation data and the model layers. In each figure, the left-hand-side plot shows the lithology data and the right-hand-side plots show the head and hydraulic conductivity data.

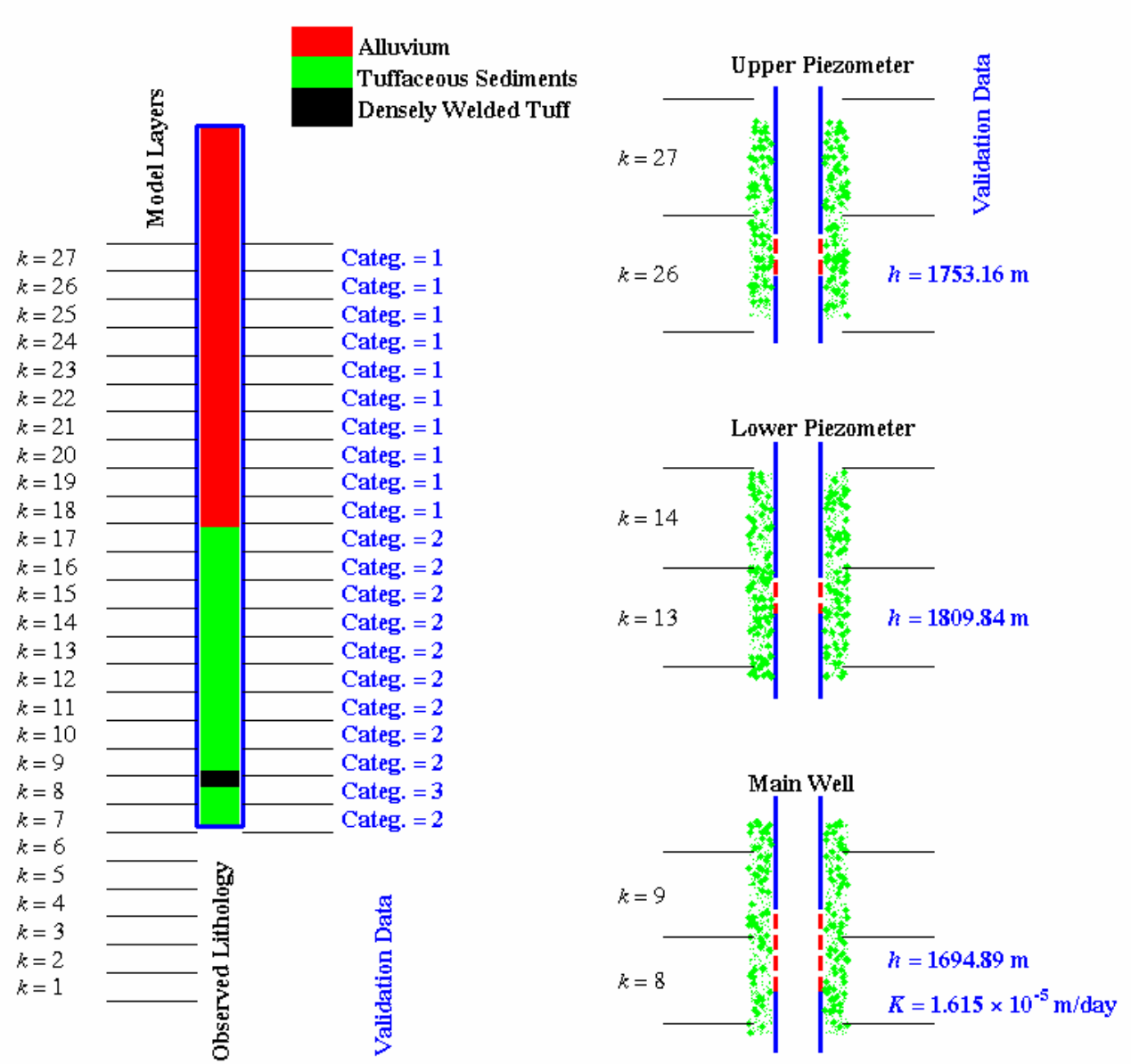


Figure 3.3. Field data from well MV-1 and conversion to validation data tied to model cells. Well screens are shown with the dashed red lines and filter pack intervals are shown with the green dots.

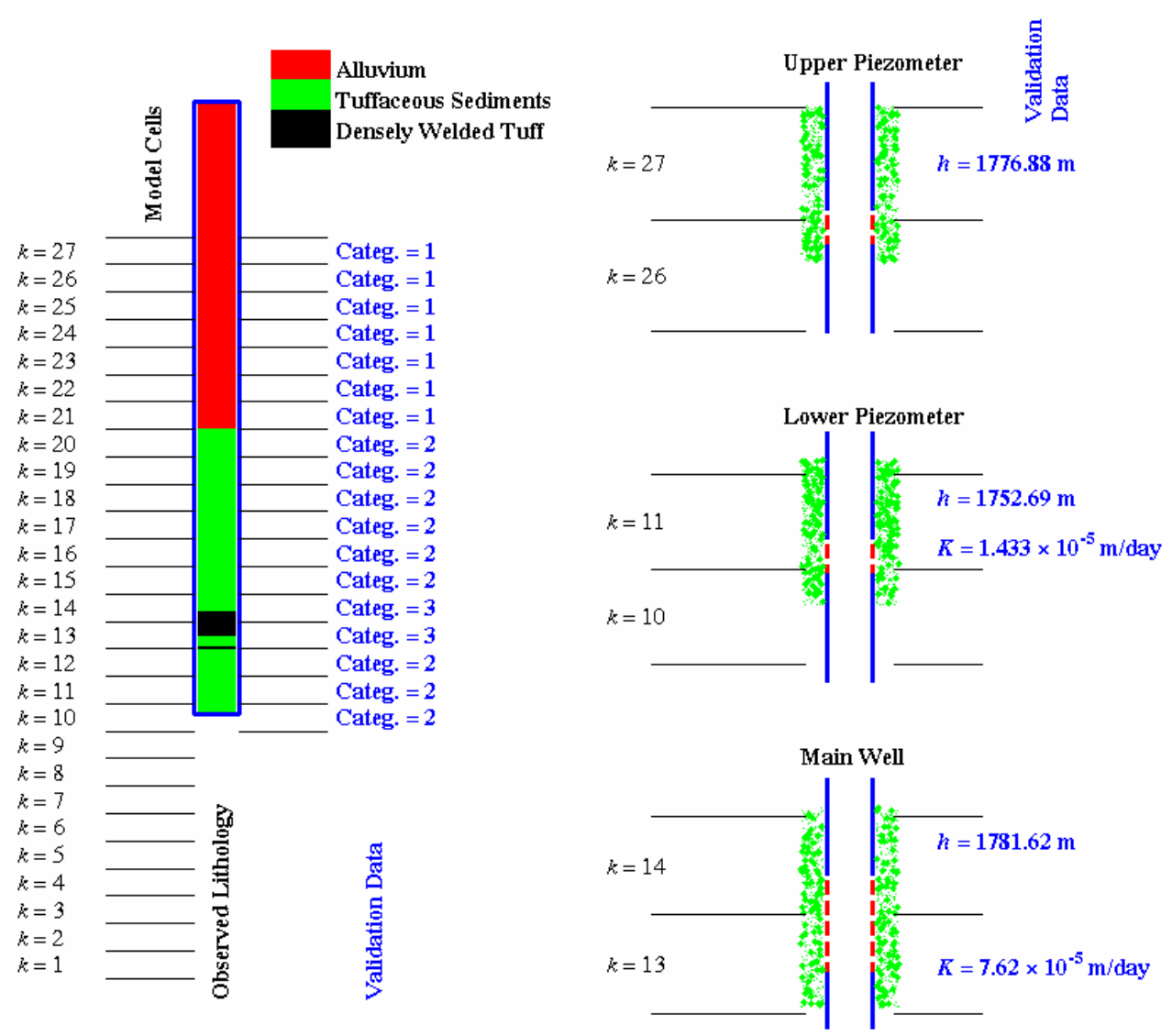


Figure 3.4. Field data from well MV-2 and conversion to validation data tied to model cells. Well screens are shown with the dashed red lines and filter pack intervals are shown with the green dots.

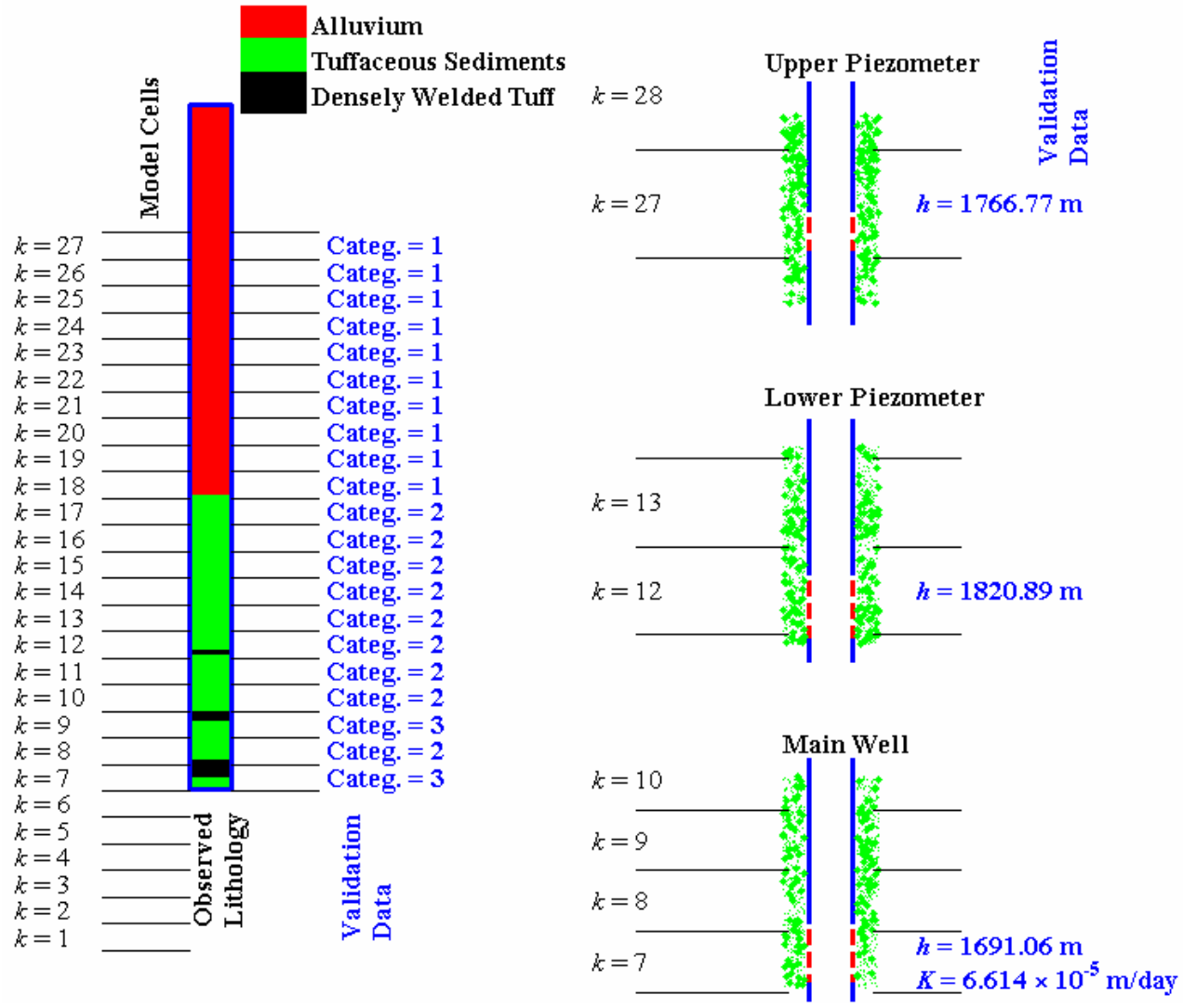


Figure 3.5. Field data from well MV-3 and conversion to validation data tied to model cells. Well screens are shown with the dashed red lines and filter pack intervals are shown with the green dots.

For MV-1, model layer 8 is densely welded tuff (flow category 1 in the 1999 Faultless model of Pohlmann *et al.*), model layers 7 and 9 through 16 are tuffaceous sediments (flow category 2) and layers 17 through 27 are alluvium (flow category 1). For MV-2, model layers 10 through 12 and 15 through 19 are tuffaceous sediments, model layers 13 and 14 are welded tuff and model layers 20 through 27 are alluvium (Figure 3.4). At the MV-3 location, layers 7 and 9 are welded tuff, layers 8 and 10 through 17 belong to the tuffaceous sediment category, and layers 18 through 27 are alluvium (Figure 3.5).

Head and hydraulic conductivity measurements are tied to model cells based on the earlier discussion and are shown in Figures 3.3 through 3.5. There are a total of nine head measurements assigned to nine cells, providing nine validation targets, and four hydraulic conductivity measurements assigned to four model cells, providing four validation targets.

In addition to the lithology, h, and K data, head gradients are computed from the measured heads and are used as validation targets. This is motivated by the fact that groundwater flows in response to gradients, not individual head values. For example, if all measured heads are much higher than modeled but gradients are the same, the model predicts the right flow directions despite underestimating heads. Horizontal gradients were calculated using the head measurements collected at roughly equivalent elevation in the three wells. Vertical gradients within a single well are also considered (Table 3.1). The head measurement locations are designated as MV-1-W, MV-1-L, and MV-1-U for the main well, lower piezometer, and upper piezometer, respectively, at MV-1. The same convention is used for the other two wells. It should be noted that the lower piezometer screen at MV-2 is deeper than the main well screen. This is different from MV-1 and MV-3 where in both cases the lower piezometer screen is shallower than the main well screen.

The gradients, $\frac{\partial h}{\partial S}$, in Table 3.1 are computed as $\frac{\partial h}{\partial S} \cong \frac{h_2 - h_1}{\Delta S}$, where S is a coordinate direction going from the first head measurement location to the second head measurement location, ΔS is the distance between the two measured heads, h_1 is the measured head at the lower elevation point, and h_2 is the measured head at the higher elevation point. The vertical gradients are calculated between adjacent measurements in a single borehole (the deepest measurement, MV-1-W, MV-2-L, or MV-3-W, and the middle one, MV-1-L, MV-2-W, or MV-3-L, and between the middle measurement and the shallowest one, MV-1-U, MV-2-U, or MV-3-U).

A total of 19 real-number validation targets (9 h values, 4 K values, and 6 $\frac{\partial h}{\partial S}$ values) are used in the validation analysis. In addition, the lithologic data provide binary-type validation targets where the category associated with each cell in the vertical profile of the three wells can be compared to the categories used in the model and the number of mismatches can then be computed.

3.2 Evaluating Calibration Accuracy for Individual Realizations (Step 3)

Step 3 of the validation process (Figure 2.1) involves using weights to evaluate the goodness of fit of each model realization using the calibration data (prevalidation data) that were used in constructing the model. Calibration of the flow model is evaluated using the average of squared differences between the measured (or observed) head h_o and the simulated head h at each of 10 straddle packer intervals in HTH-1 (Pohlmann *et al*., 1999). The root mean squared error (RMSE) is calculated for each flow realization m using the expression

$$\mathrm{RMSE}_m = \left[\frac{1}{N} \sum_{i=1}^{n} (h_m - h_o)_i^2 \right]^{0.5} \tag{3.1}$$

where N is the number of calibration targets, h_m is the simulated head for realization m, and the subscript i on the right-hand side indicates the interval at which head is measured or simulated. The RMSE ranges from 0.76 to 8.3 m, with a mean value of 1.7 m, for the full set of 500 Monte Carlo realizations.

Table 3.1. Vertical head gradients computed from the measured head values in the three wells.

#	MV-1			MV-2			MV-3			Distance ΔS (m)	Gradient $\frac{\partial h}{\partial S} \cong \frac{h_2 - h_1}{\Delta S}$
	-W	-L	-U	-L	-W	-U	-W	-L	-U		
1	h_1	h_2								250.00	4.60E-01
2		h_1	h_2							650.00	-8.72E-02
3				h_1	h_2					150.00	1.93E-01
4					h_1	h_2				650.00	-7.29E-03
5							h_1	h_2		250.00	5.19E-01
6								h_1	h_2	750.00	-7.22E-02

In a traditional stochastic numerical flow and transport model using Monte Carlo techniques, each of the realizations of flow receives equal weight. However, it is clear from the range of simulated results that some of the realizations fit the field data better than others. In an effort to honor site-specific field information throughout the modeling process, the results from those realizations that are in better agreement with the field data are given a greater relative weight in the modeling analysis than those that are in poor agreement.

The weighting procedure utilized here is the generalized likelihood uncertainty estimator (GLUE) (Beven and Binley, 1992) that extends Monte Carlo random sampling to incorporate the goodness of fit of each realization. The goodness of fit is quantified by the likelihood measure

$$L_m(\vec{\mathbf{Y}} \mid \vec{\mathbf{\Theta}}) = \left[\sum (h - h_o)_i^2\right]^{-M} \tag{3.2}$$

where $L_m(\vec{\mathbf{Y}} \mid \vec{\mathbf{\Theta}})$ is the likelihood of the vector of outputs, $\vec{\mathbf{Y}}$, for realization *m* given the vector of random inputs, $\vec{\mathbf{\Theta}}$, *h* is the simulated head at the point *i*, h_o is the observed head at that point, and *M* is a likelihood shape factor. The choice of *M* is subjective though its value defines its relative function. As *M* approaches zero, the likelihood approaches unity and each simulation receives equal weight, as in the traditional Monte Carlo analysis. As *M* approaches infinity, the simulations with the lowest RMSE receive essentially all of the weight, which is analogous to an inverse solution. In this study, the value of *M* is assumed to be unity, which is a value typically used for this type of analysis (Beven and Binley, 1992; Freer *et al*., 1996; Pohll *et al*., 2003; Morse *et al*., 2003). Each of the 500 flow realizations is weighted based on a normalized likelihood measure such that the sum of all weights is unity.

Figure 3.6 displays the calibration weights for all 500 realizations, based on using the likelihood measure of Equation (3.2) and the original (i.e., prevalidation) calibration data. The uniform weight of a traditional Monte Carlo approach (reciprocal of the number of Monte Carlo realizations, 0.002 in this case) is shown by the red line in Figure 3.6. Using the GLUE

weights to better honor the calibration data resulted in a spread of weights around the fixed value of 0.002. Realization 328 attained the highest weight, indicating that it best fits the calibration data. Put differently, the sum of squared errors for this realization was smallest among all 500 realizations. This however, may not necessarily imply good agreement as the weights convey relative performance not absolute performance. To evaluate the absolute performance, realization 328 is evaluated in terms of how the modeled results compare to the calibration data.

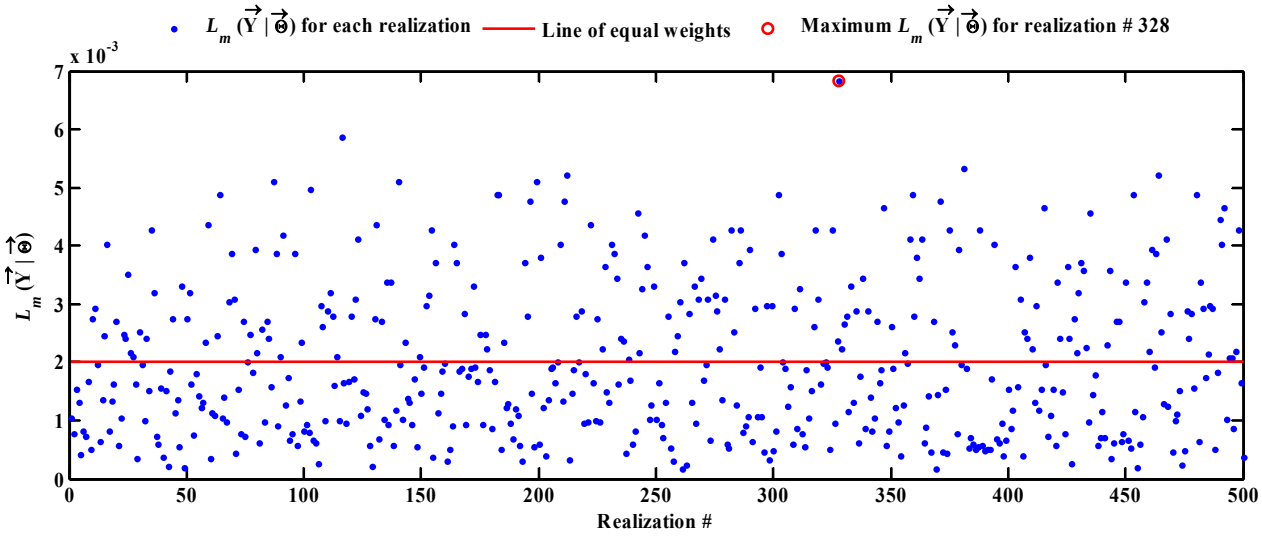


Figure 3.6. The calibration evaluation results for the model realizations with the realization having the highest likelihood measure, $L_m(\vec{\mathbf{Y}}|\vec{\mathbf{\Theta}})$, circled in red.

The correspondence between the simulated heads in the best performing realization (#328) and the observed heads at HTH-1 is good (Figure 3.7a). The data are well scattered around the unit-slope line, and it is important to note the small range of values in Figure 3.7a. Figure 3.7b shows a comparison between the modeled head profile at HTH-1 in realization 328 and the profile provided by the calibration data. Very good correspondence is observed in Figure 3.7b.

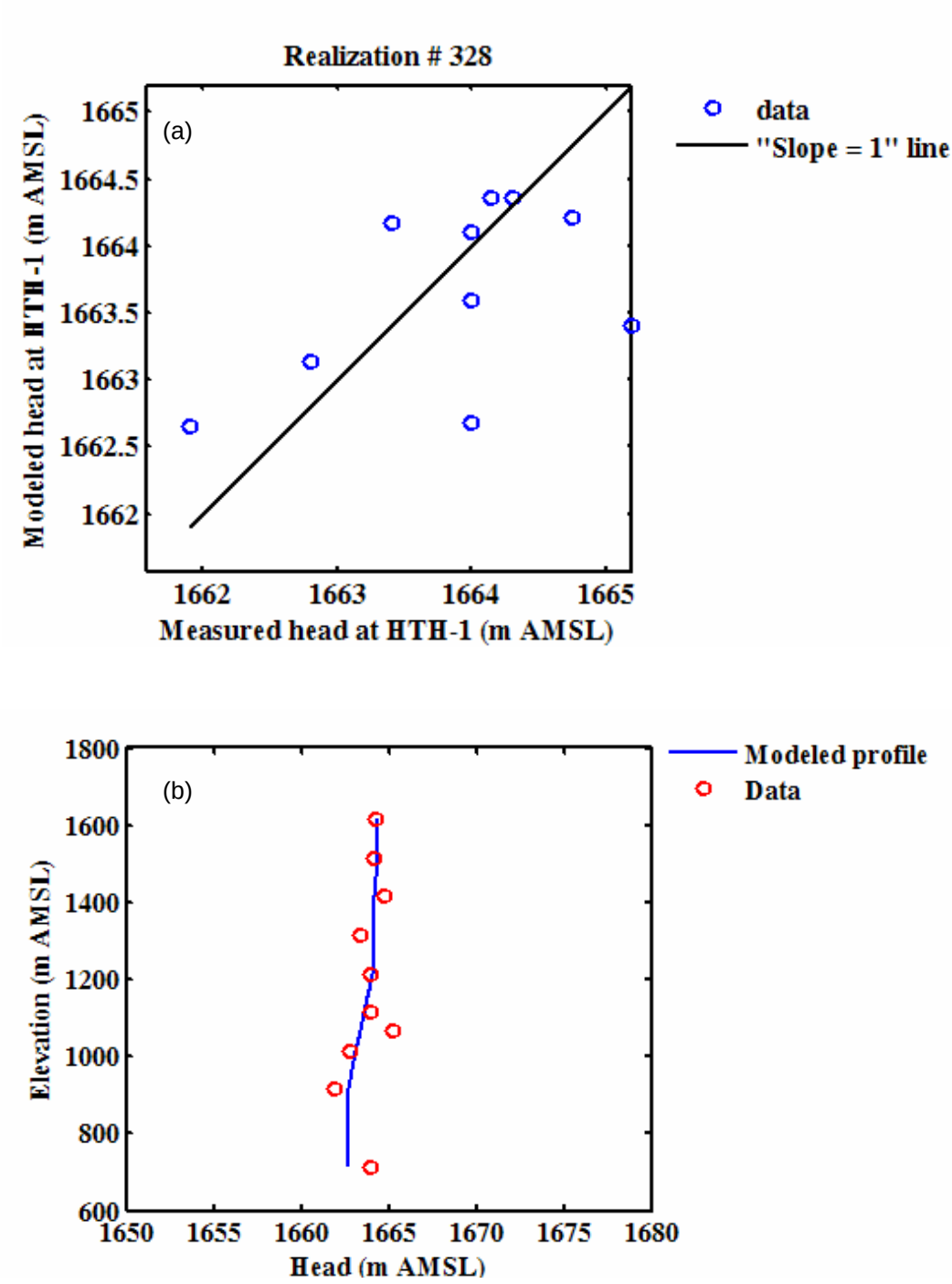


Figure 3.7. Plot of a) predicted versus observed heads at well HTH-1, and b) the modeled profile and data at HTH-1 for realization #328 that attained the highest calibration score using pre-validation data.

3.3 Using Validation Data to Evaluate the Model and Individual Realizations (Step 4)

Multiple tests of the different model components are conducted using the validation data. First, correlation-based and other goodness-of-fit measures are computed for individual realizations. Second, individual realization scores and a reference value are computed from which the P_1 criterion is obtained. The P_2 criterion is also obtained by considering the number of targets where the field observation lies within the inner 95 percent of the model-produced probability distribution. Third, the stochastic validation approach (Luis and McLaughlin, 1992) and its related hypothesis tests are conducted to obtain P_3. Hypothesis testing based on linear regression is conducted to obtain P_4. Finally, P_5 is obtained by evaluating model structure and failure possibilities.

3.3.1 Correlation-based and Other Goodness-of-fit Measures

Three measures are used here; the coefficient of determination, R^2, the index of agreement, d, and a modified index of agreement, d_1. Detailed discussion of these measures can be found in Hassan (2003a). A brief description is given here for completeness. The coefficient of determination describes the proportion of the total variance in the observed data that can be explained by the model. It ranges from 0.0 to 1.0, with higher values indicating better agreement. The coefficient of determination is calculated as follows:

$$R^2 = \frac{\sum_{i=1}^{N}(O_i - \overline{O})(P_i - \overline{P})}{\left[\sum_{i=1}^{N}(O_i - \overline{O})^2\right]^{0.5}\left[\sum_{i=1}^{N}(P_i - \overline{P})^2\right]^{0.5}} \tag{3.3}$$

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

The index of agreement, d, was developed to overcome the insensitivity of correlation-based measures to additive and proportional differences between observations and model simulations. It is expressed as (Willmott, 1981)

$$d = 1 - \frac{\sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}\left(\left|P_i - \overline{O}\right| + \left|O_i - \overline{O}\right|\right)^2} = 1 - N\,\frac{\text{MSE}}{\text{PE}} \tag{3.4}$$

The index of agreement varies from 0.0 to 1.0 and represents the ratio between the mean squared error and the "potential error" (PE), multiplied by N and then subtracted from unity. The potential error represents the largest value that $(O_i - P_i)^2$ can attain for each observed-simulated pair (Legates and McCabe, 1999). The index of agreement, d, represents an improvement over R^2, but is sensitive to extreme values owing to the squared differences.

