WATER RESOURCES RESEARCH, VOL. 49, 536–553, doi:10.1002/wrcr.20064, 2013

Parameter estimation and predictive uncertainty in stochastic inverse modeling of groundwater flow: Comparing null-space Monte Carlo and multiple starting point methods

Hongkyu Yoon,[1] David B. Hart,[1] and Sean A. McKenna[1]

Received 24 August 2012; revised 21 November 2012; accepted 14 December 2012; published 29 January 2013.

[1] Given a highly parameterized groundwater model in which the conceptual model of the heterogeneity is stochastic, a set of inverse calibrations from multiple starting points (MSPs) provide an ensemble of calibrated parameters and follow-on transport predictions. However, the multiple calibrations are computationally expensive. A recently developed null-space Monte Carlo (NSMC) method combines the calibration solution-space parameters with the ensemble of null-space parameters, creating sets of calibration-constrained parameters for input to follow-on transport predictions. The consistency between probabilistic ensembles of parameter estimates and predictions created using the MSP calibration and the NSMC approaches is examined using a highly parameterized (>1300 parameters) model of the Culebra dolomite previously developed for the Waste Isolation Pilot Plant project in New Mexico as a test case. A total of 100 estimated fields are retained from the MSP approach, and the ensemble of results defining the model fit to the data and prediction of an advective travel time are compared with the same results obtained using NSMC. We demonstrate that the NSMC fields based on a single calibrated model can be significantly constrained by the calibrated solution space, and the resulting distribution of advective travel times is biased toward the travel time from the single calibrated field. To overcome this, newly proposed strategies to employ a multiple calibration-constrained NSMC (M-NSMC) approach are evaluated. Comparison of the M-NSMC and MSP methods demonstrates that M-NSMC can provide a computationally efficient and practical solution for predictive uncertainty analysis in highly nonlinear and complex subsurface flow and transport models.

Citation: Yoon, H., D. B. Hart, and S. A. McKenna (2013), Parameter estimation and predictive uncertainty in stochastic inverse modeling of groundwater flow: Comparing null-space Monte Carlo and multiple starting point methods, *Water Resour. Res.*, *49*, doi: 10.1002/wrcr.20064.

1. Introduction

[2] Groundwater model parameters are calibrated using a limited number of direct and indirect observations of the material properties. A calibrated model often represents a simple parameterization with a small number of parameters and simplified models corresponding to a limited set of observations. This approach leads to a well-known notion that calibration is a nonunique inverse problem [*Carrera and Neuman*, 1986], i.e., a number of different parameter sets can match observations to the same degree [*Beven and Binley*, 1992]. Predictions utilizing calibrated models inevitably have uncertainty (i.e., a range of unknown errors). Recently, quantification of the uncertainty associated with model predictions has been highlighted to better understand different sources of uncertainty [e.g., *Renard et al*., 2010] and provide decision makers with increasingly accurate and precise predictive models [*Keating et al*., 2010]. However, uncertainty analysis has not been routinely used in water resources modeling, and *Pappenberger and Beven* [2006] recommend utilizing model predictive uncertainty methods in management practices for many decision policies.

[3] Over the past decades, numerous techniques have provided more efficient tools for uncertainty analysis in the context of parameter estimation. Representative techniques include the generalized likelihood uncertainty estimation (GLUE) method [*Beven and Binley*, 1992], calibration-constrained Monte Carlo and Markov chain Monte Carlo (MCMC) methods [*Harvey and Gorelick*, 1995; *Kitanidis*, 1996; *Woodbury and Ulrych*, 2000], the method of stochastic equations [*Rubin and Dagan*, 1987; *Guadagnini and Neuman*, 1999; *Hernandez et al*., 2006], and deformation methods [*Hu*, 2000; *Lavenue and de Marsily*, 2001; *Gómez-Hernández et al*., 2003] (see summaries in *Keating*

[1]Geoscience Research and Applications Center, Sandia National Laboratories, Albuquerque, New Mexico, USA.

Corresponding author: H. Yoon, Geoscience Research and Applications Center, Sandia National Laboratories, P.O. Box 5800, MS 0751, Albuquerque, NM 87185, USA. (hyoon@sandia.gov)

0043-1397/13/10.1002/wrcr.20064

et al. [2010] and *Herckenrath et al*. [2011]). These methods are computationally intensive, in particular, for strongly nonlinear subsurface processes in the areas of emerging research involving multiphase flow and reactive transport such as geological storage of CO_2, reactive transport with biogeochemical reactions, and geothermal energy; hence, practical usage of predictive uncertainty analysis has been limited [*Pappenberger and Beven*, 2006].

