SANDIA REPORT

SAND2008-5438
Unlimited Release
Printed August 2008

Development of a Deterministic Site Characterization Tool Using Multi-Model Ranking and Inference

Thomas S. Lowry, Scott C. James, Bill W. Arnold, Matthew Grace, Genetha A. Gray, Michael Ahlmann

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration under Contract DE-AC04-94AL85000.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online order: http://www.doe.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online order: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2008-5438
Unlimited Release
Printed August 2008

SNL-NUMO Collaborative Development of a Deterministic Site Characterization Tool Using Multi-Model Ranking and Inference

Thomas S. Lowry, Scott C. James, Bill W. Arnold, Matthew Grace, Genetha A. Gray,
and Michael Ahlmann
Sandia National Laboratories
Geohydrology Department, 06313
P.O. Box 5800
Albuquerque, New Mexico 87185-0735

This project was funded by the Nuclear-waste Management Organization of Japan (NUMO) under agreement number FI 061060329, entitled "Development of a Deterministic Site Characterization Tool Using Multi-Model Ranking and Inference" and was conducted by Sandia National Laboratories (SNL) for the benefit of NUMO under a joint JNC/U.S. Department of Energy (DOE) work agreement. All work was performed and completed at Sandia National Laboratories Albuquerque New Mexico and Livermore California campuses. All site-specific data were supplied to the research team by NUMO for use in this project.

Abstract

Uncertainty in site characterization arises from a lack of data and knowledge about a site and includes uncertainty in the boundary conditions, uncertainty in the characteristics, location, and behavior of major features within an investigation area (e.g., major faults as barriers or conduits), uncertainty in the geologic structure, as well as differences in numerical implementation (e.g., 2-D versus 3-D, finite difference versus finite element, grid resolution, deterministic versus stochastic, etc.). Since the true condition at a site can never be known, selection of the best conceptual model is very difficult. In addition, limiting the understanding to a single conceptualization too early in the process, or before data can support that conceptualization, may lead to confidence in a characterization that is unwarranted as well as to data collection efforts and field investigations that are misdirected and/or redundant. Using a series of numerical modeling experiments, this project examined the application and use of information criteria within the site characterization process. The numerical experiments are based on models of varying complexity that were developed to represent one of two synthetically developed groundwater sites; 1) a fully hypothetical site that represented a complex, multi-layer, multi-faulted site, and 2) a site that was based on the Horonobe site in northern Japan. Each of the synthetic sites were modeled in detail to provide increasingly informative

‘field’ data over successive iterations to the representing numerical models. The representing numerical models were calibrated to the synthetic site data and then ranked and compared using several different information criteria approaches. Results show, that for the early phases of site characterization, low-parameterized models ranked highest while more complex models generally ranked lowest. In addition, predictive capabilities were also better with the low-parameterized models. For the latter iterations, when more data were available, the information criteria rankings tended to converge on the higher parameterized models. Analysis of the numerical experiments suggest that information criteria rankings can be extremely useful for site characterization, but only when the rankings are placed in context and when the contribution of each bias term is understood.

Table of Contents

Table of Figures

List of Tables

1. Introduction

When developing a high-level nuclear waste repository, enormous efforts are spent characterizing the site, beginning from the preliminary investigation (PI), where the major features and gross suitability are analyzed, to more detailed investigations (DI), where understanding of local-scale processes is demonstrated. The bulk of this effort is spent trying to gain sufficient understanding of the site such that performance assessment can be carried out with acceptable levels of uncertainty. Because subsurface data are relatively sparse, however, reducing uncertainty can be extremely difficult, especially during the PI.

Uncertainty in site characterization arises from a lack of data and knowledge about the site and includes uncertainty in the boundary conditions, uncertainty in the characteristics, location, and behavior of major features within the investigation area (e.g., major faults, confining layers, etc.), and uncertainty in the geologic, geochemical, and hydrogeologic environments. Numerical models used to simulate conditions at the site contribute further uncertainty. By definition, models rely on assumptions and conceptualizations about the true conditions, with the hope that those assumptions and conceptualizations are of little consequence with regard to the models ability to answer useful questions and to gain understanding about the site. In addition, site characterization can involve expert elicitation and scientific analyses, which can lead to competing viewpoints and conceptualizations about the site. The aggregation of the data, model, and conceptual uncertainty represents the information gap between what needs to be known and what is actually occurring at the site.

Given the fact that uncertainty in groundwater investigations is high, inferences about a site should include all possible processes as determined through expert opinion and scientific evaluation. This follows closely Chamberlin's famous paper that describes using multiple hypothesis as a strategy for gaining maximum understanding in applied and theoretical problems [*Chamberlin*, 1965]. With regard to modeling a potential repository site, this means that multiple conceptual models formed as a result of aggregated uncertainty should remain active until one can determine which model is "best" (the term best is open to interpretation and can impact a models relative merit, as will be discussed below). To outline this strategy in the context of nuclear site investigations, this project examines the efficacy of using model selection methodologies that are based on an information theoretic approach with the objective of making better and more defensible decisions and inferences about the site.

1.1 Information Theoretic Approach

Information theoretic approaches are based on Kullback-Leibler (K-L) information, which is defined as the information, $I(f,g)$, that is lost when the true condition f is approximated by a model, g [*Kullback and Leibler*, 1951]. Mathematically, this is expressed for continuous functions as:

$$I(f,g) = \int f(x) \ln\left[\frac{f(x)}{g(x|\theta)}\right] \tag{1}$$

where f and g are n-dimensional probability distributions. Fundamentally, K-L information is a measure of the distance between conceptual reality, f, and the approximating model, g [*Burnham and Anderson*, 2001].

Akaike [*Akaike*, 1973; 1974] developed a relationship between K-L information and maximum likelihood theory that makes it possible to combine estimation (e.g., such as maximum likelihood or least squares) with model selection. Akaike defined an estimator of the expected K-L information based on the maximized log-likelihood function known as "Akaike's information criteria" (AIC) [*Burnham and Anderson*, 2001]:

$$AIC = -2\ln\left(L\left(\hat{\theta}\middle|data\right)\right) + 2k \quad \textbf{(2)}$$

where $\hat{\theta}$ are the calibrated parameters, ln[L($\hat{\theta}$ |data)] is the maximized log-likelihood function, and k is the number of calibration parameters used in the model. If one assumes a normal distribution of errors for all models in the set, then for the case of least-squares estimation, equation (2) can be expressed as:

$$AIC = n\ln(\hat{\sigma}^2) + 2k \quad \textbf{(3)}$$

where

$$\hat{\sigma}^2 = \frac{\sum \hat{\varepsilon}_i^2}{n} \quad \textbf{(4)}$$

and $\hat{\varepsilon}_i^2$ are the squared residuals between the model and the observed data.

A version of equation (3) was developed that accounts for cases when n/k > 40 [see, for example, *Hurvich and Tsai*, 1989]. Known as the corrected Akaike Information Criteria (*AICc*), it is given as:

$$AICc = n\ln(\hat{\sigma}^2) + 2k + \frac{2k(k+1)}{n-k-1} \quad \textbf{(5)}$$

Because the *AICc* asymptotically approaches the *AIC* as n gets large, it is recommended that the *AICc* always be used [*Burnham and Anderson,* 2004]. The small "*c*" in *AICc* stands for (bias) corrected.

To use the *AICc*, we assume that a set of R conceptual models have been defined. Each of the R models is fitted to the n observed data and the average of the sum of the squared residuals ($\hat{\sigma}^2$) and then the *AICc* are calculated for each model. The number of parameters, k, used to calculate the *AICc* should be the number of calibrated parameters in the model plus 1 to reflect the concept that $\hat{\sigma}^2$ is also an estimatable parameter [*Burnham and Anderson*, 2004; *Poeter and Anderson*, 2005]. The model for which the *AICc* is lowest given the available data is selected as the best model of the set of R models given the set of n observations. The best model, as determined by the *AICc* is the model that provides the highest degree of parsimony, which is defined as the tradeoff between under-fitting and over-fitting (i.e., under-parameterization and over-parameterization) [*Poeter and Anderson*, 2005].

Figure 1 shows an example of this concept with the green squares representing a single data set that are fit using two conceptual models: a 2nd and a 5th order polynomial (red and blue dotted lines, respectively). For the 2nd order model, $\hat{\sigma}^2$ is equal to 23.37 but for the 5th order model, it is equal to 0.00 (the 5th order model fits the data exactly).

However, when additional data are added (blue diamonds) $\hat{\sigma}^2$ is equal to 16.43 and 51.48 for the 2nd and 5th order models, respectively. In this case, the 5thorder model demonstrates over-fitting of the first data set, which results in its inability to simulate the additional data. Using *AICc*, the 5th order model would be penalized for being over-parameterized with respect to the available data.

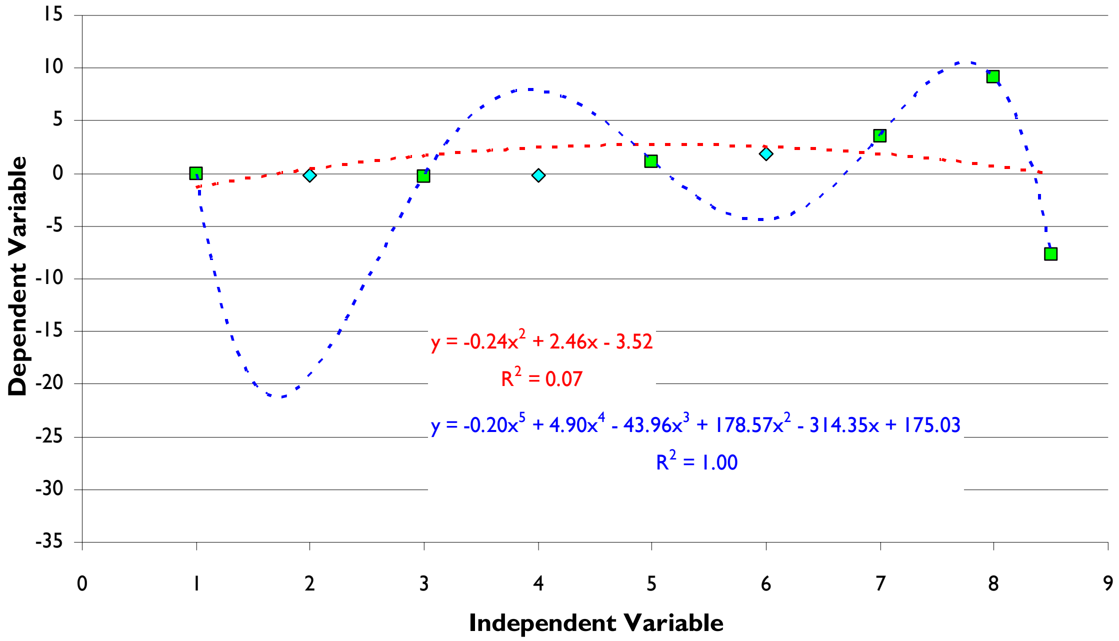


Figure 1 – The green squares represent the original data set are fit using two conceptual models: a 2nd and a 5th order polynomial (red and blue dotted lines, respectively). The blue diamonds represent data that are collected at a later date. The 2nd order model is an example of under-fitting while the 5th order model is an example of over-fitting.

Burnham and Anderson [2001] make an important point with regards to the information-theoretic approach to modeling, saying:

> "In a very important sense, we are not trying to model the data; instead, we are trying to model the information in the data."

The significance of this point reflects an underlying fact of the *AICc*; namely that the *AICc* is only comparing the *R* models to each other *for a particular data set*. In other words, if another data set is chosen, the magnitude of *AICc* values for models fitted to different data sets from the same site cannot be compared. Only the *relative rankings* of each model as compared to the set of models calibrated to the same data set are able to be compared. This stems from the fact that the *AICc* is an estimate of the K-L information and contains arbitrary constants that are affected mainly by the sample size but also by the samples themselves [*Burnham and Anderson*, 2004]. The practical aspect of this is the understanding that the *AICc* does not choose the best model based on its ability to model the data. Rather, it chooses the best model as the one that can most effectively use the information contained in the data.

Model rankings are obtained by rescaling the *AICc* values by a factor that reduces the *AICc* value of the best model to a value of 0, i.e.:

$$\Delta_i = AICc_i - AICc_{\min} \tag{6}$$

where $AICc_i$ is the *AICc* value of the i^{th} model, and $AICc_{\text{min}}$ is the *AICc* value of the best model (i.e. minimum *AICc* value). The Δ_i values provide an easy way to rank models within a single model set by providing a "strength-of-evidence" measure for model *i* versus the best model. Guidelines have been posed [*Burnham and Anderson*, 2002] for assessing the relative merits of models within a set: models with Δ_i 2 have substantial evidence, models where 4 Δ_i 7 have less support, and models where $\Delta_i > 10$ have essentially no support. As demonstrated below, we believe that the Δ_i values should not be a means for keeping or discarding a model between successive data collection efforts but rather, the should be used to indicate the relative strength of one model over another for the particular data set being modeled.

A transformation of Δ_i to $\exp(-\Delta_i / 2)$, for $i = 1, 2, \ldots, R$, provides the likelihood of model *i* given the data. Using this transformation, a 'weight of evidence' can be calculated as:

$$w_i = \frac{\exp\left(-\Delta_i/2\right)}{\sum_{r=1}^{R}\exp\left(-\Delta_r/2\right)} \tag{7}$$

Sometimes called the "Akaike weights," w_i is interpreted as the evidence that model *i* is the best model in the set, *R*. The ratio w_i/w_j is the "evidence ratio" and can be used to directly compare one model to the next and allows statements to be made such as "there is w_i/w_j times more evidence supporting model *i* over model *j*" [*Poeter and Anderson*, 2005]. It is important to note that both the Akaike weights and the evidence ratios are relative measures and should only be interpreted as being valid for the *R* models in the model set as calibrated to a single set of data.

An analogy for the evidence ratios are given by Burnham and Anderson [2002, page 79] whereby they compare an auditorium with *N* people, each of whom holds a raffle ticket. The exception is that a single person (Bob) has 3 tickets. The evidence ratio that Bob will win the raffle versus any other person is 3. While Bob has a 3 to 1 edge of winning over any other person, the evidence ratio says nothing about his probability of winning, which is dependent on the number tickets Bob holds versus the number of tickets given out. Similarly, the evidence ratio for model *i* over model *j* indicates only that model *i* has w_i/w_j more evidence of being the true model than model *j*. It does not, in any way, indicate the probability of model *i* being the true model.

1.2 Multi-Model Inference

With regard to nuclear repository site characterization, the objective is not to reproduce observed data as closely as possible, but rather it is to simulate the important processes at the site such that the sites adequacy can be determined with a reasonable amount of uncertainty. As discussed above, uncertainty is contained in the data, the conceptualizations, and even the models themselves. To account for all of these uncertainties, multi-model inference is used. When a single conceptual model is used, the uncertainty that that single model is the best model should be incorporated into estimates of its ability to fit the data [*Burnham and Anderson*, 2004]. In other words, the variance about a model's prediction needs to be independent of the selected model. This "unconditional variance" is calculated for the calibrated parameters, $\hat{\theta}$, from the best model as [*Buckland et al.*, 1997]:

$$\hat{\hat{\mathrm{v}}}\mathrm{ar}\left(\hat{\hat{\theta}}\right)=\left[\sum_{i=1}^{R} w_i \left[\hat{\mathrm{v}}\mathrm{ar}\left(\hat{\theta}_i \middle| g_i\right)+\left(\hat{\theta}_i-\hat{\hat{\theta}}\right)^2\right]^{1/2}\right]^2 \qquad \textbf{(8)}$$

where

$$\hat{\hat{\theta}}=\sum_{i=1}^{R} w_i \hat{\theta}_i \qquad \textbf{(9)}$$

is the model averaged value of parameter $\hat{\theta}_i$. The first term of equation (8) accounts for the uncertainty from the single model that is chosen while the second term accounts for the conceptual uncertainty of model *i*. Equation (9) can be used to calculate model averaged parameter values for any parameter, *k*, by letting *R* = *R'*, where *R'* is the subset of models containing parameter *k*. New model weights, $w_i^{'}$, must also be recalculated to reflect the fact that a subset of models, *R'*, is being used. Both $w_i^{'}$ and *R'* should be used in equation (8) when calculating the unconditional variance for a parameter that does not appear in the full model set. While model-averaged parameter values may be useful in some cases, Poeter and Anderson [2005] recommend not using them for groundwater problems because they are often inappropriate for use in a particular model construct (i.e. they are specific to the model for which they are calibrated and not easily generalized).

Finally, the 95% confidence interval around the predictions from model *i* can be calculated as:

$$\hat{\hat{\theta}}=\hat{\hat{\theta}}\pm 2\sqrt{\hat{\hat{\mathrm{v}}}\mathrm{ar}\left(\hat{\hat{\theta}}\right)} \qquad \textbf{(10)}$$

1.3 Other Forms

Several other forms of information-theoretic criteria exist, some which were developed from the Bayesian point of view while others which were developed from the frequentist point of view (such as the *AIC*) [see *McQuarrie and Tsai*, 1998]. For comparative purposes, and because the calculations are relatively easy once $\hat{\sigma}^2$ has been calculated, several of these additional criteria are calculated. Namely, the *BIC* [*Schwarz*, 1978], the *HQ* [*Hannan and Quinn*, 1979], and the *KIC* [*Kashyab*, 1982]. These are each given as:

$$BIC = n\ln(\hat{\sigma}^2) + k\ln(n) \qquad \textbf{(11)}$$

$$HQ = n\ln(\hat{\sigma}^2) + ck\ln(\ln(n)) \qquad \textbf{(12)}$$

and

$$KIC = n\ln\left(\hat{\sigma}^2\right) + k\ln\left(\frac{n}{2\pi}\right) + \ln\left|\mathrm{X}^T w\mathrm{X}\right| \qquad \textbf{(13)}$$

For equation (12), *c* > 2, and for equation (13), $\left|\mathbf{X}^T\mathbf{wX}\right|$ is the determinant of the Fisher information matrix, **X** is the sensitivity matrix, $\mathbf{X}^T$ is its transpose, and **w** is the weight matrix. Because this study is focused on the use of the *AICc* in the context of nuclear repository site investigations, the details and implications of each of these methods are not discussed here. However, additional information for each of these methods in the context of information-theoretic methods can be readily found in the literature [*Burnham and Anderson*, 2001; 2002; 2004; *Link and Barker*, 2006; *Ye et al.*, 2008; *Zucchini*, 2000]. More specifically, several studies exist that have suggested the use of some of these methods for the selection of groundwater models [*Carrera and Neuman*, 1986; *Neuman*, 2003; *Neuman and Wierenga*, 2003; *Ye et al.*, 2004; *Ye et al.*, 2008].

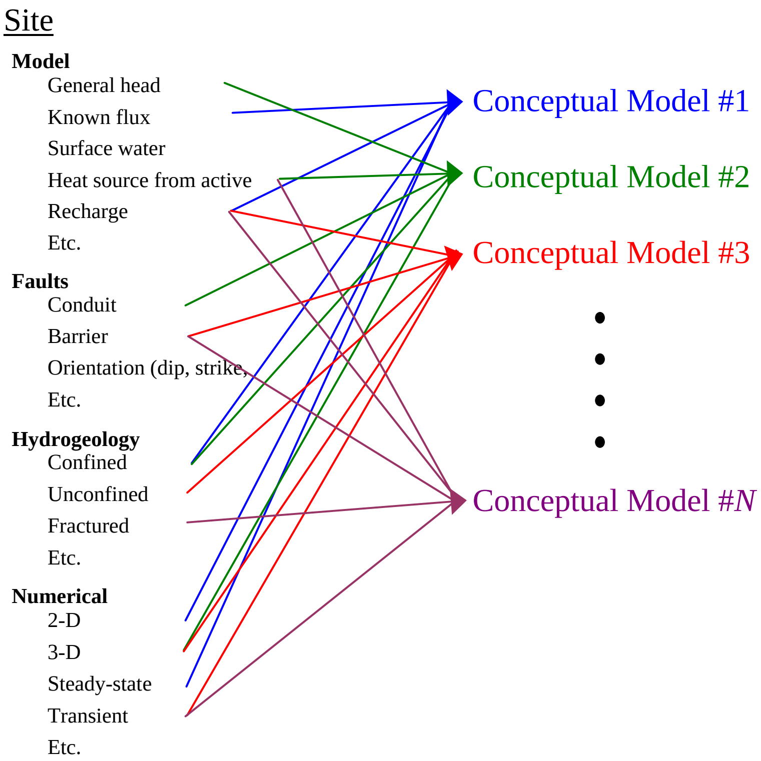


Figure 2 – Deterministic conceptual models will be formed by combining different site-processes that are representative of processes identified at the PI stage during expert field investigations and analyses.

This project consists of two Phases. The first Phase relies on a synthetically developed hypothetical site to examine how the ranking criteria behave when used in a complex, natural system and to determine the limits of its use. Successive iterations of adding observation data and calibration of the models were conducted. The second Phase utilizes data from a real-world site in Japan and concentrates on the addition of observational data only. This was done to provide insight into how the *AICc* behaved with regards to the number of available data as well as the magnitude and contribution of second degree bias (i.e. the bias that ‘penalizes’ a model for being more complex). The result of these two analyses has resulted in a recommendation list presented in the last section for using the *AICc* in the site characterization process.

2 Implementation

To implement Phase I, the project team was broken into two distinct groups: the site creation (SC) team and the modeling and analysis (MA) team. These teams worked independently to iteratively execute the seven distinct steps outlined below:

1. Synthesize a Hypothetical Site (SC Team)

For the hypothetical site (HS), the first step was to identify features and processes that should be included in the model. The goal was to create a HS that, when modeled, would show different scales of complexity and structure as well as processes that are indicative of those found in Japan (Figure 3). After considerable research and discussion, a cross-section of the Senya Fault region that is contained in the JNC H12 report [*JNC*, 2000] was selected as the template from which to build the stratigraphy.

Because the Senya Fault region is highly active and thus would not be considered as a potential location for a repository, assumptions are made for the HS model that the faults shown in the cross-section are not active and that the site qualifies as an investigation area. The Senya Fault cross-section was only used to establish contact elevations only, meaning that no attempt was made to preserve the geologic material types shown in the cross-section legend. For the HS model, we extended the approximate 5-km width of the Senya Fault cross-section to 15 km, by conceptualizing a gradual elevation decline towards the west to a constant head ocean boundary (left-hand side of the cross-section), and a sharper increase in elevation to the east, which terminates at a no-flow groundwater divide. Contact elevations on either end of the model were extrapolated to the model boundary based on the slope of the contact close to the boundary. To avoid too much complexity that could compromise our ability to analyze the rankings, the contact elevations are assumed constant in the north-south direction, which, like the east-west dimension, is also 15 km. A preliminary investigation area is defined to be a 6×6 km^2 region in the middle of the model domain. The 3-D representation is shown in Figure 3 with cross-sections shown in Figure 4 and Figure 5.

Table 1 – Key features and processes included in the hypothetical site model.