The sensitivity of R^2 and d to extreme values led to the suggestion that a more generic index of agreement could be used in the form (Willmott *et al.*, 1985)

$$d_j = 1 - \frac{\sum_{i=1}^{N} \left| O_i - P_i \right|^j}{\sum_{i=1}^{N} \left(\left| P_i - \overline{O} \right| + \left| O_i - \overline{O} \right| \right)^j} \tag{3.5}$$

where j represents an arbitrary power (i.e., a positive integer). The original index of agreement d given in Equation (3.4) becomes d_2 using this notation. For $j = 1$, the modified index of agreement, d_1, has the advantage that errors and differences are given their appropriate weighting, not inflated by their squared values.

The above three measures are applied to the CNTA model using the validation data. The computations for head data, hydraulic conductivity data, and head gradients are performed separately because the different data sets have varying orders of magnitudes and varying units.

The R^2 values are computed for each realization using the three data sets (heads, conductivities, and head gradients). Then an average R^2 value is obtained for each realization by averaging the three values of the different data sets (Figure 3.8). The highest value attained in each case is circled with red. These high values start from 0.6 and get close to unity (good agreement). However, as indicated above, this measure is insensitive to additive and proportional differences between observations and model predictions. These realizations will be closely evaluated later to see whether the high R^2 values indicate good agreement or are impacted by additive or proportional differences. Overall, Figure 3.8 indicates that most of the realizations attain values for R^2 less than 0.5 for all three data sets.

The index of agreement, d, and the modified index, d_1, are shown in Figures 3.9 and 3.10, respectively, with the highest values circled in red. In both cases most of the coefficients are below 0.6. The exception is when the hydraulic conductivity data are used, resulting in a maximum d value of about 0.82 and a maximum d_1 value of about 0.75. The low values of d or d_1 indicate large deviation between the observations and the model results. As discussed above, this measure is insensitive to additive and proportional differences such that the realizations that attained the highest coefficients merit additional evaluation.

Figures 3.11 and 3.12 provide detailed comparisons for the realizations with the highest R^2, d, and d_1 that were shown in Figures 3.8 through 3.10. The field data are plotted against model predictions for these realizations and the plots are shown for each of the three data sets. For reference, a one-to-one relationship line (i.e., a unit-slope line) is shown in each plot (black line) and the best-fit line obtained using linear regression is shown in red.

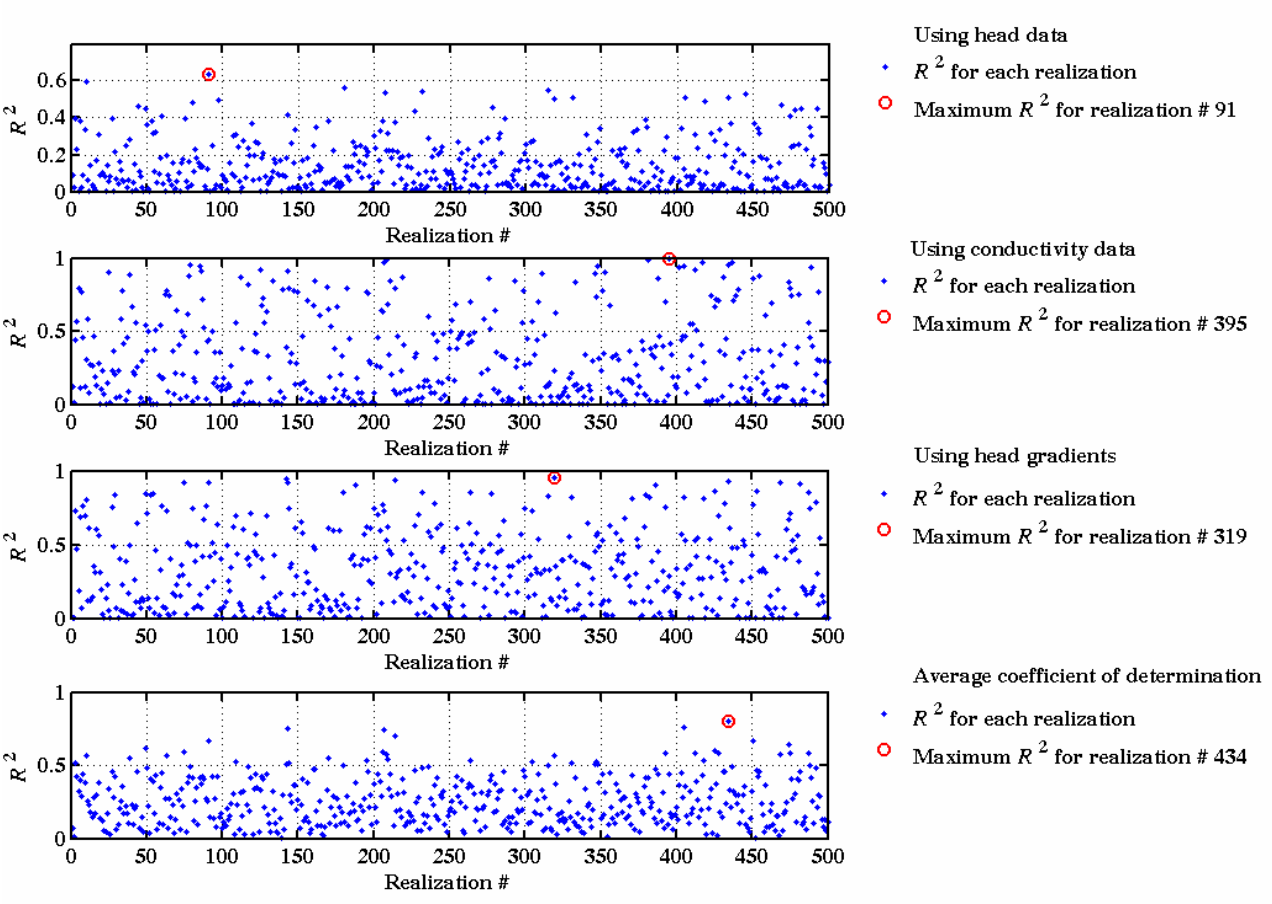


Figure 3.8. Coefficient of determination, R^2, obtained using heads, conductivities, and head gradients, with the red circle indicating the highest R^2 among all realizations. Average R^2 is also plotted.

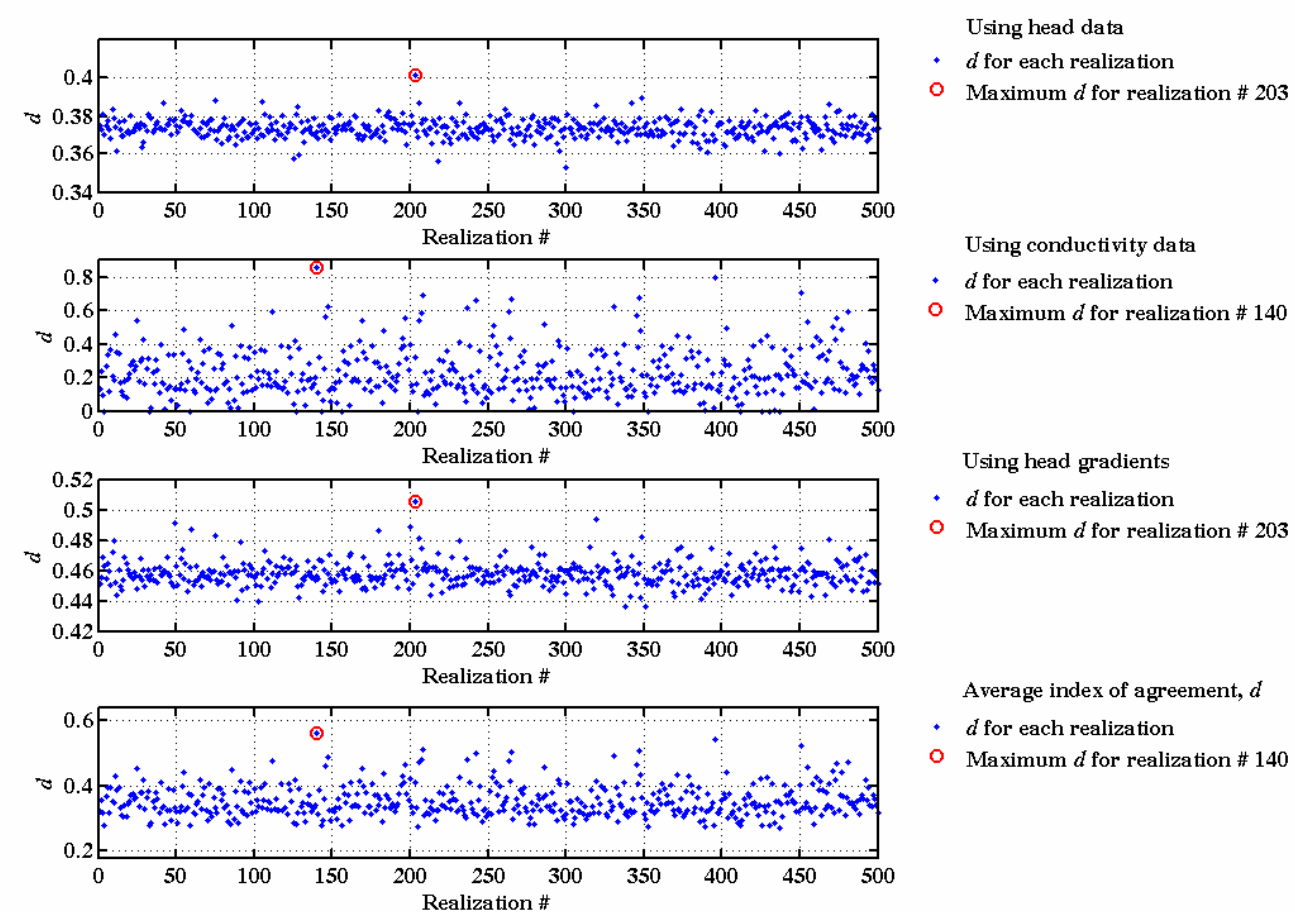


Figure 3.9. Index of agreement, *d*, obtained using heads, conductivities, and head gradients, with the red circle indicating the highest *d* among all realizations. Average *d* is also plotted.

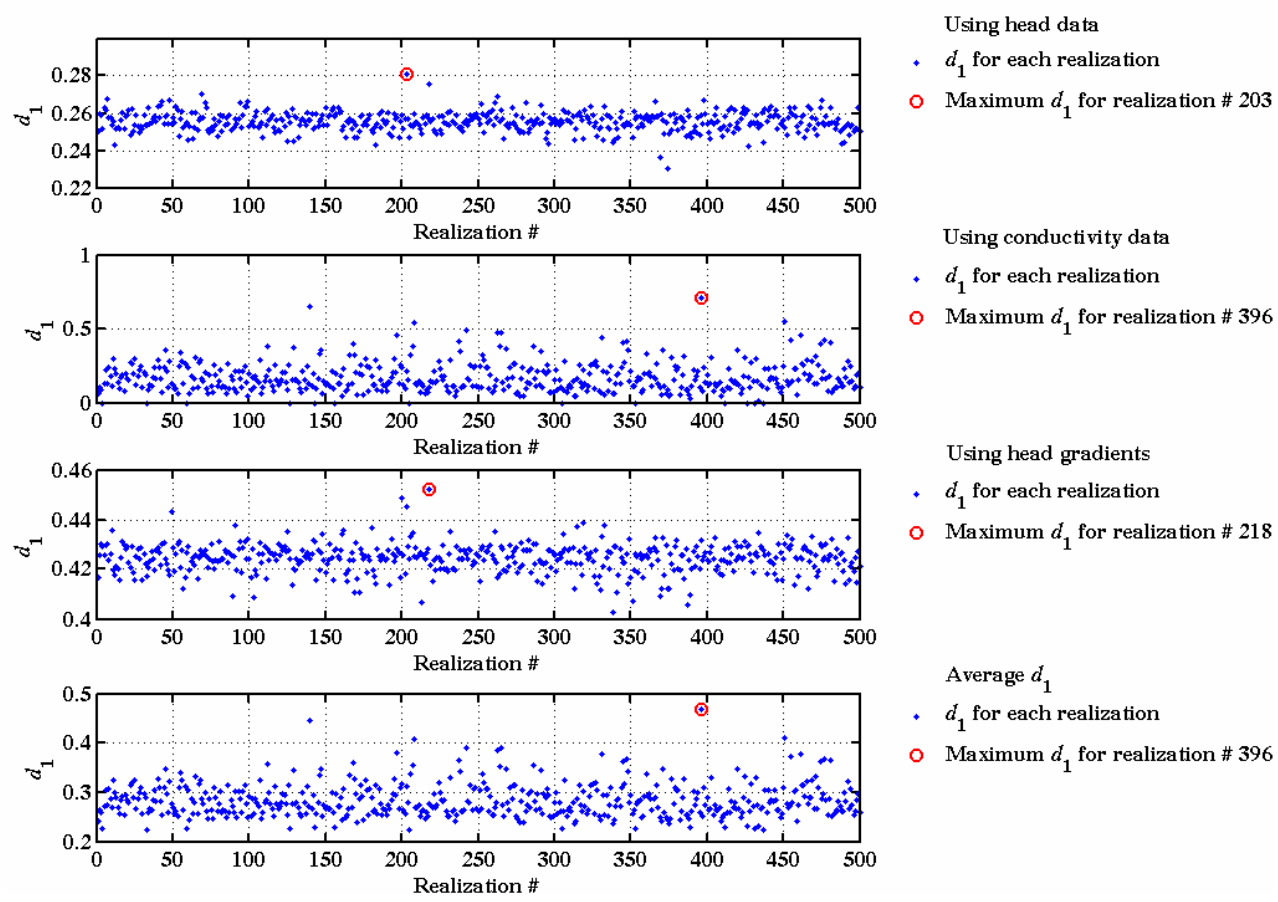


Figure 3.10. Modified index of agreement, d_1, obtained using heads, conductivities, and head gradients, with the red circle indicating the highest d_1 among all realizations. Average d_1 is also plotted.

Realizations 91, 276, and 214 that attained the highest R^2 values for heads, hydraulic conductivity, and gradient comparisons, respectively, are shown in the left-hand side of Figure 3.11. Although the linear relations in the head and head gradient plots seem good, the relation dramatically deviates from the unit-slope line. In particular, realization 214 had an R^2 very close to 1.0 for the head gradient case (Figure 3.8), but the line fitting the observed-modeled relation has a slope of almost zero (Figure 3.11), which is very far from the desired slope of 1.0. Similar results are shown for the other realizations attaining highest d or d_1 values as shown from Figure 3.11. Figure 3.12 shows the comparison between the observed data and the modeled results for the three realizations with the highest average R^2 (left-hand-side plots), d (middle plots), and d_1 (right-hand-side plots). Very similar patterns to those in Figure 3.11 are observed for these three realizations.

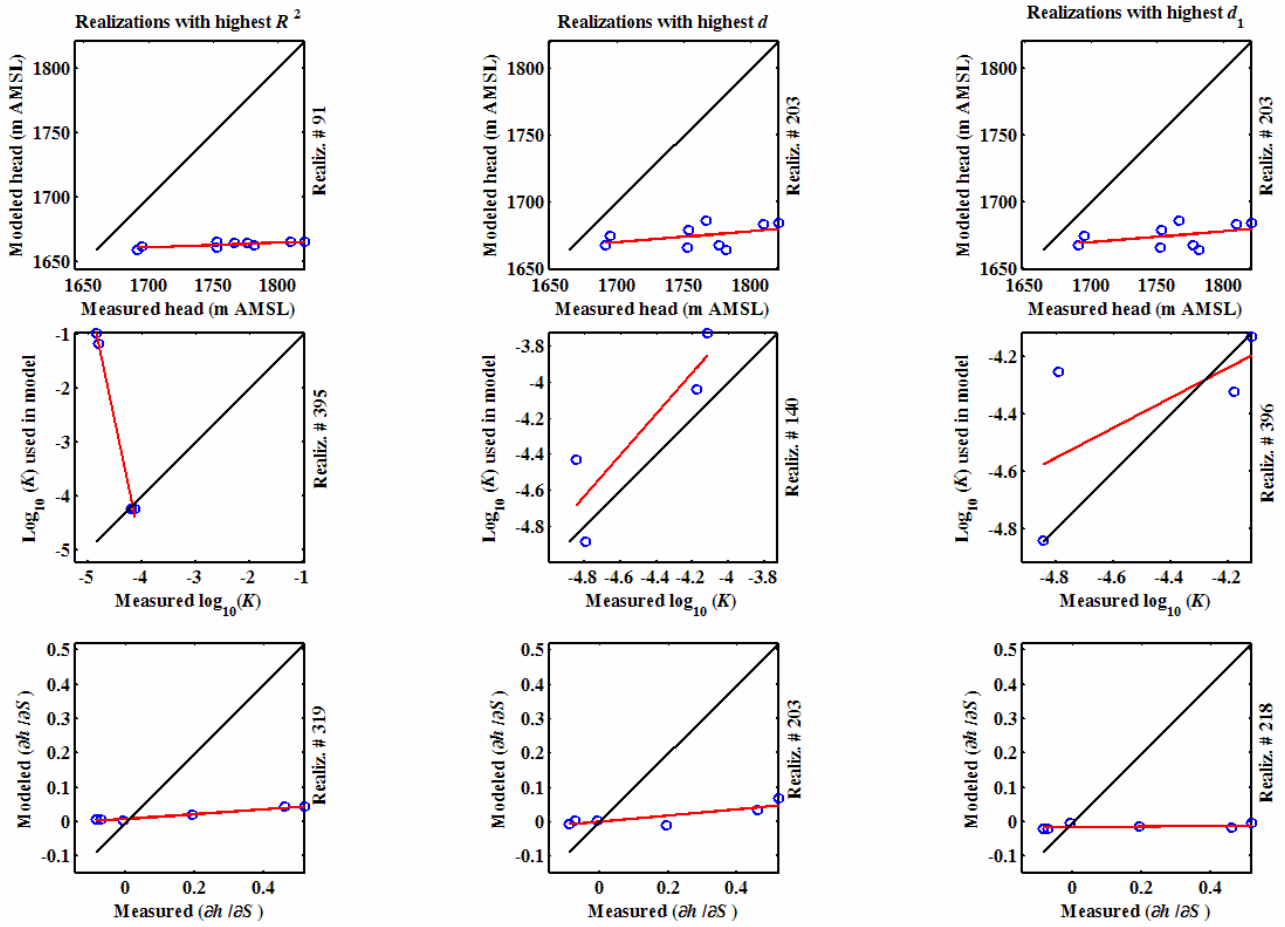


Figure 3.11. Observed versus modeled heads (m AMSL), conductivities (m/d), and head gradients (dimensionless) for the realizations that attained highest R^2, d, and d_1. Shown also are the best-fit line (red) and the one-to-one ratio line (black).

The analysis of the goodness of fit measures indicates that all model realizations are deviating from the observed data. There are good correlations between the model and the observations for the heads and the gradients. These correlations mean that when the heads increase in the data they do so in the model and vice versa, but the range of values is dramatically different. For example, measured heads increase from about 1,680 m (AMSL) to about 1,825 m whereas the model range is from 1,627 to about 1,686 m. So the observed heads are in a much higher and wider range compared to the modeled heads. Similarly, the measured vertical gradients are much higher and, in the case of the comparison between the alluvium and tuffaceous sediments, in a different direction than modeled gradients. The hydraulic conductivity ranges in the field data and in the model are similar, but the measured K values tend to be lower than the simulated values.

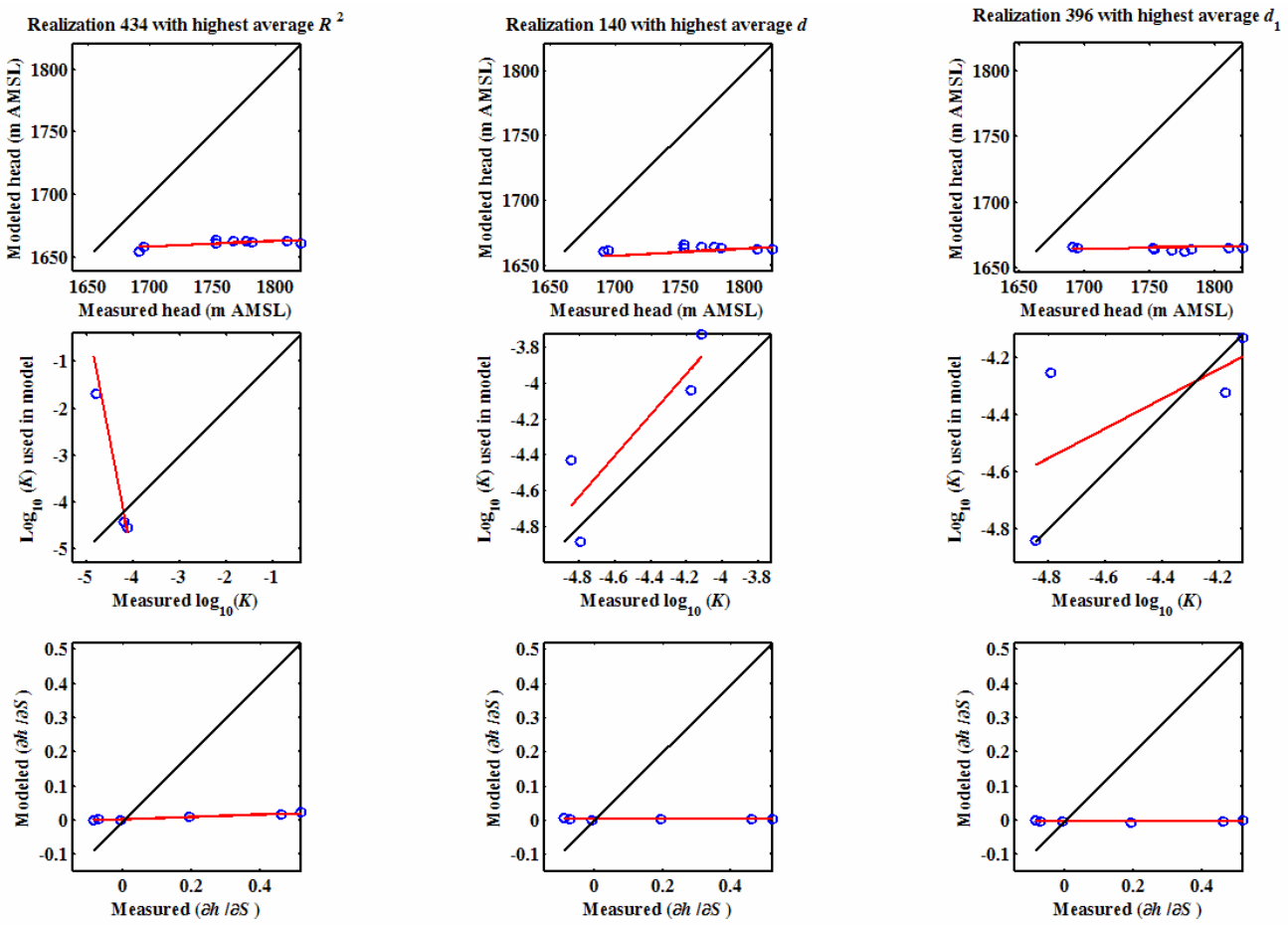


Figure 3.12. Observed versus modeled heads (m AMSL), conductivities (m/d), and head gradients (dimensionless) for the three realizations that attained highest average R^2, d, and d_1. Shown also are the best-fit line (red) and a one-to-one ratio line (black).

3.3.2 Realization Scores, S_j, Reference Value, RV, and First Criterion, P_1

The P_1 criterion is obtained by computing the number of realizations with scores, S_j, above a reference value, RV. For the general case of having N validation targets, the RV and the individual scores, S_j, will depend on the sum of squared deviations between each observation, O, and the corresponding $P_{2.5}$ or $P_{97.5}$. The parameters $P_{2.5}$ and $P_{97.5}$ are the 2.5th and the 97.5th percentiles as used in (input) or produced by (output) the CNTA model. The reference value and the realization score can be computed as

$$RV = \exp\left(-\sum_{i=1}^{N} \min\left[(O_i - P_{2.5_i})^2, (O_i - P_{97.5_i})^2\right] / \sum_{i=1}^{N} [P_{97.5_i} - P_{2.5_i}]^2\right) \quad (3.6)$$

$$S_j = \exp\left(-\sum_{i=1}^{N} [O_i - P_{ji}]^2 / \sum_{i=1}^{N} [P_{97.5_i} - P_{2.5_i}]^2\right) \quad \text{for} \quad j = 1, \ldots, NMC \quad (3.7)$$

where O_i is the field observation for validation target i, $P_{2.5_i}$ and $P_{97.5_i}$ are the 2.5th and the 97.5th percentiles of the model distribution for validation target i, and P_{ji} is realization j prediction of the model for validation target i.

For CNTA, 23 validation targets are available. These are nine head measurements, four hydraulic conductivity measurements, and six inferred vertical head gradients. For each one of these targets, the stochastic CNTA model provides a distribution of values, as each realization of the model has different values for these targets. Using Equations (3.6) and (3.7), the realization scores and the reference value are computed and compared (Figure 3.13). The value of P_1 from the figure is found to be only 1.0 percent (=5/500). Only five realizations attained scores higher than RV. As will be indicated from the analysis of P_2, many of the validation targets fall outside the middle 95 percent of the target distribution produced by the model.