[4] Parameterization of strongly heterogeneous spatial fields represents a particular challenge in groundwater flow and transport models. Pilot points as a device for spatial parameterization have been commonly used in model calibration [*Doherty*, 2003; *Alcolea et al*., 2006; *Doherty et al*., 2010a]. In this approach, model parameters are determined at selected points (i.e., pilot points), and the rest of model domain is interpolated from the pilot-point values based on a measured or assumed covariance [*de Marsily et al*., 1984, 1999]. Recent developments in the pilot-point method exploit the fact that, in many highly parameterized estimation problems, a minimum number of linear combinations of parameters account for the majority of the sensitivity of the overall calibration to the observed data. A notion that even a perfectly calibrated highly parameterized model (e.g., perfect fit to the observed data) can still have sufficient uncertainty was highlighted by analyzing a travel time arrival in a synthetic heterogeneous aquifer with various levels of model parameterization and calibration [*Moore and Doherty*, 2005]. In particular, *Moore and Doherty* [2005] provides a theoretical basis of the role of the calibration null space that is comprised of all differences between the simplified (or regularized) parameter field represented in a model and the complexity of the natural world. Since the calibration null space is comprised of the combinations of parameters whose sensitivity to the model output during calibration is negligible, these combinations in the null space often pertain to the fine-scale information such as local heterogeneity resulting from conceptual, temporal, and spatial simplifications. A notable technique to utilize this concept is the null-space Monte Carlo (NSMC) approach [*Tonkin and Doherty*, 2009] where, based on the parameter sensitivities during calibration, the null space of the calibration parameter field is identified. Stochastically generated parameter fields are then projected through the null space; the perturbation of the null-space components does not significantly change the overall calibration, and the solution-space components are retained. NSMC allows for providing a large number of calibration-constrained parameter sets in a computationally efficient manner, and the NSMC approach is not limited to any particular model as demonstrated for both groundwater and watershed management problems in *Tonkin and Doherty* [2009].

[5] *Keating et al*. [2010] compared the results of NSMC with those of MCMC for a highly nonlinear groundwater model using a simplified surrogate model, and both techniques provided similar results. Only the NSMC method possesses the computational efficiency necessary to explore the predictive uncertainty of the highly parameterized groundwater model, suggesting that the NSMC method can serve as a practical tool for predictive uncertainty analysis in strongly nonlinear environmental problems. *Herckenrath et al*. [2011] also applied NSMC for a saltwater intrusion model based on the Henry problem that addresses density-dependent groundwater flow with a diffused saltwater front against a flowing freshwater field within a confined aquifer. A synthetic model selected from stochastic hydraulic conductivity fields was used to create a set of observations (i.e., heads and concentrations), which were then used to calibrate a model. Predictive uncertainty variances from the NSMC method and linear analysis were similar, calling for further comparison of the two methods in the future. Since NSMC is constrained by a calibrated parameter field, the performance of calibration-constrained NSMC fields is strongly dependent upon the starting point for the Jacobian (i.e., sensitivity) matrix and the proportion of parameters that lie in null space [*Tonkin and Doherty*, 2009; *Doherty et al*., 2010b]. Hence, it is necessary to explore the efficacy of a gradient-based calibration method on predictive uncertainty of the highly parameterized model.

[6] Areas of emerging subsurface research are increasingly centered on nonlinear problems, and these are complicated by natural material properties exhibiting strong heterogeneity. Because of the nonuniqueness of the inverse problem and model uncertainty, multiple parameter sets are preferred over a single parameter set that "best" calibrates the model [e.g., *McKenna et al*., 2003]. Several methods such as GLUE [*Beven and Binley*, 1992; *Beven*, 2006] and the formal Bayesian approach [*Woodbury and Ulrych*, 2000; *Neuman*, 2003] have addressed model uncertainty with a set of parameter values for predictive analysis. Recently, the dynamically dimensioned search (DDS) was used to find a set of high-quality calibration solutions from numerous random initial solutions or parameter sets using a nongradient-based local search algorithm [*Tolson and Shoemaker*, 2007]. However, the efficiency of DDS in subsurface problems is unclear, and, furthermore, it is not suitable for highly parameterized models. For a practical means of addressing uncertainty, the multiple starting point (MSP) method has seen extensive application to find multiple sets of pilot-point parameters by calibrating multiple initial random fields, as generated through geostatistical simulation and resulting in an ensemble of calibrated fields for follow-on flow and transport predictions [see *RamaRao et al*., 1995; *Capilla et al*., 1998; *Keating et al*., 2010]. *Skahill et al*. [2009] demonstrated that, using the secant Levenberg-Marquardt method for local search, MSP with a clustering method (i.e., multilevel single linkage) can avoid local minima efficiently, resulting in optimization performance similar to global optimization algorithms such as the Shuffled Complex Evolution [*Duan et al*., 1992, 1993] and the Covariance Matrix Adaptation Evolutionary Strategy [*Hansen and Ostermeier*, 2001; *Hansen et al*., 2003]. However, the MSP approach with highly parameterized models remains computationally expensive, as a separate calibration process is required for each ensemble member. A number of questions remain regarding the consistency between MSP and other ensemble approaches for quantifying uncertainty in model predictions such as travel time analysis. In particular, it is unclear if the more recently developed NSMC method can replicate the parameter and predictive uncertainty associated with MSP calibrations and be used as a proven standard method for uncertainty analysis.