Feature / Process	Reason to Include in Model
Faults	The faults are conceptualized to be conduits, barriers, or mixed conduit/barrier to flow. Potentially, conceptual models that more closely match the treatment of faults in the hypothetical model should rank higher. The ordering of the conceptual model ranks may also provide a means for inferring the true nature of the faults.
Variable recharge	Recharge is conceptualized to increase toward the east in the higher elevations and decrease in the west at the lower elevations. Spatial variations in groundwater head could show up in the calibrated conceptual models and in turn, help infer the recharge conditions.
Thermal gradient	Thermal gradients are common in Japan and thus are included in the hypothetical model.
Thermal intrusion	The hypothetical model includes a granitic intrusion that is conceptualized to be much warmer then the surrounding rock.
Complex geology	By using a complex geologic system as the hypothetical model, it is hoped that we can identify how complex the numerical models need to be to rank highly.
Heterogeneous hydrogeology	This is included by developing 3-D varying, spatially correlated random hydraulic conductivity fields in some of the geologic layers. This is a common approach to modeling heterogeneous systems and thus was included in the hypothetical model.

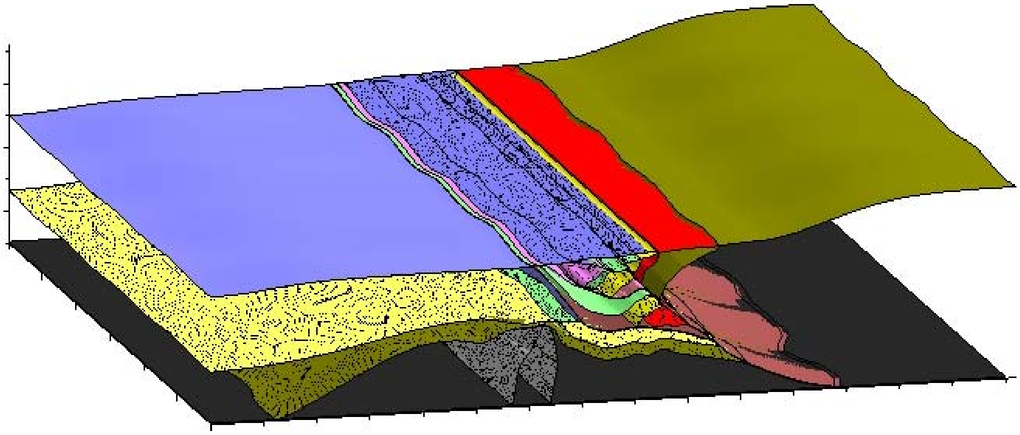


Figure 3 – Three-dimensional view of the hypothetical site model.

2. Model Flow Conditions over Time (SC Team)

Because a thermal gradient and an intrusion feature are included in the hypothetical model, the Finite Element Heat and Mass [FEHM - *Zyvoloski et al.*, 1999] code was selected to model the site. FEHM is able to model 3-D, time-dependent, multiphase, multi-component, non-isothermal, reactive flow through porous and fractured media. It can accurately represent complex, 3-D geologic media and structures and their effects on subsurface flow and transport. FEHM has been used to simulate groundwater and contaminant flow and transport in deep and shallow, fractured and un-fractured porous media by the US Department of Energy (DOE) and is the primary code used for modeling the saturated zone for the Yucca Mountain Project (YMP).

Despite its name, FEHM is a finite volume code that uses as part of its input a finite element mesh (the finite volume mesh is calculated internally by FEHM). The meshing process itself was extremely time consuming, requiring a mesh that was detailed enough to capture fine-resolution three-dimensional groundwater flow dynamics while conforming to the contact surfaces of the various units. Difficulties arose in the meshing process due to the numerous “pinch-outs” and discontinuities in the geology, which caused the meshing algorithm to produce meshes that resulted in negative volumes when FEHM converted the finite element mesh to a finite volume mesh. Due to these difficulties, a new strategy was developed; create a mesh by first meshing a two-dimensional cross-section in the east-west direction and then projecting the two-dimensional mesh in the north-south direction (Figure 4 and Figure 5). In addition, the only contact elevations that were preserved are along the faults, with the balance of the volume treated as a single meshing volume. Parameters were mapped to the mesh by overlaying the meshed volume with the stratigraphy volume and assigning a material ID number to each node based on where it falls in the stratigraphy. While this approach does represent a compromise to the initial objective, high resolution in the area of interest was maintained.

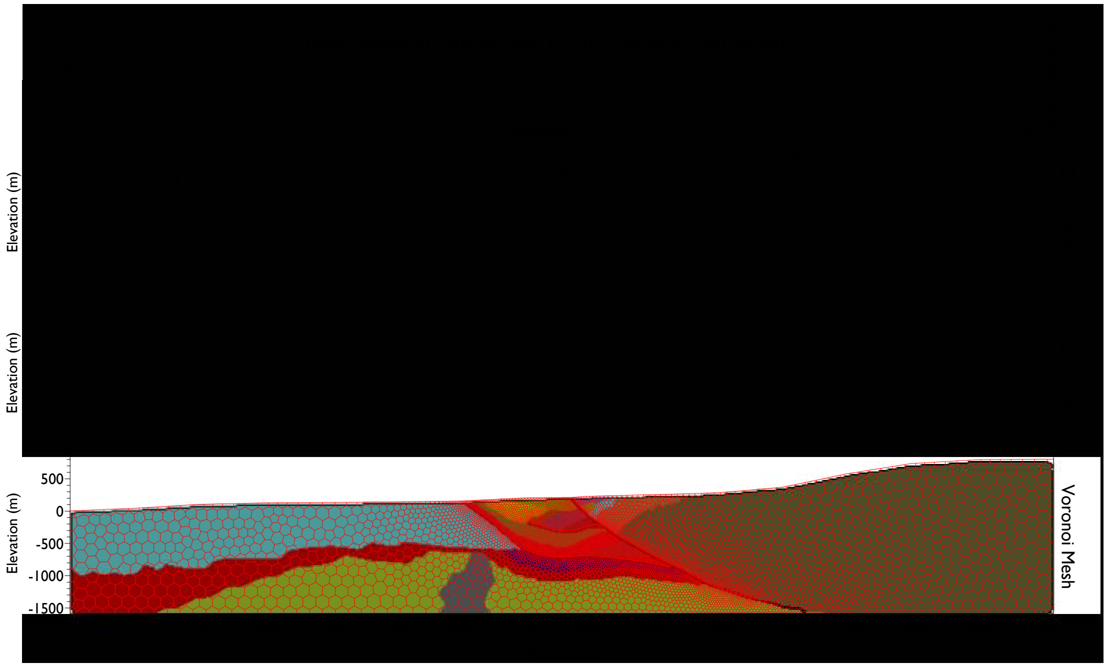


Figure 4 – Cross section of the HS model showing the stratigraphy and the finite element and finite volume meshes (DeLauney and Voronoi meshes).

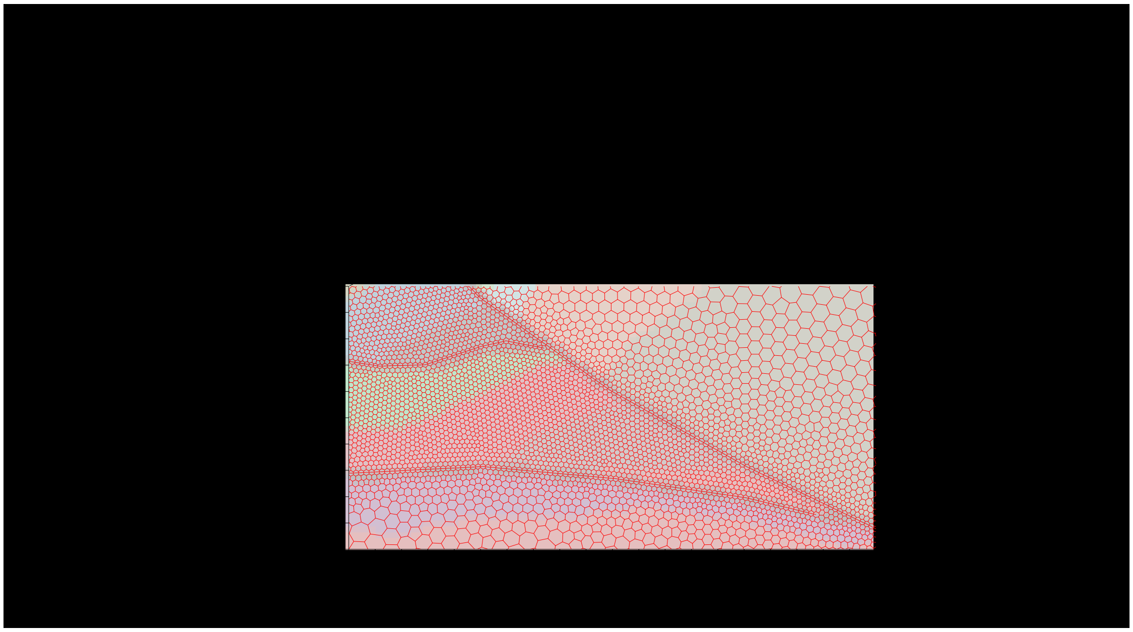


Figure 5 – Close-up of the DeLauney and Voronoi meshes. The only structures preserved in the mesh are the faults. Nodal values are assigned based on the geologic unit in which it falls.

3. Collect Data from the Site (SC Team)

Data supplied to the MA team consisted of a general description of the site, rainfall data, assumed boundary conditions, geologic setting, and borehole measurements that included both static and temporal head levels as well as stratigraphic contact elevations. Increasingly descriptive data were given to the MA team over three iterations.

4. Form the Conceptual and Numerical Models (MA Team)

With each data set, the MA team formed a set of conceptual models. The conceptual models defined the stratigraphic and hydrostratigraphic conditions, the important features, structures, and processes, as well as the boundary conditions at the site (as best determined from the supplied data). To maintain consistency, the numerical models encompassed the same domain and orientation.

5. Calibrate and Rank the Numerical Models (MA Team)

The MA team calibrate the numerical models to supplied observational data using the parameter estimation code, PEST [*Doherty*, 2007]. The models were then ranked and their Akaike weights and posterior model probabilities were calculated.

6. Analyze Results (MA Team)

Analysis of the results involved identifying features and processes that were highly ranked and determining areas within each model (in both parameter space and real space) with the highest uncertainty.

7. Acquire More Data and Repeat the Process (SC Team)

Steps 3-6 were repeated three times to establish a ranking trend for each of the models.

3 Phase I

3.1 Testing MMRI with Analytical Solutions

Anticipating that the development of the hypothetical site was going to take some time, the MA team began the by implementing the MMRI process using a series of analytical solutions that describe 1-dimensional groundwater transport under various initial and boundary conditions [*van Genuchten and Alves,* 1982]. The governing equation for this system is:

$$R\frac{\partial c}{\partial t} = D\frac{\partial^2 c}{\partial x^2} - u\frac{\partial c}{\partial x} \qquad \textbf{(14)}$$

where c [M/L^3] is the concentration of solute, t [t] is time, x [L] is the spatial reference, R is the retardation factor [–], D [L^2/t] is the dispersion coefficient, and u [L/t] is the advective groundwater velocity. Van Genuchten and Alves [1982] developed 12 different analytical solutions (A1 to A12) to this equation based on different initial and boundary conditions with solutions varying in the number of parameters (from 6 to 8). Each solution represents a unique conceptualization of equation (14) and are analogous to the conceptualizations one would make during preliminary and detailed investigations of a potential repository site.

For this test, a single "true" solution (solution A9) was selected from the set of 12 with $R = 1.5$, $D = 0.005$ m^2/s, and $u = 0.001$ m/s. The initial condition represents a uniform positive concentration across the entire domain ($C(x,0) = 1.0$). The left hand boundary condition describes a declining mass flux over time ($C(0,t) = C_a + C_b e^{\;\lambda t}$, with $C_a = 0.05$, $C_b = 0.95$, and $\lambda = \quad 2\times10^{\;5}$) while the right-hand boundary is a no-flux boundary projected at infinite ($\partial C/\partial x(L = \quad ,t) = 0$). The spatial and temporal model domains are 0 x 25 m, 0 t 100,000 s, respectively.

Five iterations were conducted at time 500, 1,000, 10,000, 50,000, 100,000 seconds as illustrated in Figure 6. These comprise the calibration data for the other solutions. Successive iterations include data from the next timestep and all 12 solutions (A1–A12) were calibrated to those data using PEST. Thus, at the first timestep, each of the 12 solutions was fit to the concentration profile across all x as calculated by A9 at $t = 500$ s. For the second iteration, calibration was against the concentration profiles at $t = 500$ and 1,000 s, and so on.

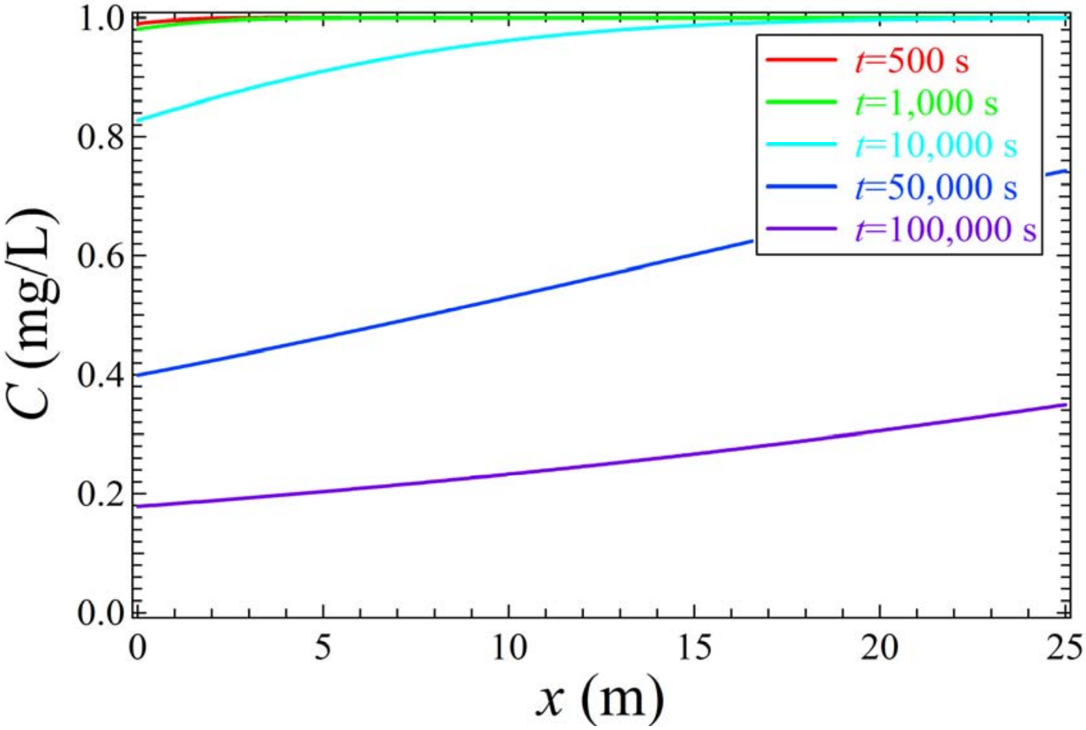


Figure 6 – The concentration profiles of solution A9 at each timestep.

Figure 7 shows the results of this exercise by plotting the ranking of each solution for each iteration along with its w_i value. The colored lines trace each models ranking across the 5 iterations. The *AICc* values are negative due to the fact that the logarithm of the residuals, which are all less then 1 in this case, results in negative values for the first order bias term of the *AICc*. For the first three iterations, rankings are seemingly random. By the fourth and fifth iteration, the rank changes are more orderly because there are enough data to adequately inform the calibration process. From the second iteration onward, the delta value of the true solution A9 continues to decrease. Likewise, the mean of the deltas as well as the standard deviation increase with each iteration because additional calibration data result in unsupported models with much higher *AICc* values with respect to A9. While it is not possible to compare *AICc* across iterations because of increasing calibration data, *n*, Figure 7 suggests that poorly performing models (e.g., A9 is ranked worst by the second iteration) should not be immediately eliminated from further consideration.

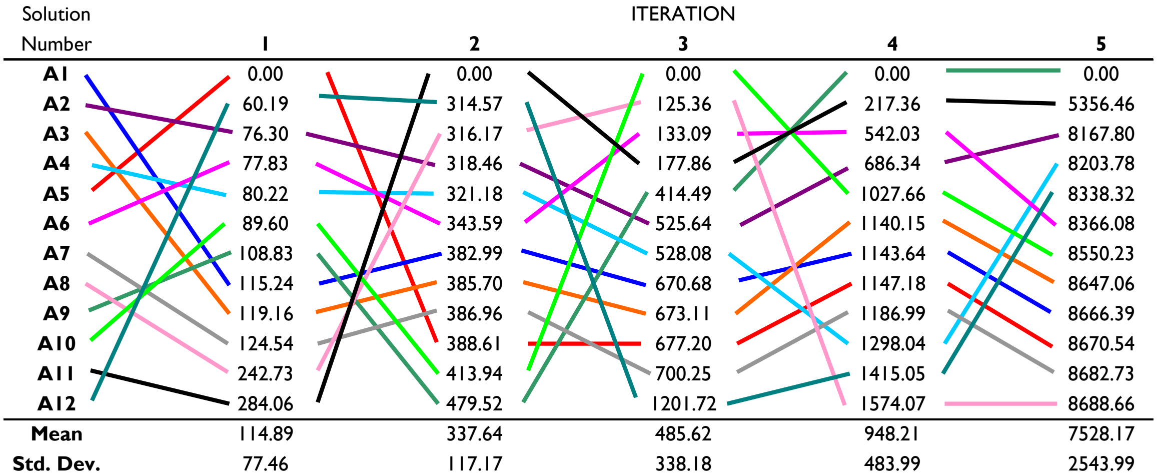


Figure 7 – The Δ_i values of each of the 12 analytical models over successive iterations. Note how the mean and standard deviation of the Δ_i values increase with each successive iteration.

3.2 Applying MMRI to the Hypothetical Site

For each iteration, data generated from the HS were supplied to the MA team and included contact elevations and material types at each contact point, screen elevations, hydraulic head, land surface elevation, *x* and *y* coordinates, geologic material descriptions and/or 'estimated' conductivity values in selected boreholes, as well as information regarding rainfall in the area. With the first data set, a set of alternative conceptual models are constructed, calibrated, and ranked using the *AICc*. This process is repeated as subsequent data sets are supplied, and models are adjusted as appropriate to coincide with the improved understanding of the HS. Starting with the second data set, models simulate the new observations to evaluate predictive capability, before model calibration is performed. Calibration and model ranking results for all of the models from a given data set are reported at the end of each section.

3.3 First Data Set

For the first data set, the MA team received data on 7 boreholes (Table 2), a general site description, and the 40 year averaged rainfall amounts for three rainfall stations (Table 3). The general site description was:

> The 15×15 km^2 regional area transitions from forested highlands to rural, small agricultural use in the low-lands. Beginning at about 300 m in elevation, the vegetation mix is mixed hardwoods consisting of fir, cedar, cottonwood, and ash in equal amounts and trending towards mostly fir with some cedar for elevations above 600 m. Small agriculture dominates below elevations of 300 m, with the primary crops being rice (40%), wheat (20%), and vegetables (40%). The area is sparsely populated by small farm houses and rural dwellings.

> Rainfall generally increases as one moves from the ocean towards the mountains due to orographic mountain effects. Three rainfall gauges exist in the area (see the Rainfall worksheet).
>
> The area is underlain by marine sediments consisting mainly of mudstones and some shale's. Above that are alternating strata of volcanic and alluvial deposits that have been truncated and folded due to past seismic activity. Several faults cross the area and two faults have surface expressions running south to north at x = 6,075 and 7,750 m inland from the ocean.

Using these data coupled with the general site description, the MA team constructed several conceptual models. The first model was based on a one-dimensional Darcy equation that assumes transmissivity varies only in the horizontal (x) direction based on 1 to 5 distinct zones of hydraulic conductivity. The second model was based on an analytical solution to the two-dimensional Laplace equation that assumes a linear water table [*Toth*, 1963]. The third set of models are parameterized variations of a MODFLOW model. Each of these models are described in more detail below.

The average yearly rainfall rates were plotted as a function of x and a linear regression line was fitted that indicates that rainfall rates, $R(x)$ [m/day], range from $1.94\times10^{\;3}$ m/day at the west boundary to $2.60\times10^{\;3}$ m/day at the east boundary (Figure 8):

$$R(x) = 4.55\times10^{-8}x + 1.94\times10^{-3} \qquad \textbf{(15)}$$

Table 2 – Well location and head measurements supplied for the first iterations. Note that heads are "point-in-time" measurements (snapshots) and do not represent long-term static water levels.

Borehole Name	Easting [m]	Northing [m]	Screen Elevation [m]	MODFLOW Layer	Head [m]	Land Surface [m]
BH-5	4,120	10,120	1,447.5	23	63.4	135.4
BH-14	10,637	9,533	279.5	11	330.8	312.4
BH-18	5,667	6,694	29.0	9	109.0	146.0
BH-22	7,434	7,131	410.0	13	186.7	213.2
BH-31	4,413	1,717	251.6	11	80.1	137.9
BH-33	511	13,957	1,475.8	23	8.8	13.9
BH-34	12,426	6,019	212.0	11	390.8	655.2

Table 3 - Monthly rainfall data from the first data set.

Month	Station 1 [m]	Station 2 [m]	Station 3 [m]
January	0.065	0.069	0.051
February	0.142	0.121	0.124
March	0.144	0.119	0.136
April	0.131	0.094	0.088
May	0.110	0.124	0.103
June	0.097	0.096	0.065
July	0.074	0.058	0.057
August	0.054	0.059	0.042
September	0.034	0.033	0.029
October	0.013	0.013	0.008
November	0.016	0.016	0.013
December	0.032	0.032	0.028
Total [m/yr]	**0.913**	**0.835**	**0.742**
Location	**Station 1 [m]**	**Station 2 [m]**	**Station 3 [m]**
Easting (m)	12,566	7,036	2,318
Northing (m)	2,408	12,026	5,769

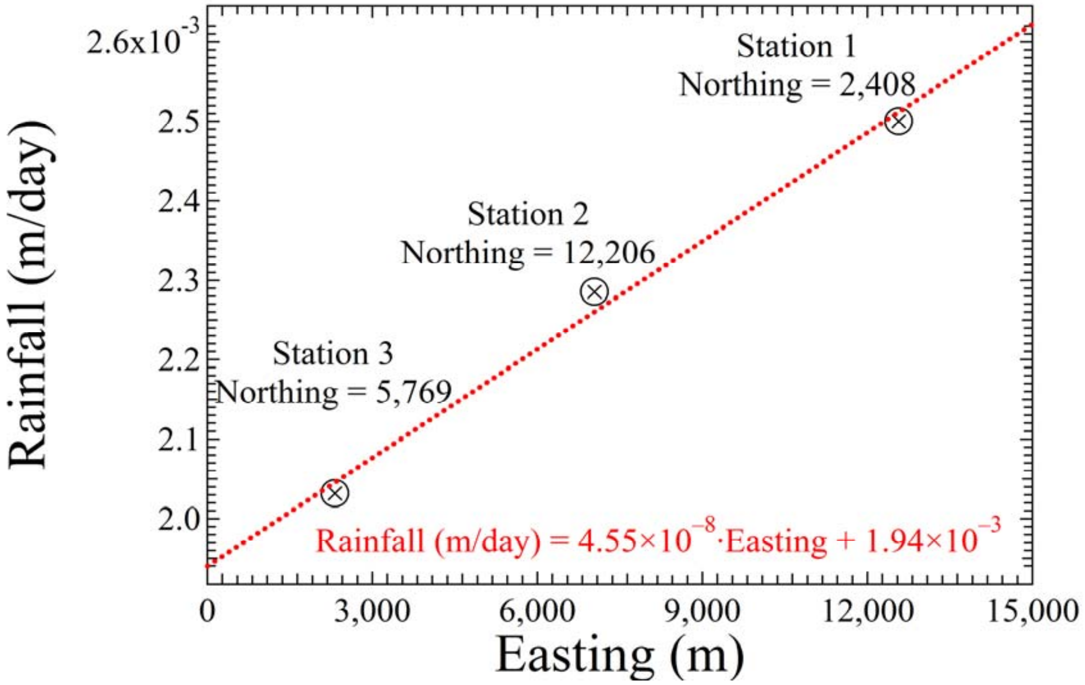


Figure 8 - Data collected from the three rain gages and the best linear fit to the average rainfall rate showing increasing rainfall from west to east.