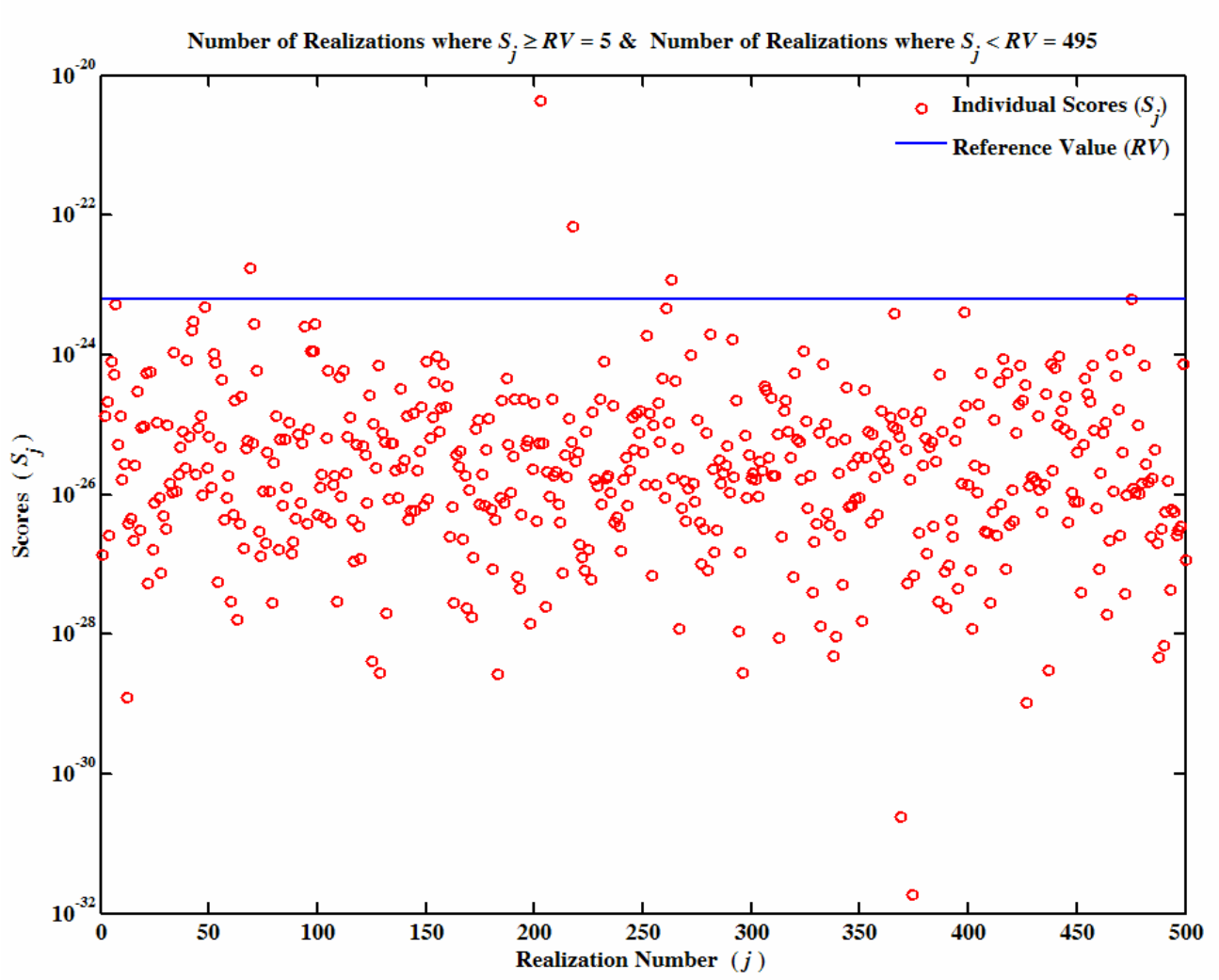


Figure 3.13. Realization scores, S_j, relative to the reference value, RV, for the CNTA model with 19 validation targets. The P_1 value here is 1.0 percent (=5/500).

Since P_1 is found to be less than 30 percent, the next step in the decision tree (Figure 2.2) is to check P_2, the number of validation targets where the field observation lies in the inner 95 percent of the model distribution of that target (i.e., between the 2.5th and the 97.5th percentiles) relative to the total number of targets. All of the nine head targets (Figure 3.14) fall outside the model distribution and are much higher than the heads predicted by the model. The differences between the highest values predicted by the model and the observed heads range from 10 m to about 130 m (Figure 3.14).

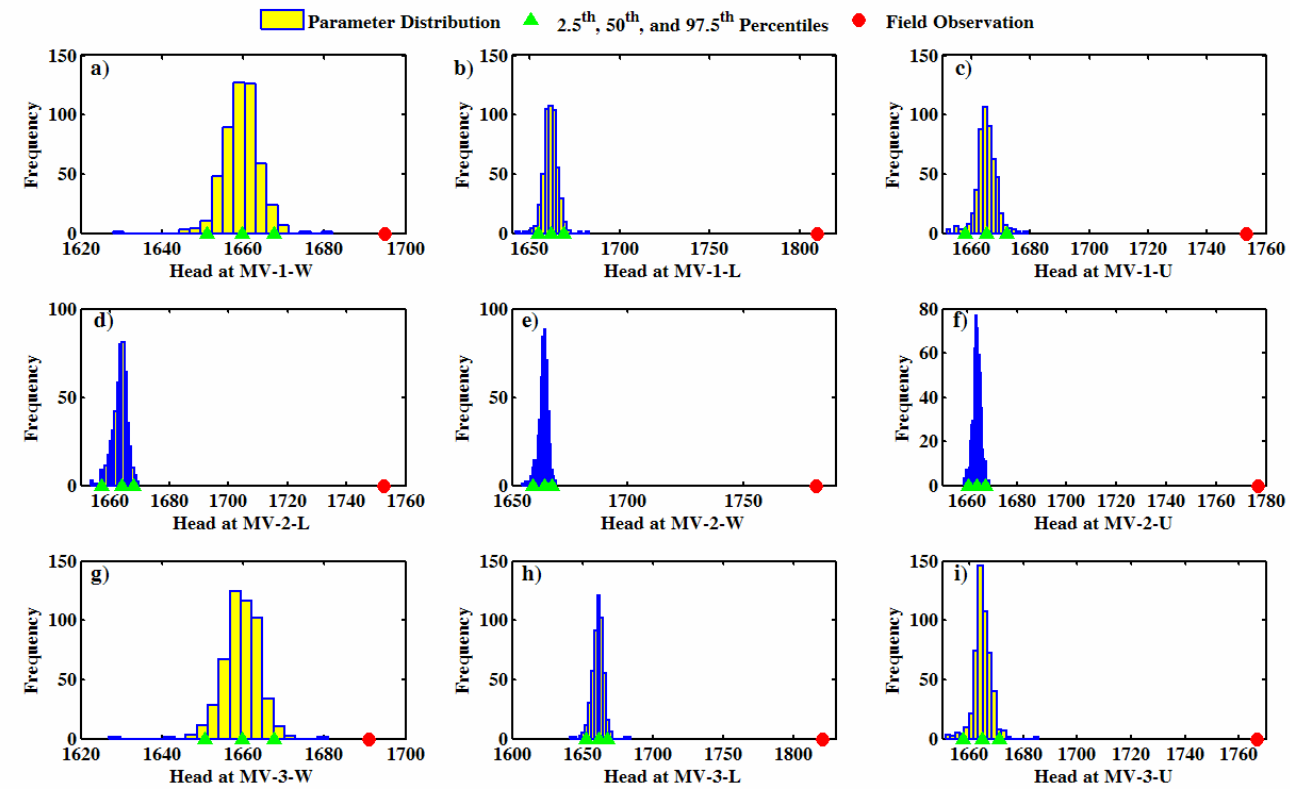


Figure 3.14. The nine head observations (red circles) relative to the distributions produced by the model at each of their respective locations. The 2.5th, 50th and 97.5th percentiles of the model heads are also shown (green triangles).

Similar to the head targets, the hydraulic conductivity targets are plotted on the histograms of the model output (Figure 3.15). All of the measured hydraulic conductivity values fit within the inner 95 percent of the model distributions. Most of the values match the peak density (i.e., the value with the highest frequency) of the model distributions. From these plots and those in Figures 3.11 and 3.12 it can be concluded that the overall range of the hydraulic conductivity used in the model was reasonable and the field observations at the three wells validate the hydraulic conductivity ranges used in the model.

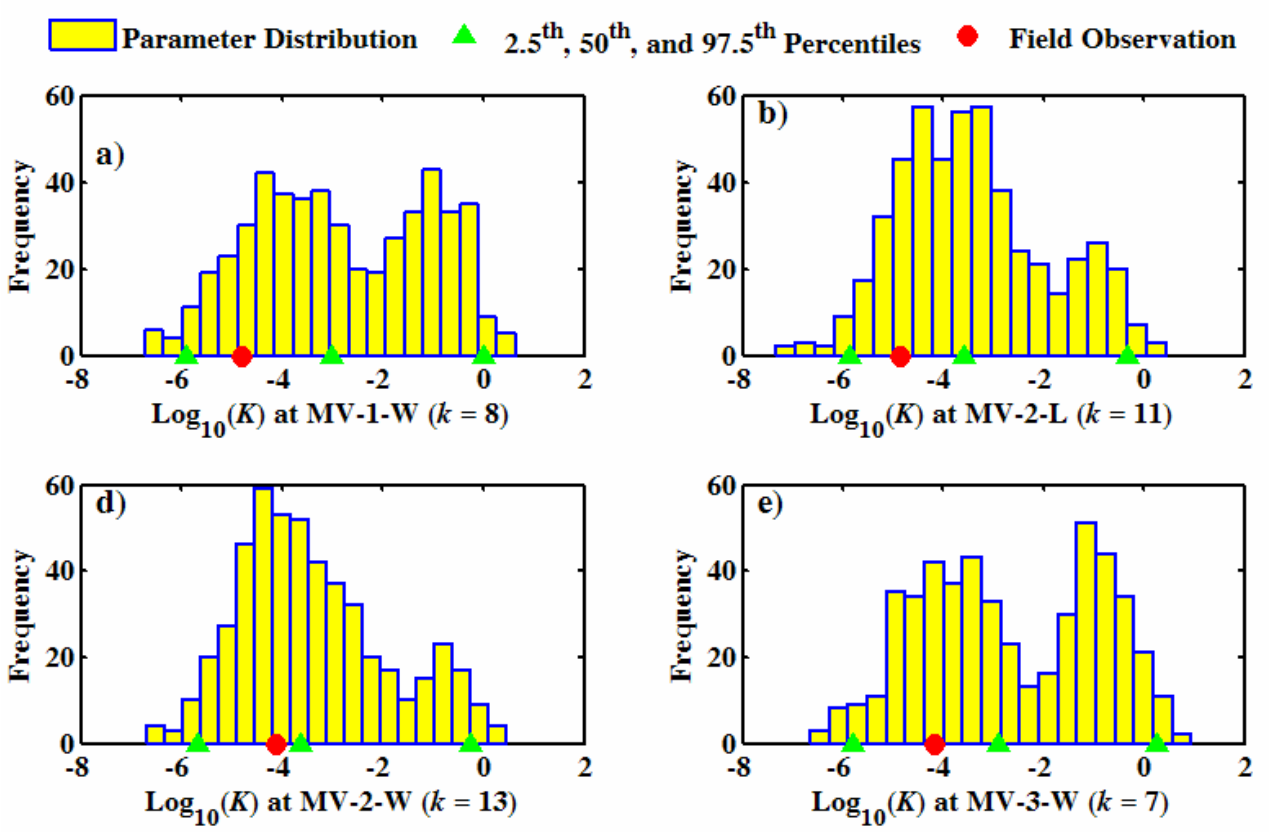


Figure 3.15. The four hydraulic conductivity observations (red circles) relative to the distributions used in the model at each of their respective locations. The 2.5th, 50th and 97.5th percentiles of the model conductivities are also shown (green triangles).

The results of the hydraulic conductivity comparisons demonstrate the limitations of adhering to a strict comparison of data to model-cell assignments and the importance of having a broader view of the model components in light of the validation data. The comparison between the new hydraulic conductivity values and those assigned to the corresponding model cells was reasonably good; the field values were near the mode of the distribution for each cell. That apparent validation is in fact misleading because it ignores the hydrostratigraphic unit assignment. Each of the main well strings for the CNTA wells was completed in a densely welded tuff because the densely welded tuffs are considered to present the fastest groundwater travel pathways. The measured hydraulic conductivities in the MV wells were focused on sections that encountered the densely welded tuff. However, when compared to the model (Figure 3.15), it was compared to the conductivity assigned to the location of the screen/filter in each well whether this location was considered densely welded tuff in the model or not. Therefore, a more fair comparison is of the measured conductivities to the welded tuff conductivity distribution used in the original CNTA model. This comparison is shown in Figure 3.20.

The original model (Pohlmann *et al.*, 1999) employed a lognormal distribution for the densely welded tuff conductivity with a $\log_{10}$ mean of -0.87 and a $\log_{10}$ standard deviation of 0.632. This lognormal distribution is shown with the blue line in Figure 3.16. The four

conductivity measurements are superimposed on this distribution. Comparison of the measured hydraulic conductivities with the distribution of *K* used in the model for the densely welded tuff category reveals that the new data are significantly lower than the low end of the distribution used. The measured values are at least an order of magnitude lower than the lowest possible value used in the model. In fact there is between three and four orders of magnitude difference between the measured hydraulic conductivity values and the mode of the distribution used in the model. The apparent match suggested by the location-to-cell comparison of the validation analysis (Figure 3.15) occurs because the majority of realizations assigned the well-screen cell locations to the category of tuffaceous sediments, which had a much lower distribution of *K*.

The remaining six targets that belong to the vertical head gradients are plotted with the model histograms in Figure 3.17. All of the six gradient targets fall outside the inner 95 percent of the model distribution for these targets. The magnitude of the field gradients is much higher than the model-produced range, except for $(\partial h/\partial S)_4$, which is close to the 2.5th percentile. For some targets (e.g., $(\partial h/\partial S)_2$ and $(\partial h/\partial S)_6$), the field gradient is opposite in direction to the gradient obtained in most of the model realizations.

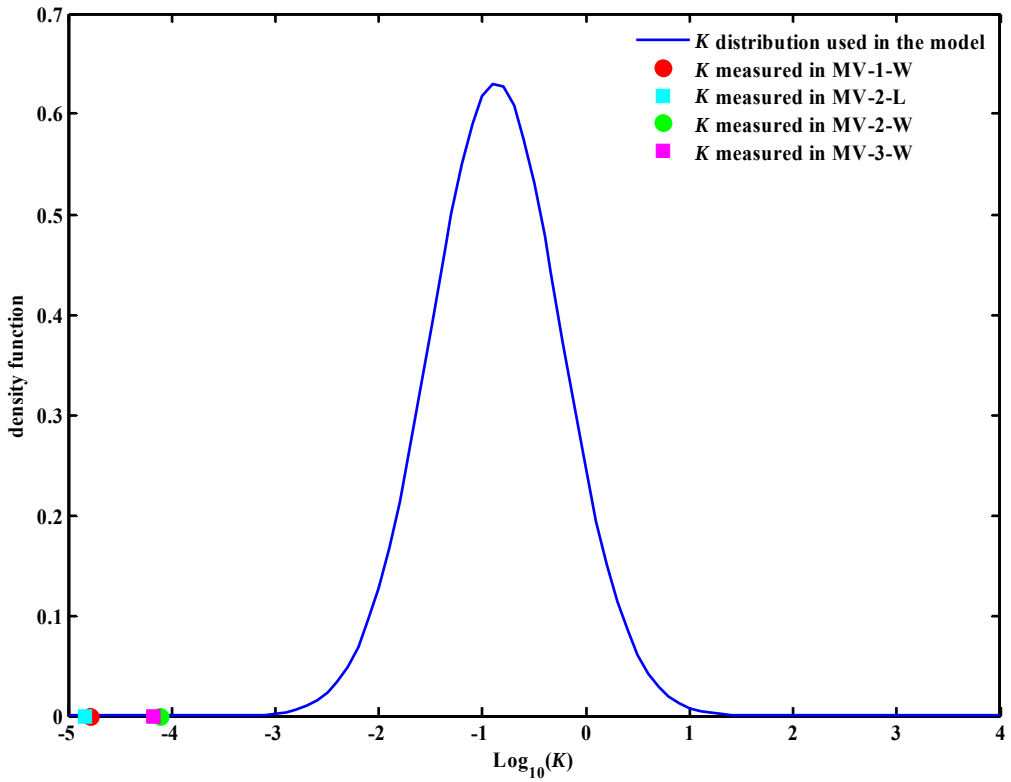


Figure 3.16. Conductivity distribution for the densely welded tuff that was used in the original CNTA model (Pohlmann *et al*., 1999) and relation to the measured *K* values of the densely welded tuff encountered in the three wells.

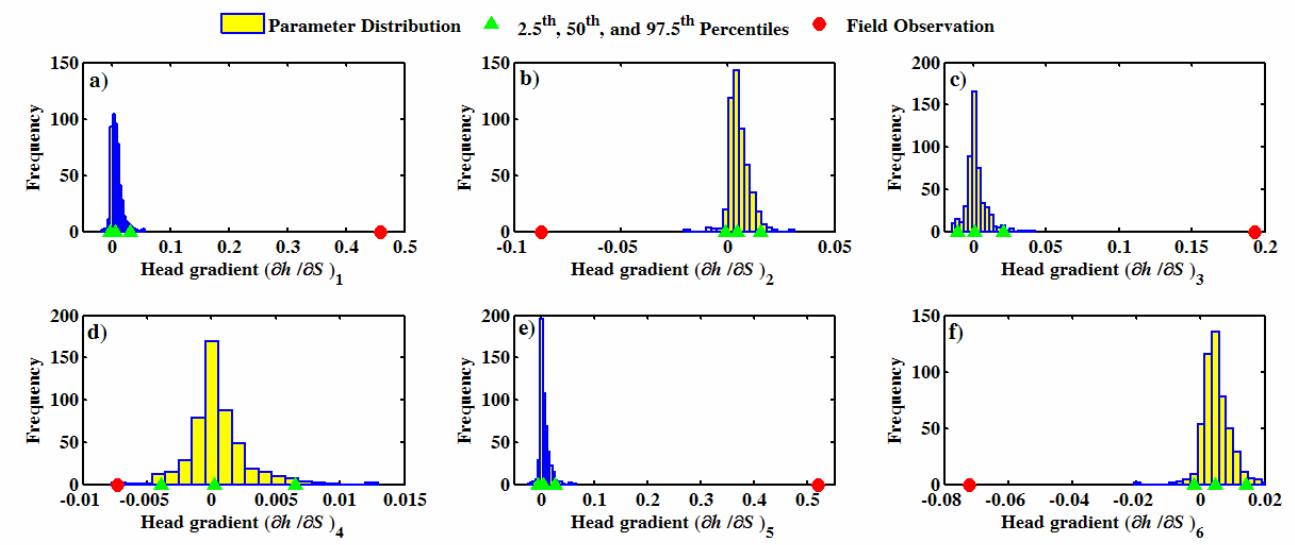


Figure 3.17. Similar to previous figures but for head gradients $(\partial h / \partial S)_1$ through $(\partial h / \partial S)_6$.

Combining all targets together, only four out of 19 targets fall between the 2.5th and 97.5th percentiles of the model distributions for these targets. This gives a P_2 value of 21 percent, which is less than the threshold of 40 percent in the decision tree (Figure 2.2). Accordingly and based on the decision tree, the other three criteria, P_3, P_4, and P_5, should be evaluated so that a final conclusion could be made about the model validation process.

3.3.3 Applying the Stochastic Validation Approach of Luis and McLaughlin (1992), P_3

This approach is applied here using the head data only. Details of the approach can be found in Luis and McLaughlin and also in Hassan (2003a) and (DOE, 2004). A brief description of the aspects related to the application to the CNTA model is presented here for completeness. The approach is based on the assumption that the flow model is used for predicting the distribution of hydraulic head in space, which describes the large-scale flow behavior of the system. Another assumption is that the observations made for the purpose of model validation are small-scale observations collected at sparse points in space and are assumed to be consistent with the steady-state assumption of the model. Both of these assumptions are met in the CNTA model and thus the analysis can be applied to the model.

Under these assumptions, the differences between predicted and measured head values can be attributed to the following three error sources: (1) measurement errors, which represent the difference between the true values and measured values of hydraulic head; (2) spatial heterogeneity, which represents the difference between the large-scale trend (or smoothed head) that the model is intended to predict and the true small-scale, actual values of head; and (3) model error, which represents the difference between the model prediction and the actual smoothed trend. Figure 3.18 shows a schematic representation of these error sources, where an actual, fluctuating (due to heterogeneity) head distribution, h_j, with a large-scale trend, $\overline{h}_j$, is shown in conjunction with a hypothesized stepwise distribution representing model prediction, $\hat{h}_j$.

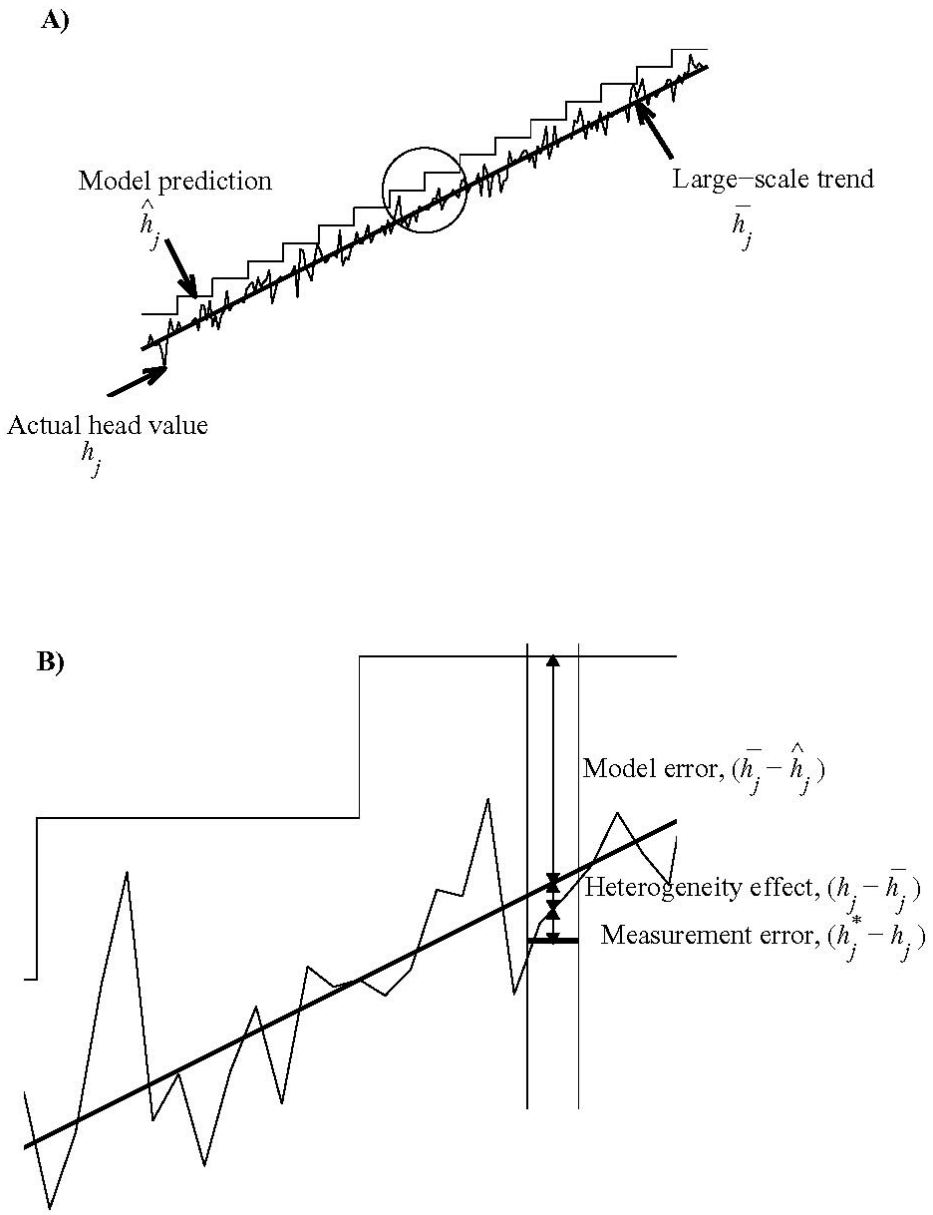


Figure 3.18. Schematic representations of the actual head distribution, large-scale trend, and stepwise model prediction (A), and the decomposition of the measurement residual into three error sources or components (B).

The j^{th} measurement residual, ε_j, observed at location $\mathbf{x}_j$ (for $j = 1, ..., N$), where N is the total number of head measurements used for validation, can be written in terms of three components of the error or the mismatch. This leads to the equation

$$\varepsilon_j = [h_j^* - h_j] + [h_j - \bar{h}_j] + [\bar{h}_j - \hat{h}_j(\hat{\eta})] \qquad (3.8)$$

where the first term between the square brackets represents measurement error, the second bracketed term represents the effect of geologic heterogeneity, and the last term represents the model error. In (3.8), $h_j = h(\mathbf{x}_j)$ is the true head value at $\mathbf{x}_j$ and $\bar{h}_j = \bar{h}(\mathbf{x}_j)$ is the smoothed

value of the large-scale trend or the expected value of h_j. Equation (3.8) defines the separate errors contributing to the differences between measurements and predictions.

If the model is valid, the hypothesis that the model prediction is equal to the smoothed, large-scale values should be accepted. This is equivalent to accepting that the model error term in (3.8) is zero. In statistical terms, the following null hypothesis is considered:

$$\begin{aligned} &H_0 : \text{Model error is negligible, } \hat{h}_j(\hat{\eta}) = \overline{h}_j \\ &H_1 : \text{Model error is significant, } \hat{h}_j(\hat{\eta}) \neq \overline{h}_j \end{aligned} \tag{3.9}$$

Luis and McLaughlin (1992) proposed few tests that can capture the different aspects of model evaluation. They proposed a quantitative approach to determine whether statistics such as the sample mean and covariance of the residuals are consistent with hypothesis H_0 in (3.9). When the hypothesis is true, it can be shown that the desired measurement residual variance can be written as

$$\sigma^2_{\varepsilon_j} = \sigma^2_{h^*} + \sigma^2_{h_j} \tag{3.10}$$

where $\sigma^2_{\varepsilon_j}$ is the measurement residual variance, $\sigma^2_{h^*}$ is the measurement error variance (human error, device error, etc.), and $\sigma^2_{h_j}$ is the head variance stemming from geologic heterogeneity. The head variance, $\sigma^2_{h_j}$, in (3.10) plays a key role in this approach since it defines how much variability one should expect around the model's predictions when the model structure and measurements are both perfect. In other words, this variance establishes a type of lower bound on the model's ability to predict point values of head (Luis and McLaughlin, 1992). The head variance can be derived from the results of the flow model and evaluated at each node of the discretized domain. Equation (3.10) can then be used to evaluate the measurement residual variance under the assumption that H_0 is correct. One can thus test the assumption that the mean residual is zero and use the mean squared residual (Equation 3.10) to test the null hypothesis H_0 in Equation (3.9).

Mean Residual Test

If the null hypothesis is true (i.e., the model is predicting correctly the desired large-scale trend), a sample mean computed from many measurement residuals should be close to zero. This implies a test of the following form:

$$\begin{aligned} &H_0 : \text{Mean residual is negligible, } \overline{\varepsilon}_j = 0 \\ &H_1 : \text{Mean residual is significant, } \overline{\varepsilon}_j \neq 0 \\ &\text{Test statistic : } m_\varepsilon = \left| \frac{1}{N} \sum_{j=1}^{N} \frac{\varepsilon_j}{\sigma_{\varepsilon_j}} \right|. \end{aligned} \tag{3.11}$$

The null hypothesis, H_0, is true if $m_\varepsilon < v$, where v is a test threshold selected to give the desired two-sided type I error probability (or significance level, α). The null hypothesis, H_0, in Equation (3.11) is equivalent to H_0 in Equation (3.9). If it is assumed that m_ε is normally distributed (based on the central limit theorem), the threshold value may be obtained from a standard normal probability table (Luis and McLaughlin, 1992).

Using the nine head measurements from MV-1, MV-2, and MV-3, this hypothesis test is conducted for each individual realization of the CNTA model. First, the 500 realizations are used to compute the head variance at the locations of the nine head measurements. These variances are denoted as $\sigma^2_{h_j}$ in Equation (3.10), where j = 1, 2, …, 9. The measurement error variance term, $\sigma^2_{h^*}$, needed in Equation (3.10) represents the errors associated with the field observations. To find this value, assume that there is a 95-percent confidence that the true head at any of the measured head locations in the three wells is within ± 0.3 m (i.e., ± 1.0 ft) of the observed head. If it is further assumed that a normal distribution applies, then the 95 percent confidence interval means that the interval from [the measured head value - 1.96 σ_{h^*}] to [the measured head value + 1.96 σ_{h^*}] is equivalent to 0.3 × 2 = 0.6 m. This implies that 1.96 σ_{h^*} = 0.3, thereby giving a value of 0.02343 for the measurement error variance, σ_{h^*}. Equation (3.9) is then used to obtain $\sigma^2_{\varepsilon_j}$ at each of the nine locations where head is measured.