[7] Here, we examine the consistency between probabilistic ensembles of parameter estimates and predictions using the MSP and NSMC approaches. A total of 100 estimated

fields are retained from the MSP approach and the ensemble of results defining the model fit to the data, the reproduction of the variogram model and prediction of an advective travel time are compared with the same results obtained using the NSMC method. A highly parameterized model of the Culebra dolomite previously developed for the Waste Isolation Pilot Plant (WIPP) project in New Mexico is used as the test case [*Hart et al*., 2008, 2009]. The specific objectives of this work are to evaluate the impact of single calibration-constrained NSMC sampling on predictive uncertainty of advective travel time with an application to the Culebra dolomite and to test multiple calibration-constrained NSMC (M-NSMC) approaches for effective predictive uncertainty characterization by developing practical strategies of selecting the calibrated fields. In particular, from five different sets of starting points for the NSMC ensemble, the resulting NSMC fields are used to evaluate the M-NSMC approach for effective sampling of uncertainty over a wide range of parameter solution space. The strategy of selecting the calibrated fields for input to the NSMC method is discussed from the perspective of developing a practical means of addressing uncertainty for the case where having hundreds of calibrations from multiple starting fields is not available.

2. Culebra Groundwater Flow Model and Parameterization

[8] Details of conceptual model development with parameterization at the WIPP site are given by *Hart et al*. [2009]. In this section, the site-specific pilot-point parameterization is briefly discussed. The Culebra dolomite lies approximately 450 m above the Waste Isolation Pilot Plant (WIPP) in southeastern New Mexico and is a relatively thin, approximately 8 m, and laterally extensive, strongly heterogeneous (greater than 6 orders of magnitude in transmissivity) confined unit. As the most transmissive unit near the WIPP repository, the Culebra presents a potential groundwater migration pathway for radionuclides away from the WIPP site and has been the subject of extensive characterization activities. The Culebra has a long history in the development of stochastic inverse models, particularly for techniques that apply pilot points for parameterization, including development of the MSP approach for generating multiple calibrated fields [*RamaRao et al*., 1995; *LaVenue et al*., 1995], the sequential self-calibration approach [*Capilla et al*., 1998], and techniques for heterogeneous field estimation in the presence of a strong trend [*Rubin and Seong*, 1994; *Seong and Rubin*, 1999]. The Culebra has also motivated the development of new stochastic inverse approaches [e.g., *Zimmerman et al*., 1998] and served as an international test case for these approaches [e.g., *Larrson et al*., 1996]. The most recent stochastic inverse calculations performed as part of the Compliance Recertification Application are documented by *U.S. Department of Energy* [2009] and *Hart et al*., [2009]. These calculations are the source of the 100 MSP fields used in this work, and the calculation approach is briefly summarized here.