3.3.1 Darcy Models

The Darcy model envisioned the HS as a one-dimensional system with flow in the negative *x*-direction (east to west). Although this reduction eliminates two spatial degrees of freedom, some hydrologic properties of the original system are retained, such as infiltration and potential fault locations. To incorporate the piecewise nature of the flow system, the model domain is separated into a series of zones (referred to as *T*-zones) with different transmissivities. Model conceptualizations are formed by altering the number and width of each *T*-zone. Figure 9 illustrates an example of five equally-spaced *T*-zones and the measured head data along the *x*-axis (head values were projected onto the one-dimensional domain by assuming that the *y* value is the same for each borehole).

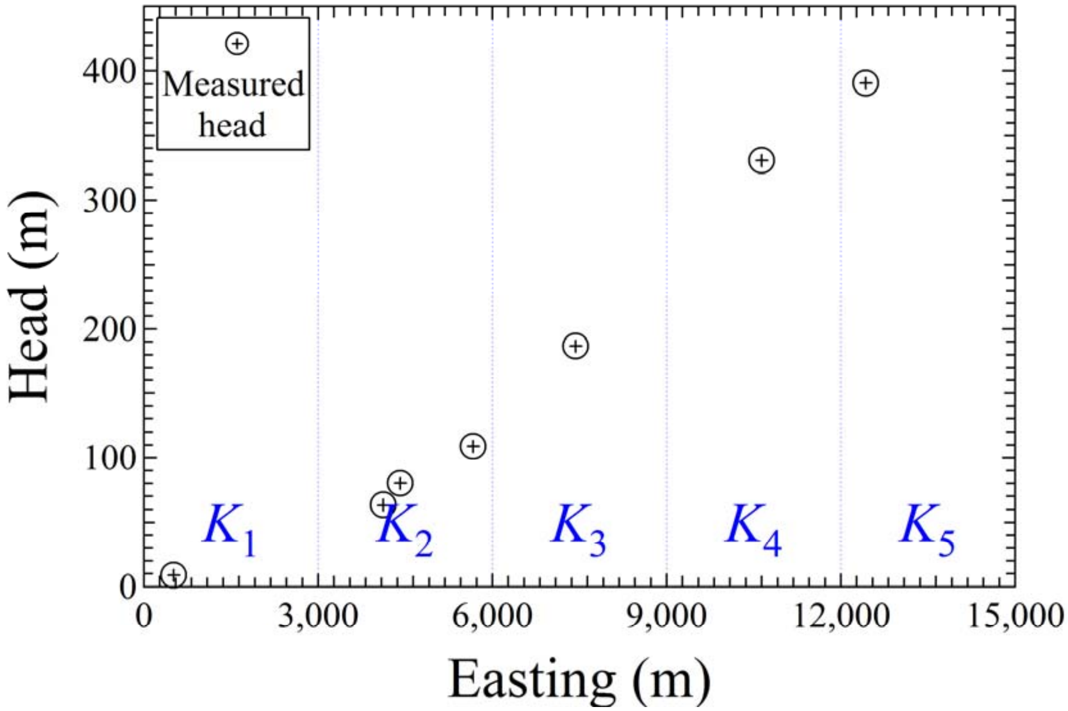


Figure 9 - Example of a 5 T-zone conceptualization with measured head values.

Assuming constant fluid density, flow through the system is described by Darcy's Law (steady state):

$$q(x) = -T(x)\frac{\mathrm{d}h(x)}{\mathrm{d}x} \tag{16}$$

where $q(x)$ is the specific discharge [m^2/day], $T(x)$ is the transmissivity [L^2/t], and $h(x)$ is the hydraulic head [L]. The left-hand boundary ($x = 0$) is modeled as a constant head boundary with a value of zero; i.e., $h(0) = 0$. The right hand boundary is modeled as a no-flow boundary on the east face ($x = L = 15{,}000$ m). Specific discharge is determined exclusively from rainfall rates $R(x)$.

The Darcy model is solved numerically with the following finite-difference approximation:

$$h(x + \Delta x) = h(x) - \frac{q(x)}{T(x)}\Delta x \tag{17}$$

where $\Delta x = 1\,\mathrm{m}$.

The calibration parameters are the transmissivities of each *T*-zone. Six different models are developed and are labeled as 1D-*j*, where *j* denotes the number of serial layers (followed by an optional subscript ‘f’ if Faults A and B are represented). Models 1D-1 through 1D-5 divide the domain into equally-spaced regions, while 1D-5_f contains three equally-spaced regions and two 100-m layers that represent Faults A and B. Using PEST [*Doherty*, 2007], these models were calibrated to the head data given in Table 3.

3.3.2 Toth Model

The next conceptual model considers a two-dimensional, steady-state groundwater model that is based on the Laplace equation:

$$\frac{\partial h(x,z)}{\partial x^2} + \frac{\partial h(x,z)}{\partial z^2} = 0 \qquad \textbf{(18)}$$

where *h*(*x*, *z*) is the hydraulic head [L] at the coordinates *x* and *z*. This model can be interpreted as a vertical cross section parallel to the *x*-axis of the hypothetical site. To solve this equation Toth [1963] assumes a linear sloping water table with slope *c*, and no-flow boundaries at (i) $x = 0$ and $x = s$ for $0 \le z \le z_0$ and (ii) $z = 0$ for $0 \le x \le s$, as illustrated in Figure 10. The solution is given as:

$$h(x,z) = z_0 + \frac{cs}{2} - \frac{4cs}{\pi^2} \sum_{m=0}^{\infty} \frac{\cos\left[(2m+1)\pi x / s\right] \cosh\left[(2m+1)\pi z / s\right]}{(2m+1)^2 \cosh\left[(2m+1)\pi z_0 / s\right]} \qquad \textbf{(19)}$$

where *s* is the domain length [L], and z_0 is the minimum saturated thickness [L]. For our purposes, the infinite series is calculated with *m* = 10, which is enough to minimize truncation error to insignificant levels (<0.02 m).

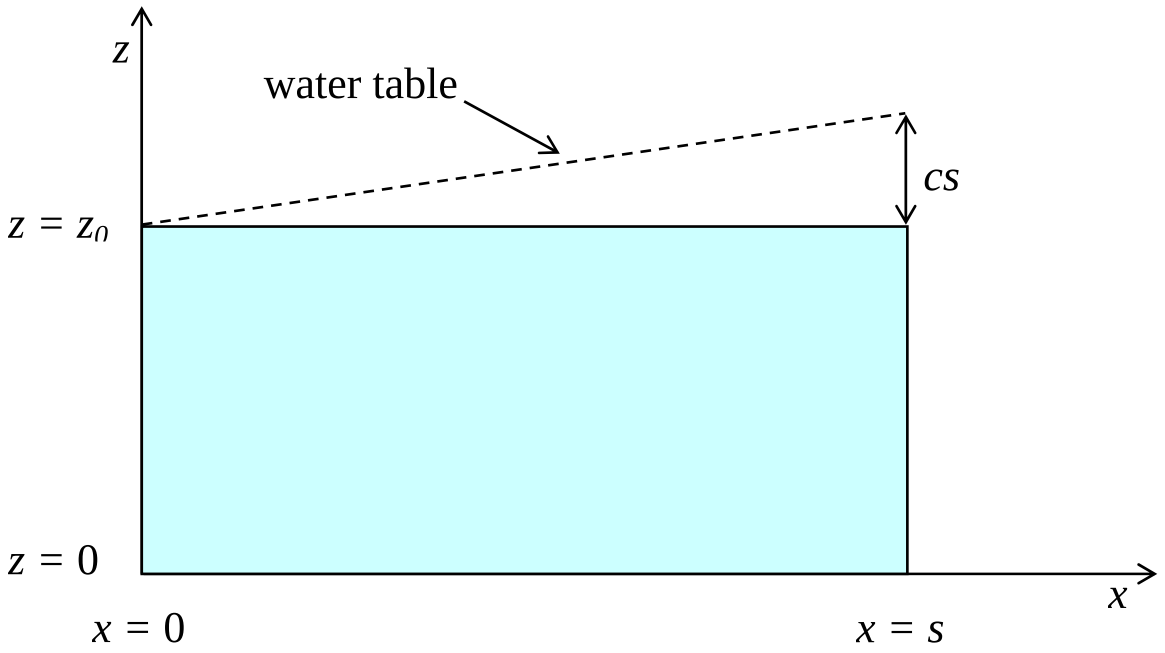


Figure 10 – Two-dimensional model with a linearly varying water table.

Calibration for the Toth model was done using the **lsqcurvefit** data-fitting function that is part of the Matlab optimization toolkit.

3.3.3 MODFLOW Models

The MODFLOW model was constructed using MODFLOW-2000 [*Harbaugh et al.*, 2000]. The domain was divided into 100-m^3 finite-difference cells comprising 150 rows, 150 columns, and 28 layers, resulting in 630,000 total cells with 516,000 active cells, as shown in Figure 11 (top). A zero-head boundary is assigned to the west face while the bottom and east boundaries are designated as no-flow boundaries. The domain is modeled as unconfined with wetting and drying capabilities activated. Recharge is applied to the top-most active layer. Given the limited dataset available for calibration, only three geologic units are considered, the host or background hydraulic conductivity (red) and the two faults (blue) (Figure 11 bottom).

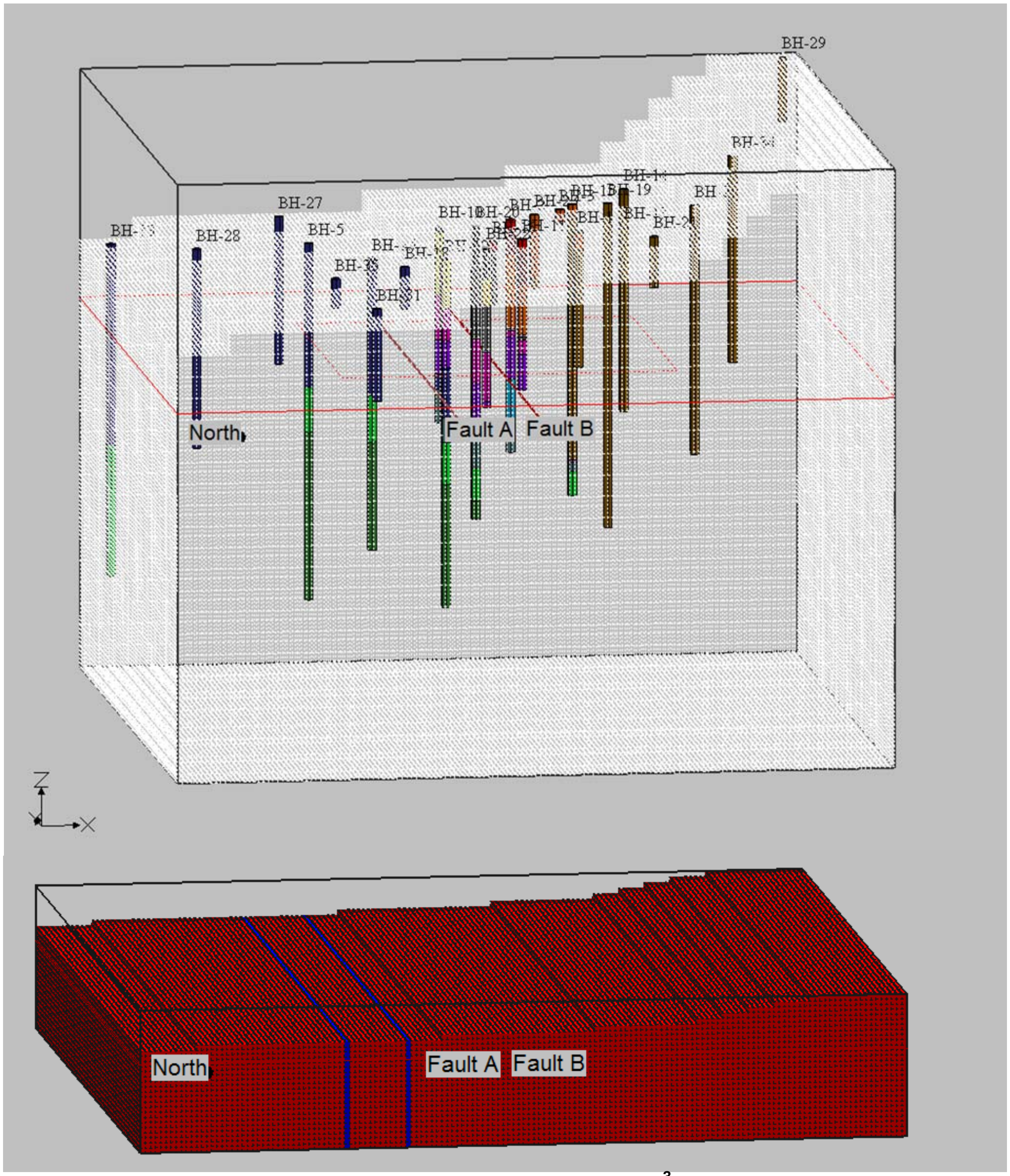


Figure 11 – MODFLOW-2000 grid with 516,600 active 100-m^3 cells. In the top figure, the *z*-axis is magnified by a factor of 4.5. In the bottom figure the host rock is shown in red and faults are shown blue.

Building from a single MODFLOW model, alternate conceptual models were developed and calibrated to the head data and ranked according to the Akaike Information Criterion [*Hill and Tiedeman*, 2007; *Poeter and Anderson*, 2005; *Poeter and Hill*, 2008]. Models were developed with increasing numbers of parameters ranging from one to seven. Table 4 lists the alternate conceptual models developed for this exercise. These models are labeled as MOD-*j*, where *j* is the model index.

Table 4 – Alternate conceptual model descriptions.

Model description and estimated parameters	Number of parameters
MOD-1 Uniform hydraulic conductivity	1
MOD-2 Host rock and Fault B hydraulic conductivities	2
MOD-3 Host rock and Faults A and B hydraulic conductivities	3
MOD-4 Host rock and Faults A and B hydraulic conductivities and vertical anisotropy for host rock	4
MOD-5 Host rock and Faults A and B hydraulic conductivities plus recharge multiplier	4
MOD-6 Host rock and Faults A and B hydraulic conductivities and unique horizontal anisotropies for Faults A and B	5
MOD-7 Host rock and Faults A and B hydraulic conductivities and unique vertical anisotropies for the host rock and Faults A and B	6
MOD-8 Host rock and Faults A and B hydraulic conductivities, unique vertical anisotropies for the host rock and Faults A and B, and horizontal anisotropy for the host rock	7

3.4 Calibration with PEST

With the exception of the Toth model, all models were calibrated with PEST 11.4 [*Doherty*, 2007; *Watermark Computing*, 2003; 2004; 2006]. The PEST (parameter estimation) software is based on the robust, widely-applicable, and well-established Levenberg-Marquardt (LM) optimization algorithm [*Press et al.*, 1992, pp. 678 to 683]. It searches for the local minima of the weighted sum-of-squared residuals (WSSR) between a set of observation data and those calculated by each of the numerical models. For this study, the WSSR is designated as Φ and is calculated as:

$$\Phi(\mathbf{p}) = \sum_{i=1}^{n} w_i \left[h_{obs}^{i} - h_{sim}^{i}(\mathbf{p})\right]^2 \qquad \textbf{(20)}$$

where n is the number of observation data points (borehole heads), h_{obs}^{i} and h_{sim}^{i} are observed and simulated heads at the i^{th} borehole location, respectively, $\mathbf{p}$ is the set of estimation parameters, and $w_i \geq 0$ is the corresponding weight (w_i = 1 for all models in this study).

3.5 Results

3.5.1 Darcy Models

For each of the six Darcy models, optimal values for the hydraulic conductivities are summarized in Table 5 and illustrated in Figure 12. Note that model 1D-5_f (three equally-spaced zones and two 100-m faults) does not yield a significant improvement compared to models 1D-3 and 1D-4. Aside from the two 100-m faults, model 1D-5_f has the same zonation as model 1D-3. Moreover, only one of the faults contains a data point (BH-5 in Fault A), which limits the improvement that this model yields compared to 1D-3. Figure 12 shows that the zonation of models 1D-3 to 1D-5$_f$ neatly divides the head data into regions that are quadratic (in head) with the best fit for five equally-spaced layers in 1D-5. Thus, each zone is accurately fit by a single value of transmissivity. Table 6 contains the calibrated values of the transmissivities. Due to its low Φ (and RMSE), 1D-5 is the best calibrated model (*AICc* cannot be calculated for 1D-5 and 1D-5$_f$ because the second order correction has an indeterminate form).

Table 5 – Results from Darcy Model calibrations for the first data set. For models 1D-5 and 1D-5$_f$, values of *AICc* are not reported because the second-order bias involves division by zero.

		h_{sim} [m]					
Borehole	h_{obs} **[m]**	**1D-1**	**1D-2**	**1D-3**	**1D-4**	**1D-5**	**1D-5$_f$**
BH-5	63.4	140.4	92.4	65.7	62.3	67.6	62.7
BH-14	330.8	277.5	329.7	322.0	330.8	330.8	323.9
BH-18	109.0	182.7	120.3	112.9	122.3	109.5	112.1
BH-22	186.7	223.7	147.3	196.2	180.5	186.7	196.9
BH-31	80.1	148.8	98.0	69.7	74.3	76.0	74.2
BH-33	8.8	19.5	12.8	9.1	7.0	6.2	8.4
BH-34	390.8	295.3	391.6	393.6	390.8	390.8	394.3
Φ	—	29,500	2,860	300	250	40	210
RMSE	—	64.9	20.2	6.6	6.0	2.4	5.5
AICc	—	65.4	56.1	54.4	95.1	—	—

Table 6 – Estimated parameters for the Darcy-flow model calibrated with the first data set. For model 1D-5$_f$, T_4 and T_5 are the transmissivities for Faults A and B, respectively.

	Transmissivities [m^2/day]				
Model	T_1	T_2	T_3	T_4	T_5
1D-1	884	–	–	–	–
1D-2	1,340	255	–	–	–
1D-3	1,890	436	220	–	–
1D-4	2,450	623	312	238	–
1D-5	2,780	893	419	261	273
1D-5$_f$	2,050	229	2,050	229	773,000

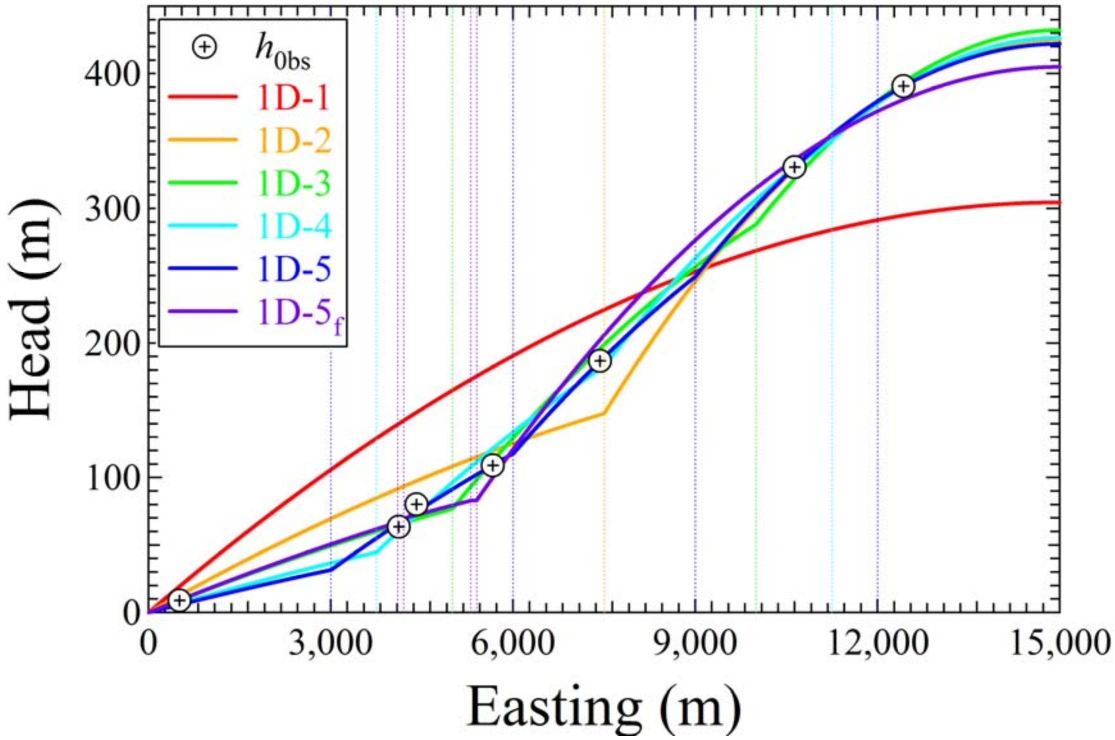


Figure 12 – Head values for calibrations performed for serial geologic units with unique transmissivities. The colored vertical lines designate the location of the serial geologic units for each of the six partitioning scenarios.

3.5.2 Toth Model

For the Toth model, the predicted heads for the calibrated value of c = 0.028 are shown in Table 7. Interestingly, this model produces a better fit (i.e., lower RMSE) than the corresponding one-parameter Darcy-flow model 1D-1, which may be related to the increased dimensionality of the Toth model.

Table 7– Calibrated heads for the Toth model for the first data set.

Borehole	h_{obs} **[m]**	h_{sim} **[m]**
BH-5	63.4	116.1
BH-14	330.8	299.0
BH-18	109.0	159.7
BH-22	186.7	209.3
BH-31	80.1	124.7
BH-33	8.8	17.5
BH-34	390.8	347.0
Φ	—	10,800
RMSE	—	39.4
AICc	—	54.2

3.5.3 MODFLOW Models

Calibrated head data for the first data set are presented in Table 8. With the exception of MOD-4, all models calibrate to reasonably similar head values, although none are particularly close to the observations. Note that MOD-5 provides a non-unique solution to the flow problem (i.e., values of recharge and hydraulic conductivity can vary in a commensurate manner with little effect on the model outcome). Thus, MOD-5 does not differ significantly from MOD-3, but it does yield a significantly different estimated hydraulic conductivity for Fault A. Overall, this suggests that these conceptual models may be too complex for the given amount of data. Table 9 lists the calibrated parameters for the MODFLOW models.