To conduct the hypothesis test according to Equation (3.11), the test statistic, m_ε, is computed for each realization, where ε_j is obtained as the difference between the current realization head prediction and the measured head for each measurement location j = 1, 2, …, 9. This test statistic is compared to the critical value of the standard normal variate, Z, at exceedence probability of 0.975. This is based on a two-tail test at a 95-percent confidence level or a 5 percent significance level. The results of this hypothesis testing are shown in Figure 3.19a. For all realizations, the test statistic, m_ε, is greater than the critical Z value and thus the null hypothesis (Equation [3.9] or [3.11]) is rejected for all realizations. This indicates that the model prediction of the heads do not represent the large-scale trend inferred from the field measurements.

Mean Squared Residual Test

If one assumes that measurement residuals conform to a particular probability distribution, it would be expected that a certain percentage would lie outside confidence bounds derived from this distribution. If, for example, that distribution is normal, the interval $h_j = \hat{h}_j \pm 1.96\sigma_{\varepsilon_j}$ defines a 95-percent confidence interval around the predicted value $\hat{h}_j$, where σ_{ε_j} is obtained from Equation (3.10). If a significant number of the measurements h^*_j lie outside this interval, the null hypothesis H_0 is rejected. A more convenient version of the same concept relies on the following mean-squared error test (Luis and McLaughlin, 1992):

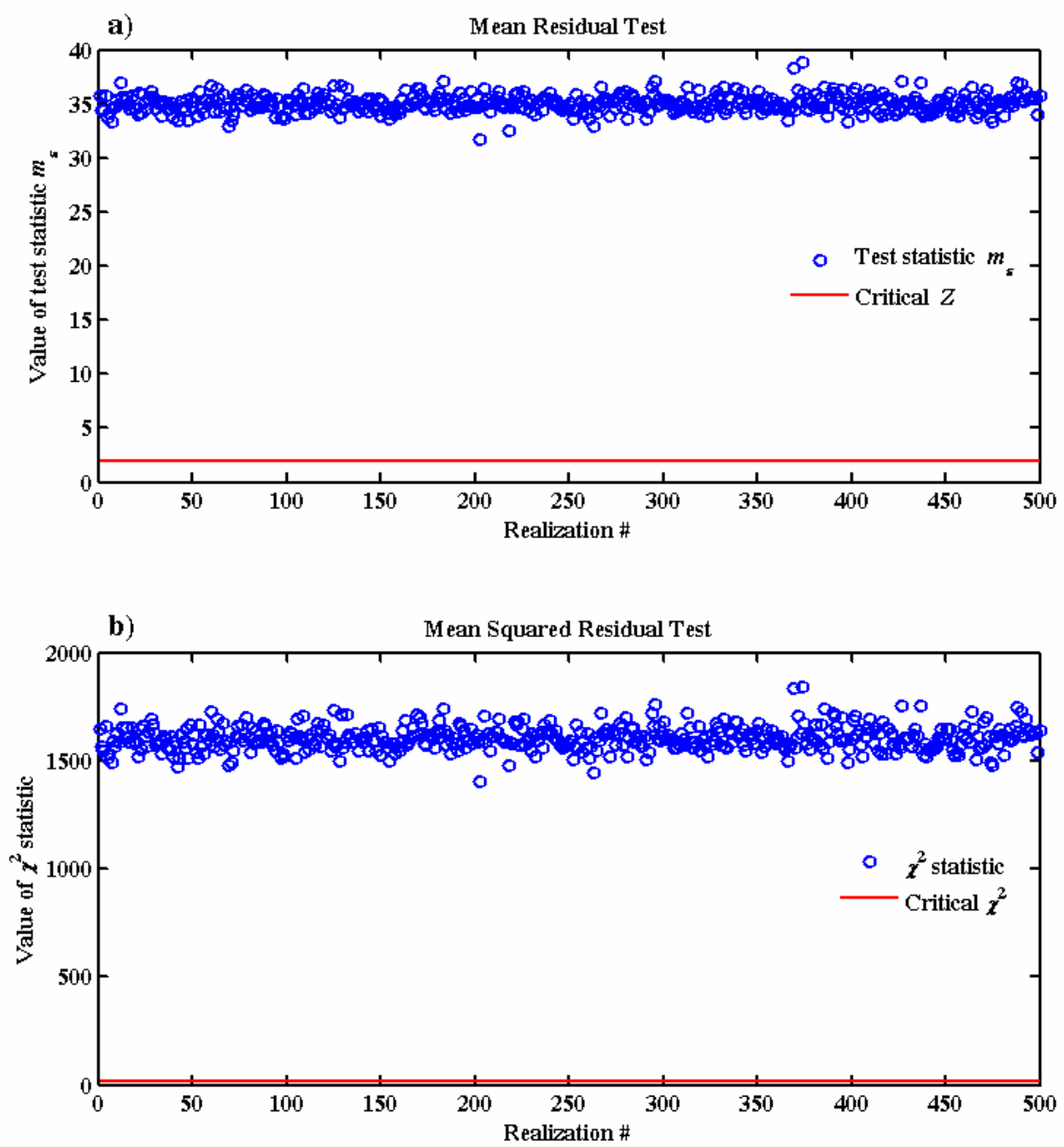


Figure 3.19. Results of the hypothesis testing formulated according to the stochastic validation approach of Luis and McLaughlin (1992).

$$\text{Decide that } H_0 \text{ is true if: } \chi^2 = \frac{1}{N}\sum_{j=1}^{N}\frac{\varepsilon_j^2}{\sigma_{\varepsilon_j}^2} < v \qquad (3.12)$$

where v is a test threshold selected to give the desired significance level. If the hypothesis is true and the measurements are sufficiently far apart for the residuals to be uncorrelated, normally distributed random variables, the test statistic χ^2 follows a chi-squared probability distribution with N degrees of freedom. With only nine head measurements at CNTA, it is difficult to determine whether this assumption is met or not. However, the test will be applied to the model using the head data assuming the impact of the assumption would be relatively small.

Equation (3.12) is used for each realization to obtain the test statistic χ^2. Then the critical value of the test is obtained from a chi-squared distribution at a significance level of 5 percent and 9 degrees of freedom. The results of this test are shown in Figure 3.19b. The test statistic χ^2 is much greater than the critical value for all realizations. In fact the difference is so large that the impact of the uncorrelated residual assumption would be negligible on these results. The null hypothesis in Equation (3.9) or (3.11) is also rejected for all model realizations.

Consistent with the previous analyses and tests regarding the head measurements, this analysis indicates that the model predictions for the heads are not valid. None of the model realizations was acceptable based on the two hypothesis tests conducted for the residuals. The results of this analysis provide an insight into the performance of model realizations and constitute the criterion P_3 needed in the decision tree of Figure 2.2.

3.3.4. Hypothesis Testing on Linear Regression Line, P_4

A linear regression analysis of calculated against measured data provides a method to evaluate empirically the quality of the data-model fit. Bias in the model and uncertainty in the input and measured data would be expected to affect both the slope of the regression line and the standard error of the regression. There are several techniques for fitting a straight line through x-y data pairs using regression analysis. The most common regression analysis in general is the Ordinary Least Squares (OLS) regression of a dependent variable against an independent variable.

If the model predictions represent the field conditions (expressed by the validation data), the regression line should have a slope of 1.0 and an intercept of zero. Based on this linear regression, one needs to statistically test the assertion that the slope of the regression line is unity and that the intercept of the line is zero. Hypothesis testing can be used for this purpose with the null and alternative hypotheses expressed as

$$\begin{aligned} H_0 &: \text{Slope} = 1 \\ H_1 &: \text{Slope} \neq 1 \end{aligned} \qquad (3.13)$$

The test statistic is ((Slope-1) ÷ standard deviation of the slope). This statistic is to be compared to the critical value of the t-distribution at (N - 2) degrees of freedom (N is the number of data points) and at the α level of significance, $t(N-2,1-0\cdot 5\alpha)$. If the absolute value of the test statistic exceeds the critical value, the null hypothesis is rejected.

In a similar manner, the null hypothesis of a zero intercept can be examined. Assuming b is the intercept of the linear regression line, the intercept hypothesis test is formulated as

$$\begin{aligned} H_0 &: b = 0 \\ H_1 &: b \neq 0 \end{aligned} \tag{3.14}$$

The test statistic is ((b-0) ÷ standard deviation of the intercept). This statistic is to be compared to the critical value of the t-distribution at (N - 2) degrees of freedom and at the α level of significance, $t(N-2,1-0\cdot 5\alpha)$. If the absolute value of the test statistic exceeds the critical value, the null hypothesis is rejected. Failing to reject both null hypotheses does not necessarily mean the model is free of biases, it only means that this analysis fails to identify any bias (Flavelle, 1992).

Figures 3.20 and 3.21 exhibit the testing results for the slope and the intercept, respectively. For the slope results, the unit-slope hypothesis is rejected for all realizations using the head and head gradient data. For conductivity regression analysis, 474 realizations had a regression line slope that is statistically not significantly different than 1.0. That is, the unit slope hypothesis testing was accepted for these realizations. For the head zero intercept tests, the null hypothesis is rejected for all realizations, whereas 476 of the 500 hydraulic conductivity zero-intercept tests were accepted, and 377 of the 500 head gradient zero intercept tests were accepted.

It is important to look at multiple tests and evaluate the different aspects of each model realization in different ways. Some of the tests may give a false indication about performance, but the collective results of multiple tests will increase the odds that the correct decision about model performance is reached. The results of the hypothesis testing on the regression line lead to the fifth measure, P_5, needed in the decision tree process (Figure 2.2) for model validation.

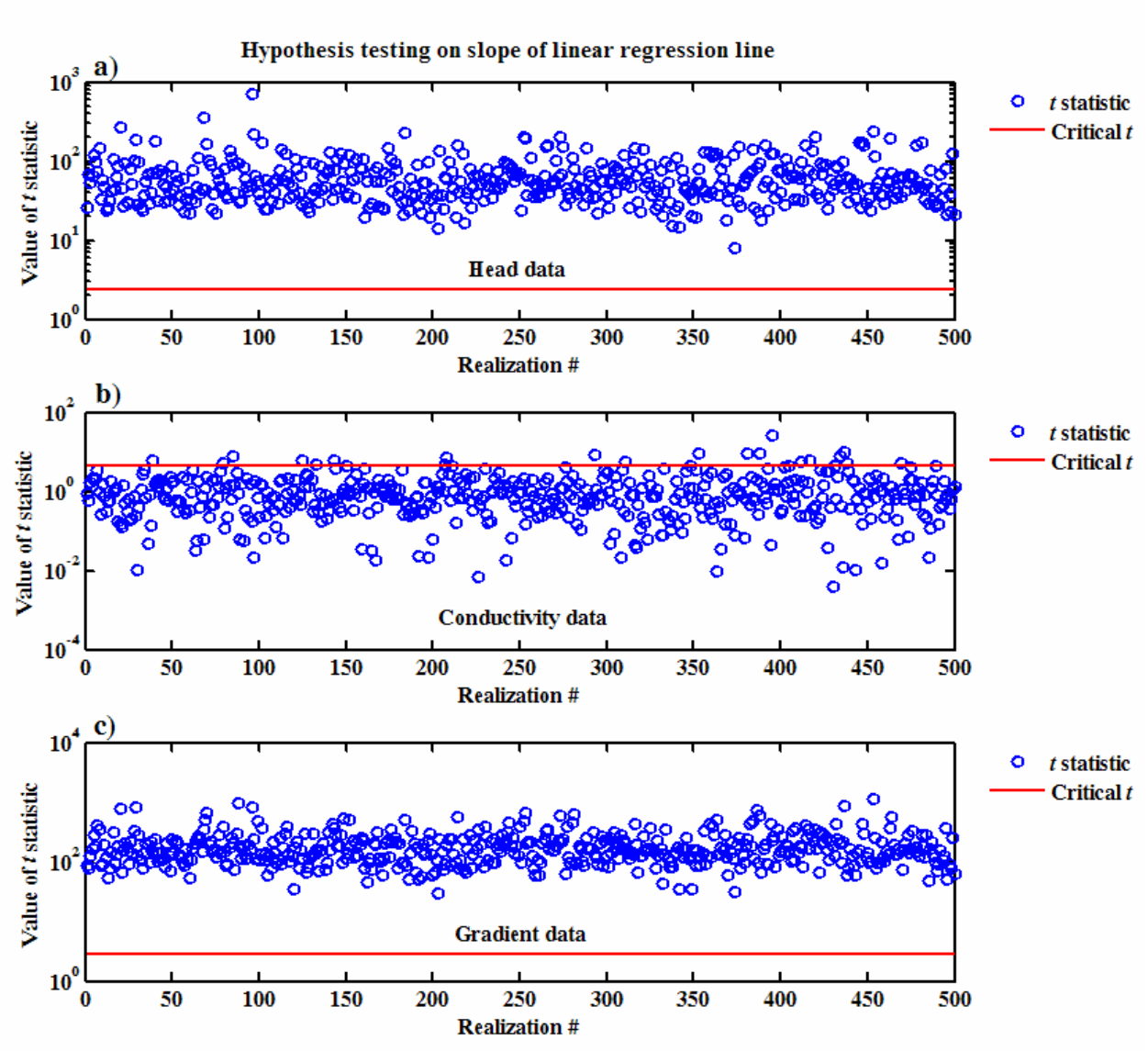


Figure 3.20. Results of hypothesis testing on the slope of the linear regression line using head data (a), hydraulic conductivity data (b), and head gradients (c).

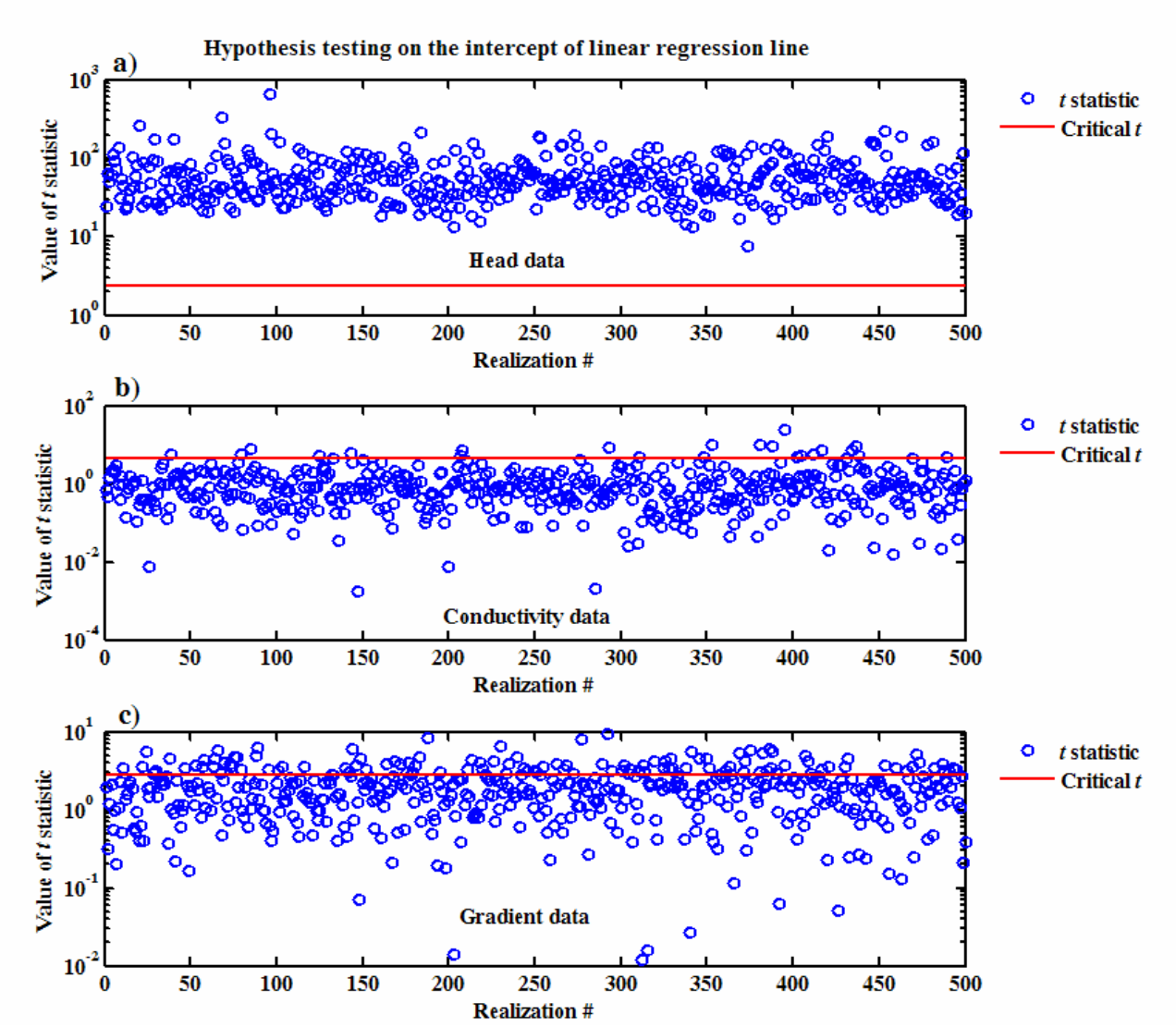


Figure 3.21. Results of hypothesis testing on the intercept of the linear regression line using head data (a), hydraulic conductivity data (b), and head gradients (c).

3.3.5 Testing Model Structure and Failure Possibility, P_5

Three types of analyses are performed here. First, the percentage of realizations where a model layer is assigned to each category is calculated. These percentages are compared to the field lithology obtained from the validation wells. If a model cell is assigned the correct category more than 50 percent of the time (i.e., in more than 250 realizations), then it is acceptable. If the percentage is less than 50 percent but is the highest percentage (see for example $k = 20$ and 22 at MV-2, Table 3.3), it can still be acceptable.

Table 3.2 displays the lithology comparison for MV-1. All layers at the MV-1 location were assigned lithology categories in the CNTA model based on new data from MV-1 except layer 18, which was most frequently considered tuffaceous sediment in the model, but which the validation data indicate should be alluvium cells. For this layer, 31 percent of the realizations assigned this cell as alluvium. Although layer 18 is considered unacceptable based on the 50-percent criterion mentioned above, it was still assigned the correct lithology in about one third of the model realizations. This layer in the model is within the zone of uncertainty in the alluvium-volcanic contact (Table 3.3). The model was constructed recognizing that the true alluvium boundary was likely to be between layers 15 and 19.

The lithology comparison results for MV-2 are shown in Table 3.3. The majority of model realizations assigned layers 13, 14, and 21 to categories different than those found in MV-2. They were assigned the correct category in only 15 percent, 16.4 percent, or 17 percent of the model realizations, respectively. This makes these cell assignments unacceptable. On the other hand, although model layers 20 and 22 do not meet the 50-percent criterion, they were assigned the correct lithology (Category 2 for layer 20 and Category 1 for layer 22) in about 48 percent and 40 percent of the model realizations, respectively. These are larger proportions than assigned to either of the other individual categories. They thus can be considered acceptable. The higher percentage of densely welded tuff assigned in the upper part of the volcanic section at MV-2 reflects the fact that this horizon was above the elevation of the conditioning data used to constrain the proportions of Categories 2 and 3 near the Faultless site. A much smaller percentage of densely welded tuff was observed in the volcanic section near the Faultless site, as compared to elsewhere in Hot Creek Valley. This lower percentage was honored by the conditioning data, but where conditioning data are absent (below the nuclear test horizon and at shallow elevations), the higher, regionally observed percentage (about 45 percent) of densely welded tuff was used as a conservative approach.

Table 3.2. Comparison between model lithology and field lithology at MV-1.

Well	Model Layer (k)	Percentage of realizations where the model cell belongs to each category			Field Lithology	Acceptable
		Alluvium (Category 1)	Tuffaceous Sediments (Category 2)	Densely Welded Tuff (Category 3)		
MV-1	27	100.00	0.00	0.00	1	Yes
	26	100.00	0.00	0.00	1	Yes
	25	100.00	0.00	0.00	1	Yes
	24	100.00	0.00	0.00	1	Yes
	23	100.00	0.00	0.00	1	Yes
	22	100.00	0.00	0.00	1	Yes
	21	100.00	0.00	0.00	1	Yes
	20	100.00	0.00	0.00	1	Yes
	19	61.80	22.80	15.40	1	Yes
	18	31.20	42.00	26.80	1	No
	17	10.80	63.00	26.20	2	Yes
	16	2.60	74.00	23.40	2	Yes
	15	1.20	79.40	19.40	2	Yes
	14	0.40	82.20	17.40	2	Yes
	13	0.20	79.40	20.40	2	Yes
	12	0.20	77.60	22.20	2	Yes
	11	0.00	74.60	25.40	2	Yes
	10	0.00	71.20	28.80	2	Yes
	9	0.00	69.80	30.20	2	Yes
	8	0.00	64.80	35.20	3	No
	7	0.00	55.00	45.00	2	Yes

Table 3.3. Comparison between model lithology and field lithology at MV-2.

Well	Model Layer (*k*)	Percentage of realizations where the model cell belongs to each category			Field Lithology	Acceptable
		Alluvium (Category 1)	Tuffaceous Sediments (Category 2)	Densely Welded Tuff (Category 3)		
MV-2	27	100.00	0.00	0.00	1	Yes
	26	100.00	0.00	0.00	1	Yes
	25	100.00	0.00	0.00	1	Yes
	24	100.00	0.00	0.00	1	Yes
	23	72.00	14.40	13.60	1	Yes
	22	40.40	31.20	28.40	1	Yes
	21	17.00	39.60	43.40	1	No
	20	6.20	48.40	45.40	2	Yes
	19	1.60	54.60	43.80	2	Yes
	18	0.20	62.20	37.60	2	Yes
	17	0.00	69.40	30.60	2	Yes
	16	0.00	75.00	25.00	2	Yes
	15	0.00	82.80	17.20	2	Yes
	14	0.00	83.60	16.40	3	No
	13	0.00	84.80	15.20	3	No
	12	0.00	83.60	16.40	2	Yes
	11	0.00	83.00	17.00	2	Yes
	10	0.00	82.20	17.80	2	Yes

For MV-3, the analysis indicates that only layers 7 and 9 can be considered unacceptable, as they were assigned Category 3 only 40 percent and 29.8 percent of the time in the model, respectively, and were assigned Category 2 60 percent and 71.2 percent of the time (Table 3.4). The lithology data from MV-3 indicate that these two layers belong to Category 3, densely welded tuff. Combining the results from the three wells, it can be concluded that the model lithology (or structure) is acceptable. Given that the interface between the alluvium and the tuffaceous sediments was taken as uncertain in the model, and given that densely welded tuff zones were infrequently encountered in the field, the results shown in Tables 3.2 through 3.4 show an overall good correspondence between the model and the lithologic validation data.

The lithology comparisons represent a second example of the importance of having a general view of the model and data and not relying heavily on the quantitative analysis and cell-to-data comparisons. The CNTA model structure was based on probabilities of occurrence of three hydrostratigraphic categories: alluvium, tuffaceous sediments, and densely welded tuff. These assignments were made based on spatial statistics gathered from boreholes and conditioned on borehole observations. At Well MV-3, densely welded tuff was

only encountered at the very bottom of the borehole. The model cell corresponding to that location was assigned to the tuffaceous sediment category in 60 percent of the realizations and to the welded tuff category in 40 percent of the realizations. The field data invalidated the lithology assigned in the model because most realizations called for tuffaceous sediments. This cell-by-cell comparison misses the larger issue of the overall proportions of each hydrogeologic category assigned by virtue of the spatial statistics. The original model hydrostratigraphy was based on borehole observations from throughout Hot Creek Valley. It was noted that the occurrence of densely welded tuff appeared to be smaller in the immediate vicinity of the Faultless test, but the proportion of densely welded tuff prescribed in the model was held to the higher regional value because those units were recognized to represent the critical pathways. Comparison of the aggregated hydrostratigraphy observed in the new wells with the proportions simulated in the model (Figure 3.22) clearly shows that the model over-represented densely welded tuff in the model domain.

Table 3.4. Comparison between model lithology and field lithology at MV-3.

Well	Model Layer (*k*)	Percentage of realizations where the model cell belongs to each category			Field Lithology	Acceptable
		Alluvium (Category 1)	Tuffaceous Sediments (Category 2)	Densely Welded Tuff (Category 3)		
MV-3	27	100.00	0.00	0.00	1	Yes
	26	100.00	0.00	0.00	1	Yes
	25	100.00	0.00	0.00	1	Yes
	24	100.00	0.00	0.00	1	Yes
	23	100.00	0.00	0.00	1	Yes
	22	100.00	0.00	0.00	1	Yes
	21	100.00	0.00	0.00	1	Yes
	20	100.00	0.00	0.00	1	Yes
	19	100.00	0.00	0.00	1	Yes
	18	69.20	20.40	10.40	1	Yes
	17	30.00	52.80	17.20	2	Yes
	16	9.20	72.20	18.60	2	Yes
	15	2.80	79.80	17.40	2	Yes
	14	0.60	82.40	17.00	2	Yes
	13	0.20	82.00	17.80	2	Yes
	12	0.00	81.00	19.00	2	Yes
	11	0.20	80.60	19.20	2	Yes
	10	0.00	74.00	26.00	2	Yes
	9	0.00	70.20	29.80	3	No
	8	0.00	64.00	36.00	2	Yes
	7	0.00	59.20	40.80	3	No

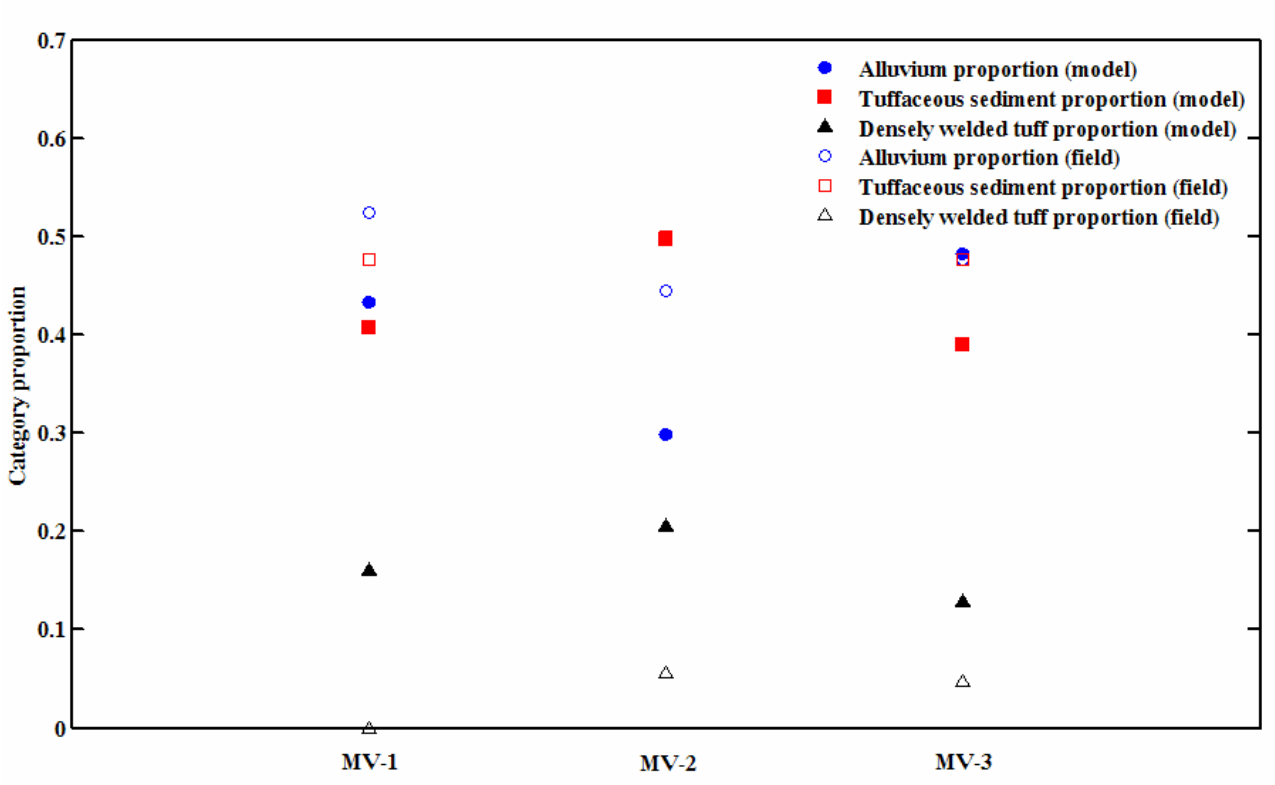


Figure 3.22. Proportions of different hydrostratigraphic assignments in the model and the results revealed by the validation data.