[9] The two-dimensional (2-D) Culebra model domain is 28.4 km × 30.7 km and discretized with 87,188 cells, each 100 m × 100 m. Parameter zones are developed for each of the parameters including transmissivity (T, m^2/s), horizontal anisotropy (A) for transmissivity, storativity (S, dimensionless), and recharge (R, m/s) based on the conceptual model of the Culebra flow [*Hart et al*., 2008]. An example of T zones is shown in Figure 1a. For T zones, the stochastic zones (zones 0 and 1) are located in the central portion of the domain where zone 0 defines more heavily fractured dolomite as identified in core analysis and corresponding to dual-porosity responses during pressure testing, and zone 1 is less fractured dolomite corresponding to single-porosity responses to hydraulic testing. The locations where zone 0 or 1 is present are only known at the well locations. Therefore, the zonation of the central portion of the domain is treated stochastically with geostatistical indicator simulation used to generate realizations of the distribution of these two zones. The geostatistical realizations are coupled with geologic information to create a large number of "seed" fields that are conditional to the static data and need to be calibrated to head and drawdown observations. The site-specific conceptual model including zones, multiple linear regression to estimate T, and the location of some of the zone boundaries was developed based on more than 3000 oil and gas well logs and other site characterization results including pumping and tracer tests [*Hart et al*., 2009]. In addition, zone 2 represents a high T region on the western side of the domain where the Culebra is closer to the ground surface and becoming unconfined, zone 3 on the eastern side where the transmissivity of the Culebra is extremely low due to halite cement within the pore spaces, and zone 4, approximately 15% of the domain, is a no-flow zone (i.e., inactive model domain, Figure 1). Anisotropy is the ratio of T in the north-south to the east-west, and anisotropy zones are the same as T zones. For storativity fields (Figure 1b), zones are divided into confined (zone 0), transition (zone 1), unconfined (zone 2), static confined (zone 3), and no-flow zones (zone 4). For most of the highly heterogeneous T regions (zones 0 and 1), storativity is assigned to a single confined zone. A small recharge zone represents a groundwater divide in the southwest portion of the model domain, characterized by a simple line.

[10] After zones for all parameters are created, pilot points are placed in the model domain as shown in Figure 1. Parameter values at pilot points are either fixed or variable, and two placement approaches of pilot points are employed. The pilot points are located using a regular triangular grid across the majority of the model domain with additional pilot points located between pumping and observation wells. The measured T values from single-well pumping and slug tests were used as fixed points, and extra points are placed along the northern and southern boundaries to limit the impact of fixed boundary conditions on transient pumping and steady-state model results (Figure 1a). For anisotropy, there are no fixed pilot points, and pilot points are located coarsely, compared with T fields. For storativity, in the confined zone, the only variable pilot points are placed along transient pumping test connectivity lines, and the remaining pilot points are assigned a fixed value. In the unconfined and transition zones, variable pilot points are used (Figure 1b). For the recharge zone with only four pilot points, one fixed point by the western domain boundary is used, and others are assigned as variable. A total of 141 fixed pilot points (61 for T, 79 for S, and 1 for R) are used, and a total of approximately 1300 parameters are estimated.