Table 8 – Calibrated head values for each conceptual MODFLOW model for the first data set. For models MOD-6 through MOD-8, *AICc* is not reported because the second-order bias is either negative or involves division by zero.

		h_{sim} [m]							
Borehole	h_{obs} [m]	**MOD-1**	**MOD-2**	**MOD-3**	**MOD-4**	**MOD-5**	**MOD-6**	**MOD-7**	**MOD-8**
BH-5	63.4	139.3	112.4	119.0	125.0	114.0	117.1	126.8	114.2
BH-14	330.8	273.6	277.4	283.9	274.2	278.2	278.2	270.9	276.5
BH-18	109.0	185.4	207.3	209.2	176.8	206.6	204.4	189.2	203.4
BH-22	186.7	222.8	236.7	240.7	196.6	236.8	235.5	224.9	235.4
BH-31	80.1	151.4	122.4	126.6	133.8	121.3	125.2	134.4	120.7
BH-33	8.8	17.6	14.2	15.1	14.5	14.5	14.9	16.8	15.1
BH-34	390.8	290.4	290.8	298.3	316.6	292.0	292.4	286.1	290.0
Φ	—	31,400	29,200	29,000	20,100	28,800	28,900	29,400	28,600
RMSE	—	67.0	64.6	64.4	53.6	64.2	64.2	64.9	64.0
AICc	—	65.9	72.4	86.3	125.7	128.3	—	—	—

Table 9 – Estimated parameters for the MODFLOW models calibrated with the first data set. Optimized parameters are in bold type, while all other parameters are fixed.

Model	K_{host}	K_{FA}	K_{FB}	V_{host}	V_{FA}	V_{FB}	H_{host}	H_{FA}	H_{FB}	R_{mult}
MOD-1	**0.42**	0.42	0.42	1	1	1	1	1	1	1
MOD-2	**0.52**	0.52	**0.018**	1	1	1	1	1	1	1
MOD-3	**0.49**	**26,800**	**0.020**	1	1	1	1	1	1	1
MOD-4	**0.48**	**35.0**	**21.2**	**150**	1	1	1	1	1	1
MOD-5	**0.026**	**11.2**	**9.60**	1	1	1	1	1	1	**0.052**
MOD-6	**0.49**	**23.7**	**0.021**	1	1	1	1	**0.1**	**8.70**	1
MOD-7	**0.46**	**977**	**0.040**	**0.080**	**402**	**0.29**	1	1	1	1
MOD-8	**0.51**	**141**	**0.019**	**0.0032**	**6.96**	**1.17**	**1.70**	1	1	1

3.6 First Data Set Summary

Table 10 presents the *AICc*, Δ_i, and w_i values (calculated according to equations (5), (6), and (7)) for the full ensemble of conceptual models (Darcy, Toth, and MODFLOW models) for the first data set. An interesting result is the ranking of the Toth model with the largest weight, with a model probability of 43%, followed closely by 1D-3. The fact that this model incorporates head variations along the *z*-axis might be the reason it gives a slightly better fit (in terms of RMSE) than Darcy models 1D-1 and 1D-2. With only 7 data points for the calibration, these rankings favor models with a smaller number of parameters that can provide a reasonable fit, which could explain why the Toth model also outperforms the MODFLOW models. All but one of the Darcy models fit the observations better than the Toth model (i.e., they have a lower RMSE), but their *AICc* is penalized by the 2nd order bias term. This suggests that the amount of information contained in the 7 data points is not enough to support a larger number of parameters.

Table 10 – Model ranking statistics for the 15 models calibrated with the first data set. The best fit model (Toth) is bolded. For models 1D-5, 1D-5$_f$ and MOD-6 through MOD-8, *AICc* is not reported because the second-order bias is either negative or involves division by zero.

Model	***k***		**RMSE**	***AICc***	$_i$	***w$_i$***
1D-1	2	29,500	64.9	65.4	11.2	1.6E-03
1D-2	3	2,860	20.2	56.1	1.9	1.7E-01
1D-3	4	3,040	6.6	54.4	0.2	4.0E-01
1D-4	5	250	6.0	95.1	40.9	5.7E-10
1D-5	6	40	2.4	NA	NA	NA
1D-5$_f$	6	210	5.5	NA	NA	NA
Toth	**2**	**10,800**	**39.4**	**54.2**	**0.0**	**4.3E-01**
MOD-1	2	31,400	67.0	65.9	11.6	1.3E-03
MOD-2	3	29,200	64.6	72.4	18.1	5.0E-05
MOD-3	4	29,000	64.4	86.3	32.1	4.7E-08
MOD-4	5	20,100	53.6	125.7	71.5	1.3E-16
MOD-5	5	28,800	64.2	128.3	74.0	3.6E-17
MOD-6	6	28,900	64.2	NA	NA	NA
MOD-7	7	29,400	64.9	NA	NA	NA
MOD-8	8	28,600	64.0	NA	NA	NA

3.7 Second Data Set

For the second phase of model development, 13 additional head measurements were given to the MA team (Table 11). These data were used in the calibration of the MODFLOW conceptual models introduced in the previous sections. Material descriptions and some estimated hydraulic conductivities as a function of depth for more boreholes were also provided. Analysis of these descriptions revealed the presence of a distinct horizontal layer in the host rock between screen elevations of 500 m to 900 m.

Table 11 – Well location and head measurements used in the calibration of the second data set. Note that heads are "point-in-time" measurements (snapshots) and do not represent long-term static water levels.

Borehole Name	Easting [m]	Northing [m]	Screen Elevation [m]	MODFLOW Layer	Head [m]	Land Surface [m]
BH-5a	4,120	10,120	NA	6	69.9*	135.4
BH-5b	4,120	10,120	100.0	10	67.7	135.4
BH-5c	4,120	10,120	300.0	12	69.4	135.4
BH-5d	4,120	10,120	650.0	15	67.1	135.4
BH-5e	4,120	10,120	900.0	18	65.1	135.4
BH-5f	4,120	10,120	1,200.0	21	65.8	135.4
BH-5g	4,120	10,120	1,447.5	23	63.4	135.4
BH-12	6,713	8,097	1,442.5	23	197.5	180.9
BH-14	10,637	9,533	279.5	11	330.8	312.4
BH-18	5,667	6,694	29.0	9	109.0	146.0
BH-22	7,434	7,131	410.0	13	186.7	213.2
BH-31	4,413	1,717	251.6	11	80.1	137.9
BH-33	511	13,957	1,475.8	23	8.8	13.9
BH-34a	12,426	6,019	212.0	11	390.8	655.2
BH-34b	12,426	6,019	NA	4	426.3*	655.2
BH-41	5,613	5,351	799.8	16	97.3	146.0
BH-44	7,948	9,781	NA	7	194.3*	226.4
BH-48	5,926	8,871	2.6	8	118.5	148.9
BH-52	10,034	5,510	424.4	13	297.6	284.6
BH-82	5,455	4,037	1,247.0	21	77.3	145.4

*water level measurement

3.7.1 Second Data Set Model Verification

Before the data in Table 11 were used to further calibrate the conceptual models from the previous section, a verification was performed. Using the calibrated model parameters resulting from the first data set, simulations were performed that calculated the additional heads reported in the second data set. This process evaluates the predictive capability of the calibrated models, and is related to methods in cross validation [e.g., *Foglia et al.*, 2007]. Because the Darcy model is one-dimensional, the series borehole data are eliminated from model verification, yielding 13 head observations (BH-5f and BH-34b are retained). For these simulations, and RMSE are calculated just as in the case of direct calibration, except with an expanded number of observations. The *AICc* and corresponding weights are also calculated, where *n*, the number of observations, is 13. These results are reported in Table 12. The model weights from verification with the second data set will be compared to the model weights from calibration with this data set to investigate the additional insight that prediction adds to model development.

Table 12 – Results for the 15 models verified with the second data set. The data for the best fit model, (in terms of *AICc* weights) 1D-3, is bold.

Model	***k***		**RMSE**	***AICc***	$_i$	w_i
1D-1	2	54,100	64.5	113.5	30.5	2.3E-07
1D-2	3	8,800	26.1	93.5	10.4	5.3E-03
1D-3	**4**	**2,800**	**14.8**	**83.0**	**0.0**	**9.9E-01**
1D-4	5	4,200	18.0	93.7	10.7	4.7E-03
1D-5	6	3,100	15.4	97.0	14.0	8.9E-04
1D-5_f	6	2,700	14.3	95.2	12.2	2.2E-03
Toth	2	23,900	42.9	102.9	19.9	4.7E-05
MOD-1	2	52,100	63.3	116.5	33.5	5.3E-08
MOD-2	3	63,500	69.9	123.4	40.4	1.7E-09
MOD-3	4	64,900	70.6	129.3	46.3	8.9E-11
MOD-4	5	77,300	77.1	139.0	56.0	7.0E-13
MOD-5	5	62,900	69.6	136.3	53.3	2.6E-12
MOD-6	6	61,300	68.7	146.4	63.3	1.7E-14
MOD-7	7	57,300	66.4	161.1	78.1	1.1E-17
MOD-8	8	66,400	71.5	189.0	106.0	9.5E-24

Comparing Table 12 with Table 10, the Toth and almost all of the MODFLOW models (except for model MOD-4) maintain a relatively constant RMSE from calibration to prediction, indicating a degree of model stability. In contrast, all but one of the Darcy models change substantially. Note that the model with the lowest RMSE is not the same from model calibration (1D-5) to subsequent prediction (1D-5_f). In terms of *AICc*, model 1D-3 now has the largest weight, which is due to a balance between RMSE and the small number of parameters used for calibration. Figure 13 illustrates the predictive capability of the calibrated Darcy models from the previous section by combining the head curves from Figure 12 and the new heads from the second data set (Table 11).

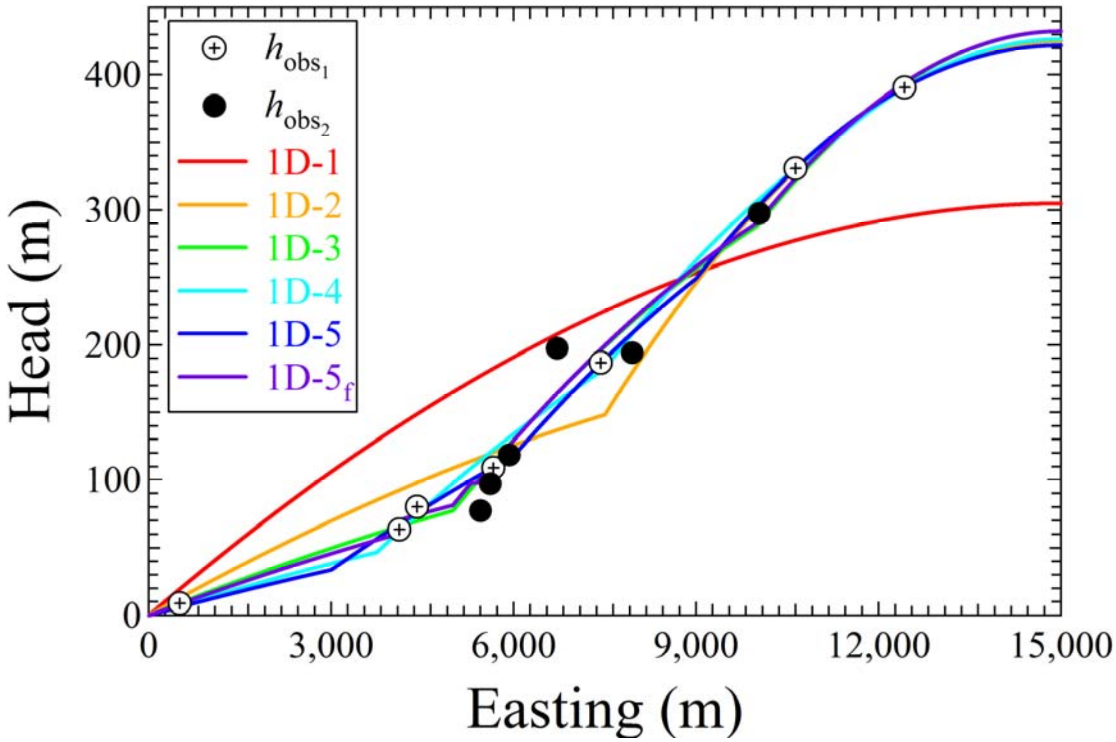


Figure 13 – Hydraulic head curves from Figure 12, with the measured heads from the first and second data sets. Based on differences between new heads from the second data set, the calibrated Darcy models from the first data set have a limited predictive ability.

3.7.2 Updates to the Darcy Models

With the additional head observations in the second data set, the number of possible zones was increased, which resulted in the number of models increasing from 6 to 14, with a maximum of 10 zones of transmissivity. Because there are no head observations in the interval $12,426 < x \le 15,000,$ when the number of layers was greater than 5, they were equally spaced between $0 \le x \le 12,426.$ This adjustment was done because any zones contained within $12,426 < x \le 15,000$ would not affect model calibration efforts.

3.7.3 Toth Model

No changes were made to the Toth model for use in the second data set. Interestingly, it outperformed all of the MODFLOW models in calibration (Table 10) and prediction (Table 12).

3.7.4 Updates to the MODFLOW Model

An additional estimation parameter was included in each of the conceptual models listed in Table 4 (yielding a maximum of 8 parameters). This parameter resulted from partitioning the hydraulic conductivity of the host rock into two independent regions: K_{host} for vertical layers 14 through 17 (corresponding to screen elevations of 500 m to

900 m), and K_{host} for all other 24 layers. Based on the material description of the host rock in the second data set, K_{host} was added to accommodate a horizontal region of higher hydraulic conductivity. Aside from the larger set of borehole data and the inclusion of the estimation parameter K_{host}, the eight conceptual models were solved with the same MODFLOW-2000 and PEST coupling routine as described previously.

3.8 Results

3.8.1 Darcy Models

Calibrated head data for the second data set are presented in Table 13. Compared to the results in Table 10, the average RMSE increased by approximately 3 m. Interestingly, the model with the largest number of zones did not result in the best calibrated model. Compared to the first data set, the RMSE for all models decreased. Table 14 contains the calibrated values of the transmissivities.

Table 13 – Results from Darcy model calibrations for the second data set.

		***h*$_{\text{sim}}$ [m]**								
Borehole	***h*$_{\text{obs}}$ [m]**	**1D-1**	**1D-2**	**1D-3**	**1D-4**	**1D-5**	**1D-6**	**1D-7**	**1D-8**	**1D-10**
BH-5a	63.4	127.8	93.8	60.5	52.0	62.3	58.8	69.7	62.5	63.4
BH-12	197.5	189.3	139.0	161.8	155.6	154.4	153.9	175.0	158.4	166.3
BH-14	330.8	252.6	329.5	323.8	324.4	328.1	327.0	334.0	333.7	331.5
BH-18	109.0	166.3	122.1	108.4	116.9	109.8	112.4	102.6	111.6	109.6
BH-22	186.7	203.7	149.5	195.2	179.8	187.2	186.4	199.8	189.5	197.0
BH-31	80.1	135.5	99.4	64.2	65.0	71.8	69.1	71.6	63.5	63.4
BH-33	8.8	17.8	13.0	8.4	5.3	4.4	8.2	8.8	9.2	8.8
BH-34	390.8	268.8	390.5	393.1	390.7	390.7	391.5	390.8	391.5	391.5
BH-41	97.3	165.1	121.1	105.4	114.8	108.3	110.7	98.6	109.2	106.1
BH-44	194.3	213.2	181.5	217.3	206.6	208.8	207.9	199.9	207.8	197.9
BH-48	118.5	172.2	126.4	122.1	126.8	117.2	120.8	121.3	123.0	125.9
BH-52	297.6	245.3	302.1	292.7	302.6	301.7	301.6	292.9	294.0	298.3
BH-82	77.3	161.3	118.4	96.8	108.5	103.7	105.5	86.9	102.0	95.9
		49,600	8.800	2,700	3,800	3,000	3,300	1,000	2,800	1,900
RMSE	—	61.7	26.0	14.4	17.1	15.2	15.8	8.7	14.7	11.9
AICc	—	112.4	93.4	82.3	92.4	96.8	108.2	108.4	147.8	350.5

Table 14 – Estimated parameters for the Darcy-flow model calibrated for the second data set. For model 1D-5$_f$, T_4 and T_5 are the transmissivities for Faults A and B, respectively.

	Transmissivities [m^2/day]									
Model	T_1	T_2	T_3	T_4	T_5	T_6	T_7	T_8	T_9	T_{10}
1D-1	970	–	–	–	–	–	–	–	–	–
1D-2	1,320	258	–	–	–	–	–	–	–	–
1D-3	2,050	418	228	–	–	–	–	–	–	–
1D-4	3.250	577	324	202	–	–	–	–	–	–
1D-5	3,950	786	426	268	203	–	–	–	–	–
1D-6	2,110	2,110	691	429	316	244	–	–	–	–
1D-7	1,950	1,410	4,000	308	647,000	172	280	–	–	–
1D-8	1,867	1,450	7,540	503	450	656	179	303	–	–
1D-10	1,950	1,790	1,640	543,000	352	455	64,400	186	213	302

3.8.2 Toth Model

Calibration results for the Toth model are reported in Table 15. Given that this model only has one calibration parameter, the fact that RMSE increased by only 2 m compared to the calibration with the first data set is interesting. It may be possible to improve the calibration of this model by incorporating oscillatory variations in the water table boundary condition, thereby adding two additional parameters (amplitude and frequency of oscillations) for calibration [*Toth*, 1963].

Table 15 – Calibrated heads for the Toth model from the second data set.

Borehole	$\boldsymbol{h_{\text{obs}}}$ **[m]**	$\boldsymbol{h_{\text{sim}}}$ **[m]**
BH-5f	63.4	116.1
BH-12	197.5	189.0
BH-14	330.8	299.0
BH-18	109.0	159.7
BH-22	186.7	209.3
BH-31	80.1	124.7
BH-33	8.8	17.5
BH-34b	390.8	347.0
BH-41	97.3	158.2
BH-44	194.3	223.8
BH-48	118.5	167.0
BH-52	297.6	282.3
BH-82	77.3	153.7
Φ	—	22,300
RMSE	—	41.4
AICc	—	99.2

3.8.3 MODFLOW Models

Calibrated head data for the second data set are presented in Table 16. As before (see Table 8), all models calibrate to similar head values, but with few exceptions, most do not come particularly close to the observations. The *AICc* values show that MOD-1 is the best ranked model followed closely by MOD-2. The average RMSE for the second data set is approximately 10 m smaller than the first, demonstrating that the combination of an increased number of head values and the inclusion of the additional parameter K $_{\text{host}}$ improved the ability of the MODFLOW models to represent the true model. For all cases, note the consistent difference between K_{host} and K $_{\text{host}}$ of approximately one order of magnitude or more, supporting the continued use of the host rock partitioning scheme.

Table 16 – Calibrated head values [m] for each alternate conceptual MODFLOW model.

		h_{sim} [m]							
Borehole	**h_{obs} [m]**	**MOD-1**	**MOD-2**	**MOD-3**	**MOD-4**	**MOD-5**	**MOD-6**	**MOD-7**	**MOD-8**
BH-5f	63.4	115.4	91.5	110.2	108.8	110.2	116.2	101.8	91.6
BH-12	197.5	179.2	197.3	173.9	165.0	175.3	181.4	166.3	195.8
BH-14	330.8	237.2	242.9	237.3	240.5	240.5	247.0	250.3	242.7
BH-18	109.0	164.3	184.8	163.8	169.6	167.3	169.4	166.3	179.5
BH-22	186.7	194.1	209.0	190.7	187.0	192.8	198.6	194.5	206.6
BH-31	80.1	128.1	99.3	119.9	120.3	120.6	126.2	110.5	97.3
BH-33	8.8	13.7	11.2	13.7	13.3	13.7	14.5	12.0	11.8
BH-34b	390.8	251.8	254.2	253.7	260.8	257.4	263.9	272.2	254.8
BH-41	97.3	160.5	183.1	154.7	146.9	156.1	161.1	158.0	178.0
BH-44	194.3	208.2	218.5	214.9	241.6	220.8	221.6	266.4	214.4
BH-48	118.5	170.5	189.4	171.7	181.5	175.8	177.3	181.5	184.0
BH-52	297.6	229.5	237.2	226.9	222.5	229.0	236.7	227.7	237.1
BH-82	77.3	77.0	78.5	60.0	88.1	53.9	72.7	64.1	82.0
Φ	—	48,100	50,400	46,700	47,900	46,400	44,100	45,200	47,600
RMSE	—	60.8	62.2	60.0	60.7	59.8	58.2	59.0	60.5
AICc	—	115.5	120.4	125.0	132.8	132.4	142.1	158.0	184.7

All estimated parameters are reported in Table 17. For all models, note the consistent difference between K_{host} and $K\ \ _{\text{host}}$ of approximately one order of magnitude or more, supporting the use of the host rock partitioning scheme. The reduced RMSE supports the inclusion of a horizontally conducting feature in layers 14 through 17 ($K\ \ _{\text{host}}$).

Table 17 – Estimated parameters for the second data set. Optimized parameters are bold (all other parameters are fixed).

Model	K_{host}	$K\ \ _{\text{host}}$	K_{FA}	K_{FB}	V_{host}	V_{FA}	V_{FB}	H_{host}	H_{FA}	H_{FB}	R_{mult}
MOD-1	**0.22**	**1.84**	0.22	0.22	1	1	1	1	1	1	1
MOD-2	**0.42**	**1.64**	0.42	**0.018**	1	1	1	1	1	1	1
MOD-3	**0.083**	**2.37**	**1,240**	**0.28**	1	1	1	1	1	1	1
MOD-4	**0.063**	**2.48**	**1,380**	**0.54**	**2.20**	1	1	1	1	1	1
MOD-5	**0.060**	**2.16**	**413**	**0.23**	1	1	1	1	1	1	**0.89**
MOD-6	**0.093**	**2.19**	**18.5**	**0.29**	1	1	1	1	**84.8**	**0.72**	1
MOD-7	**0.14**	**2.35**	**23,200**	**0.10**	**7.61**	**390**	**0.0004**	1	1	1	1
MOD-8	**0.038**	**3.11**	**159**	**0.035**	**0.025**	**0.57**	**57.1**	**0.66**	1	1	1

3.9 Second Data Set Model Ranking

Table 18 summarizes the calibration results for the second data set. According to the *AICc*, the highest ranking model is the Darcy model with three calibration parameters (1D-3). This model does not have the lowest   (or RMSE), but *AICc* weights the other models unfavorably due to second order bias. Comparing calibrated results in Table 18 to verification results in Table 12, in terms of absolute values, the RMSE decreases slightly for the Darcy and Toth models and more significantly for the MODFLOW models. Interestingly, from prediction to calibration, *AICc* model weights remain approximately the same for almost all models that appear in Table 12. The ranking of the models in this data set differs from the previous ranking in that, because of the larger number of observations, all models now have a calculable weight although one model (1D-3) is vastly superior to all others. Comparing Table 18 to Table 10 (first data set calibration), the RMSE and Δ_i values for the MODFLOW models decreased on average, while the opposite occurs for the Darcy and Toth models. This suggests that the MODFLOW models are responding well to the increased number of observations and parameters, and increasing the complexity of the parameterization scheme as more data become available may continue to improve the calibration and ranking results of these models.

Table 18 – Model ranking statistics for the 20 models calibrated with the second data set. The best fit model, 1D-3, is bold.