The second step in testing model structure is to comparison of the measured head variance, hydraulic conductivity variance, and gradient variance with the model predicted variances. This gives an overall idea of how the model structure compares to what is found from the validation data. The model predictions for the 19 validation targets are analyzed for each realization. The variance of the nine head values, σ_h^2, is obtained for the measured heads and for the modeled heads of each realization. Similarly, the four hydraulic conductivity values measured in the validation wells are used to compute $\sigma_{\log K}^2$, and a similar value is computed for each realization. The variance of the head gradients is also obtained from the six values inferred from the head measurements and a corresponding value is computed for each model realization. The results are plotted in Figure 3.23.

Ideally, each field point would plot within the cloud produced by the model realizations. This is, however, not the case for the CNTA model. The values computed based on field data (red circles) fall far from the plots of all model realizations, indicating a significant deviation between model structure and that inferred from the validation data. The field points fall far from the clouds created by model realizations in all three plots of Figure 3.23. Therefore, this second aspect of model structure indicates unacceptable model performance.

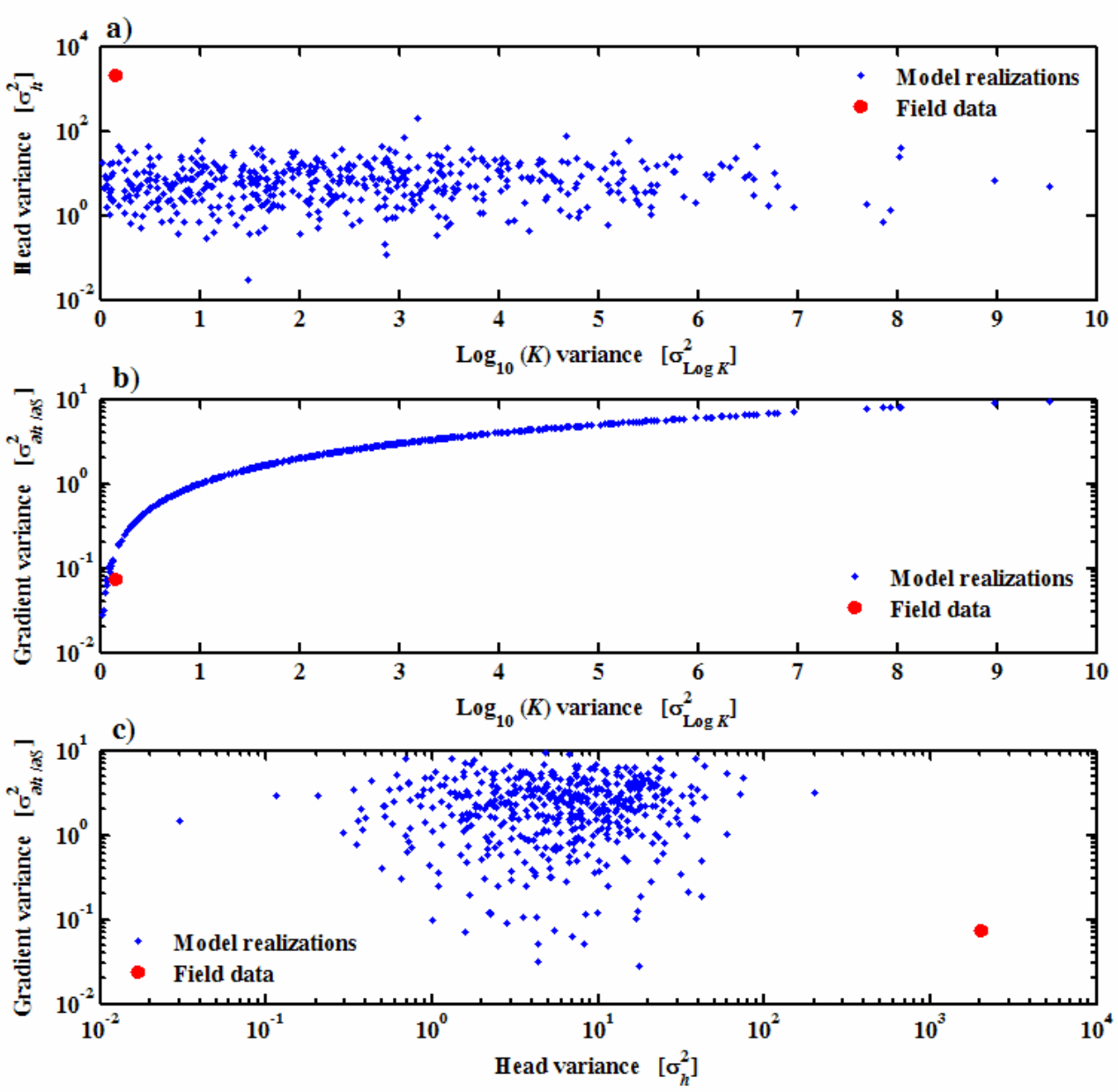


Figure 3.23. Relations between head, hydraulic conductivity, and gradient variances as obtained from the model and the validation data.

A third and final check of model structure and failure possibilities is for the presence of radionuclides (e.g., tritium) above background levels in the wells. Based on the analysis of tritium in samples collected from the three wells, no evidence is found of radionuclides above natural background.

3.4 Linking Calibration and Validation Analyses and Developing Composite Scores for Individual Realizations (Step 5)

The calibration and validation analyses performed in the previous sections can be categorized into two types. One type is applicable to individual realizations (e.g., goodness of fit measures, stochastic validation approach) and the other is applicable to the model as a whole (e.g., P_1 and P_2 measures, model structure test through variance relations). According to the decision tree (Figure 2.2), if P_1 is less than 30 percent and P_2 is less than 40 percent, which is the case here, then one evaluates the other measures (P_3, P_4, and P_5) and uses subjective judgment to determine whether the model would require major revision or whether it is a matter of a low number of acceptable realizations such that a revision could be made using the prevalidation data only (i.e., right-hand-side loop of the validation process, Figure 2.1).

The first type of analysis pertaining to the individual realizations is used to develop a composite score for each realization and determine the number of acceptable realizations. This number along with P_1, P_2, and the variance results (Figure 3.23) will guide the decision regarding the model assessment.

An acceptable realization in a perfect world would have a high calibration weight (the GLUE weights shown in Figure 3.6), values for the goodness-of-fit measures R^2, d, and d_1 as close to 1.0 as possible, accepted hypothesis testing on the aspects related to the residuals and the linear regression line, and lithology matching what the validation data indicated. To quantify these aspects, the following scoring system is used.

1. The calibration weight is divided by the maximum GLUE weight attained. This gives the realization with the maximum weight a score of 1.0 and all other realizations scores less than 1.0.
2. The goodness-of-fit results for different data sets are used as obtained, because R^2, d, and d_1 have values between zero (worst performance) and 1.0 (best performance).
3. The results of hypothesis testing are binary-type results (i.e., the null hypothesis is either accepted or rejected). These are converted to a binary [0, 1] system. A score of zero is given if the hypothesis is rejected and a score of 1.0 is given if the hypothesis is accepted.
4. The lithology results are scored based on the number of model cells matching the lithology validation data (Figures 3.3 through 3.5) relative to the total number of cells where lithology validation data are available. This gives a maximum score of 1.0 if the realization is matching all of the lithology validation data.

Based on this scoring system, the maximum score that is possible is 19.0 (Table 3.5). Table 3.5 displays the different tests and the scoring system for the first 15 realizations. These realizations attained scores ranging from 3.4 to about 7.4. None of the 500 realizations exceeds 8.0 on this composite score measure (Figure 3.24). Though a satisfactory score is an arbitrary assessment, it is clear from the results that the model would require major revision rather than simply production of additional realizations using prevalidation data.

Table 3.5. Example of the scoring system used to develop a composite score, showing results from 15 of the 500 realizations.

Realization #	$L_m(\vec{Y}\|\vec{\Theta})$	R^2			d			d_1			Residual tests (Luis and McLaughlin, 1992)		Hypothesis testing on regression line [Slope = 1.0 test]			Hypothesis testing on regression line [Intercept = 0.0 test]			Lithology	Total Score
		head data	conductivity data	gradient data	head data	conductivity data	gradient data	head data	conductivity data	gradient data	m_ε	χ^2	head data	conductivity data	gradient data	head data	conductivity data	gradient data		
1	0.1519	0.0926	0.119	0.0039	0.3749	0.1359	0.4516	0.25	0.0612	0.4166	0	0	0	1	0	0	1	1	0.8	**5.8576**
2	0.1121	0.0219	0.0121	0.0012	0.3733	0.2411	0.4553	0.2594	0.1041	0.4259	0	0	0	1	0	0	1	1	0.85	**5.8564**
3	0.2256	0.396	0.4387	0.7263	0.3807	0.0962	0.4691	0.2592	0.0892	0.4304	0	0	0	1	0	0	1	1	0.8167	**7.3281**
4	0.193	0.2272	0.5654	0.4682	0.37	0	0.4632	0.2513	0	0.4291	0	0	0	1	0	0	1	1	0.7333	**6.7007**
5	0.0621	0.002	0.7926	0.1883	0.3748	0.2684	0.4545	0.2632	0.1283	0.4256	0	0	0	1	0	0	1	1	0.8	**6.7598**
6	0.1204	0.3857	0.1147	0.6848	0.378	0.3377	0.4648	0.2616	0.2216	0.4299	0	0	0	1	0	0	1	1	0.7833	**7.1825**
7	0.1055	0.0239	0.7672	0.071	0.3732	0.1207	0.4593	0.2676	0.1272	0.4302	0	0	0	1	0	0	1	1	0.6833	**6.4291**
8	0.2435	0.0884	0.4599	0.7628	0.3686	0.3634	0.4507	0.2579	0.24	0.4236	0	0	0	1	0	0	1	1	0.7833	**7.4421**
9	0.0737	0.3354	0.1185	0.6998	0.3836	0.2056	0.4719	0.2577	0.1803	0.4263	0	0	0	1	0	0	1	1	0.8167	**6.9695**
10	0.4011	0.5989	0.3033	0.8082	0.379	0.1863	0.4795	0.2537	0.1175	0.436	0	0	0	1	0	0	1	1	0.65	**7.6135**
11	0.4293	0.0645	0.5847	0.1203	0.3741	0.4563	0.4521	0.2559	0.3005	0.4198	0	0	0	1	0	0	1	0	0.8	**6.2575**
12	0.2865	0.0009	0.0808	0.6102	0.3616	0.3498	0.4436	0.243	0.2351	0.4178	0	0	0	1	0	0	1	1	0.7667	**6.796**
13	0.0932	0.143	0.2767	0.1196	0.3772	0.3427	0.4601	0.2516	0.1974	0.4211	0	0	0	1	0	0	1	1	0.6833	**6.3659**
14	0.1975	0.039	0.2664	0.1457	0.3729	0.1709	0.4499	0.2526	0.1348	0.42	0	0	0	1	0	0	1	1	0.7	**6.1497**
15	0.3581	0.0263	0.4676	0.3516	0.3695	0.1151	0.4469	0.2515	0.1008	0.4157	0	0	0	1	0	0	1	1	0.7167	**6.6198**

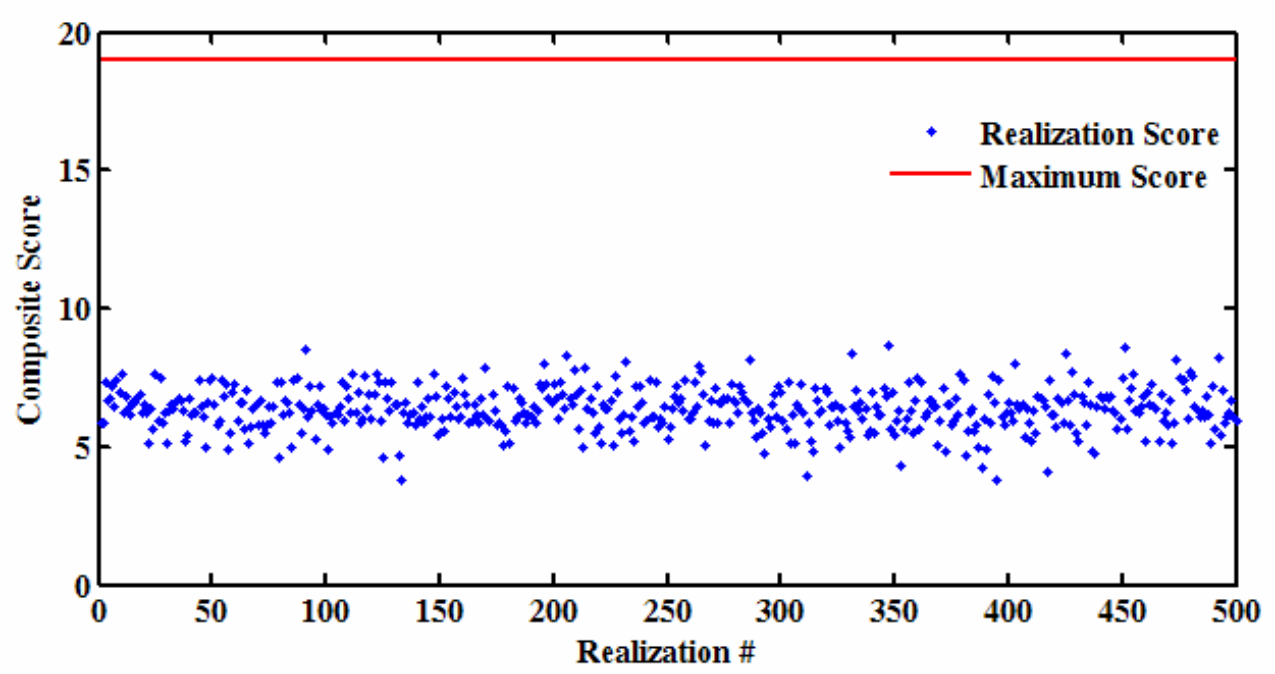


Figure 3.24. Composite score for all model realizations, including those presented in Table 3.5.

3.5 Final Assessment of Model Adequacy (Step 6)

Based on the results of the quantitative analysis and according to the validation process, the outcome is that the model would need major revision to adequately represent the validation data. As stated earlier in the discussion of the validation process, when the number of unacceptable realizations is very large compared to the total number of model realizations, the model either has a major deficiency or incorrect input distributions. Thus, there could be situations of poor model performance attributable to incorrect distributions of input parameters that could be corrected by simply adjusting those distributions. In the current case, though, the results of the validation tests indicate a major deficiency in the model in regard to head values and some flow directions.

In addition to the composite score for individual realizations, the tests performed for the whole model such as P_1, P_2, and the variance relations, indicate that the model has a major deficiency. The first criterion, P_1, was about 1 percent and the second criterion, P_2, was about 18 percent. Both did not meet the minimum threshold as per the decision tree (Figure 2.2). The variance relations plotted in Figure 3.26 indicate a significant deviation between the variances obtained from the model realizations and the field-based variances. These measures are consistent with the composite score findings and thus support the decision of the need for major revision for the model.

The validation analysis followed a rigorous quantitative process that strictly compared the new data to parameter values used in the CNTA CAU (Corrective Action Unit) model at those locations. By the measures prescribed in the CADD/CAP, the model fails to meet validation criteria. This means that the new data do not match the data assigned to the

particular model cells occupied by the wells in a sufficient number of realizations. This rigorous, quantitative approach, however, does not allow broader analysis of the model as a whole to determine what the fundamental failures are and their consequences. Two examples were discussed earlier in this section that pertain to the hydraulic conductivity comparisons (Figure 3.16) and the hydrostratigraphic assignments in the model (Figure 3.22). These examples demonstrated the limitations of adhering to a strict comparison of data to model-cell assignments.

If the prescribed process resulted in overall validation of the model, meaning that the field data were found to correspond to the cell-based assignments of the model realizations, adequacy of the overall model structure and performance could be easily presumed. In the present case, however, the lack of validation of the distributions assigned to individual cells provides little information regarding what aspects of the model are incorrect (and correct) and their significance. Before decision makers can determine if the validation results meet regulatory objectives (Step 7), a broader analysis of the model is required.

A less quantitative, but more hydrogeologically based, analysis of the CNTA CAU model relative to the data provided by the MV wells is presented in the following sections. This is done by examining the basic components and assumptions that comprise the CNTA model and assessing how those would be handled in light of the new data. The discussion concludes by examining the likely impact of including the processes and properties identified as needing revision, using a very simplified, modified model.

4.0. HOLISTIC CNTA MODEL EVALUATION

Rather than testing individual model realizations in regard to data targets for model cells, the following analysis examines the conceptual model, how it was implemented, and how that would change based on the new data. The interaction between the various components is also considered, as the ultimate problem of determining how far, how fast, and how much contaminant migration may occur is not the result of any single model aspect. The focus below is necessarily on flow model components because the new data predominantly pertain to flow parameters.

4.1 Assumption of Steady State

The original flow model was based on hydrogeologic conditions prior to the Faultless test under the assumption that transport over the long term would be controlled by these factors rather than short-term effects of the test. In addition, flow was considered to be at steady state owing to the large size of the Hot Creek Valley hydrologic system and the absence of excessive groundwater withdrawals. No new data call into question the steady-state nature of Hot Creek Valley as a whole, but the MV well data indicate that the hydraulic impacts of the Faultless test may be more persistent in both time and space than previously assumed. In particular, though it was recognized that aquifers were still recovering from the nuclear test within the down-dropped block defined by the most prominent faults, the head data from MV-1 suggest that an overpressured zone at the test elevation persists in that direction, outside the northeastern bounding fault. Though not at pretest conditions, the system may be at a quasi-steady state with respect to the overpressured zone. Water levels monitored in well UC-IP-1S over the last several decades have remained relatively stable despite being very elevated compared to assumed pretest conditions. This was previously ascribed to trapping of high pressures within test-related faults. The elevated heads observed

at the test horizon in wells MV-1 and MV-3 suggest that the elevated pressure may be primarily trapped by the very low *K* of the hydrostratigraphic unit. The timescales for decay of that pressure pulse are very long. Initial testing of a simplified transient model (discussed later) indicates that this pressure pulse may need hundreds of years to completely decay.

Though the CAU model predicts contaminant transport for 1,000 years into the future, the timescale over which monitoring and resource management decisions can be practically made is on the order of decades. Given the length of time expected for the nuclear test pressure pulse to dissipate, a revised flow model may retain a steady-state assumption, but would be updated to reflect the quasi-steady-state hydraulic system post-nuclear test, rather than conditions before Faultless. In other words, the assumption would be made that transport over the next management timescale would be controlled by the hydrogeologic conditions following the Faultless test.

4.2 Impact of Faults

Structural features, such as faults activated by the nuclear test, were not explicitly included in the CNTA model due to lack of information regarding their subsurface locations and hydraulic characteristics. The model was successfully calibrated (using pre-Faultless data) without representing the faults. No new data specifically regarding the faults were collected, but representing the larger set of hydraulic head data may now require the inclusion of faults, likely as barriers to groundwater flow. A revised flow model would begin by attempting to replicate the flow system without speculating about fault locations and properties, but should include such features if required for calibration.

4.3 Model Scale

The original CNTA model was constructed at a scale intermediate between the scale of the near-cavity environment and the scale of regional groundwater flow. Very low groundwater velocities led to a reduction in model size (Pohll *et al*., 2003) for calculation of the contaminant boundary. The 2003 model domain easily encompasses MV wells such that no modification of domain size would be required.

4.4 Identification of Hydrostratigraphic Units

Three categories of hydrostratigraphy were simulated in the Faultless model: Quaternary alluvium, Tertiary tuffaceous sediments (bedded tuffs, and partially welded tuffs), and Tertiary rhyolites and densely welded tuffs. Each of these units was encountered, but no additional hydrogeologic units were identified during drilling of the MV wells.

4.5 Spatial Representation of Hydrostratigraphic Facies

The geometry of the Quaternary alluvium category was delineated using thicknesses of alluvium in northern Hot Creek Valley estimated by Healey (1968). The approach used to configure this boundary in the model is described in detail by Pohlmann *et al*. (1999). Uncertainty in the location of the base of the alluvium was included in the model such that at unknown points (distant from wells), the base was picked for an individual realization within a 150-m vertical interval centered on the estimated location. Lithologic logs from MV-1 and MV-2 show that the alluvium is thicker in those locations than expected. The uncertainty in alluvium depth was broad enough, however, that some realizations simulated alluvium to the depths encountered in the wells. The alluvium encountered at MV-3 matched the model predictions very well. These results suggest that the approach to simulating the geometry of

the alluvium category is robust and that the uncertainty bounds used are broad enough. A revised model would benefit from the additional conditioning points of the new wells, but the overall representation of the alluvium would not change significantly.

The Tertiary volcanic section was divided into Category 2 (tuffaceous sediments, bedded tuffs, and partially welded tuffs) and Category 3 (rhyolites and densely welded tuffs) on the basis of geophysical log signatures. The procedure is described in Pohlmann *et al*. (1999) and relies on the high electrical resistivity of densely welded tuffs. The model domain below the alluvium was populated with Categories 2 and 3 based on a spatial correlation structure developed through analysis of the well log data. A vertical variogram was calculated, but the large horizontal spacing between wells prevented development of a horizontal variogram. The horizontal correlation length was estimated using a 10:1 horizontal to vertical anisotropy ratio. The vertical correlation length was 325 m, while the horizontal was 3325 m. The new data indicate that the values used in the model were conservative in that they amply covered correlations observed in the new wells.

The multiple realizations of the three-dimensional maps of hydrogeologic categories were populated in the volcanic section by adhering to relative proportions of Categories 2 and 3. The target values of the simulation proportions of the two categories (61 percent for Category 2 and 39 percent for Category 3) represented a compromise between values determined from the entire CNTA dataset (throughout Hot Creek Valley) and values determined from wells in the immediate Faultless area. In the entire dataset, the volcanic rocks are evenly divided between Category 2 and 3, whereas the local dataset was dominated by Category 2 (73 percent of the volcanic section penetrated by boreholes). It was assumed that the pattern of Categories 2 and 3 observed throughout Hot Creek Valley would be present in the lower part of the domain, below the emplacement well where there were no observations. This can be seen in the category assignments at the new well locations (Tables 3.2, 3.3, and 3.4) by the significant increase in Category 3 simulated in the cells below the original conditioning data. The new data support the absence of densely welded tuff in the volcanic section at and above the nuclear test elevation at MV-1 and MV-3. Figure 4.1 presents comparisons between the model assignments and the MV data in terms of the proportions of the three categories above (Figure 4.1a) and below (Figure 4.1b) the Faultless working point. In both cases, the simulated proportions of the densely welded tuff are overestimated. In addition, though one densely welded tuff interval was intercepted near the base of wells MV-1 and MV-3, the new data indicate that there is less Category 3 at depth than assumed in the original model. A model conditioned on the new data would simulate fewer layers of densely welded tuff immediately below the cavity than the original model. Realizations with densely welded tuff were the only ones that resulted in significant transport.

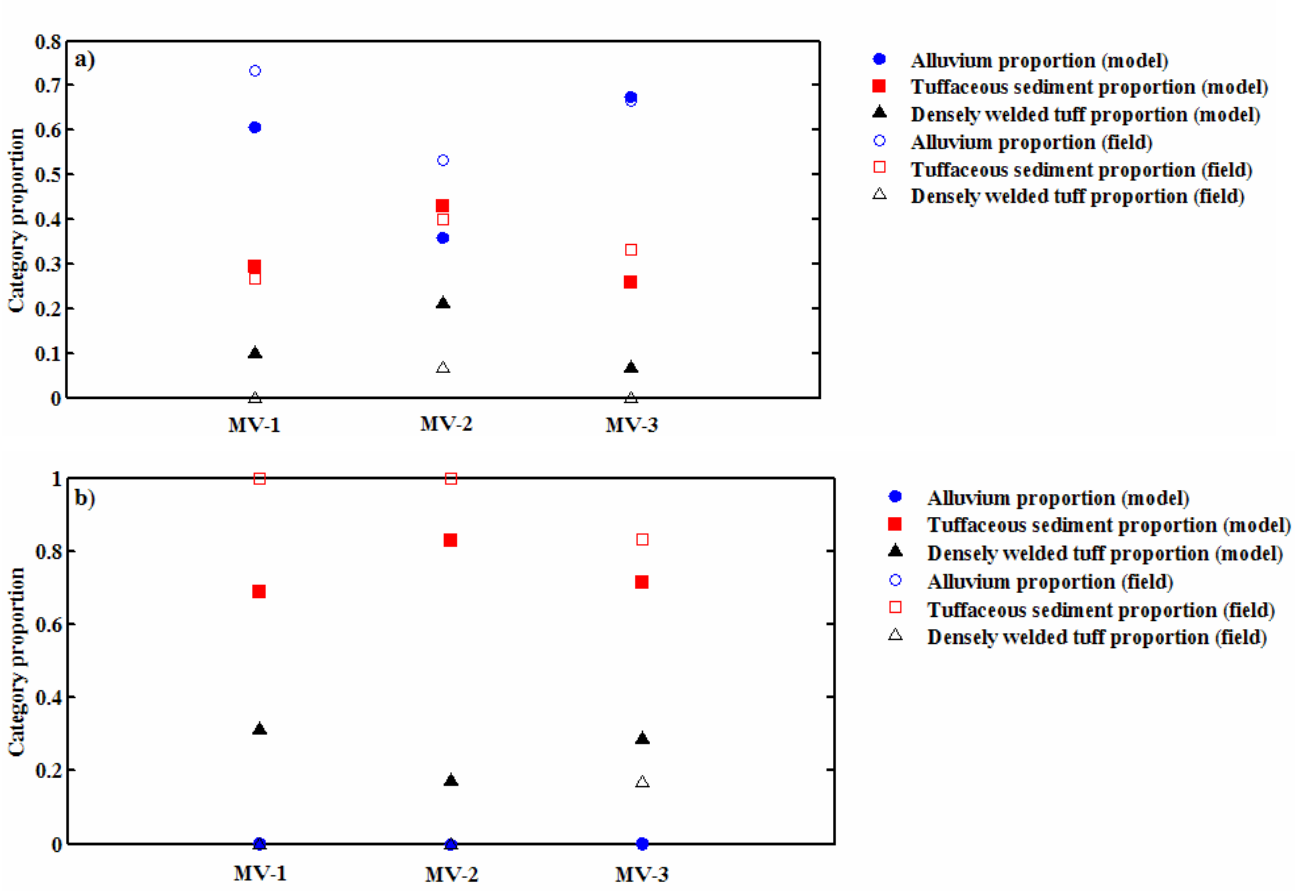


Figure 4.1. Comparisons of simulated and observed proportions of Categories 1, 2, and 3 a) above and b) below the working point in the three MV wells.