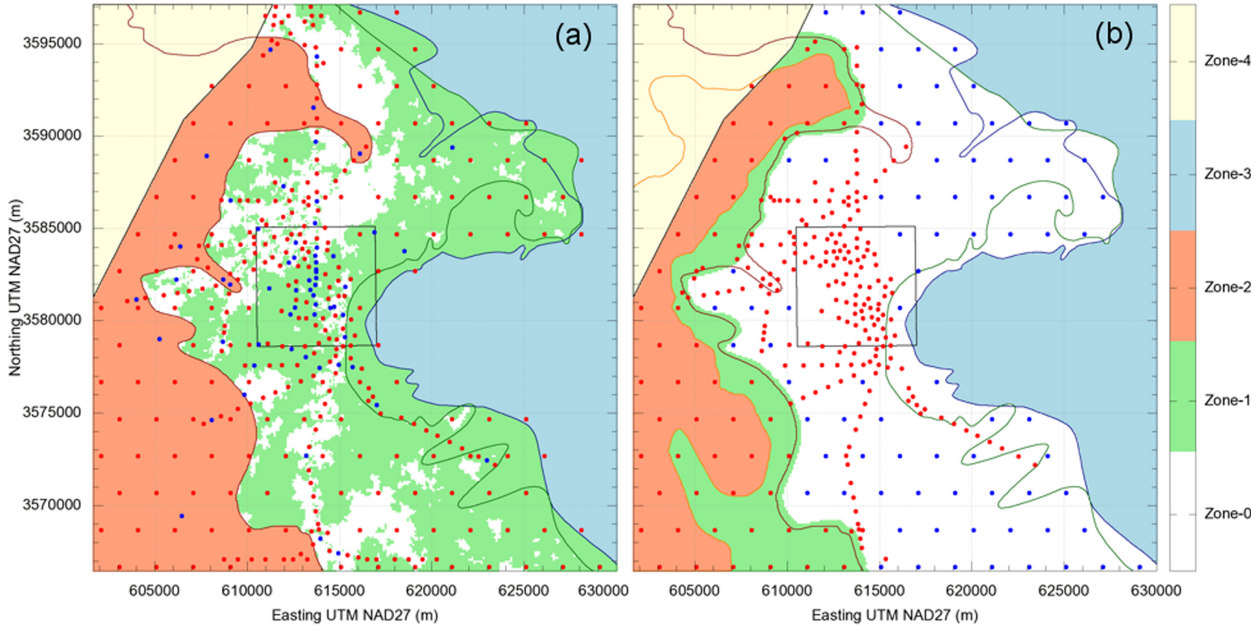

Figure 1. An example of (a) transmissivity (T) zone map and locations of pilot points and (b) storativity (S) zones and locations of pilot points. Blue points are fixed values, and red points are variable parameters. For T, zones 0 and 1 are the stochastic zones. For S, only pilot points along transient pumping tests and points in the unconfined zones (zones 1 and 2) are variable, and the remaining points are fixed.

Because of the stochastic nature of the transmissivity fields, unique zones are associated with each field, as defined by *Hart et al*., [2008], and the exact number of pilot-point locations used is unique to each seed field.

[11] Steady-state freshwater head equivalents were measured in 44 wells in 2007. Additionally, drawdown observations from nine pumping tests conducted over a 20 year period with a total of 65 observation wells are used. The pumping and recovery periods last from 2 to 18 months. In total, there are 1380 head and drawdown observations. Boundary conditions are fixed heads on the south, east, and portions of the north and west boundaries. The remaining portions of the north and west boundaries are no-flow boundaries. For each calibrated field, advective particle tracking is employed to calculate the travel time from a fixed starting point over the repository to the land-withdrawal boundary (i.e., a legal boundary around the WIPP site) as shown by a square in the center of model domain (Figure 1).

3. Parameter Estimation and Uncertainty Analysis

3.1. Calibration

[12] Model calibration and predictive uncertainty analysis were performed using the parameter estimation package PEST [*Doherty*, 2010; *Doherty and Hunt*, 2010a] with the pilot-point method. In particular, a special version of Parallel PEST, BeoPEST [*Hunt et al*., 2010b], was used to improve computational efficiency. Parameter estimation was performed using the truncated singular value decomposition (TSVD) technique, "SVD-assist" methodology [*Tonkin and Doherty*, 2005], to minimize the objective function. The objective function (Φ) is the weighted sum of squared errors between the observed data and simulated values as follows:

$$\Phi = \left(\mathbf{h} - \mathbf{X}\underline{\mathbf{p}}\right)^t \mathbf{Q} \left(\mathbf{h} - \mathbf{X}\underline{\mathbf{p}}\right), \quad (1)$$

where $\mathbf{h}$ is the observation data, $\mathbf{X}$ is the sensitivity of the simulated equivalent of each observation with respect to each parameter, $\underline{\mathbf{p}}$ is the vector of current parameter values, the diagonal matrix $\mathbf{Q}$ represents the square of the weight, and a superscript t is the transpose operation. For nonlinear models, a parameter upgrade vector, $\Delta\underline{\mathbf{p}}$, is iteratively calculated using the Gauss-Newton method with the Levenberg-Marquardt parameter [*Seber and Wild*, 1989] as follows:

$$\Delta\underline{\mathbf{p}} = (\mathbf{X}^t\mathbf{Q}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Q}\mathbf{r} \quad (2)$$

where $\mathbf{r}$ is the residual vector.

[13] For highly parameterized models where parameters are not uniquely estimable, $\mathbf{X}^t\mathbf{Q}\mathbf{X}$ is not invertible. With TSVD, the parameter set at the pilot points is decomposed into two orthogonal subspaces. The decomposition is based on the primary components from singular value decomposition of the sensitivity matrix whose linear combinations are called superparameters. The calibration solution space that is estimable by the calibration data set (i.e., superparameters) is determined by the TSVD truncation threshold or the number of parameters to be calibrated as defined by the user [*Doherty et al*., 2010b]. The calibration null space is defined by calibration insensitive (i.e., not

estimable) parameter combinations. An advantage of using TSVD for highly parameterized models is that it is numerically stable [*Tonkin and Doherty*, 2005] and computationally efficient. In the SVD-assist technique, the number of parameters to be calibrated is reduced dramatically into a smaller set of superparameters, instead of computing sensitivities with respect to all variable parameters. In this work, the threshold value of 10^{-4} was used as a default truncation threshold, and the maximum number of singular values was set equal to the number of adjustable parameters so that the number of superparameters was determined by the threshold value. The number of superparameters with the threshold value of 10^{-4} was generally 10% of the original number of parameters.