Model	k		RMSE	*AICc*	$_i$	w_i
1D-1	2	49,500	61.7	112.4	30.1	2.8E-07
1D-2	3	8,800	26.0	93.4	11.2	3.7E-03
1D-3	**4**	**2,700**	**14.4**	**82.3**	**0.0**	**9.9E-01**
1D-4	5	3,800	17.1	92.4	10.1	6.2E-03
1D-5	6	3,000	15.2	96.8	14.5	6.9E-04
1D-5$_f$	6	2,500	13.9	94.5	12.2	2.2E-03
1D-6	7	3,300	15.8	108.2	25.9	2.3E-06
1D-7	8	1,000	8.7	108.4	26.1	2.1E-06
1D-8	9	2,800	14.7	147.8	65.6	5.7E-15
1D-10	11	1,900	11.9	350.5	268.2	5.7E-59
Toth	2	22,300	41.4	99.2	16.9	2.1E-04
MOD-1	3	48,100	60.8	116.5	34.2	3.6E-08
MOD-2	4	50,400	62.2	123.4	41.2	1.1E-09
MOD-3	5	46,800	60.0	129.3	47.0	6.1E-11
MOD-4	6	47,900	60.7	139.0	56.7	4.8E-13
MOD-5	6	46,400	59.8	136.3	54.0	1.8E-12
MOD-6	7	44,100	58.2	146.4	64.1	1.2E-14
MOD-7	8	45,200	59.0	161.1	78.8	7.5E-18
MOD-8	9	47,600	60.5	189.0	106.7	6.5E-24

3.10 Multi-model Parameter Estimations

If we consider the host rock hydraulic conductivity K_{host} from all 8 MODFLOW models, which is the complete set of models that contain K_{host} as a parameter, we can calculate the multi-model averaged value using equation (9). Table 19 contains values of K_{host} and w'_j for all models and the resulting estimated parameter $\bar{K}_{\text{host}}$. Because one of the models (MOD-1) is weighted much larger than all of others, the resulting value of $\bar{K}_{\text{host}}$ is basically the same as that for MOD-1.

Table 19 – Host rock hydraulic conductivity K_{host} for each MODFLOW model, renormalized weights, and the estimated model-averaged host rock hydraulic conductivity $\bar{K}_{\text{host}}$.

Model	K_{host}	w'_j
MOD-1	0.22	0.97
MOD-2	0.42	0.030
MOD-3	0.083	1.6E-03
MOD-4	0.063	1.3E-05
MOD-5	0.060	4.9E-05
MOD-6	0.093	3.2E-07
MOD-7	0.14	2.0E-10
MOD-8	0.038	1.7E-16
	$\bar{K}_{\text{host}}$	**0.22**

3.11 Third Data Set

This data set expands the description of the site, most notably with the inclusion of time-dependent information. In addition to static head values from BH-6 and BH-30 (BH-30 is a multi-completion well) and static water levels from BH-6b, BH-16, and BH-70, which are listed in Table 20, this data set also contains partial time series information between 1971 and 2000 for BH-25 (measured at $z =$ 1,222.1 m) and BH-34 (measured at $z =$ 212.0 m) as shown in Figure 14. It also reports an important characteristic about rainfall recharge, namely that recharge is zero west of x = 7,750 m. Moreover, the new data set includes transient rainfall information from Station 1 between 1971 and 2000 as shown in Figure 15. Because previous head data are “point-in-time” measurements (snapshots and not static water levels), they were not considered in the calibration of transient models.

Table 20 – Static head data used to calibrate the two-dimensional transient, unconfined model.

Borehole Name	*x* [m]	*z* [m]	Head [m]
BH-6a	4,917	NA	93.3*
BH-16	7,790	NA	194.4*
BH-70	13,605	NA	470.4*
BH-6b	4,917	1,310.7	77.0
BH-30a	8,213	95.2	233.2
BH-30b	8,213	93.7	234.9
BH-30c	8,213	223.7	229.7
BH-30d	8,213	266.7	227.1
BH-30e	8,213	428.8	221.6
BH-30f	8,213	625.6	219.8
BH-30g	8,213	698.43	219.6
BH-30h	8,213	800.1	225.8
BH-30i	8,213	975.4	227.2
BH-30j	8,213	1,163.3	226.3
BH-30k	8,213	1,251.0	229.6

*Water-level measurement

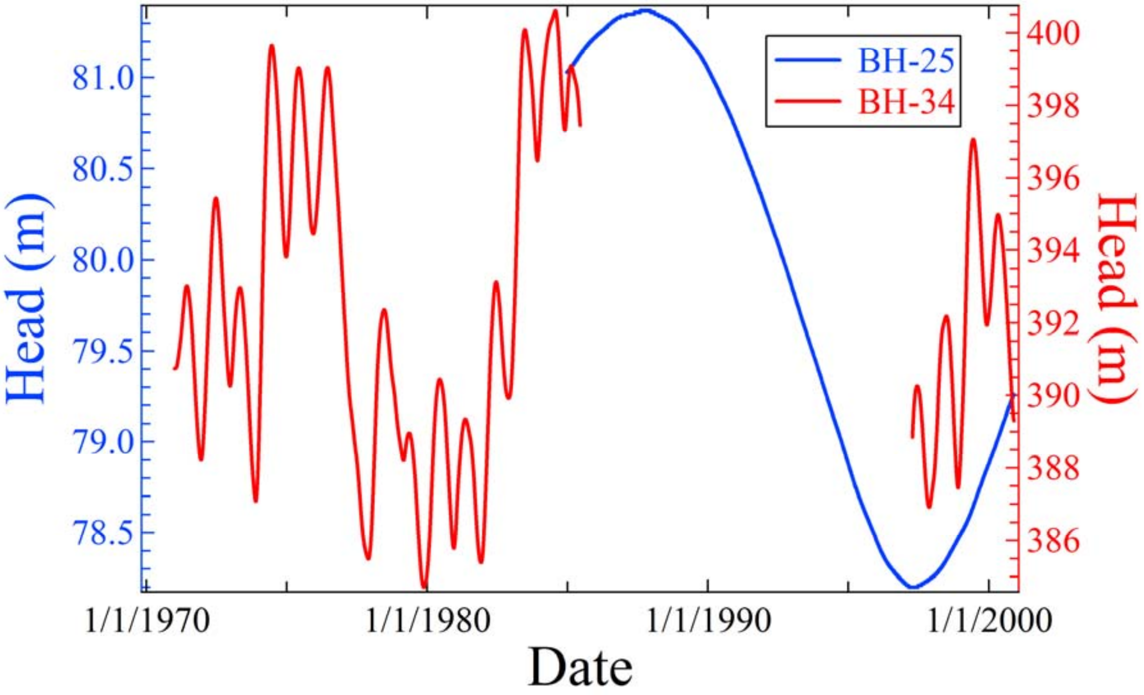


Figure 14 – Transient head data for BH-25 and BH-34.

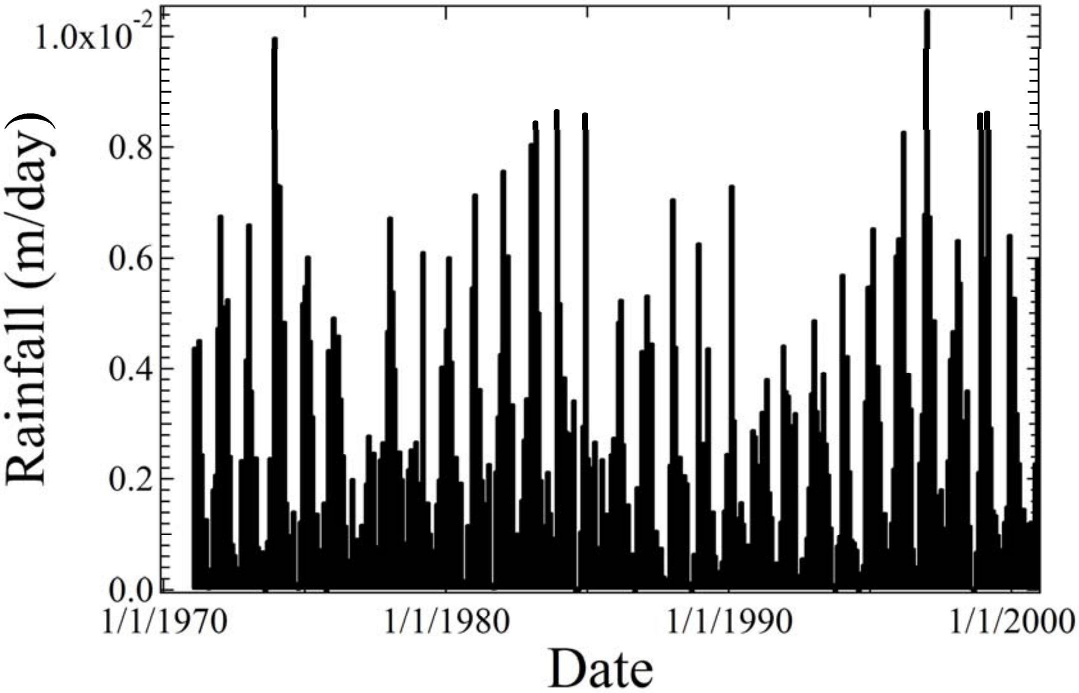


Figure 15 – Rainfall data at Station 1 (*x* = 12,566 m).

Based on the expanded material descriptions (i.e., geologic units) from borehole data, a horizontal low-hydraulic-conductivity fault consisting of cemented material is present at screen elevations of 200 to 300 m. The data set also revealed that the geology, water table, and recharge are homogeneous in the *y*-direction, meaning that they vary minimally in the north-south direction, making a two-dimensional (*x* and *z*) representation is appropriate.

3.11.1 Third Data Set Model Verification

The model verification includes the static heads from the second data set and BH-6b, BH-16, BH-70, and BH-30k from Table 20 (because the Darcy models are one-dimensional, only one head observation from series borehole data can be included). Model 1D-3 has the highest rank. With the addition of time-dependence in this data set, only transient MODFLOW models will be developed and calibrated as described in the next section.

Table 21 – Results for the 15 models verified with the third data set. The data for the highest ranking model, (in terms of *AICc* weights) 1D-3, is bolded.

Model	*k*		RMSE	*AICc*	$_{i}$	w_i
1D-1	2	93,200	73.4	151.2	41.4	9.5E-10
1D-2	3	14,500	28.3	122.6	12.8	1.6E-03
1D-3	**4**	**5,600**	**18.0**	**109.8**	**0.0**	**9.4E-01**
1D-4	5	6,500	18.5	116.5	6.7	3.3E-02
1D-5	6	5,800	17.8	119.5	9.7	7.3E-03
1D-5$_f$	6	5,400	17.7	118.3	8.5	1.4E-02
1D-6	7	6,500	18.8	127.4	17.7	1.4E-04
1D-7	8	5,300	18.0	131.6	21.8	1.8E-05
1D-8	9	6,600	19.3	145.0	35.2	2.1E-08
1D-10	11	6,500	19.6	175.9	66.1	4.1E-15
Toth	2	38,900	47.9	133.8	24.0	5.8E-06
MOD-1	3	55,000	58.6	145.2	35.5	1.9E-08
MOD-2	4	54,500	58.3	148.6	38.8	3.6E-09
MOD-3	5	53,300	57.6	152.3	42.5	5.5E-10
MOD-4	6	56,700	59.5	158.3	48.5	2.7E-11
MOD-5	6	53,200	57.4	163.3	53.5	2.3E-12
MOD-6	7	51,300	56.6	162.6	52.9	3.1E-12
MOD-7	8	57,600	59.8	172.2	62.4	2.7E-14
MOD-8	9	52,000	57.0	180.1	70.4	5.0E-16

3.11.2 Problems with Applying MMRI to a Transient Data Set

Considering the "point-in-time" data comprising sets one and two, developing a set of transient models for calibration and ranking with the third data set proved somewhat difficult. For example, the third data set contains both static (long-term averaged) and transient data. Combining the first two data sets (and the corresponding models) with the third data set is not as simple as assuming that the "point-in-time" head data are actually static data (although this was done for verification purposes). The models constructed for data sets one and two are steady state and converting them to time-dependent models was made by the addition of a few transient-related parameters (e.g., S_y) while maintaining the original parameterization scheme. This generalization could also be applied to the Toth model by not applying the steady-state assumption to the groundwater flow equation, but maintaining the linear (or oscillating) water-table boundary condition. It is unlikely that this approach has an analytical solution, but it could be solved numerically.

Using these hypothetical transient models, calibrated models from the first data set can be introduced as steady-state approximations to the real conceptual models under consideration. Because the initial data represent a snapshot ("point-in-time"), the steady-state approximation is applied to these models to calibrate and rank them. Once time-dependent observations are available, the transient versions of the models are used. Although this approach was not followed in this exercise, the idea of iteratively calibrating and developing models in the context of information criteria and evolving (transient) data is one of the novel aspects of this work. Keep in mind that computational expense may limit the feasibility of calibrating a set of three-dimensional transient models and reduction from three to two dimensions may be needed to spare computational expense. If such a computational restriction exists, it would be necessary to relate the three-dimensional models from previous data sets to two-dimensional transient representations of the same models. All these areas represent significant areas of future research.

3.12 Summary of Phase I

This section presented the development and iterative calibration and ranking of three types of models: a series of one-dimensional Darcy flow models, a two-dimensional Toth model, and three-dimensional MODFLOW models. The first data set contained 7 head observations that represented the sparseness of data one would in during a preliminary site investigation. Calibration parameters for the Darcy and MODFLOW models were mostly restricted to the transmissivity and hydraulic conductivity, respectively. As a reflection of the size of this data set, calibration efforts favored the simpler models, especially the Darcy model 1D-5, which had an RMSE of less than 3 m. The MODFLOW models did not attain satisfactory calibration results or model rankings, while the Toth model performed reasonably well, especially given that it only has a single parameter for calibration, specifically, the slope of the linear water table. Although somewhat incomplete because *AICc* was incalculable for all models, the corresponding relative weights favored a multi-model average.

The second data set provided 19 head observations for calibration, allowing for the inclusion of additional calibration parameters. For the Darcy models, the maximum number of parameters increased from 5 to 10 (the number of transmissivity zones), and the maximum number of parameters for the MODFLOW models increased from 7 to 8. Calibration with this data set increased the average RMSE of the Darcy and Toth models by approximately 3 m and 2 m, respectively, while the average MODFLOW RMSE decreased by approximately 10 m.

Investigating the predictive ability of the models calibrated with the first data set against the new data in the second data set identified the MODFLOW models as the most stable on average with respect to variations in RMSE from calibration to prediction. In addition to model weights, predictive ability should also be considered when deciding which models advance to the calibration and ranking iteration, especially when these models are developed for predictive purposes.

Based on the performance of the MODFLOW model from the first to second data sets, increasing the number of observations and the complexity of the parameterization scheme will produce superior results. Similarly, allowing the Toth model water table to vary in an oscillatory manner will enable the model to represent fluctuations in the water table of the HS and increase the number of calibration parameters from 1 to 3. Despite its simplicity, the Toth model always outperformed the MODFLOW models (and some of the Darcy models) in terms of RMSE and *AICc*, and could be included in future iterations if possible.

4 Phase II

Based on the lessons learned from Phase I, the tasks of Phase II were modified from the original statement of work to examine the variability and degree of usefulness of using the *AICc* for multi-model ranking and inference across successive data collection efforts. Phase II began by creating a ground-truthing (GT) model that was based on the latest version of the JAEA Horonobe model [*Ota et al.*, 2007]. Several different conceptual models were created along with their respective numerical models. Sample data from the GT model of the Horonobe site were then 'collected' and the conceptual models were calibrated to those data. The changes from the statement of work concerned how the simulations were analyzed such that the analysis now looks at:

1. The use of both the WSSR and the *AICc* in evaluating the appropriateness of a model
2. The variability of multi-model averaged parameter values and predictions across different randomly sampled sets
3. The magnitude and importance of each of the bias terms and how those bias terms can inform model selection

Items 1-3 were applied to develop lessons and a methodology for application in the field.

4.1 Development of Ground-Truthing Model

The GT model was based on the latest version of the Horonobe model as described in [*Ota et al.*, 2007] (Figure 16). The model, and input and output files were supplied to the Phase II team for the 'Case 0' or base-case simulation. The original model was created using DTransu-3D-EL, which is a mass transport code developed by JAEA that simulates discrete or single continuum media, unsaturated liquid flow, thermal vapor diffusion, and advection-diffusion for transport. The JAEA model was ported to FEMWATER so that the model could interface with the Groundwater Modeling System (GMS) [*GMS*, 2008], which made the development of the conceptual models an easier task (see below). While care was taken to maintain the site specific features of the JAEA model, reproducing the JAEA model exactly was not a priority.

The 78,792 node finite-element grid used in the JAEA model was able to be directly ported to FEMWATER so that the original mesh and stratigraphy were maintained (Figure 17). A coordinate change was made by rotating the model grid -20.87^o around node #1 of the original mesh, and then translating it by $X = X_0 + 2257.24$m and $Y = Y_0 + 5893.89$m, where X_0 and Y_0 are the (X,Y) coordinates from the original mesh. This translation provided a convenient reference such that the lower left hand corner of the finite-difference grid of the conceptual models could be given an (X,Y) coordinate of (0,0).

The lateral and bottom boundaries of the JAEA model were maintained, with no-flow on the north, south, east, and bottom boundaries and constant head on the western boundary.

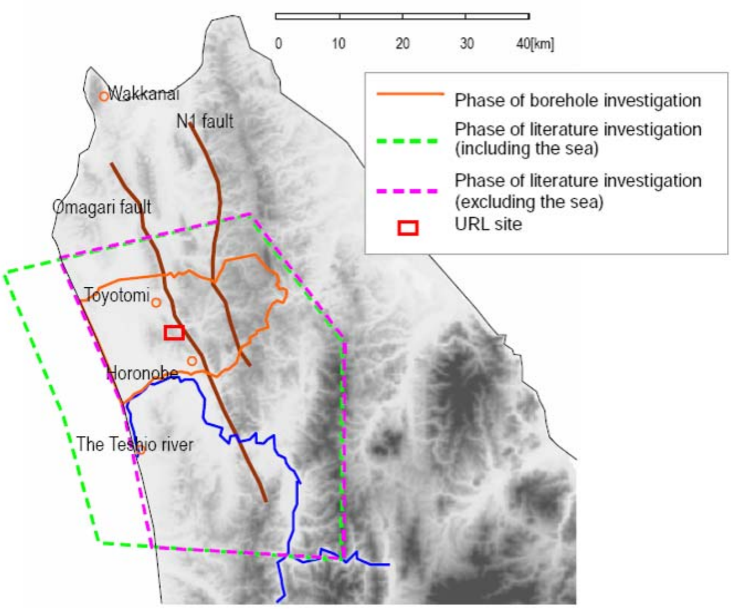

0
10
20
30
40[km]
Phase of borehole investigation
Phase of literature investigation (including the sea)
Phase of literature investigation (excluding the sea)
URL site
Wakkanai
N1 fault
Omagari fault
Toyotomi
Horonobe
The Teshio river

Figure 16 - Site domain for JAEA Horonobe model.

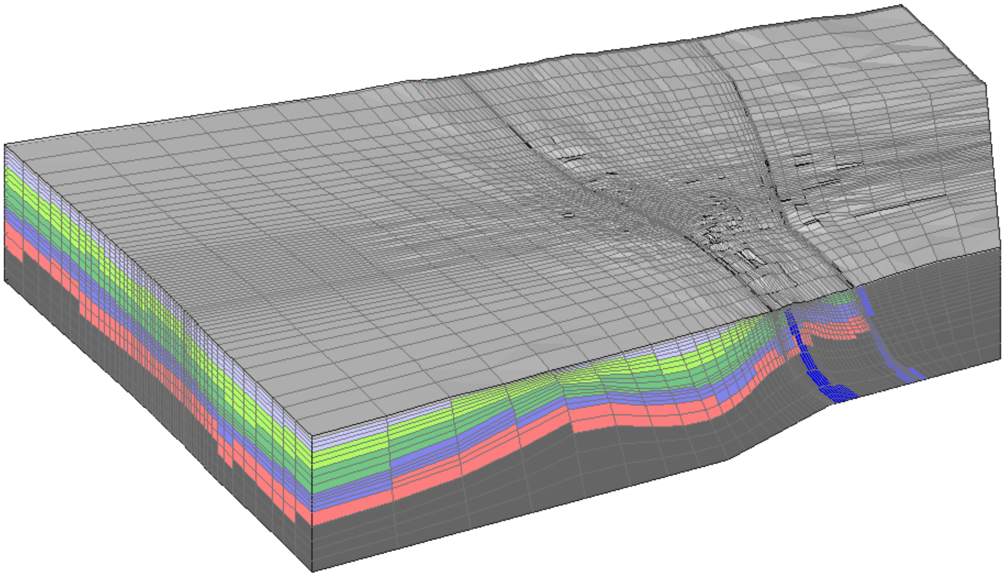


Figure 17 - Horonobe model finite element mesh.

To allow for faster convergence of the conceptual models during calibration, the top boundary was modified from the non-linear saturated/unsaturated boundary with variable recharge and discharge to a linear prescribed recharge boundary. To establish the spatial distribution and magnitude of the prescribed recharge, the spatial distribution of recharge and discharge from the JAEA model were plotted onto the top layer of the mesh (Figure 18). Discharge areas (colored blue in Figure 18) were given a value of zero recharge while recharge areas (colored red in Figure 18) were scaled appropriately to maintain the net total recharge from the JAEA model.

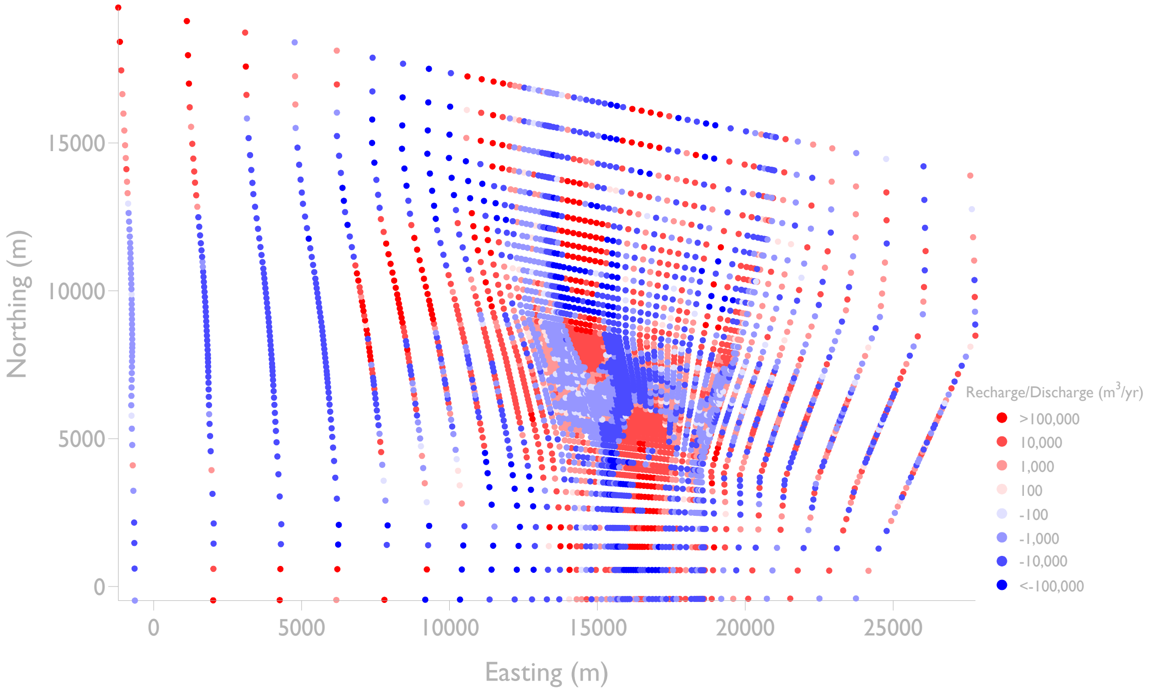


Figure 18 - Recharge (red) and discharge (blue) areas as simulated by the JAEA Horonobe model.