4.6 Hydraulic Conductivity Distributions

Extensive hydraulic testing was performed in wells in the CNTA prior to the Faultless test. These tests were conducted in open-hole intervals or through perforated casing using inflatable straddle packers and were designed to characterize vertical variations in hydraulic head, hydraulic gradient, relative specific capacity, storage coefficient, groundwater velocity, and hydrochemistry. The packer test data formed the basis for characterizing the spatial distribution of *K* in the CNTA flow model (Pohlmann *et al.*, 1999). The distribution for alluvium was based on nine measurements, 23 for Category 2, and 26 for Category 3. The means and distributions followed the trend expected from the conceptual model, with the highest *K* values in Category 3 (representing densely welded tuffs), the lowest in Category 2 (representing tuffaceous sediments), and the alluvium values intermediate.

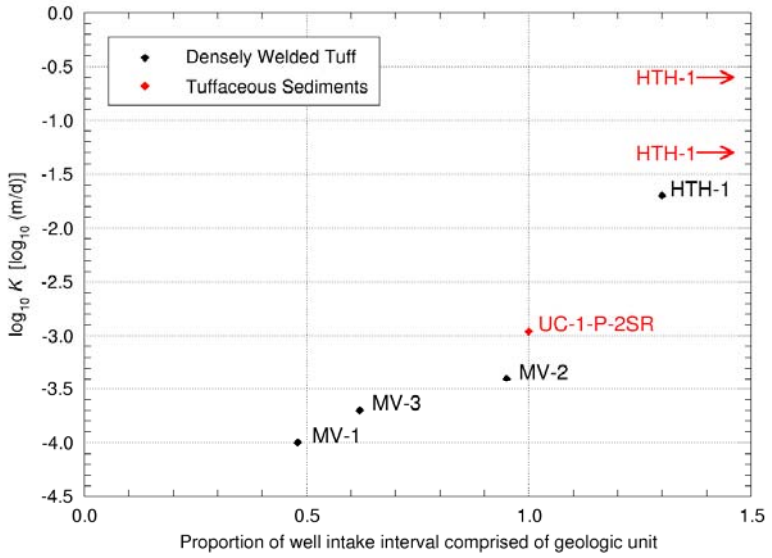
Hydraulic tests have only been performed in the three main well intervals in the MV wells, all three of which are completed in densely welded tuffs. The lower piezometer in MV-2 was also tested, but its screen was open to the tuffaceous sediments section. Qualitatively, drilling records of borehole fluid balance indicate that the alluvium was somewhat permeable and that the tuffaceous sediments were of very low permeability, consistent with expectations. However, the *K* estimated for the densely welded tuff from the aquifer tests is much lower

than the distribution applied in the model (Figure 3.16). As noted previously, very little densely welded tuff was encountered in the Faultless area such that the packer test data for densely welded tuff was predominantly from tests elsewhere in the valley. The results from the MV wells suggest that the conceptual model of densely welded tuff as highly conductive, fractured pathways may be in error in the immediate Faultless area. Apparently, not only is densely welded tuff less common than elsewhere in the valley, when it does occur, its conductivity more closely resembles values previously considered as representative of the tuffaceous sediments. A revision of the Faultless flow model would face a decision regarding whether to maintain the conceptual model but spread the distribution of densely welded tuff *K* to include lower values, or abandon the distinction between tuffaceous sediments and densely welded tuff.

A more detailed analysis of the hydraulic conductivity values from the MV wells has been performed to address the issue of why they are so much lower than *K* values estimated from earlier wells near Faultless. The main difference between the *K* values estimated from the MV wells and the results from the original packer tests is that a considerably larger volume of water was used in the constant-rate MV tests. Even though the pumping period was relatively short for MV-1, it was treated as a slug test in the analysis. In contrast, the original packer tests were conducted by injecting or swabbing through small-diameter tubing into the testing interval. In addition, the MV tests were conducted through slotted casing with gravel packs, while the packer tests were conducted through shot-perforated casing that was cemented in place. The design of the MV tests should provide a more representative hydraulic response of the tested formation than the packer tests.

Another aspect to consider in the comparison is the length of the tested interval (Figure 4.2). In the case of HTH-1, the packed interval was always smaller than the thickness of the unit tested, such that there is confidence that the tested unit contributed/accepted the majority of the stress. For MV-1 and MV-3, the slotted interval was quite a bit longer than the thickness of the densely-welded tuff, with the remainder of the interval comprised of non-, partly, or moderately welded tuff. Thus, the *K* values calculated from these tests are likely to be impacted by the influence of these lower-*K* rocks; the less densely welded tuff in the slotted interval, the lower the estimated *K*. There is a linear correlation between *K* and the proportion of intake interval open to the densely welded tuff for the MV wells (Figure 4.2). However, the slotted interval of MV-2 matched the thickness of the densely welded tuff, so the interval length alone cannot account for the lower *K* values in the MV wells.

Another issue related to the interval length is that the *K* values used for validation were calculated using the full thickness of the aquifer, not the length of the slotted interval, which would be more consistent with the original packer tests. The values in Figure 4.2 were calculated using the intake interval length, not the full aquifer thickness. Table 4.1 exhibits the *K* values calculated using the full aquifer thickness, the screen length, and just for comparison purposes, the thickness of the densely welded tuff alone. As the length is reduced, the estimated *K* value increases, but in all cases the results are still lower than any of the other volcanic rocks in the Faultless area. Despite the issues raised above, the fact still stands that the recent tests in the MV wells show that the densely welded tuff has very low *K* in the Faultless area, even lower than the few *K* values reported for the tuffaceous sediments. Based on these results, a case could be made to eliminate the higher *K* category in the model.



Notes: *K* values calculated from *T* values using intake interval length
MV well intake length = 50 m
HTH-1 packer interval length in densely welded tuffs = 18.4 m
HTH-1 tuffaceous sediments thickness is many times greater than packer interval length
UC-1-P-2SR *K* value calculated from recovery associated with cavity filling

Figure 4.2. Conductivity values plotted against interval lengths for the MV wells and the preexisting wells at Faultless. The MV well intake length is 50 m and the HTH-1 packer interval in the densely-welded tuff is 18.4 m. Tuffaceous sediment thickness is many times greater than the HTH-1 packer intervals in that unit. Conductivity for UC-1-P-2SR is calculated from water level recovery associated with resaturation of the cavity and rubble chimney subsequent to the Faultless test.

Table 4.1. Conductivity values from the MV wells computed using different lengths.

Hole Name	Rock Unit	K (m/d)	ln K (ln m/d)	log K (log m/d)	Length (m)	T (ft²/d)	T (m²/d)
DWT thickness:							
MV-1 main	dwt	2.0e-04	-8.51	-3.69	23		
MV-2 main	dwt	3.9e-04	-7.84	-3.41	47		
MV-3 main	dwt	2.2e-04	-8.43	-3.66	30		
Slotted interval length:							
MV-1 main	dwt	9.5e-05	-9.26	-4.02	48.8		
MV-2 main	dwt	3.9e-04	-7.84	-3.41	49.7		
MV-3 main	dwt	2.2e-04	-8.44	-3.67	49.4		
Full aquifer thickness:							
MV-1 main	dwt	1.6e-05	-11.04	-4.79	289	5.0e-02	4.6e-03
MV-2 main	dwt	7.6e-05	-9.48	-4.12	256	2.1e-01	2.0e-02
MV-3 main	dwt	6.7e-05	-9.61	-4.17	159	1.1e-01	1.1e-02

4.7 Groundwater Flow Directions

Based on hydraulic head measurements scattered throughout Hot Creek Valley, groundwater flow in the alluvium was believed to flow toward the south. Similarly, measurements in Hot Creek and surrounding valleys formed the basis for an assumption of generally northward flow in the deeper volcanic system. Head was assumed to decrease with depth (recharge conditions) in the north part of Hot Creek Valley, and increase with depth (discharge conditions) in the south part. The immediate CNTA area in the model was located in the zone of transition between these vertical gradient regions, with downward flow in the immediate test area and upward gradients to the south.

Head data from the MV wells and HTH-2 are used to solve the three-point problem for flow direction determination at each screen level. The upper piezometer measurements provide flow direction in the alluvium, which can be compared to the gradient in the model based on the mean head distribution of the stochastic model realizations (Figure 4.3). Multiple three-point problems are solved by choosing different combinations of the four wells. The results using the three MV wells are shown in Figure 4.3a, using MV-1, MV-2, and HTH-2 wells are shown in 4.3b, using MV-1, MV-3, and HTH-2 wells are shown in 4.3c, and using MV2, MV-3, and HTH-2 wells are shown in 4.3d.The model produces a mostly southerly-southwesterly flow direction in the alluvium layer but the MV heads (and the combination with HTH-2) indicate more of an easterly-southeasterly direction. Similarly, Figures 4.4 and 4.5 display the flow directions for the intermediate screen (open to the tuffaceous sediments at the general elevation of the nuclear test) and the lower screens (open to a densely welded tuff), respectively. The direction from the MV well data in the nuclear-test layer is almost reversed compared to the model. However, in the densely welded tuff unit, the flow directions indicated by the wells and by the model are almost the same, pointing to the north-northwest direction.

The elevated heads at the nuclear test horizon in the MV wells are likely to represent a trapped pressure pulse caused by the test and low *K* barriers in the test vicinity. Therefore, the flow directions indicated by the MV wells are representative of local (both in space and time) directions, whereas the model was designed to represent regional conditions. A revised model considering the trapped pressure pulse and the presence of some low-*K* barriers may be capable of matching the flow directions inferred from the MV wells.

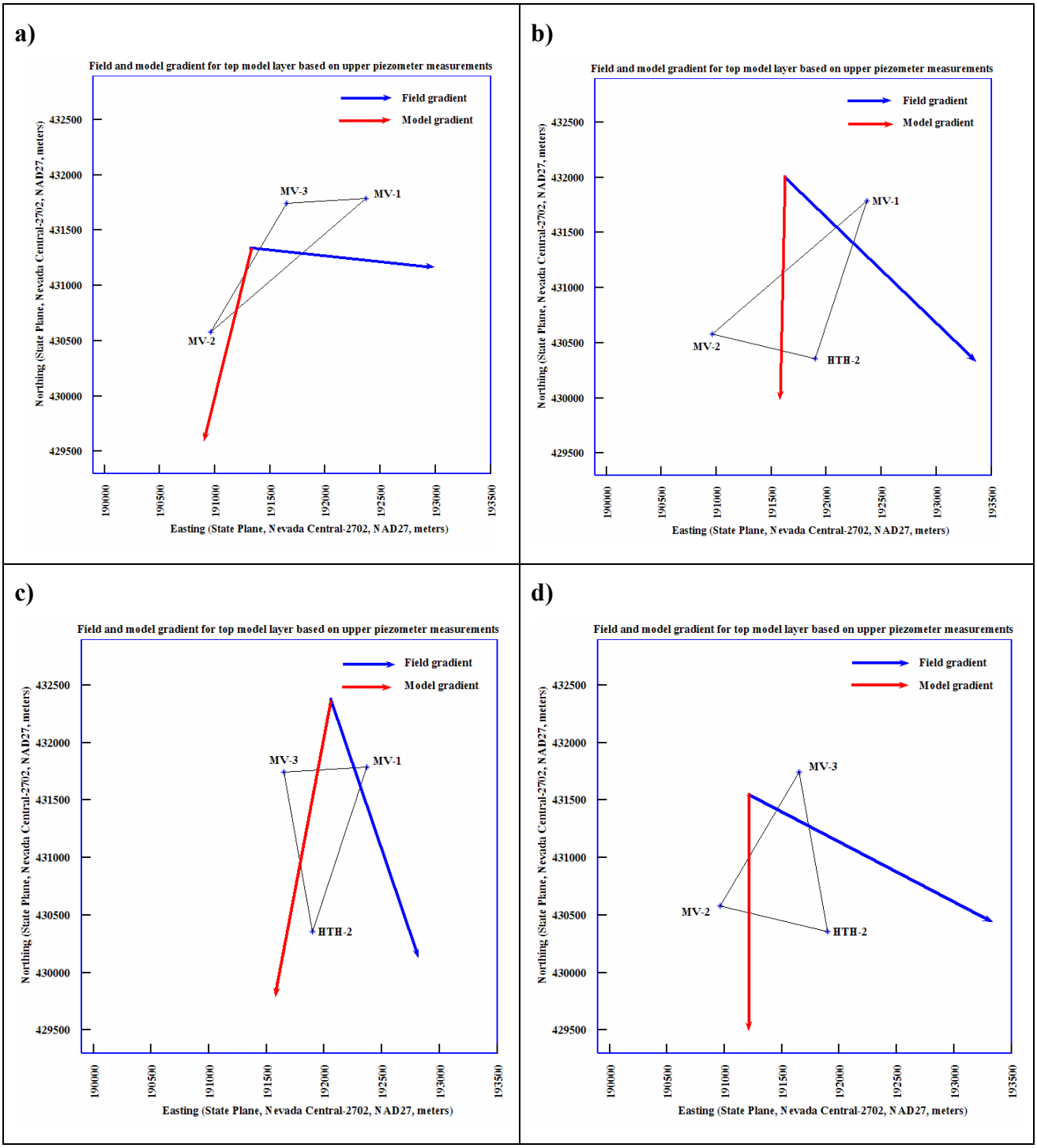


Figure 4.3. Solution of the three-point problem for flow direction at the upper screen level using field data from MV wells (MV-1-U, MV-2-U, MV-3-U) and HTH-2 compared to the flow direction using the mean heads from the model.

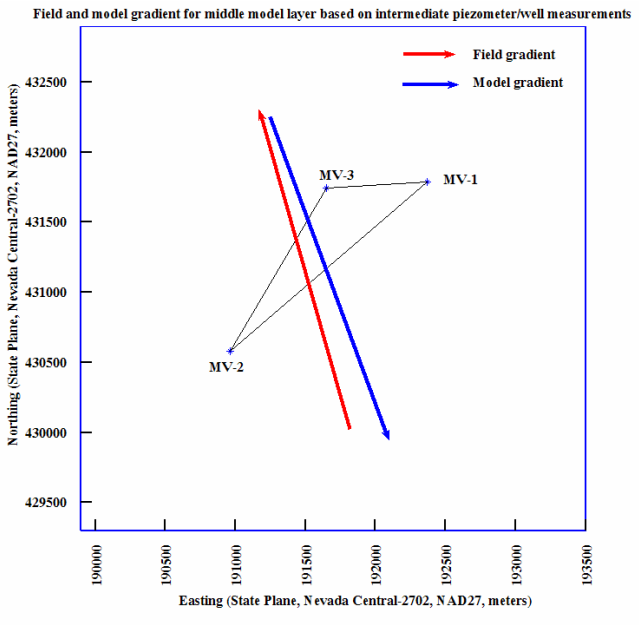


Figure 4.4. Solution of the three-point problem for flow direction at the intermediate screen level (approximate nuclear test elevation) using field data from the MV wells (MV-1-L, MV-2-W, MV-3-L) compared to the flow direction using the mean heads from the model.

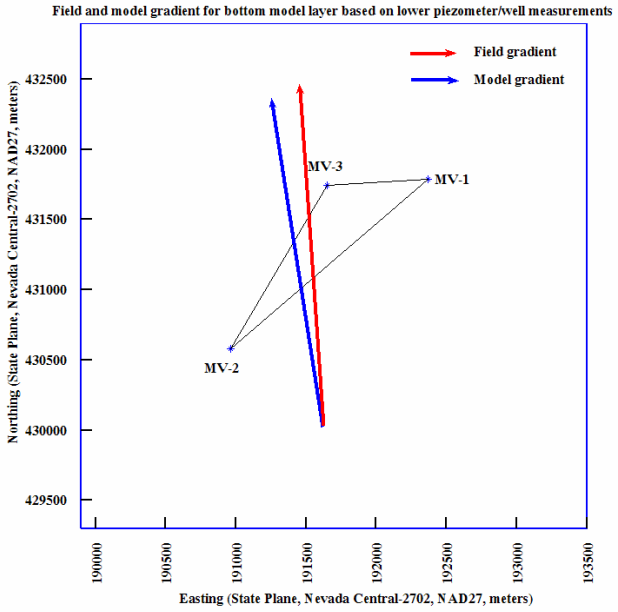


Figure 4.5. Solution of the three-point problem for flow direction at the lower screen level (densely welded tuff) using field data from the MV wells (MV-1-W, MV-2-L, MV-3-W) compared to the flow direction using the mean heads from the model.

4.8 Transport Parameters

The CNTA transport model included distributions of porosity, assignments of retardation and matrix diffusion, and assumptions regarding the apportioning and release characteristics of radionuclides. The MV wells did not collect data pertaining to these transport parameters, though water samples were collected from the wells to assess the absence of transport predicted by the model. Qualitatively, the abundance of tuffaceous sediments and low *K* values encountered in the wells is consistent with both significant reactive surfaces for sorption reactions, and presence of porous material available for matrix diffusion.

4.9 Implications of the Validation Results and Expected Outcomes of a Revised Model

The CNTA flow model was based on hydrogeologic conditions prior to the Faultless test, under the assumption that transport over the long term would be controlled by these conditions rather than the relatively short-term effects of the test (Pohlmann *et al*., 1999). Furthermore, flow was considered to be at steady state owing to the large size of the Hot Creek Valley hydrologic system and the absence of excessive groundwater withdrawals such that significant temporal fluctuations in regional water levels were not expected under current climatic conditions. Local structural features such as faults were not explicitly included due to the lack of information regarding their subsurface locations and hydraulic characteristics.

The validation data indicate that the assumptions regarding the hydraulic impact of the nuclear test and faults are wrong. Given the observation that heads in all three validation wells are much higher than modeled based on pre-test conditions, it seems that the near-field conditions impacted by the test and the down-dropped block have major impacts on the heads in the wells. The fact that these impacts have persisted indicates the possibility that they are long-term and not just temporary impacts as perceived in the original modeling study.

Model revision to incorporate nuclear-test effects on the flow system is a major effort that may or may not be needed to meet the regulatory objectives of the site. This will be decided as Step 7 in the validation process, a decision for DOE and NDEP. A simplified, preliminary three-dimensional model incorporating some of the near-field features is presented here to test some hypotheses regarding the elevated heads observed in the MV wells. In the following, the geometry of this preliminary revised model is described and the model features are discussed. Boundary and initial conditions are also presented, followed by a discussion of some preliminary results and their implications for the CNTA closure process.

4.9.1 Description of Preliminary Revised Model

The model domain used in Pohll *et al*. (2003) for calculation of the contaminant boundary is used here. The model domain is 3.6 km long on each side (approximately twice the length of the land withdrawal area), is centered over UC-1, and is aligned in the north-south direction (Figure 3.1). The domain covers the same 1,350-m vertical section included in the 1999 and the 2003 model. This model domain easily encompasses the MV wells such that no modification of domain size is required.

The domain is assumed to be composed of two lithologic layers, the alluvium layer and the volcanic rock layer, which includes the tuffaceous sediments and the densely welded tuff together as one unit. For simplicity and for the purpose of this preliminary testing, no spatial heterogeneity is considered within the two layers. Several faults, as identified by

McKeown *et al*. (1968) are included around the Faultless cavity. However, the downward extent of these features and their dip/strike are not known, and as such, they are considered to be vertical and to extend across the entire model thickness.

Due to the presence and the size of these features, the original finite-difference code (MODFLOW) used in Pohlmann *et al*. (1999) and Pohll *et al*. (2003) could not be used here. Instead, the FEFLOW (Diersch, 1998) finite-element code is used for the flow simulations of this preliminary model. FEFLOW is a finite-element simulation package available from the WASY Institute for Water Resources Planning and Systems Research Ltd., developed for two-dimensional and three-dimensional density-dependent flow, mass, and heat transport processes in groundwater. It is well-suited for incorporating features such as faults, shear zones, and discrete fracture zones. Being a finite-element code, FEFLOW provides the ability to refine the domain discretization around important features in the model. The faults around the Faultless cavity were discretized and incorporated in the FEFLOW-based model.

The preliminary model consists of 13 layers (Figure 4.6). The uppermost three layers represent the alluvium stratum with hydraulic conductivity equal to 0.0487 m/day. The remaining 10 layers represent the volcanics with hydraulic conductivity equal to 8.49×10^{-4} m/day. These values of hydraulic conductivity are equal to the mean log hydraulic conductivity values used for Categories 1 and 2, respectively, in the 1999 and the 2003 models.

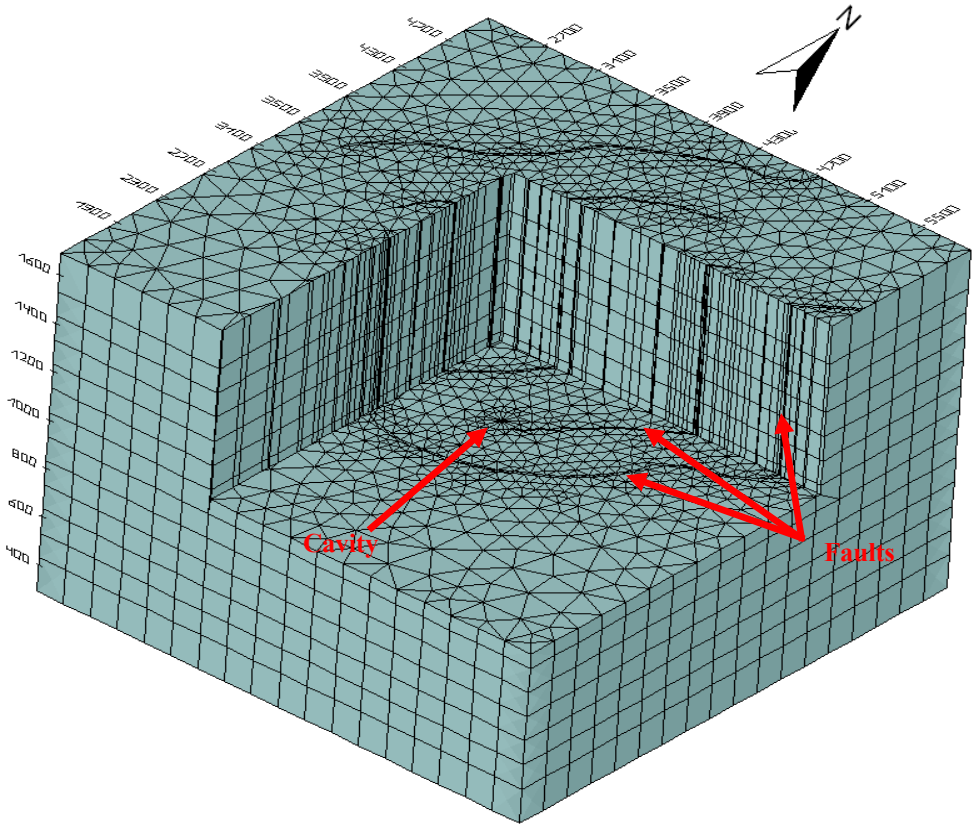


Figure 4.6. The finite-element mesh of the revised model showing the location of the faults and the test cavity.

Other than the inclusion of faults and the change from finite differences to finite elements for the method of flow simulation, this preliminary revision of the groundwater flow model is configured in much the same way as the original model. The northern, southern, and bottom faces are specified head boundaries, and the east and west faces are no-flow boundaries (this configuration is based on the predominantly north-south flow patterns simulated in the 1999 model that were based on regional data analysis) (Figure 4.7). Although the heads from MV wells indicate different flow directions in the upper and middle model layers, these may be local and may be impacted by bounding faults. Therefore, the boundary conditions were not changed in this preliminary testing. The aim is to incorporate the least amount of change in the 1999 model as a first step to explain the reason for the elevated heads in the MV wells.

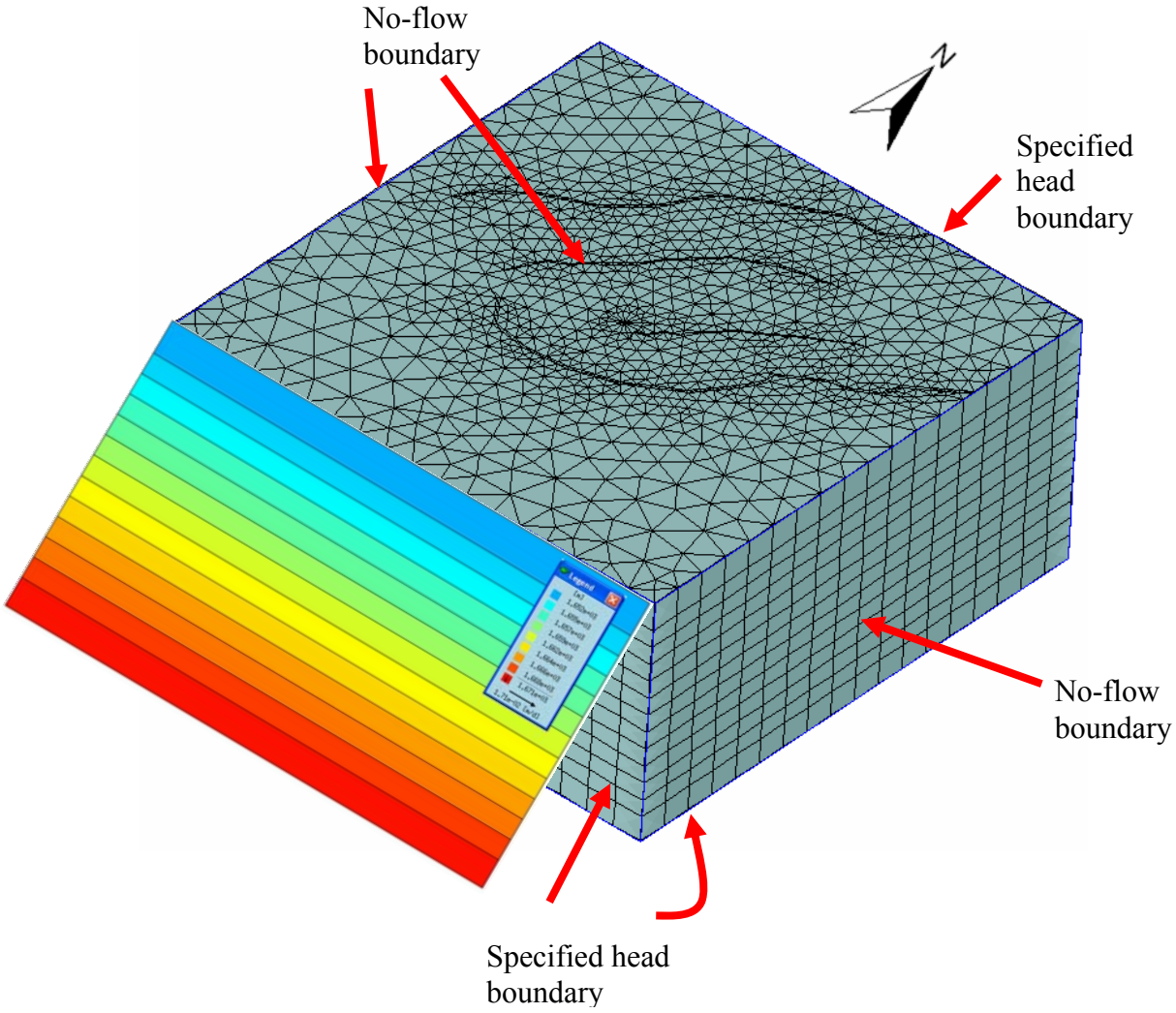


Figure 4.7. Schematic diagram showing the boundary conditions for the revised model.