The FEMWATER model was simultaneously calibrated to the same borehole data for which the JAEA model had been calibrated (Table 22). The FEMWATER model was calibrated using PEST [*Doherty*, 2007] by changing a recharge multiplier as well as the hydraulic conductivity for each geologic unit. Head values were interpolated from the finite element grid to the actual borehole location using tri-linear interpolation. Where applicable (see below), the allowable range of hydraulic conductivities used during calibration was ± 150% of those given in Table 4.2.2-1 of Ota et al. [2007]. This was necessary to account for any differences that might occur due to the different recharge boundary conditions. The FEMWATER model was able to mimic the JAEA model very well while some bias was present against the borehole data (Figure 19). The calibrated FEMWATER model is what served as the GT model for Phase II.

Table 22 - Coordinate data of the HDB wells used in the ground truthing model calibration.

Borehole	North latitude			East longitude			UTM (zone54)		Elevation	Borehole length	FEMWATER Model Coordinates	
	Degree	Minute	Second	Degree	Minute	Second	X	Y	(m)	(m)	X (m)	Y (m)
HDB-1	45	02	24.03	141	51	52.83	4987758.42	568102.26	69.10	720	9295.45	4287.95
HDB-2	44	59	48.38	141	55	10.38	4983002.94	572478.75	42.53	720	13671.94	-467.52
HDB-3	45	02	39.50	141	51	26.64	4988229.85	567524.24	58.19	520	8717.43	4759.39
HDB-4	45	03	16.35	141	52	30.29	4989381.79	568904.23	63.61	520	10097.42	5911.32
HDB-5	45	02	57.75	141	52	47.09	4988811.81	569278.06	78.77	520	10471.25	5341.34
HDB-6	45	02	39.09	141	51	38.64	4988219.99	567786.85	60.21	620	8980.05	4749.52
HDB-7	45	02	51.30	141	50	39.32	4988583.10	566485.23	43.75	520	7678.42	5112.63
HDB-8	45	03	00.05	141	52	09.31	4988873.70	568450.84	70.05	470	9644.03	5403.23
HDB-9	45	03	36.20	141	51	00.45	4989973.23	566932.89	97.19	520	8126.09	6502.76
HDB-10	45	03	31.92	141	53	38.11	4989878.52	570382.40	50.83	550	11575.59	6408.05
HDB-11	45	02	08.72	141	52	09.10	4987289.75	568463.27	66.85	1020	9656.46	3819.29

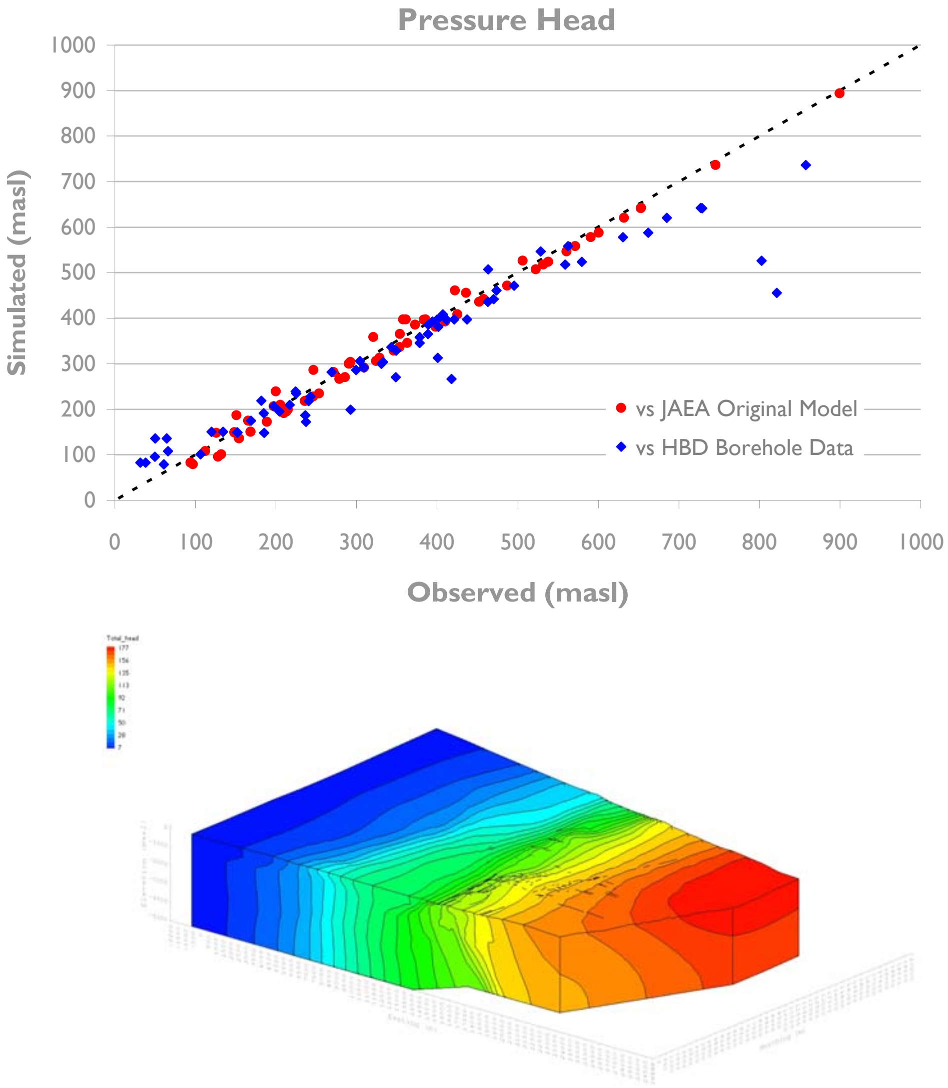


Figure 19 - Calibration results of the FEMWATER model against the JAEA Horonobe model and the HBD borehole data (top) and the resulting head distribution (bottom).

4.2 Conceptual Models

To create the conceptual models, a series of seven separate conceptualizations were formed and then modeled using MODFLOW. A common MODFLOW grid was used for all seven conceptual models that was comprised of 48x33x52 finite difference cells of 600 x 600 x 100m dimensions (Figure 20).

The seven different conceptual models were constructed by varying the number of unique hydraulic conductivity zones and allowing or dis-allowing anisotropy. Each of the models are described in Table 23 and illustrated in Figure 21 (Eight conceptual models were originally developed but model 5 was lost to a hard-drive malfunction. After analyzing the remaining 7 models, it was determined that recreating Model 5 would not benefit the analysis).

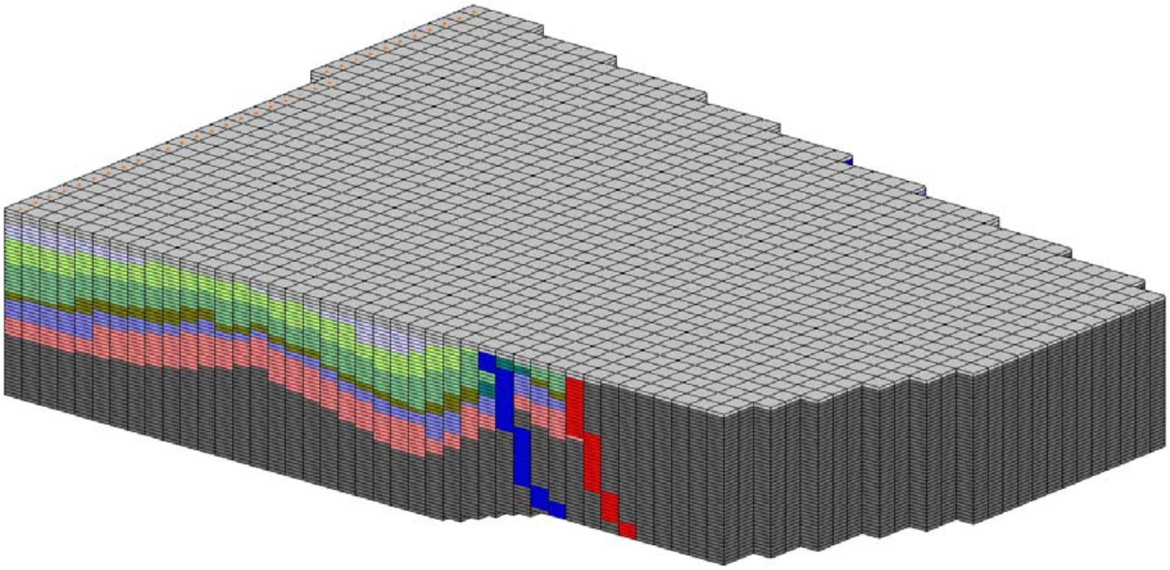
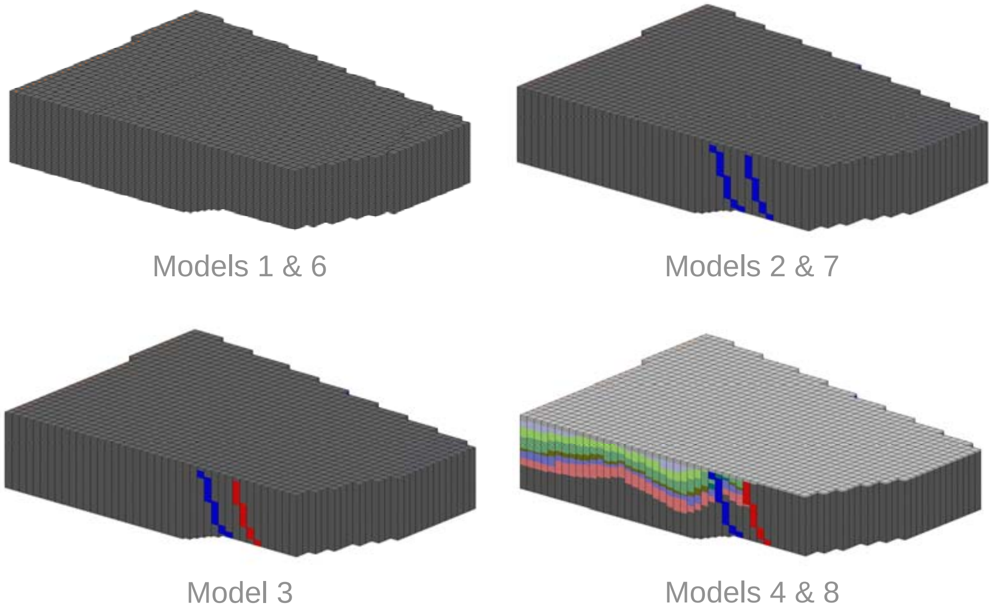


Figure 20 - MODFLOW grid for the conceptual models.

Table 23 - MODFLOW conceptual models for the Horonobe site.

Model #	Description	# of Parameters	Comments
1	Homogeneous	2	Entire domain treated as a single, homogeneous unit.
2	Faults / non-faults	3	Domain split into 2 K-zones; areas that represent faults and areas that don't.
3	2 Faults / non-faults	4	Domain split into 3 K-zone; one zone for each fault, and one for non-fault areas.
4	11 materials	13	Each of the 11 geologic units has a unique K value.
6	Model1 + α_{xz}	3	Same as Model 1 but with vertical anisotropy.
7	Model 2 + α_{xz}	5	Same as Model 2 but with vertical anisotropy for each material.
8	Model 4 + α_{xz}	22	Same as Model 4 but with vertical anisotropy for each material.



Models 1 & 6

Models 2 & 7

Model 3

Models 4 & 8

Figure 21 - MODFLOW conceptual models for the Horonobe site.

4.3 Generating Observation Data

Ten iterations of observation data are sampled from the FEMWATER GT model. Each successive iteration doubles the number of data points from the previous iteration by adding the same number of randomly sampled points to the previous iteration. In this way, ‘old’ observation points are preserved as would occur during a site investigation where data and observations are accumulated with successive field studies over time. Each iteration contains 2^k observation points, where k = 3..10 is the iteration number. Thus, iteration #3 has 8 data points while iteration #10 has 1024 data points (iterations 1 and 2 did not contain enough data points to allow for a multi-model comparison). One hundred data sets were generated with a data set defined as a single ensemble of iterations 3..10. The observation points are randomly selected within a rectangular region of the model domain as shown in Figure 22, which plots the data points of the first 5 data sets for iterations 2 through 10 (iteration 2 is included for illustrative purposes only but is not used in any of the analyses). All of the conceptual models described in the previous section were calibrated to each individual iteration / data set combination, which resulted in over 500,000 model runs. Due to the lengthy simulation time involved in calibrating its 22 parameters, Model #8 was calibrated to the first 25 data sets only.

4.4 Assessing Model Ranks

The first analysis of Phase II examines the behavior of the multi-model approach as a function of the data set. As mentioned above (see section 1.1), information-criteria cannot be compared for models that are calibrated to different sets of data, which introduces the risk of prematurely eliminating a model when it may in fact be an excellent performer as more data are collected. This condition was demonstrated in Phase I with the MODFLOW models where the models that were mathematically and structurally closest to the GT model did not rank high due to the amount of data that were available. The *AICc* values for these models were heavily influenced by the second order bias term, which means that the model complexity was too great for the amount of information contained in the data. Much can be learned from these models and thus they should not be eliminated in the early, data scarce, iterations. Thus, how can we know prior to collecting more data whether or not a model will rise in its *AICc* ranking? To answer this question, we examine how a model with limited amount of data compares to itself if there were an infinite amount of data.

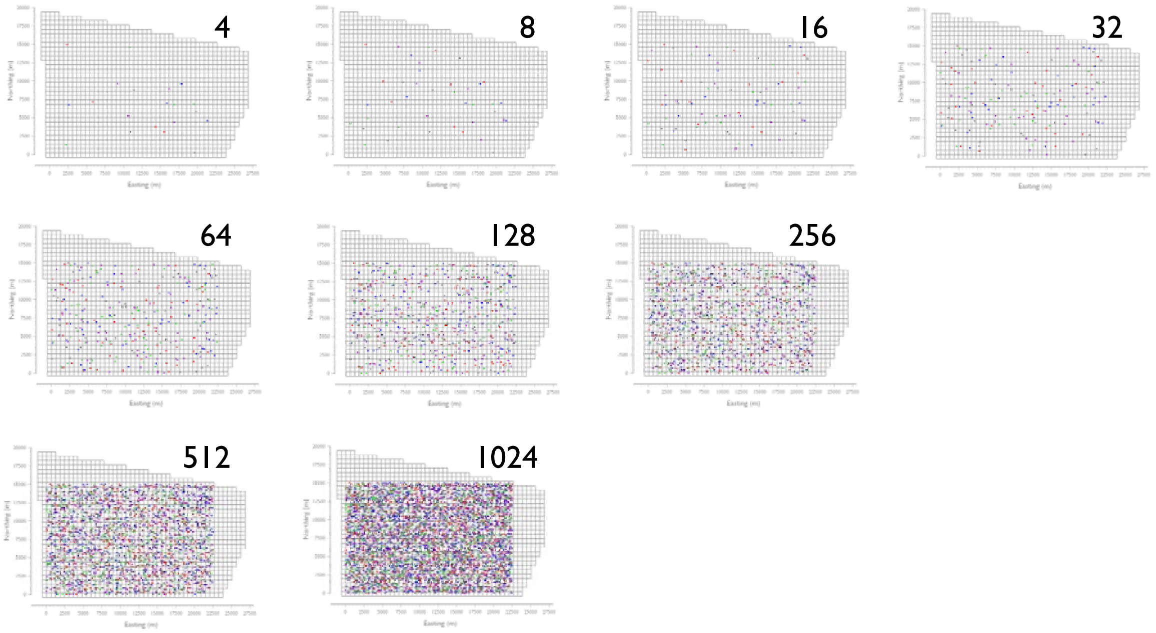


Figure 22 - The placement of the randomly selected data points of each iteration for the first 5 data sets. The point color differentiates the different data sets. The number to the upper right of each plot shows the number of data points per data set for each iteration.

One way to think of site data is to regard each successive collection of data as a separate realization of a random variable that is based on an unknown operating model [*Zucchini*, 2000]. The operating model in this case would be the real-world conditions at the Horonobe site that produced the measured or observed data. Given an initial data-set, several conceptual models are generated and then ranked. However, if the initial data set is discarded and another data set of the same size is sampled and the rankings are recalculated using the same models, chance exists for the rankings to change. If rankings are highly variable as a function of the data set, then a conservative approach should be taken when deciding to eliminate a model. Conversely, if the rankings are stable, then one can assume that the rankings calculated using a single data set are indicative of the ensemble ranking across all data sets.

We have given this approach the name of ‘auto-ranking’ to describe the process of comparing a conceptual model against itself across different sets of data. In order for this approach to be valid, we cannot use an information criteria comparison. However, we can use the analogous metric of discrepancy [*Zucchini*, 2000], or more specifically, the *discrepancy due to approximation* (DDA), the *discrepancy due to estimation* (DDE), and the *overall discrepancy* (OD). The DDA is defined as sum of the squared residuals between the GT model and the best approximating model where the best approximating model is the calibrated model if an infinite number of data were available. Mathematically, this is represented as:

$$\Delta\left(GT, g_{\theta_0}^{i,\infty}\right) = \int_1^N \left(h^{GT} - h^{i,\infty}\right)^2 dx \qquad \textbf{(21)}$$

where h^{GT} is the head from the GT model, and $h^{i,\infty}$ is the head predicted by model *i* after calibration to an infinite number of data points, *N*. The term $g_{\theta_0}^{i,\infty}$ represents the best approximating model *i* using the set of calibrated parameters θ_0.

The DDE is defined as the discrepancy between $g_{\theta_0}^{i,\infty}$ and $g_{\hat{\theta}}^{i,k}$, where $g_{\hat{\theta}}^{i,k}$ is the calibrated model *i* based on the data from iteration *k* (*k* = 3..10). The term, $\hat{\theta}$, represents the set of calibrated parameters for model *i* and dataset *k*. This is represented as:

$$\Delta\left(g_{\theta_0}^{i,\infty}, g_{\hat{\theta}}^{i,k}\right) = \int_1^{2^k} \left(h^{i,\infty} - h^{i,k}\right)^2 dx \qquad \textbf{(22)}$$

Examination of equation (21) shows that the DDA is a function of the operating model (i.e. the GT model in this case) and not of the data. The DDE on the other hand is a function of the underlying data set. In descriptive terms, the DDA is the minimum discrepancy possible for each model while the DDE is the discrepancy that arises from the lack of information that is a result of the model being calibrated to a dataset that contains less than *N* data points. The OD is simply the sum of the DDA and the DDE.

Relating this to information criteria, the *AICc* can be thought of as a way to estimate the *expected* OD (EOD), where the EOD is defined as the average OD for a model over all data set combinations of a given size [*Zucchini*, 2000]. In simple terms, the *AICc* cannot be compared across different data sets since it contains in its genesis terms that describe the DDA, which are conceptual and operation model specific. In order to compare a model to itself across different data sets, we must compare the DDE values and this cannot be done without knowing the DDA. While there is no real way to calculate this directly (at least not without calibrating to an infinite number of data points) we can estimate it by plotting the OD for each model as a function of the number of data, fitting a curve to that plot, and then extrapolating that curve to infinite.

Since a cell can only be calibrated to one head value at a time, we define *N* as the number of active, variable head cells in the MODFLOW finite difference grid (60664 cells). In reality, other data types, such as recharge, groundwater flux rates, and the like should be included in the infinite data set but since this exercise is using only observed head values as the calibration metric we limit *N* to the highest number of observed heads possible. The OD is calculated as:

$$OD = \frac{\sum_{1}^{2^k} \left(h^{GT} - h^{i,k}\right)^2}{2^k} \qquad \textbf{(23)}$$

which is simply the mean of the sum of the squared residuals (MSSR). Figure 23 shows the results of this exercise. In all cases, the OD virtually matches the DDA from the seventh iteration (128 data points) onward. In this case, the OD tends to *increase* as more data are added to the calibration process. This indicates that the conceptual models are incapable of matching a key element of the GT model, which is reflected by a spatial bias in the fits and thus a worsening of the *average* OD as more data points are added (Figure 24). The more simple models (1 and 6) reflect this trend more greatly than the complex models. The two most complex models (4 and 8) have the lowest OD's as well as the flattest trend, meaning that from the first data set, the more complicated models are performing closer to their DDA (i.e. their theoretical potential) than the simple models. If a model contained the necessary features and processes to accurately reproduce the observational data, than the DDA should *decline* as more data are added to the calibration process.

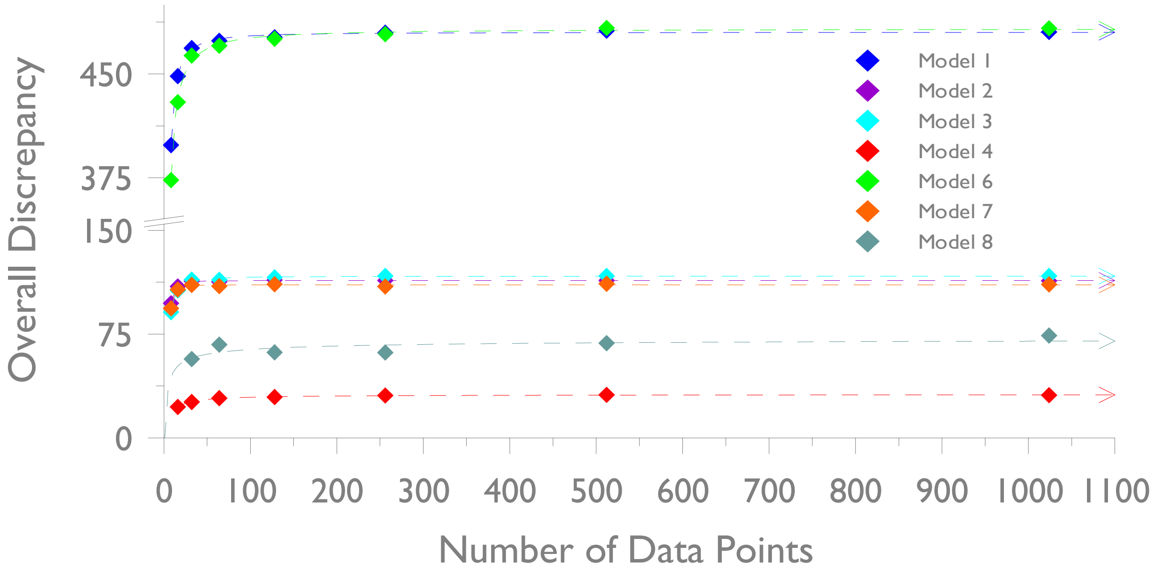


Figure 23 - The overall discrepancy and the fitted models for extrapolation to estimate the DDA (dotted lines). The fitted models are all of the form OD = A * n^{B} + C, where n is the number of calibration data, and A, B, and C are model parameters. The DDA is calculated by setting n = 60664.

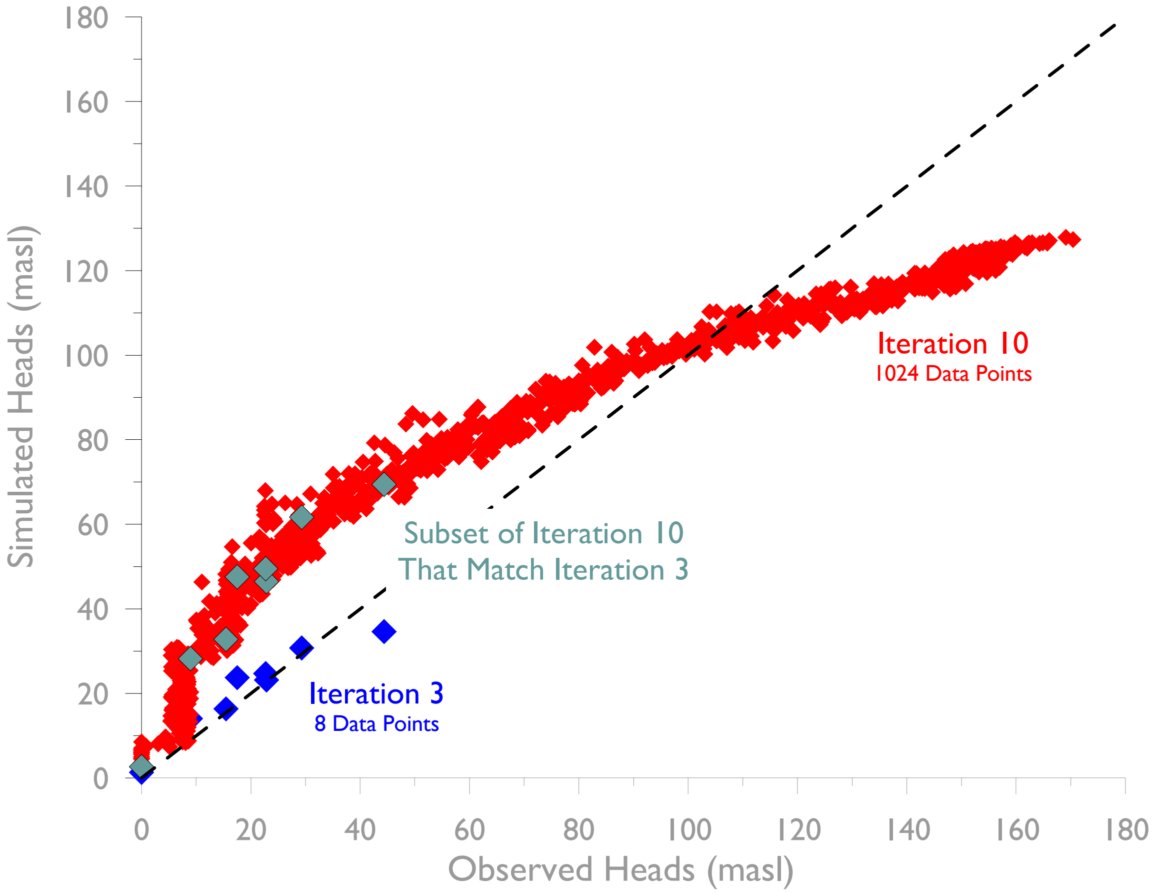


Figure 24 - As more data are added to the calibration process, the model bias becomes more evident (red diamonds versus blue diamonds). The green diamonds represent the 8 calibration points from Iteration #3 and the fit that resulted from using all 1024 data points. In this case, the OD for Iteration #3 is less than for Iteration #10.

Subtracting the DDA from the OD results in the DDE (Figure 25). Recall that this represents each models performance as a function of the amount of information contained in the data. With the exception of model 8, all the models are very close to their DDA values by the 9th or 10th iteration. The fact that model 8 cannot approach an asymptotic value after calibrating to 1028 data points is an indication that the type of data (i.e. head data) probably do not contain the right type of information for model 8 to reach its DDA. For utilization in the field, this type of situation could point to a need to try and inform the model with additional data of a different type (as opposed to gathering more of the same type of data).

From these results we conclude that *no model be eliminated in the early iterations* but rather that the early iterations be used to examine the DDE of each model. Coupled with an informational criteria ranking such as the *AICc*, this will supply the characterization team with a method for examining the relative strengths and weaknesses of each model. Thus, *a model should only be eliminated if its DDE is close to zero and it also ranks poorly using the AICc*. This combined analysis will help with the premature elimination of a model based solely on its *AICc* ranking.

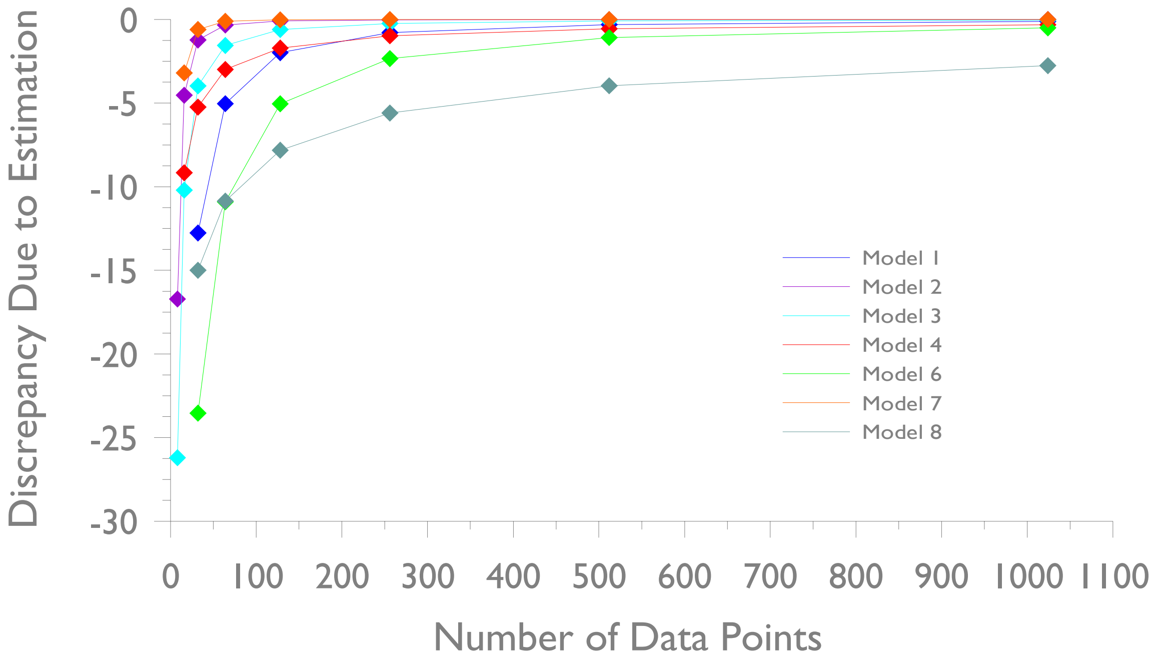


Figure 25 - The discrepancy due to estimation (DDE) for each model.

4.5 Multi-Model Averaging

4.5.1 Bounding Parameter Values

One of the primary goals for nuclear waste site characterization is to identify the key features and processes that will be important to assessing the performance of the site as well as to estimate parameter values associated with those processes. It has been stated that the use of multi-model averaged parameters in a specific conceptual model is rarely useful for groundwater problems since the averaged parameter values do not account for specific processes associated with a specific model, making use within that model inappropriate [*Poeter and Anderson*, 2005]. However, averaging parameters and then calculating the variance and confidence intervals can be useful for bounding parameter values at the site.

To illustrate this, the multi-model effective recharge is calculated using a 'bootstrap' approach. The bootstrap approach calculates *R* multi-model averaged parameter values for a single iteration and dataset combination by successively omitting one model from the calculation. When a model is eliminated from the group, the weights are recalculated to sum to 1. Effective recharge is used because it is the only parameter that is common amongst all seven conceptual models. For this example, we repeat the process 10 times using 10 randomly chosen data sets picked from the set of 100 (data sets 8, 16, 27, 36, 48, 65, 67, 68, 73, 97) to ensure that the result is not a function of a single data set. Figure 26 plots the multi-model average values and the ±95% confidence interval (the confidence interval is clipped at zero) for data set 16, which is representative of the 9 other datasets. The dotted blue line that is above the other 95% upper confidence interval is the result when Model 4 is eliminated from the calculation. In other words, the uncertainty in our parameter estimation becomes much greater when Model 4 is not used. This is because

Model 4 has a high Akaike weight and thus when it is included, the recharge value is heavily weighted towards the Model 4 value and the corresponding variance is low.

At a higher level, Figure 26 shows that a third type of uncertainty exists in the modeling process, which *is uncertainty in choosing the set of conceptual models*. If after a couple of iterations of data collection and modeling, none of the parameter estimations from the conceptual models agree well with data or expert opinion, then, the bootstrap method should be used to examine the third type of uncertainty by extending the ±95% confidence intervals to be the maximum and minimum values calculated for each iteration.

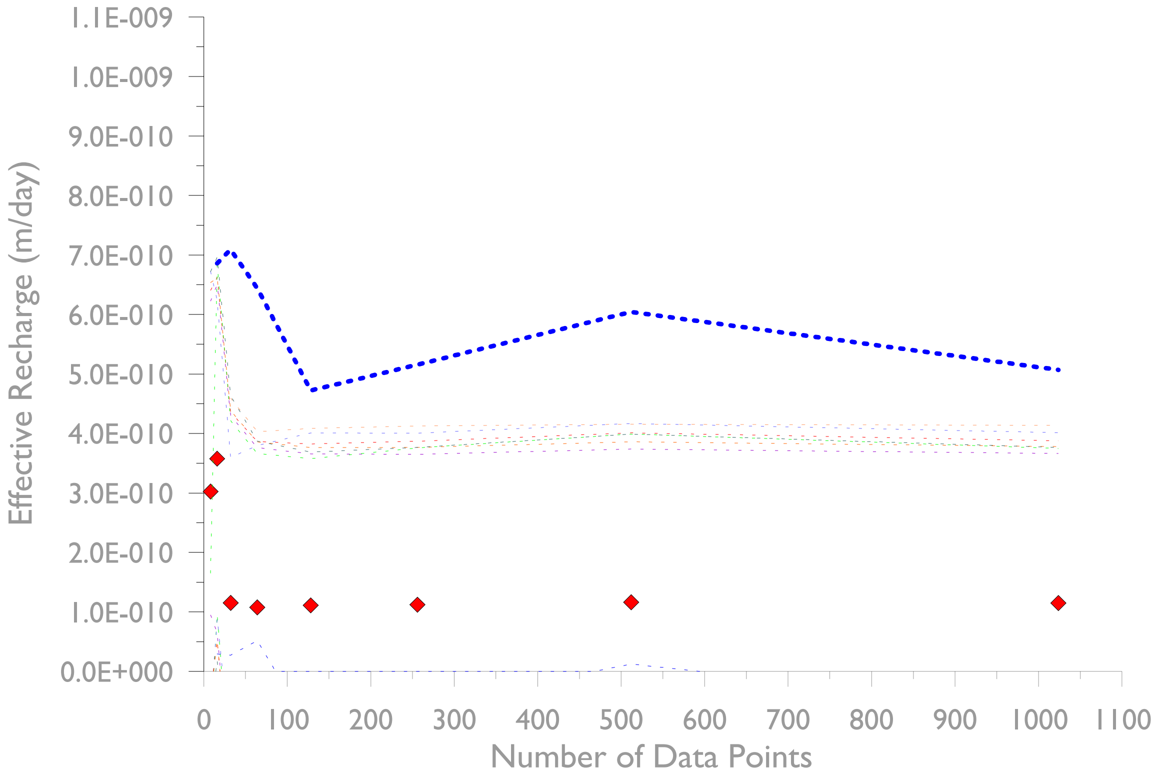


Figure 26 - The multi-model average effective recharge (diamonds) and the ±95% confidence intervals calculated using the bootstrap method for dataset 16. The heavy dotted blue line is the upper bounds when Model 4 is eliminated from the calculation.

4.5.2 Trends in Multi-Model Parameter Values

By comparing a parameter value that is calculated by a single model over successive data sets, insight into the relevance of that parameter can be gained. This is done through comparison of the multi-model average of a single model against itself by calculating its *AICc* ranking for each iteration across all data sets. In order for this to be valid, an assumption is made that if the model is close to its DDA, than the data-set specific constants that are inherent in the *AICc* (and which prevents direct comparison of the *AICc* across different data sets) should be similar enough to allow for multi-model averaging across all data sets.

Figure 27 shows a plot of the multi-model averaged parameter values and the arithmetic averaged parameters along with the 95% confidence interval for model #1. Model #1 is within 99% of its DDA for iterations 5-10. Predictably, iteration #5 is also the point that we see the arithmetic average coincide with the average calculated using equation (9). For the lower iterations where data are more sparse, the arithmetic average is lower than the multi-model average and the confidence interval is narrower.

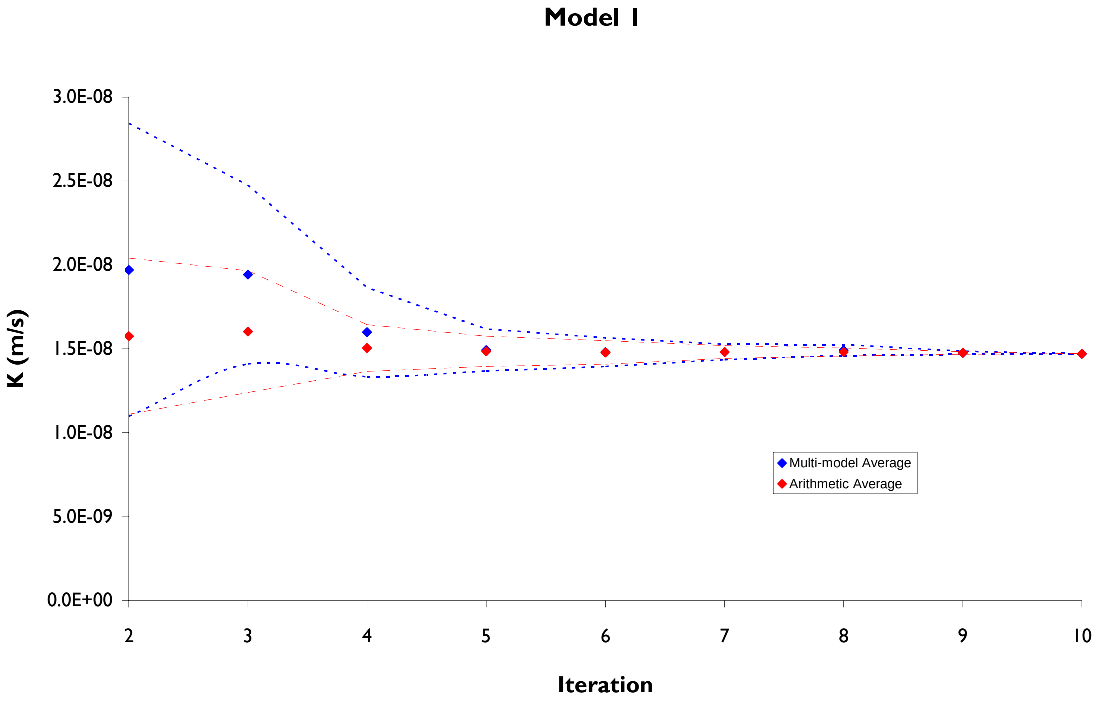


Figure 27 - Multi-model average, arithmetic average, and their corresponding 95% confidence intervals.

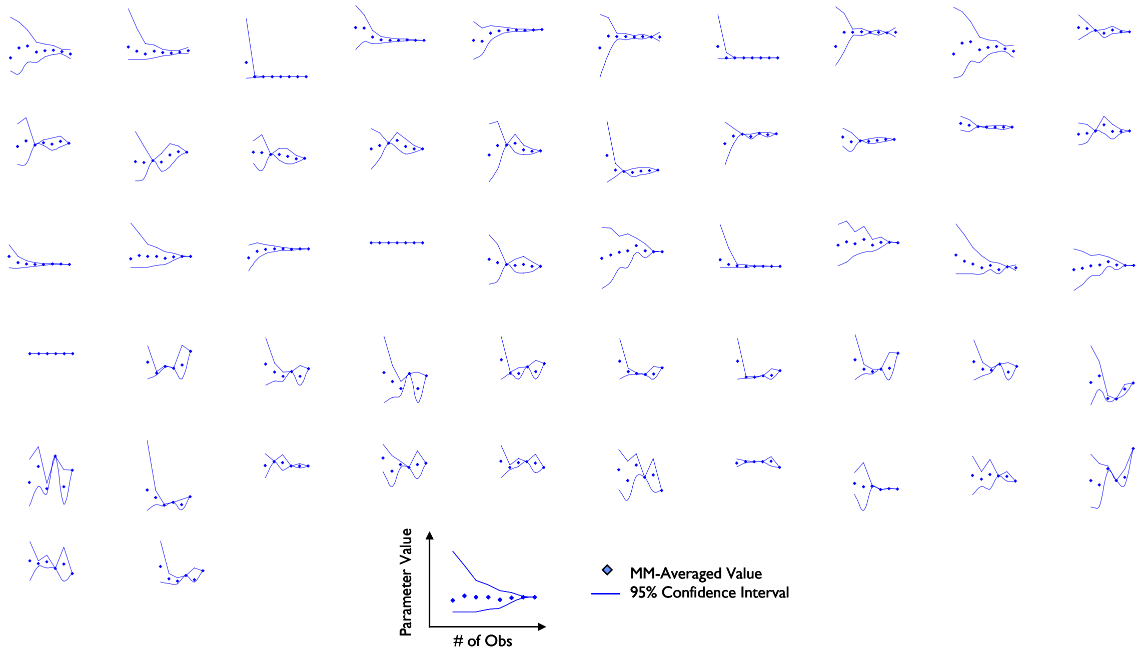


Figure 28 - Multi-model averaged parameters for each conceptual model showing both the parameter value and the 95% confidence interval as a function of the number of data points.

Continuing this analysis across all models, data sets, and parameters, information about the relevance of a particular parameter can be gained. Relevance is defined as a parameters importance with respect to the available data. If a parameter is relevant to a certain kind of data, then obtaining more of those data will help narrow the uncertainty of that parameter. If it is irrelevant, then more of the same data will not reduce that parameters uncertainty.

Figure 28 shows a plot of the multi-model averaged parameters for each of the models across successive iterations and shows how the trend in uncertainty changes with each iteration. Direct comparison of the uncertainties is not needed and thus the axes for each plot have been removed for clarity. Three trends in Figure 28 can be identified. The first is identified by a variance trend that is very large with small numbers of observations but quickly declines to almost zero as the number of observations increases (Figure 29a). Examination of the PEST calibration output files shows that trends of this type indentify parameters that have either very low sensitivity or very high sensitivity. If a parameter has low sensitivity, then it can be fixed for future calibrations. If it shows high sensitivity, then it may require different data to become relevant. The second trend shows either a random pattern of both the variance and parameter value or a straight line from start to finish (Figure 29b). These patterns suggest parameters that are ‘irrelevant’ to the simulation. If a parameter is shown to be irrelevant then the model(s) containing that parameter may be over-parameterized and could require different types of data to perform better. The final trend shows a smoothly converging confidence interval with increasing numbers of observations (Figure 29c) and indicates parameters that are relevant to the simulation.

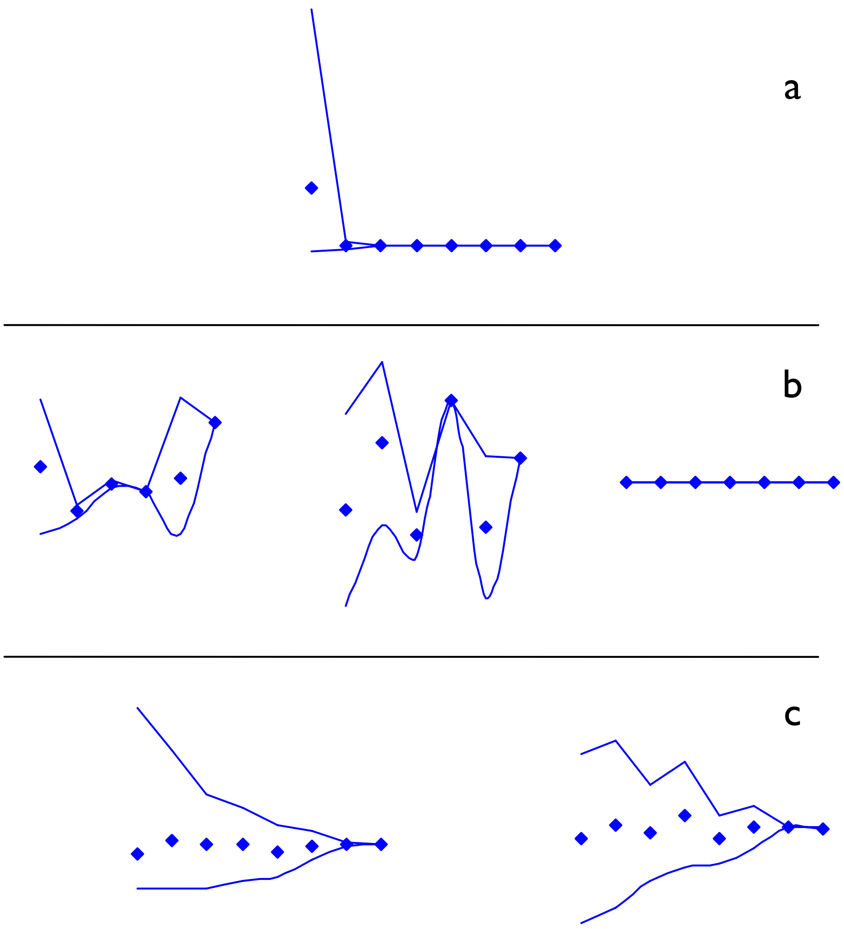


Figure 29 - The trending patterns for auto-ranked models.

For repository site characterization, plotting the parameter trends over successive efforts of data collection can be useful in determining which processes are important amongst the different conceptualizations. New conceptualizations can be formed from the trending information by combining relevant parameters into a single model. Models that contain mostly irrelevant parameters could be eliminated or saved until different types of data are available for which to calibrate to.

4.6 Bias Contribution

Information criteria rankings such as the *AICc* are different from minimizing errors between prediction and observation in that information criteria factors in the amount of information available from the data in relation to the complexity of the model. A simple calibration that relies only on minimizing the sum of the squared errors can result in the acceptance of a model that may fit the data very well but that does not represent the true conditions at the site (and thus by association is also a poor predictor of conditions at the site).

Figure 30 is a contour plot of the first term (i.e. the goodness of fit term) of the *AICc* equation (equation (5)) as a function of the weighted sum of squared residuals (WSSR) and the number of observations (n) for an arbitrary range of n and WSSR. The plot indicates that the goodness of fit term for the *AICc* has a relatively small contribution to the *AICc* when the number of observations are small. This makes sense in that the amount of information that is contained in the data is also small and thus the *AICc* should be dominated by the first and second order bias terms. As n increases, the sensitivity to the WSSR increases. For moderate values of n (i.e. halfway between 'low' and 'high' in Figure 30) there isn't much improvement if the WSSR is decreased from a very high value to a moderately high value. However, as the WSSR decreases to low values, the AICc becomes more and more sensitive to the WSSR.