Boundary heads are estimated using head measurements made in the straddle-packed intervals of the CNTA exploratory wells UCe-20, UCe-18, andHTH-1. As in the 1999 model, the process uses HTH-1 as the starting point, as it is located much closer to Faultless than the

other wells. Estimates of head at the elevations of the top and base of the model at the location of HTH-1 were calculated in the 1999 model by vertical extrapolation of HTH-1 head values. Similarly, the horizontal gradients at the top and base of the model are determined using the uncertainty in the linear regression of head data from UCe-20, UCe-18, and HTH-1. The heads of the remaining nodes on the specified head boundaries are obtained by linear interpolation between the heads on the edges. Figure 4.8 shows the bottom boundary condition of the model where the heads are linearly interpolated from 1,668.6 m south to 1,652.4 north.

The simulations are performed in two stages. First, the model is run to steady state using homogeneous isotropic properties for the two layers, the alluvium layer and the volcanic rock layer. The main objective of running the steady-state condition is to obtain the initial head condition for the transient case.

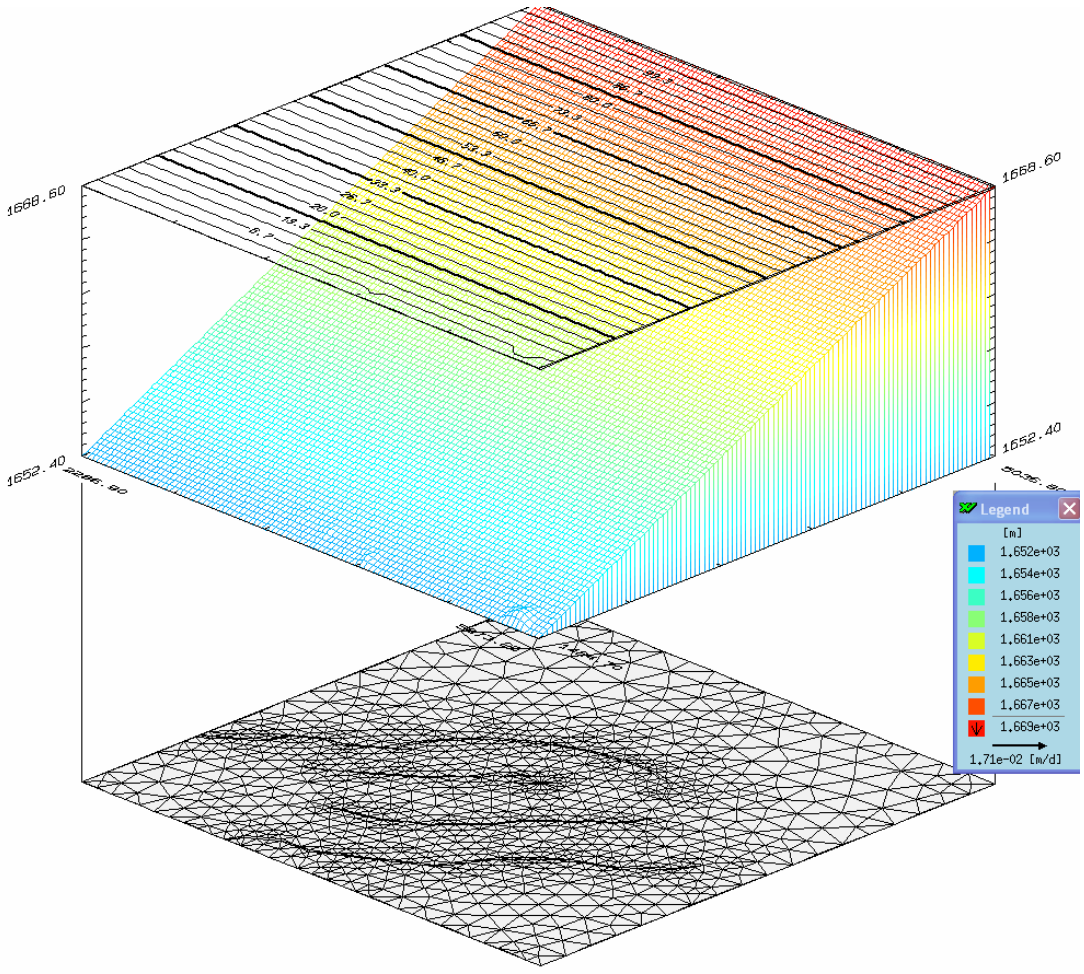


Figure 4.8. Specified head boundary conditions for the model base.

Second, the model is run in a transient mode to investigate the impact of faults on an elevated pressure pulse that is imposed on the domain as an initial condition. This pressure pulse is due to the Faultless test and is assumed to impact three model layers at, above, and

below the working point. Close to the cavity, the nuclear test caused a drop in water level, as observed in UC-1-P-2-SR. Away from the cavity, the pressure pulse caused rises in the water levels in wells HTH-1, HTH-2, and UC-1-P-1S. To honor these observations, the initial head distribution is assumed to be about 1,200 m AMSL within a circle surrounding the cavity and to gradually increase radially outward to a maximum value of about 1,900 m at a radial distance of about 900 m from the working point. Figure 4.9 displays the initial pressure distribution in the middle layer (passing through the Faultless working point) which is also imposed on the layer above and the layer below. The initial heads in the remaining layers are assigned the steady-state head values.

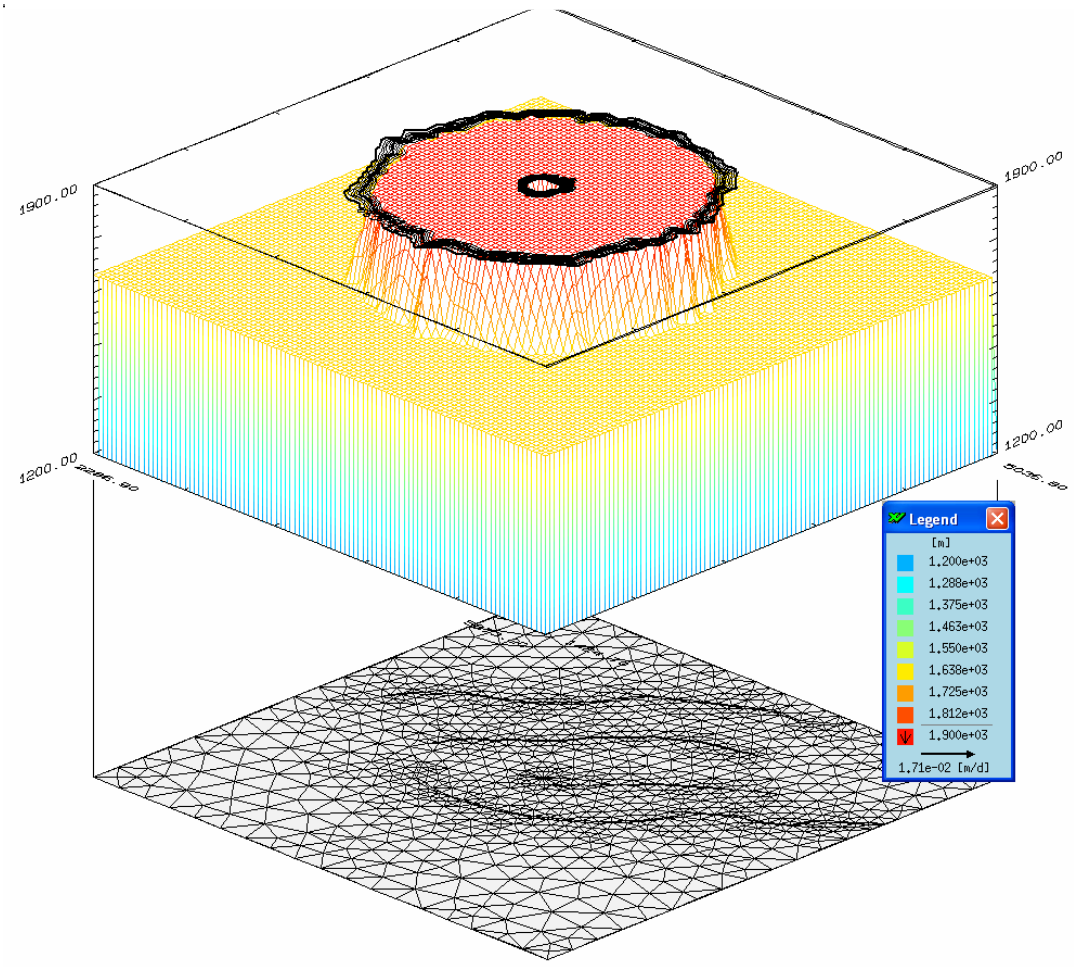


Figure 4.9 Initial head distribution at the model layer passing through the test cavity.

4.9.2 The Steady-state (Calibration) Results

The steady-state values of hydraulic head in HTH-1 and HTH-2 are used to evaluate the goodness of fit (i.e., calibration) of the model and its ability to represent the steady-state conditions prior to the Faultless test. Heads at the MV well locations are also monitored. The distribution of the 19 steady-state head observation points is as follows: nine points at different screen levels in HTH-1, one point at HTH-2 and three points for each of the MV wells. The data from the MV wells are not used in this calibration. They are simply observed to evaluate the differences in the model results at the MV wells locations between the steady-state case and the transient case.

The comparisons between the measured heads and the modeled heads are shown in Table 4.2. Similar to the previous models, this model accurately simulates the heads in both HTH-1 and HTH-2. However, it could not reproduce the elevated head values in the MV wells. As can be seen from Figure 4.10, the hydraulic heads in all layers under steady-state conditions range from 1,662 to 1,670 m. It is evident from the steady-state results that the high heads at the MV wells cannot be readily obtained by simply adjusting the parameters of the steady-state model.

Table 4.2. Comparison between the measured heads and the modeled heads using the revised simplified model under the steady-state conditions.

Data							Model Steady		Error	
#	Screen Elev.		Layer Elev	Name	Slice	Measured Head	Obs point	H	Squ Err	Err
1	1,619.37	1,573.37	1,610	HTH1	1	1,664.30	1	1,664.43	0.02	-0.13
2	1,542.37	1,481.87	1,480	HTH1	2	1,664.15	6	1,664.43	0.08	-0.28
3	1,405.67	1,375.17	1,380	HTH1	3	1,664.75	7	1,664.43	0.10	0.32
4	1,326.37	1,308.17	1,280	HTH1	4	1,663.40	8	1,664.43	1.06	-1.03
5	1,268.47	1,228.87	1,180	HTH1	5	1,664.00	9	1,664.43	0.18	-0.43
6	1,161.77	1,131.37	1,080	HTH1	6	1,665.20	10	1,664.39	0.66	0.81
7	1,027.67	1,006.37	980	HTH1	7	1,662.80	11	1,664.35	2.40	-1.55
8	933.17	914.97	880	HTH1	8	1,661.90	13	1,664.31	5.81	-2.41
9	738.37	715.37	680	HTH1	10	1,664.00	17	1,664.22	0.05	-0.22
10	1,655.56	1,500.00	1,610	HTH2	1	1,667.10	2	1,664.26	8.07	2.84
11	987.92	884.90	880	MV-1L	8	1,809.84	14	1,663.33	21,465.77	146.51
12	1,630.44	1,545.70	1,610	MV-1U	1	1,753.16	3	1,666.25	7,554.04	86.91
13	757.49	633.74	680	MV-1W	10	1,694.89	18	1,661.42	1,120.11	33.47
14	847.27	771.07	780	MV-2L	9	1,752.69	16	1,663.89	7,885.80	88.80
15	1640.66	1,571.17	1,610	MV-2U	1	1,776.88	4	1,664.71	12,582.56	112.17
16	994.49	892.99	980	MV-2W	8	1,781.62	12	1,664.30	13,764.92	117.32
17	946.40	834.24	880	MV-3L	8	1,820.89	15	1,663.37	24,811.61	157.52
18	1,655.98	1,568.80	1,610	MV-3U	1	1,766.77	5	1,666.18	10,118.55	100.59
19	768.09	593.44	580	MV-3W	10	1,691.06	19	1,660.63	925.80	30.43

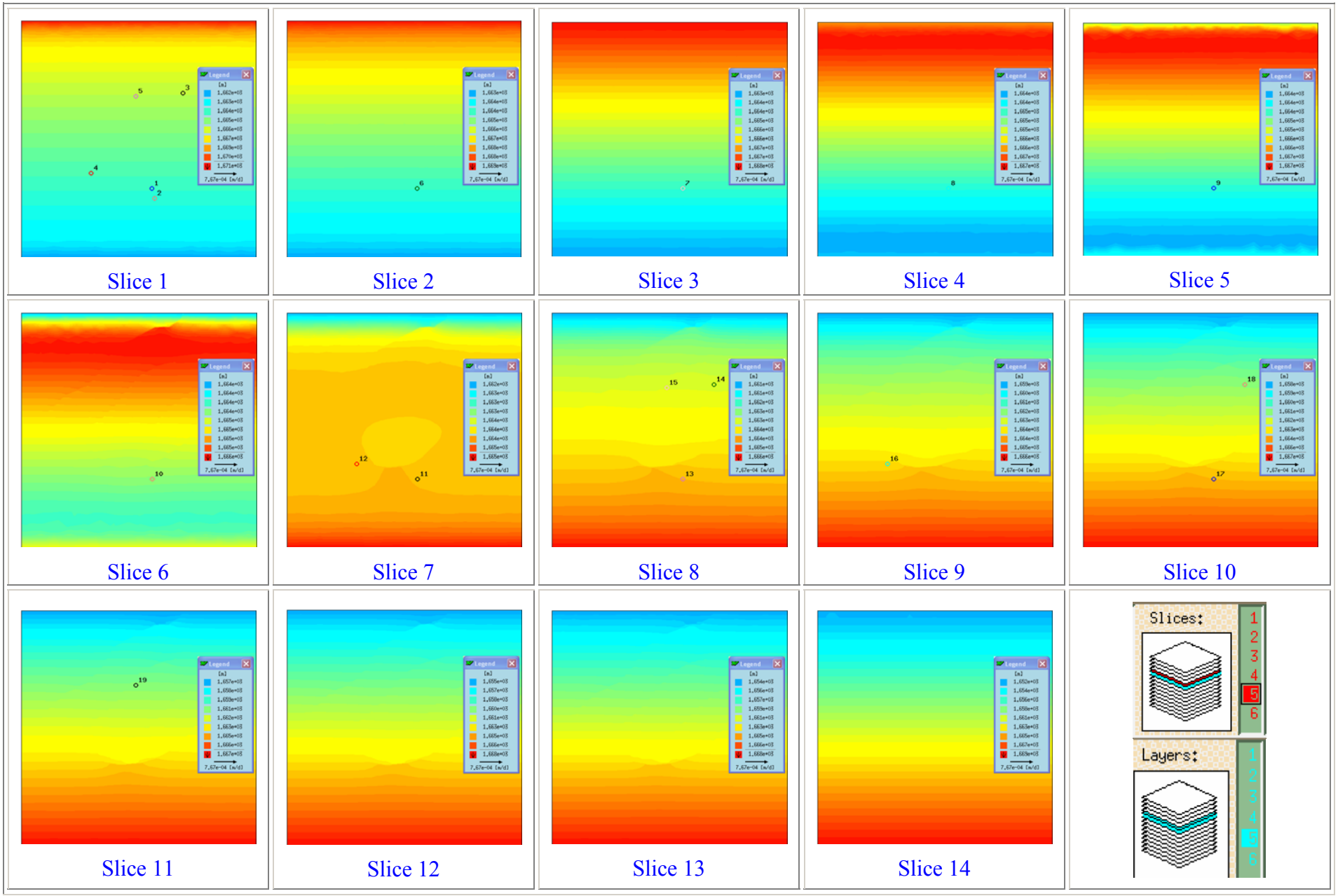


Figure 4.10. Steady-state head results of the revised model.

4.9.3 The Transient State Initial Results

The main objective of the transient case is to investigate the effect of the faults, and the effect of the elevated pressure pulse caused by the Faultless test. The fault locations are identified based on McKeown *et al*. (1968), and the hydraulic conductivity of these faults is assumed to be 1×10^{-8} m/day. There are no data to support the extent or the magnitude of the pressure pulse. Therefore, the location of the pressure pulse is assumed to be in the middle layer containing the cavity. Two hypotheses for the initial pressure conditions were evaluated. In the first hypothesis, the head values just after the test at the locations of wells HTH-1, HTH-2, UC-1-P-1S, and UC-1-P-2SR were used as initial conditions for the head at the cavity layer. Initial heads in between these wells were linearly interpolated. Figure 4.11 shows the initial heads under the first hypothesis. This hypothesis failed to simulate the high-head values in the MV wells. That is because wells HTH-1, HTH-2, UC-1-P-1S, and UC-1-P-2SR and the linear interpolation between them did not generate initial high head values in the locations of the MV wells. Consequently, the initial head at the location of the MV wells were at the steady-state values, which were significantly lower than the measured heads.

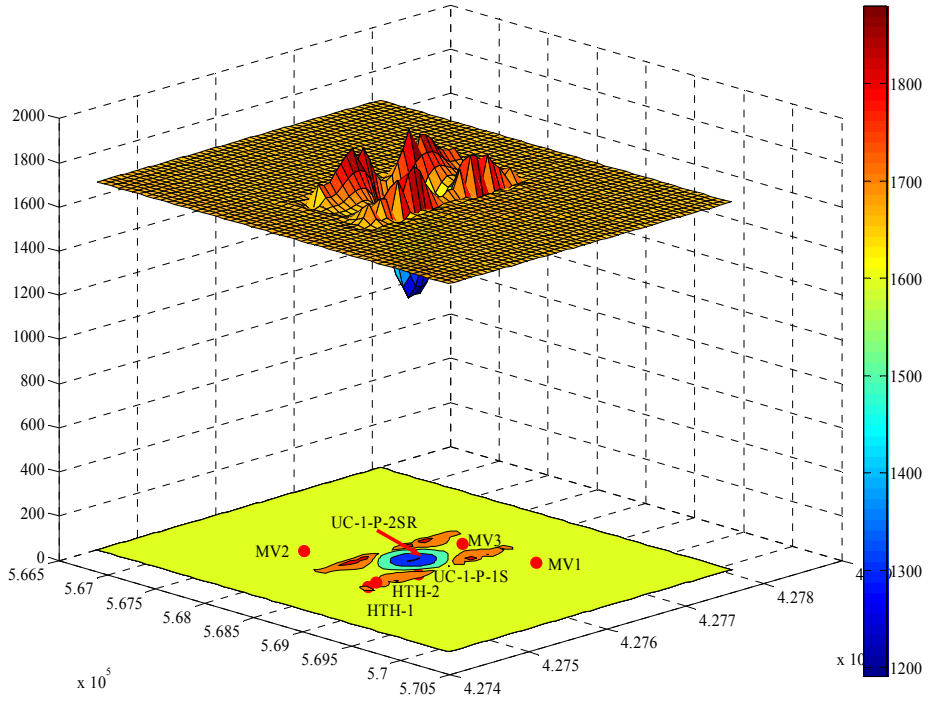


Figure 4.11. Simulation of the pressure pulse (first hypothesis) and the initial head distribution**.**

In the second hypothesis, a circle centered on the cavity with a radius of about 900 m (to pass through MV-2, the farthest well from the cavity) is used to simulate the pressure pulse. Two observations at UC-1-P-1S after the Faultless test (2/2/1968 and 2/11/1968) showed that water was flowing from the well at rates of 0.32 L/s and 0.95 L/s, respectively. Therefore, the maximum head value is assumed to be 1,900 m (about 62 m

above land surface level at UC-1-P-1S). The initial head used under this hypothesis is shown in Figure 4.9. This second hypothesis succeeded at simulating the high head values in the MV wells. Therefore, a manual calibration is used to obtain the parameter values that best simulate the observations. The objective function used in the calibration was the sum of squared error at each observation point.

The calibration parameters in the manual calibration were 1) the storage compressibility in both alluvium and volcanic rocks, and 2) the hydraulic conductivity and the storage compressibility in the three layers adjacent to the cavity (referred to here as the cavity zone). It is assumed that the properties of the volcanic rocks in the cavity zone are different from the host rock as a result of the nuclear test. Figure 4.12 shows the extent of this zone that extends laterally from the cavity to the boundary of the down-dropped block. The hydraulic conductivity anisotropy ratio and conductivity values of the alluvium and host volcanic rocks were held constant at values representative of the 1999 and 2003 models. The faults were assigned a constant hydraulic conductivity value equal to 1×10^{-8} m/day.

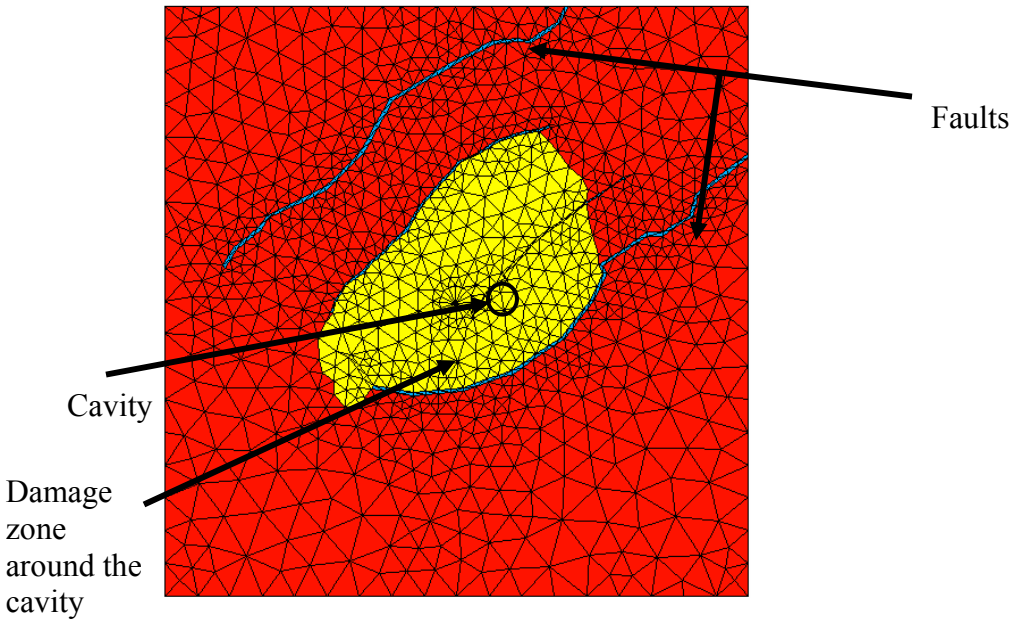


Figure 4.12. The lateral extent of the cavity zone where values of hydraulic parameters may have been impacted by the Faultless test.

The objective function was used in the calibration to minimize the sums of squared errors between the current measurements in the observation wells and the model prediction. Seven wells (with 35 observation points) were used in the calibration process. Table 4.3 gives the summary of the wells used in the calibration and the elevation of their screens. One challenge in the calibration process was interpreting composite head measurement in wells HTH-1, HTH-2, UC-1-P-1S, and UC-1-P-2SR. That is because there are multiple, or very long, screened intervals in these wells. FEFLOW does not produce composite heads, only heads simulated in each model layer. Therefore, the average model results were compared to the single composite value. Figure 4.13 shows

the distribution of the calibration wells and their intervals in relation to the revised model layers. The best parameters obtained from the manual calibration of the revised model under transient conditions are summarized in Table 4.4.

Table 4.3. Summary of the wells used in the calibration processes.

ID #	Well name	Data		Model	
		Screen Elev. (m)		Slice #	Slice Elevation (m)
	HTH-1	1,619.37	1,573.37	1	1,690
2	HTH-1	1,542.37	1,481.87	2	1,530
3	HTH-1	1,405.67	1,375.17	3	1,430
4	HTH-1	1,326.37	1,308.17	4	
5	HTH-1	1,268.47	1,228.87	5	1,230
6	HTH-1	1,161.77	1,131.37	6	1,130
7	HTH-1	1,027.67	1,006.37	7	1,030
8	HTH-1	933.17	914.97	8	930
9	HTH-1	738.37	715.37	10	730
10	HTH-2	1,655.56	1,500	1	1,690
11	HTH-2			2	1,530
12	UC-I-P-1S	1,533.33	1,000	2	1,530
13	UC-I-P-1S			3	1,430
	UC-I-P-1S			4	1,330
15	UC-I-P-1S				1,230
16	UC-I-P-1S			6	1,130
17	UC-I-P-1S			7	1,030
18	UC-I-P-2SR	885	1,360	4	1,330
19	UC-I-P-2SR			5	1,230
20	UC-I-P-2SR			6	1,130
21	UC-I-P-2SR			7	1,030
22	UC-I-P-2SR			8	930
23	UC-I-P-2SR			9	830
24	MV-1W	757.49	633.74	10	730
25	MV-1L	987.92	884.9	8	930
26	MV-1U	1,630.44	1,545.7	1	1,690
27	MV-1U			2	1,530
28	MV-2W	994.49	892.99	8	930
29	MV-2L	847.27	771.07	9	830
30	MV-2U	1,640.66	1,571.17	1	1,690
31	MV-2U			2	1,530
32	MV-3W	768.09	593.44	10	730
33	MV-3L	946.4	834.24	8	930
34	MV-3U	1,655.98	1,568.8	1	1,690
35	MV-3U			2	1,530

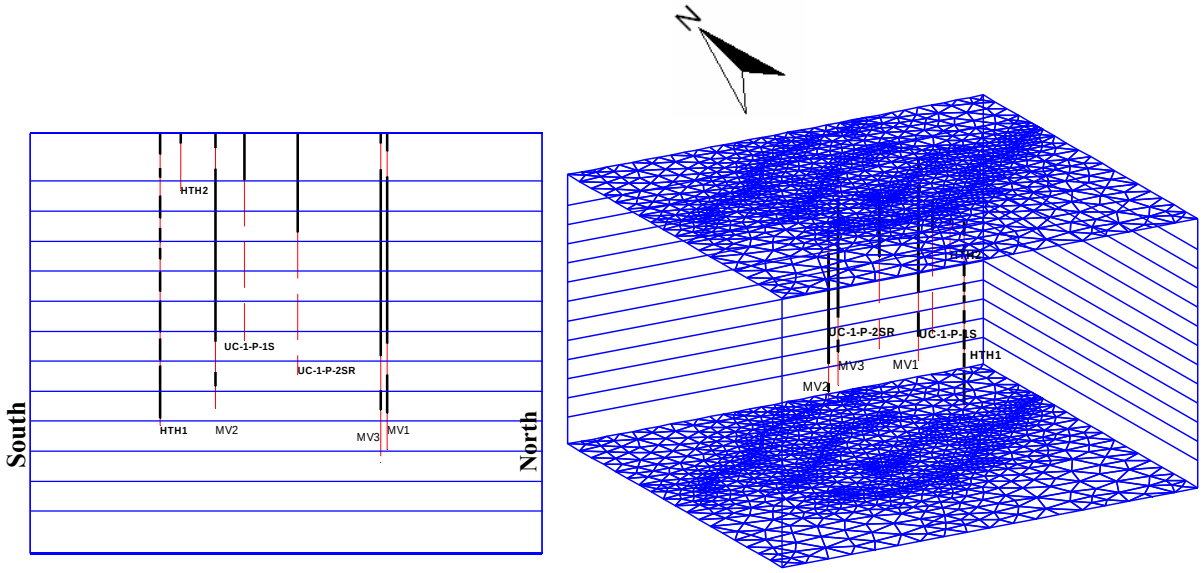


Figure 4.13. The distribution of open well intervals within the model layers.

Table 4.4. Parameter values for the optimal solution obtained during calibration of the revised model under transient conditions.

	Hydraulic conductivity x-direction (m/day)	Anisotropy ratio in y-direction (m/day)	Anisotropy ratio in z-direction (m/day)	Storage compressibility
Alluvium	4.87E-02	1.0	5.3	1.00E-04
Volcanic rocks	8.49E-04	1.0	5.3	1.00E-04
Cavity zone	5.28E-04	1.0	5.3	9.00E-04

The comparison between the measured heads and the model results is summarized in Table 4.5. Figure 4.14a shows the initial head distribution in slice 8 (cavity layer) at the beginning of the simulation, whereas Figure 4.14b shows the head distribution in the same layer after 40 years (i.e., the current conditions). It is evident from the figure that simulating the faults as flow barriers can retain the high pressure pulse within their boundaries. As a result, the error in well MV-1 decreased to 12 m from 146 m in the original steady-state model, the error in well MV-2 decreased from 88 m to 34 m, and the error in well MV-3 decreased from 157 m to 3.6 m. These results indicate that including the faults as hydraulic boundaries, and including the pressure transient from the nuclear test, may allow for simulation of the heads observed in the MV wells, especially in the lower region of the model. A more detailed model with refined discretization, recalculated boundary conditions, and incorporation of more detailed near-field conditions may thus be able to reproduce the head distribution observed in the MV wells.

Table 4.5. Comparison between observed heads and simulated heads for the transient model 40 years after the test.

Field Data					Transient Model			
#	Well	Screen Elev. (m)		Measured Head (m)	Obs. #	Model Head (m)	Sq Error	Error
1	HTH-1	1,619.37	1,573.37	1,668.13	1	1,664.32	14.52	3.81
2	HTH-1	1,542.37	1,481.87	1,668.13	2	1,664.40	13.94	3.73
3	HTH-1	1,405.67	1,375.17	1,668.13	3	1,664.53	12.99	3.60
4	HTH-1	1,326.37	1,308.17	1,668.13	4	1,664.74	11.52	3.39
5	HTH-1	1,268.47	1,228.87	1,668.13	5	1,680.48	152.42	-12.35
6	HTH-1	1,161.77	1,131.37	1,668.13	6	1,698.13	899.70	-29.99
7	HTH-1	1,027.67	1,006.37	1,668.13	7	1,712.64	1,980.78	-44.51
8	HTH-1	933.17	914.97	1,668.13	8	1,713.04	2,017.09	-44.91
9	HTH-1	738.37	715.37	1,668.13	9	1,684.57	270.31	-16.44
10	HTH-2	1,655.56	1,500.00	1,667.36	10	1,,664.19	10.05	3.17
11	HTH-2	1,655.56	1,500.00	1,667.36	11	1,664.27	9.58	3.10
12	UC-1-P-1S	1,533.33	1,000.00	1,755.61	12	1,671.34	7,101.77	84.27
13	UC-1-P-1S	1,533.33	1,000.00	1,755.61	13	1,671.74	7,034.34	83.87
14	UC-1-P-1S	1,533.33	1,000.00	1,755.61	14	1,672.35	6,933.06	83.26
15	UC-1-P-1S	1,533.33	1,000.00	1,755.61	15	1,714.50	1,689.70	41.11
16	UC-1-P-1S	1,533.33	1,000.00	1,755.61	16	1,768.19	158.21	-12.58
17	UC-1-P-1S	1,533.33	1,000.00	1,755.61	17	1,839.48	7,034.01	-83.87
18	UC-1-P-2SR	885.00	1,360.00	1,685.34	18	1,672.05	176.76	13.29
19	UC-1-P-2SR	885.00	1,360.00	1,685.34	19	1,711.18	667.71	-25.84
20	UC-1-P-2SR	885.00	1,360.00	1,685.34	20	1,753.80	4,686.91	-68.46
21	UC-1-P-2SR	885.00	1,360.00	1,685.34	21	1,747.73	3,892.26	-62.39
22	UC-1-P-2SR	885.00	1,360.00	1,685.34	22	1,647.72	1,415.04	37.62
23	UC-1-P-2SR	885.00	1,360.00	1,685.34	23	1,691.29	35.41	-5.95
24	MV-1W	757.49	633.74	1,694.89	24	1,665.55	861.13	29.35
25	MV-1L	987.92	884.90	1,809.84	25	1,784.78	627.95	25.06
26	MV-1U	1,630.44	1,545.70	1,753.16	26	1,670.34	6859.81	82.82
27	MV-1U	1,630.44	1,545.70	1,753.16	27	1,670.45	6840.61	82.71
28	MV-2W	994.49	892.99	1,781.62	28	1,798.71	291.90	-17.09
29	MV-2L	847.27	771.07	1,752.69	29	1,691.04	3801.22	61.65
30	MV-2U	1,640.66	1,571.17	1,776.88	30	1,666.72	12,135.01	110.16
31	MV-2U	1,640.66	1,571.17	1,776.88	31	1,666.82	12,112.98	110.06
32	MV-3W	768.09	593.44	1,691.06	32	1,665.77	639.48	25.29
33	MV-3L	946.40	834.24	1,820.89	33	1,865.53	1,992.46	-44.64
34	MV-3U	1,655.98	1,568.80	1,766.77	34	1,670.87	9,196.62	95.90
35	MV-3U	1,655.98	1,568.80	1,766.77	35	1,671.09	9,154.66	95.68

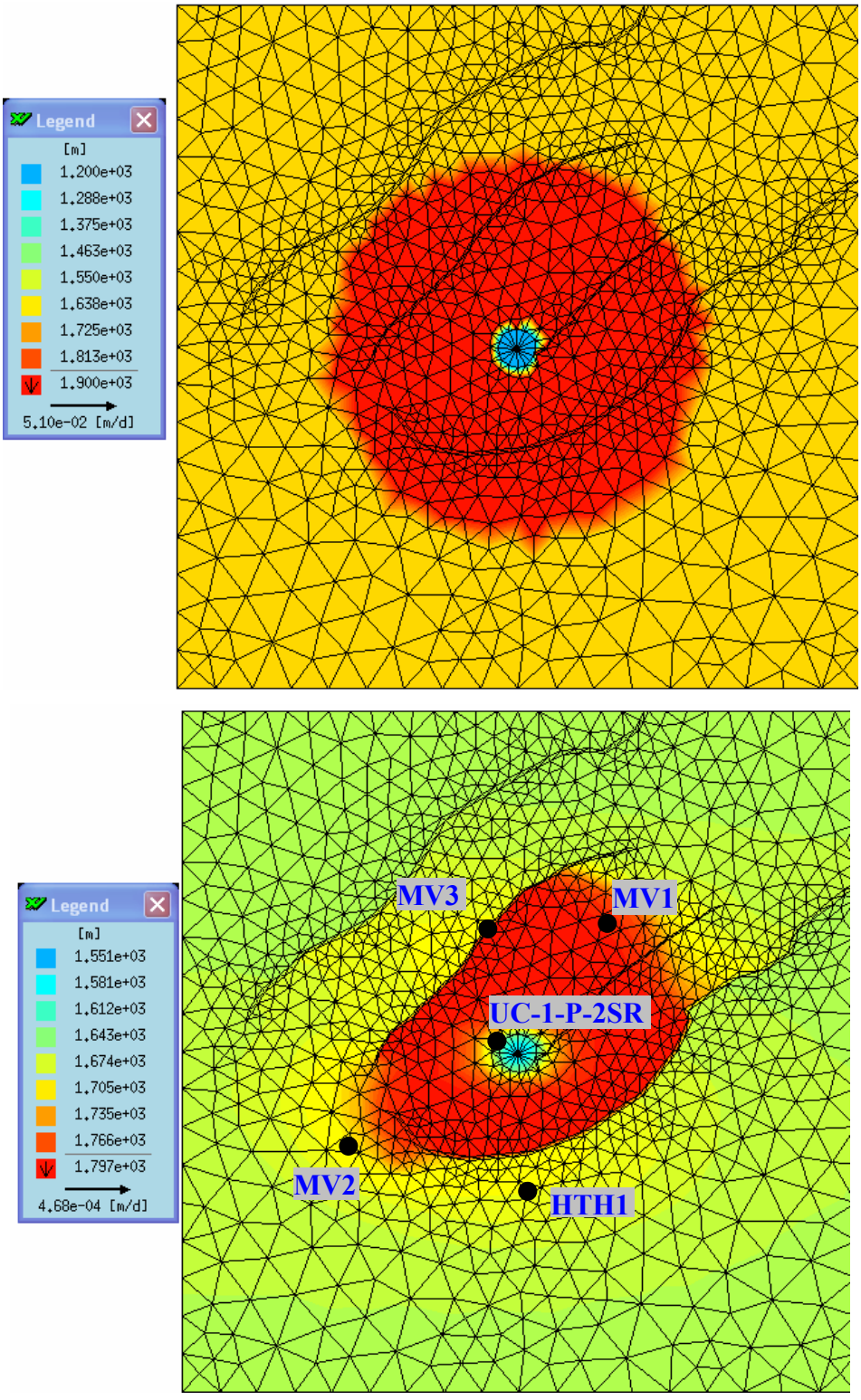


Figure 4.14. Distribution of head simulated in slice 8 passing through the cavity at a) time zero (immediately after the test), and b) 40 years after the test.

Figure 4.15 shows the head distribution simulated in the alluvium, 40 years after the test. The high pressure pulse has impacted the heads in the alluvium and is still maintained at high levels within the inner area surrounded by the faults. However, there remain major discrepancies between the model and the observed heads in the alluvium. The field data show that the head values at HTH-2 and MV-2 are 1,667 m and 1,776 m, respectively, which implies that there is a large gradient (0.11) from the west to the east. However, the boundary conditions that were interpolated from wells throughout Hot Creek Valley (Pohlmann *et al.*, 1999; Pohll *et al.*, 2003) suggest a low gradient (i.e., 0.002) from north to south. The revised model could not match the high head values in the alluvium observed at the MV wells, though the average error in these wells reduced slightly from 101 m (in the steady-state model) to 73 m in the revised model. These results suggest that the faults observed around Faultless may not only play an important role in persistence of the nuclear-test pressure pulse in the volcanic section, but may also act as natural hydraulic barriers, dividing the alluvial aquifer into compartments of similar head separated by zones with very high gradients.

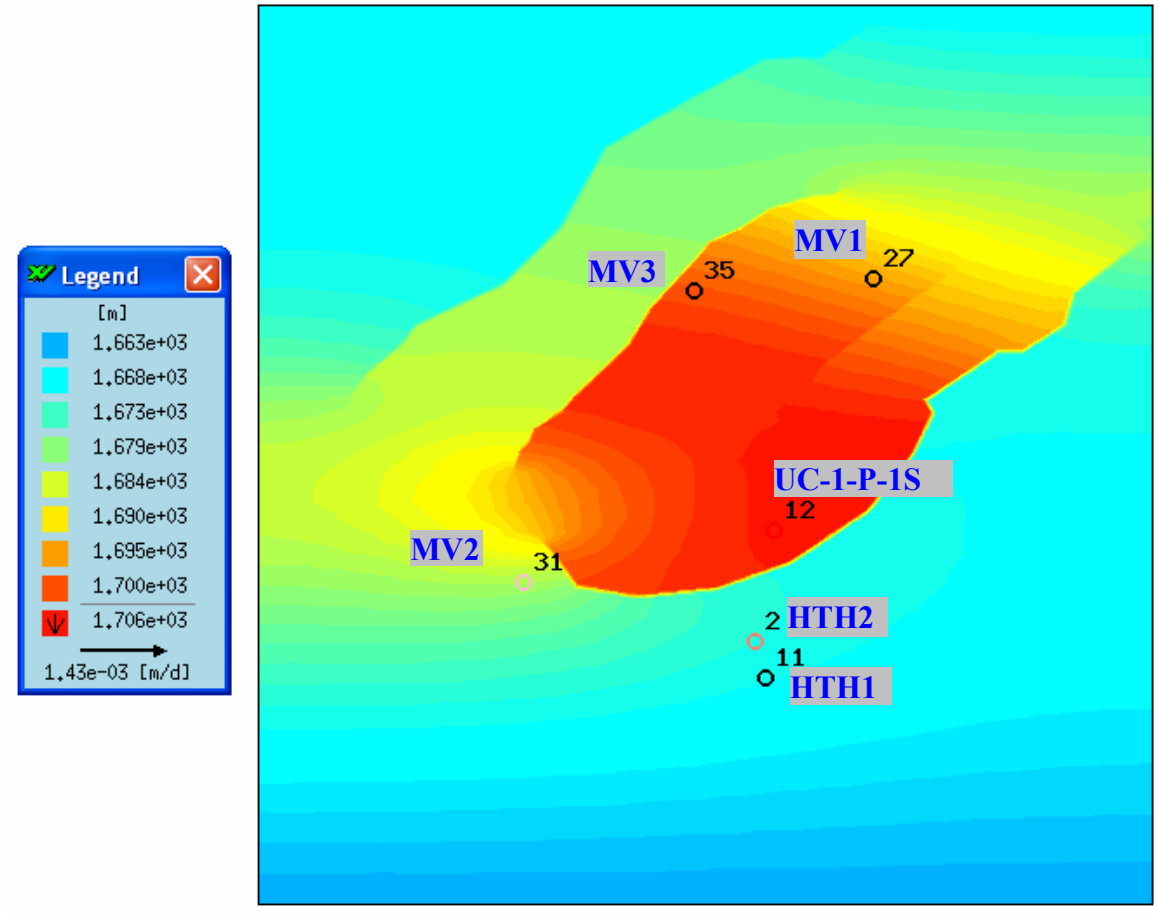


Figure 4.15. The head distribution simulated in the alluvium 40 years after the test.

5.0 SUMMARY AND CONCLUSIONS

Three monitoring/validation wells were installed at CNTA in 2005. The wells provided new data (Lyles *et al*., 2006; DOE, 2006) that are used for the validation process of the groundwater flow and transport model at Faultless. Data sets regarding lithology, hydraulic head, hydraulic conductivity, and water chemistry are used. The validation process for the Faultless stochastic groundwater model detailed in Hassan (2003a, 2004b) and in DOE (2004) is implemented and the step-by-step procedure is followed using the new data sets.

The collected validation data fall into two broad categories. One category contains the data pertaining to the model input parameters and the other category pertains to the model-produced output. Resistivity logs and the resulting lithology profiles that identify the presence and location of different geologic units (alluvium, tuffaceous sediments, and densely welded tuff) and the hydraulic conductivity data belong to the first category. Measured heads in the three wells and the "inferred" gradients from these measurements belong to the second category.

These data sets are analyzed and are tied to the model cells at their analogous locations so that comparisons between data and model values could be made. This resulted in 60 model cells having lithologic data (i.e., identified flow categories), 9 cells having measured heads, 4 cells having measured conductivities, and 6 cell pairs where vertical head gradients are known from the head data. The lithologic data provided binary type or categorical type data (either category 1, 2, or 3), whereas the other data sets provided real-number data values. Therefore, 19 real-number validation targets were available plus 60 categorical validation targets.

The validation process (Figure 2.1) was then followed step by step. First, the calibration accuracy evaluations using the GLUE weights were performed (Step 3). Step 4 involved performing different tests using the validation data and developing the acceptance criteria measures (P_1 through P_5). Step 5 was then conducted where the calibration and validation results were linked and a composite score was developed for each model realization. Steps 6 and 7 related to making the decision about the model in light of the resulting measures P_1 through P_5, the overall tests of the model, the realization scores, and the decision chart of Figure 2.2.

The calculated measures P_1 and P_2 are very low (1 percent and 18 percent, respectively), and indicate a need for model revision. Other measures such as P_3, P_4, and P_5 also indicate a major deficiency in the model. Composite realization scores are below a selected acceptance threshold for all model realizations, supporting the failure of the validation.

Despite these results, a number of positive aspects about the model have been identified with the validation data. First, the lithology identified from the resistivity logs generally matches the lithology used in the model. The contact between the alluvium and tuffaceous sediments was taken to be uncertain in the original model and the contact identified from the three wells is invariably within the range of alluvium considered (i.e., some model realizations portrayed alluvium at the depths identified). Second, the hydraulic conductivity values obtained from aquifer tests in the three validation wells are within the distributions used in the model. In fact, all eight hydraulic conductivity validation targets fall within the inner 95 percent of the model distribution of these targets. Last, no

tritium was detected in water samples from the wells, consistent with the contaminant transport predictions for these locations.

The elevated heads in the tuffaceous sediments in the area surrounding the cavity are believed to be due to the nuclear test itself. The persistence of the high heads may be attributed to the very low hydraulic conductivity of the volcanic rocks and the down-dropped block that may have created (or accentuated) barriers to flow. These factors may not have allowed the pressure pulse around the cavity to dissipate. The original Faultless model focused on the far-field transport, intentionally neglecting nuclear test impacts that were assumed to be transient. Given the nuclear test impacts inferred from the MV well observations, a lack of agreement between the model and field data is not surprising.

The final step in the model validation process (step 7) is an assessment by the decision makers as to whether the validation results have met regulatory objectives. That assessment will be a difficult one for DOE and NDEP for CAU 443. The CAU model has clearly not been validated by the data from the MV wells. A fundamental assumption of the model, that the nuclear test impacts on the flow field were transient and unimportant over the timescales of interest, was proved wrong. The consequence of this error is that hydraulic heads are incorrect in parts of the model and flow directions misrepresented. Nonetheless, the validation data also reveal a hydrogeologic system characterized by extremely low hydraulic conductivity values and absence of units that could provide rapid contaminant flowpaths supporting the transport model finding that no far field transport is expected to occur in the 1,000 year regulatory time frame.

A simplified three-dimensional model is developed for the purpose of evaluating the effects on the flow system of the nuclear test and faults that are mapped in the vicinity of ground zero. This model is run first under steady-state conditions and is found to reasonably reproduce pre-validation head data available from HTH-1 and HTH-2. Subsequently, it is run in a transient mode but using an elevated head pulse as an initial condition created by the nuclear test. The results show that this elevated head pulse persisted for a long time, simulating much higher heads 40 years after Faultless (present time) at the MV locations, as compared to the original model. This simple model shows a potential to incorporate the near-field effects of the test and to simulate a pressure pulse that very slowly dissipates over time.

A revision of the groundwater model to create a more accurate depiction of the flow system near Faultless would need to take into consideration the near-field impacts of the test and the different possibilities of the effects of the down-dropped block on the heads and on flow directions surrounding the cavity. A major contributor to the decision to neglect near-field effects in the original model was the lack of data to support such a depiction. That lack of data remains a significant problem, as the data from the MV wells only reveal the hydraulic heads affected by Faultless, and do not provide data regarding hydraulic features (such as faults) and forces (such as pressures). A model including those features will introduce additional significant uncertainties in the absence of near-field characterization and assessment.

6.0 REFERENCES

Beven, K.J., and A.M. Binley. 1992. The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, 6, 279-298.

Chapman, J.B., T.M. Mihevc and B.F. Lyles. 1994. The application of borehole logging to characterize the hydrogeology of the Faultless site, Central Nevada Test Area. Desert Research Institute, Water Resources Center, Publication #45119, DOE/NV/10845-35, 36 p.

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

Diersch, J.J.G. 1998. Interactive, graphics-based finite-element simulation system FEFLOW for modeling groundwater flow contaminant mass and heat transport processes. FEFLOW Reference Manual, WASY Ltd., Berlin, 294 p.

Dinwiddie, G.A. 1972. Summary of recent hydrologic data, Faultless site, Hot Creek Valley, Nevada. U.S. Geological Survey letter report, 21 p.

Dinwiddie, G.A. and S.W. West. 1970. Hydrologic phenomena at the Faultless site, Hot Creek Valley, Nevada. U.S. Geological Survey letter report, 24 p.

Doherty, J. 1994. PEST. Watermark Computing, Corinda, Australia. 122 pp.

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

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

Hassan, A.E. 2003a. A Methodology for the Validation of 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 Operations 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 Operations 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.

Healey, D.L. 1968. Gravity survey of northern Hot Creek Valley, Nye County, Nevada. U.S. Geological Survey Technical Letter: Central Nevada-18, 21 p.

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

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

Lyles, B., P. Oberlander, D. Gillespie, D. Donithan, and J. Chapman. 2006. Hydrologic Data and Evaluation for Wells Near the Faultless Underground Nuclear Test, Central Nevada Test Area. Desert Research Institute, Division of Hydrologic Sciences Publication No. 45219.

McKeown, F.A., D.D. Dickey and W.L. Ellis. 1968. Preliminary report on the geologic effects of the Faultless event. U.S. Geological Survey, Central Nevada-16, USGS-474-65, 20 p.

Morse, B. S., G. Pohll, J. Huntington, R. Rodriguez Castillo. 2003. Stochastic capture zone analysis of an arsenic-contaminated well using the generalized likelihood uncertainty estimator (GLUE) methodology, *Water Resources Research, 39* (6), 1151, doi:10.1029/2002WR001470.

Pohll, G., K. Pohlmann, J. Daniels, A. Hassan and J. Chapman. 2003. Contaminant Boundary at the Faultless Underground Nuclear Test. Publication No. 45196, Desert Research Institute, Division of Hydrologic Sciences, Las Vegas, Nevada, 49p.

Pohlmann, K., J. Chapman, A. E. Hassan, and C. Papelis. 1999. Evaluation of Groundwater Flow and Transport and the Faultless Underground Nuclear Test, Central Nevada Test Area, Publication No. 49165, Desert Research Institute, Division of Hydrologic Sciences, Las Vegas, Nevada, 120p.

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

U.S. Department of Energy (DOE), 2000. United States Nuclear Tests July 1945 through September 1992. Nevada Operations Office Report DOE/NV--209-REV15, 162p.

U.S. Department of Energy (DOE), National Nuclear Security Administration Nevada Site Office. 2004. *Corrective Action Decision Document/Corrective Action Plan for Corrective Action Unit 443: Central Nevada Test Area (CNTA) - Subsurface,* Rev. 0., DOE/NV-977, Las Vegas, NV.

U.S. Department of Energy (DOE), National Nuclear Security Administration Nevada Site Office, 2006. *Well Installation Report for Corrective Action Unit 443, Central Nevada Test Area Nye County, Nevada,* Rev. 0, DOE/NV--1102, Las Vegas, NV.

Willmott, C.J. 1981. On the validation of models. *Physical Geography* 2, 184-194.

Willmott, C.J., S.G. Ackleson, R.E. Davis, J.J. Feddema, K.M. Klink, D.R. Legates, J. O'Donnell, and C.M. Rowe. 1985. Statistics for the evaluation and comparison of models. *Journal of Geophysical Research* 90, 8995-9005.

DISTRIBUTION

Jenny Chapman
Division of Hydrologic Sciences
Desert Research Institute
755 East Flamingo Road
Las Vegas, NV 89119-7363

Tim Echelard
Stoller-Navarro Joint Venture
7710 W. Cheyenne
Las Vegas, NV 89128

Kenneth A. Hoar, Director
Environment, Safety and Health Division
Nevada Site Office
National Nuclear Security Administration
U.S. Department of Energy
P.O. Box 98518
Las Vegas, NV 89193-8518

Richard Hopper
Deputy Laboratory Director
U.S. Environmental Protection Agency
Radiation and Indoor Environments
National Laboratory
4220 S. Maryland Parkway, Bldg. C
Las Vegas, NV 89119

Bruce Hurley
Hydrology Program Manager
Environment, Safety and Health Division
Nevada Site Office
National Nuclear Security Administration
U.S. Department of Energy
P.O. Box 98518
Las Vegas, NV 89193-8518

Rick Hutton
Stoller Grand Junction
2597 B 3 / 4 Road
Grand Junction, CO 81503

John Jones
Environmental Restoration Division
Nevada Site Office
National Nuclear Security Administration
U.S. Department of Energy
P.O. Box 98518
Las Vegas, NV 89193-8518

Marjory Jones
Division of Hydrologic Sciences
Desert Research Institute
2215 Raggio Parkway
Reno, NV 89512-1095

Randy Laczniak
U.S. Geological Survey
Water Resources Division
160 N. Stephanie St.
Henderson, NV 89074-8829

Tom Pauling
Office of Land and Site Management
Office of Legacy Management
U.S. Department of Energy
2597 3 / 4 Road
Grand Junction, CO 81503

Peter Sanders
Environmental Restoration Division
Nevada Site Office
National Nuclear Security Administration
U.S. Department of Energy
P.O. Box 98518
Las Vegas, NV 89193-8518

Reina Serino, Contracting Specialist
Office of Business Services
NNSA Service Center
Pennsylvania and H Street, Bldg. 20388
P.O. Box 5400
Albuquerque, NM 87185-5400

David Shafer
Division of Hydrologic Sciences
Desert Research Institute
755 E. Flamingo Road
Las Vegas, NV 89119-7363

Jacqueline Van Lier
Stoller-Navarro Joint Venture
7710 W. Cheyenne
Las Vegas, NV 89128

Janet Appenzeller-Wing, Director
Environmental Restoration Division
Nevada Site Office
National Nuclear Security Administration
U.S. Department of Energy
P.O. Box 98518
Las Vegas, NV 89193-8518

Nevada State Library and Archives
State Publications
100 North Stewart Street
Carson City, NV 89710-4285

Archives
Getchell Library
University of Nevada, Reno

DeLaMare Library/262
University of Nevada, Reno

Document Section, Library
University of Nevada, Las Vegas
4505 Maryland Parkway
Las Vegas, NV 89154

Library
Stoller-Navarro Joint Venture
7710 W. Cheyenne, Bldg. 3
Las Vegas, NV 89128

Library
Southern Nevada Science Center
Desert Research Institute
755 E. Flamingo Road
Las Vegas, NV 89119-7363

Technical Library
Nevada Site Office
National Nuclear Security Administration
U.S. Department of Energy
P.O. Box 98518
Las Vegas, NV 89193-8518

Public Reading Facility
c/o Nuclear Testing Archive
Nevada Site Office
National Nuclear Security Administration
U.S. Department of Energy
P.O. Box 98521, M/S 400
Las Vegas, NV 89193-8521

Office of Scientific and Technical Information
U.S. Department of Energy
P.O. Box 62
Oak Ridge, TN 37831-9939
(electronic copy)