The goodness of fit term and the second order bias term (third term of equation (5)) tend to offset each other and can indicate when additional model complexity should be added. The second order bias term decreases as the number of observations increase with relation to model complexity. Thus, if the reduction in the second order bias term is greater than the increase of the goodness of fit term, additional complexity can be justified. This is illustrated in Figure 31 which shows a plot of the averaged percentage

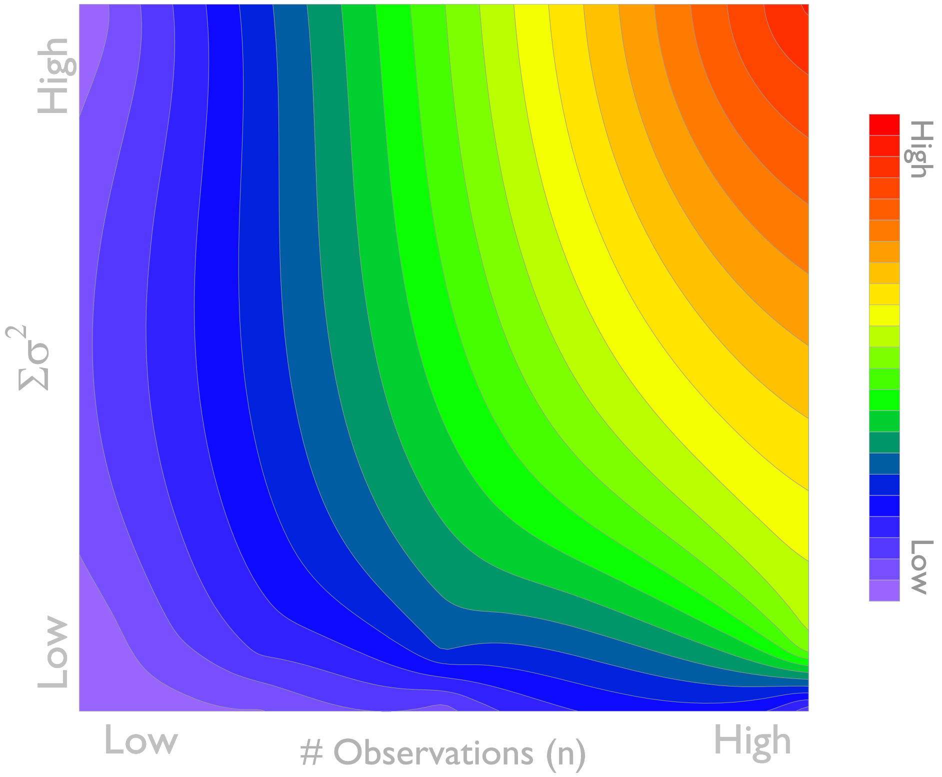


Figure 30 - The relative *AICc* value with respect to the weighted sum of the squared residuals ($\Sigma\sigma^2$) and the number of observations.

contribution to the *AICc* across all data sets of the 2^{nd} order bias term for each model as a function of the number of data. Each axis is plotted in $\log_{10}$ scale. Even for the simplest model (Model 1), the 2^{nd} order bias term contributes about 10% to the *AICc* value for iteration 3 (8 data points). For Model 4 in iteration 4 (16 data points), the 2^{nd} order bias is almost 85% of the *AICc* term. If the *AICc* criteria is blindly applied as an isolated metric, Model 4 would be omitted from consideration. However, if professional judgment indicates that Model 4 is representative of conditions at the site, and noting that the bias term represents a very large percentage of the *AICc* value, then Model 4 should be retained until the bias terms are minimal.

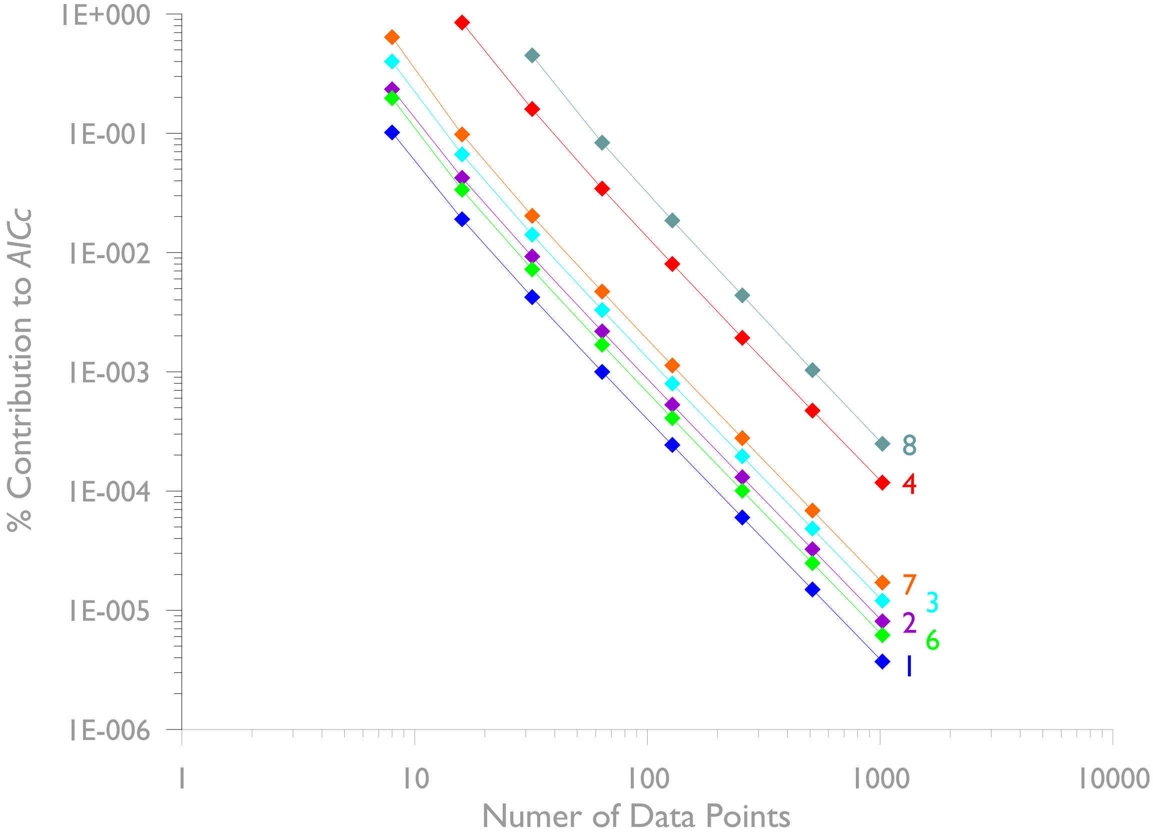


Figure 31 - The percent contribution of the second order bias term to the AICc value. The colored numbers are the conceptual model numbers.

5 Summary and Suggested Application

Nuclear repository site characterization is a unique process in that the data collection and analysis processes are inherently connected through an iterative process that can span years or even decades. Assumptions and conceptualizations that appear valid early on in this process may prove to be invalid as more data are collected. Risk of prematurely accepting or eliminating a particular assumption or conceptualization can lead to inadequate or inaccurate characterization, less defensibility, lost time, and increased costs. This project examined the application of information criteria (IC) to the repository site characterization process, in the form of the Akaike Information Criteria (*AICc*), to develop and identify procedures that could help increase the efficiency, defensibility, and accuracy of site characterization.

The application of the *AICc* is demonstrated here through a series of numerical experiments. The *AICc* provides a means for assessing both model and conceptual uncertainties, assisting to better understand site conditions as well as indentifying processes that are important to site characterization.

The concept of parsimony is introduced to describe the balance between a model's complexity and the information available to inform or calibrate the model. During hydrogeologic characterization, available information typically comprises surface observation data, well logs, pump tests, geophysical measurements, and the like. With regards to the *AICc*, model complexity refers to the number of unknown parameters that require calibration, which is achieved when the weighted sum of the squared residuals between model prediction and observations are minimized.

When using the *AICc* (or any other IC metric), a distinction between observational data and parameter data needs to be made. Observational data are those data that the model(s) are trying to simulate. For groundwater problems, this could include head level measurements, groundwater flux estimations, spring flow, solute concentration, or temperature. Parameter data on the other hand are those data that provide direct estimates of model parameters. Examples of parameter data are the results from aquifer pump tests that identify transmissivity and/or storativity, grain size analyses that identify porosity, seepage flux tests that help establish recharge rates, or geologic and/or geophysical investigations that derive stratigraphy and structure.

For Phase 1 of this study, the addition of parameter and observational data were balanced in that each successive iteration of data 'collection' provided data that informed the actual parameters used in some of the models as well as additional observational data for which to calibrate the models. This balanced approach is what usually occurs during a site characterization process. Phase 2 on the other hand concentrated on the addition of observational data only. This was done to provide insight into how the *AICc* behaved with regards to the number of available data as well as the magnitude and contribution of second degree bias (i.e. the bias that 'penalizes' a model for being more complex). The result of these two analyses has resulted in the recommendations listed below for using the *AICc* in the site characterization process. It should be noted that while this study used the *AICc* as the IC metric, these procedures should be equally valid for any IC metric.

As a procedural method, the *AICc* can be realized during site characterization through implementation of the following steps:

1. Initial site investigation that includes collection of available data, literature review, and consultation with local experts.
2. Multiple conceptual model formulation.
    a. A conceptual model is defined as a model that has a unique theoretical design of the hydrologic and geologic processes at the site that is built through expert analysis of the existing data.
    b. Separate conceptual models need not differ greatly and should be purposely constructed to investigate different approaches to modeling the important processes at the site. Multiple conceptualizations can be derived from deterministic arguments, such as alternative theories about depositional environments of the sediments, deformation of rocks that make up a groundwater system, or flow characteristics of faults. They also can be derived from stochastic arguments, such as generating multiple zonations using indicator kriging or pilot point distributions. The number of conceptualizations might range from a few to a few dozen.
    c. Using the approach of point ‘b’, classes of conceptual models should be created that represent sets of similar models with slightly different theoretic approaches. In this way, *many conceptualizations can be created with little extra cost beyond the creation of a single model*.
3. Numerical model construction and calibration. Each of the conceptual models is converted to a numerical model and calibrated to the available data.
4. Calculate the relevant information criteria metrics (*AICc*, Δ_i, w_i, $\hat{\bar{y}}$, and $\hat{\bar{\text{v}}}\text{ar}(\hat{\bar{y}})$), rank, % contribution of the second order bias for each model.
5. Calculate the overall discrepancy (OD), the discrepancy due to estimation (DDA), and the discrepancy due to approximation (DDE).
    a. The OD is the mean of the sum of the squared residuals
    b. The DDA is calculated by:
        i. Plotting the OD as a function of the number of data over successive iterations of data collection.
        ii. Fitting a regression model to the OD.
        iii. Calculating the DDA with the regression model by using a large value for the number of data.
    c. The DDE is the OD minus the DDA.
    d. The DDA and DDE can only be calculated after a sufficient number of data collection iterations have occurred to allow for a representative regression of the OD.
6. Examine the model rankings. Answers to the following questions should be sought with the *primary concern of identifying which conceptualizations (and ideally which processes within those conceptualizations) clearly explain the observations*:

        i. Which predictions are most sensitive to calibration parameters and which are not?
        ii. Are the better-ranked models connected by common processes? If so, which processes?
        iii. Are the lesser-ranked models connected by common processes? If so, which processes?
        iv. Are the better-ranked models highly or sparsely parameterized?
        v. What is the contribution of the second order bias to the *AICc* for each model?
        vi. Has a model reached its DDA (this can only be answered as more data are collected)?

7. Add or eliminate models based on the following guidelines:
    a. *<u>No model should be eliminated in the early iterations</u>* since poorly ranking models could perform better as additional data are collected and they can help to define bounds on system behavior.
    b. If a model ranks poorly with respect to its *AICc*, its contribution from the second order bias is large (>5%), and its OD is not close to its DDA, then it should not be eliminated (an OD that is close to the DDA implies that the addition of more data will not improve the models ability to simulate the observational data).
    c. If a complex model initially ranks poorly due to a high contribution from the second order bias term, then it is likely that that model is lacking some fundamental attribute. Conversely, if its *AICc* is dominated by second order bias, then that model should be brought forward until enough data are available to determine its merit.
    d. To add new models:
        i. Presence of model simulation bias, such as that illustrated in Figure 24, could indicate that a key process or level of complexity is absent from that model. Examination of models with similar bias may help identify the process responsible for the bias or (in the case for models that show no bias) processes that should be included in a newly formed model.
        ii. When no direct evidence exists for adding a conceptual model, it is recommended that new models be created by making slight changes to an existing conceptual model. For example, switching recharge from a spatially uniform value to a rate that is dependent on elevation represents a new conceptual model to the model suite.
        iii. If changes to a current conceptual model are to be used as a new conceptual model in the next iteration, the old conceptual model should be retained to allow for direct comparison to the new model.
8. Collect more data and repeat the process.

The above is purposefully designated as a set of guidelines rather than procedures and embodies an important lesson; ***<u>there is no substitute for professional judgment and experience when interpreting the various calibration and IC metrics</u>***.

It should be stressed that the intent of these guidelines is to extract the greatest amount of insight from the least amount of data. To this end, these guidelines provide suggested approaches towards coaxing out more understanding about the physical processes that are governing conditions at the site. Other strategies, such as the bootstrap method described in section 4.5.1, or the auto-ranking method described in 4.5.2, can be powerful tools for examining model behavior but may or may not have applicability towards gaining insight about the site. The usefulness of these other strategies will be dependent on the amount and types of data that are available, the 'distance from reality' of the suite of conceptual models, and the complexity at the physical site. Given multiple plausible models, it can be useful to report results from individual models, including statistics that reflect model fit and parsimony, and predictions and confidence intervals on predictions. In some circumstances, a more useful analysis can be achieved by including model-averaged predictions and confidence intervals that reflect the multiple models considered. Like the modeling process itself, each of these strategies should be implemented only after more simple approaches have been judged to be inadequate with respect to gaining insight about the site.

Ultimately, the 'final' model should show no dominant spatial or temporal bias in the weighted residuals, its estimated parameter values should be reasonable (e.g., in a groundwater model, material known to be gravel should a have higher hydraulic conductivity than material known to be silt), and its complexity great enough to adequately inform the site characterization team about either the sites suitability or the next phase of the site investigation process.

Finally, successive data collection efforts can be guided by examining common processes in low-ranking models that would benefit from better or more data. However, a small portion of the field investigation budget should be set aside for less targeted data collection efforts in an attempt to uncover previously unknown processes or features. Beyond that, if accepted reasonable alternative models yield substantially different predictions of interest and it is not possible to determine the better model, additional data collection should be directed to identify likely and unlikely predictions.

6 Future Directions

At its highest level, the overarching lesson from this research is that no single approach can meet all the needs for understanding and reducing uncertainty during the site characterization process. This conclusion points to the need to explore approaches that could be used in conjunction with an IC approach to provide complimentary insight and quantification of uncertainty. In an unpublished communication to the project team for this study, John Doherty, the author of PEST [*Doherty*, 2007], states,

> '….all that we can promise, and what we must strive for, are:
> 1. That we determine a parameter field that minimizes potential predictive error, and
> 2. We quantify that error."
> 3.

To that end, several different yet related directions could be pursued that would help extend and place into context this project.

The first direction would couple the IC approach with metaheuristic techniques such as the TABU search algorithm [*Zheng and Wang*, 1996], to provide a means for significantly streamlining the model calibration and site characterization process. TABU search uses a query-based search of the solution space to determine the behavior of an objective function (e.g. minimization of the RMSE) and then identifies a list of multiple, non-unique solutions (i.e. parameter sets) that produce similar reductions to the objective function. Solution spaces could be mapped and overlaid to identify potential new conceptualizations that could be added to the suite of conceptual models as well as to quantify areas of uncertainty.

A second direction is the exploration of regularised inversion techniques using highly parameterised systems to extract the greatest amount of information from the calibration dataset. The investigation would examine the distribution and density of pilot points during calibration as well as interpolation techniques that are used to fill in spatially distributed data: both of which are important factors in the calibration process and which can greatly impact model accuracy and predictive capabilities. Results from the high parameterized analyses could be used to inform simpler models as well as the IC process itself.

A third direction could research the creation of a systematic iterative routine to couple model calibrations and data acquisition. Methods of experimental design would be applied to calculate a response surface (in head or parameter space) to determine where regions of high uncertainty exist in each model. These response surfaces would guide the collection of the next data set.

Each of these ideas would directly build on the work from this project and are given in no particular priority. The next step would be to further develop each of these ideas to determine which one would provide the most benefit. The advantage in doing any future work is that the ground-truthing models that were created at great time and expense for this project, could be re-used. Links to the IC approach would be made by including parallel IC analysis with each idea to identify complimentary as well as redundant areas of application.

7 References

Akaike, H. (1973), Information theory as an extension of the maximum likelihood principle., paper presented at Second International Symposium on Information Theory, Budapest, Hungary.

Akaike, H. (1974), A new look at the statistical model identification, *IEEE Transactions on Automatic Control AC*, *19*, 716-723 pp.

Buckland, S. T., K. P. Burnham, and D. R. Anderson (1997), Model Selection: An Integral Part of Inference, *Biometrics*, *53*, 603-618 pp.

Burnham, K. P., and D. R. Anderson (2001), Kullback-Leibler information as a basis for strong inference in ecological studies, *Wildlife Research*, *28, 10.1071/WR99107*, 111-119 pp, doi:1035-3712/01/020111.

Burnham, K. P., and D. R. Anderson (2002), *Model Selection and Multi-Model Inference: A practical Information-Theoretic Approach*, 2nd ed., Springer-Verlag, New York.

Burnham, K. P., and D. R. Anderson (2004), Multimodel Inference, Understanding AIC and BIC in Model Selection, *Sociological Methods & Research*, *33*(2), 261-304 pp, doi:10.1177/0049124104268644.

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

Chamberlin, T. C. (1965), The method of multiple working hypotheses, *Science*, *148*, 754-759 pp.

Doherty, J. E. (2007), Manual for PEST: Model Independent Parameter Estimation, Watermark Numerical Computing, Brisbane, Australia.

Foglia, L., S. W. Mehl, M. C. Hill, P. Perona, and P. Burlando (2007), Testing Alternative Ground Water Models Using Cross-Validation and Other Methods, *Ground Water*, *45*(5), 627-641 pp, doi:10.1111/j.1745-6584.2007.00341.x.

GMS (2008), Brigham Young University - Environmental Modeling Research Laboratory, edited by EMS-I, www.ems-i.com.

Hannan, E. J., and B. G. Quinn (1979), The determination of the order of an autoregression, *Journal of the Royal Statistical Society Series*, *41*(1), 190-195 pp.

Harbaugh, A. W., E. R. Banta, M. C. Hill, and M. G. McDonald (2000), The U.S. geological survey modular ground-water model-user guide to modularization concepts and the ground-water flow process, *U.S. Geological Society Open-File Report*, *00-92*.

Hill, M. C., and C. R. Tiedeman (2007), *Effective Groundwater Model Calibration*, 475 pp., John Wiley & Sons, Hoboken, New Jersey.

Hurvich, C. M., and C.-L. Tsai (1989), Regression and time series model selection in small samples, *Biometrika*, *76*(2), 297-307 pp.

JNC (2000), Supporting Report 1: Geologic Environment in Japan, Japan Nuclear Cycle Development Institute (JNC).

Kashyab, R. L. (1982), Optimal choice of AR and MA parts in autoregressive moving average models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *4*(2), 99-104 pp.

Kullback, S., and R. A. Leibler (1951), On information and sufficiency, *Annals of Mathematical Statistics*, *22*, 79-86 pp.

Link, W. A., and R. J. Barker (2006), Model weights and the foundations of multimodel inference, *Ecology*, *87*(10), 2626-2635 pp.

McQuarrie, A. D. R., and C.-L. Tsai (1998), *Regression and Time Series Model Selection*, World Scientific Publishing Company, Singapore.

Neuman, S. P. (2003), Maximum likelihood Bayesian averaging of uncertain model predictions, *Stochastic Environmental Research and Risk Assessment*, *17*(5), 291-305 pp.

Neuman, S. P., and P. J. Wierenga (2003), A Comprehensive Strategy of Hydrogeologic Modeling and Uncertainty Analysis for Nuclear Facilities and Sites, 236 pp, Nuclear Regulatory Commission.

Ota, K., H. ABE, T. Yamaguchi, T. Kunimaru, E. Ishii, H. Kurikami, G. Tomura, K. Shibano, K. Hama, H. Matsui, T. Nizato, K. Takahashi, S. Niunoya, H. Ohara, K. Asamori, H. Morioka, H. Funaki, and N. Shigeta, Fukushima, Tatsuo (2007), Hornobe Underground Research Laboratory Project Synthesis of Phase I Investigations 2001 - 2005, 414 pp, Japan Atomic Energy Agency.

Poeter, E. P., and D. R. Anderson (2005), Multi-model Ranking and Inference in Ground-Water Modeling, *Ground Water*, *43*(4), 597-605 pp.

Poeter, E. P., and M. C. Hill (2008), MMA, A computer Code for Multi-Model Analysis, in *Techniques and Methods 6-E3*, edited by U. S. Geologic Survey, p. 113, Boulder, Colorado.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992), *Fortran Numerical Recipes*, 2 ed., Cambridge University Press, Cambridge.

Schwarz, G. (1978), Estimating the dimension of a model, *Annals of Statistics, 6*(2), 461-464 pp.

Toth, J. (1963), A Theoretical analysis of groundwater flow in small drainage basins, *Journal of Geophysical Research*, *68*(16), 4795-4812 pp.

van Genuchten, M. T., and W. J. Alves (1982), Analytical solutions of the one-dimensional convective-dispersive solute transport equation, 151 pp, U.S. Department of Agriculture.

Watermark Computing (2003), Groundwater Data Utilities, User's Manual, 228 pp, Queensland Department of Natural Resources, Corinda.

Watermark Computing (2004), PEST, Model-Independent Parameter Estimation User Manual, Watermark Numerical Computing, Brisbane, Australia.

Watermark Computing (2006), Addendum to the PEST Manual, Watermark Numerical Computing, Brisbane, Australia.

Ye, M., S. P. Neuman, and P. D. Meyer (2004), Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff., *Water Resour. Res.*, *40*(5:W05113), 1-19 pp.

Ye, M., P. D. Meyer, and S. P. Neuman (2008), On model selection criteria in multimodel analysis, *Water Resour. Res., 44, W03428*, 12 pp, doi:10.1029/2008WR006803.

Zheng, C., and P. Wang (1996), Parameter structure identification using tabu search and simulated annealing, *Advances in Water Resources*, *19*(4), 215-224 pp.

Zucchini, W. (2000), An Introduction to Model Selection, *Journal of Mathematical Psychology*, *44*, 41-61 pp, doi:10.1006/jmps.199.1276.

Zyvoloski, G. A., B. A. Robinson, and Z. V. Dash (1999), FEHM Application, SC-194.

Distribution

SANDIA INTERNAL:

4	MS 0735	T.S. Lowry, 06313
1	MS 9409	S.C. James, 08757
1	MS 0778	B.W. Arnold, 06781
1	MS 9159	G.A. Gray, 08964
1	MS 9409	M. Grace, 08757
1	MS 0751	S.A. McKenna, 06311
1	MS 9409	M. Ahlmann, 08757
1	MS 0899	Technical Library, 09536