[14] With TSVD, $\mathbf{X}^t\mathbf{Q}\mathbf{X}$ is decomposed as follows:

$$\mathbf{X}^t\mathbf{Q}\mathbf{X} = [\mathbf{V}_1 \ \mathbf{V}_2]\begin{bmatrix}\mathbf{E}_1 & 0\\ 0 & \mathbf{E}_2\end{bmatrix}[\mathbf{V}_1 \ \mathbf{V}_2]^t, \tag{3}$$

where $\mathbf{E}_1$ is a diagonal matrix of the p nonzero singular values, p is the number of pretruncation eigenvalues (i.e., the number of superparameters) based on the SVD threshold value, $\mathbf{E}_2$ is a diagonal matrix of the $n-p$ zero-valued (i.e., truncated eigenvalues) singular values, n is the total number of parameters to be estimated, $\mathbf{V}_1$ is the p orthogonal unit vectors spanning the domain of the parameter solution space (corresponding to $\mathbf{E}_1$), and $\mathbf{V}_2$ is the $n-p$ orthogonal unit vectors spanning the domain of the null-space parameter (corresponding to $\mathbf{E}_2$). For pure TSVD, the estimated parameter set $\underline{\mathbf{p}}$ is given by the following equation:

$$\underline{\mathbf{p}} = \left(\mathbf{V}_1\mathbf{E}_1\mathbf{V}_1^t\right)\mathbf{X}^t\mathbf{Q}\mathbf{h} = \mathbf{G}\mathbf{h}, \tag{4}$$

where $\mathbf{G}$ is a matrix that approaches a generalized inverse of $\mathbf{X}^t\mathbf{Q}\mathbf{X}$, as the noise ($\varepsilon$) associated with measurement $\mathbf{h}$ approaches zero. Equation (4) can be expressed in terms of the resolution matrix $\mathbf{R}$, a set of true parameter values (never known) $\mathbf{p}$, and the measurement noise (ε):

$$\underline{\mathbf{p}} = \mathbf{G}\mathbf{X}\mathbf{p} + \mathbf{G}\varepsilon = \ \mathbf{R}\mathbf{p} + \mathbf{G}\varepsilon. \tag{5}$$

[15] In the absence of measurement noise, the resolution matrix represents the estimated elements of $\underline{\mathbf{p}}$ as a weighted average of true parameters ($\mathbf{p}$) multiplied by each row of $\mathbf{R}$. Through this regularization, there is a unique solution to the inverse problem, even for a highly parameterized model.

[16] In this work, the ordinary kriging system with an exponential variogram model was used to interpolate parameter values from the estimated pilot points over the entire numerical domain at the model grid. The model variogram has a smaller nugget ($0.02[\log_{10}\mathrm{T}]^2$), a large sill ($1.95[\log_{10}\mathrm{T}]^2$), and a relatively large range (9500 m) [*Hart et al*., 2009]. For parameter estimation and follow-on transport predictions, 2-D water flow and tracer transport were simulated using MODFLOW-2000 [*Harbaugh et al*., 2000] and the streamline particle tracking code DTRKMF [*Rudeen*, 2003], respectively. For the calibration process, initial values at the pilot points were set based on the conceptual model in *Hart et al*. [2008]. Briefly, the expected T value has a distinct regression equation for each zone as a function of the depth to Culebra from ground surface. Initial anisotropy was set to isotropic conditions (i.e., $A = 1.0$). Initial storativity in the confined, transition, and unconfined zones was set at 10^{-5}, 10^{-4}, and $10^{-1.5}$, respectively. Recharge was set at 10^{-11} m/s.

3.2. Predictive Uncertainty Analysis

3.2.1. Multiple Starting Points

[17] In this work, we combine the gradient-based algorithm implemented in PEST with stochastically generated "starting" fields to find a set of equally plausible calibrated solutions. Each calibration result cannot guarantee the best solution, and the determination of the best solution for predictive uncertainty is rather subjective. Instead, a set of inverse calibrations from MSPs provide an ensemble of calibrated parameters and follow-on transport predictions.

[18] A schematic of MSP and NSMC methods is shown in Figure 2. For MSP, 200 stochastically generated "seed" fields underwent calibration, and a total of 100 fields that had the best calibration were selected for predictive uncertainty analysis. As mentioned previously, the Culebra has motivated the development of new stochastic inverse approaches. For example, *Zimmerman et al*. [1998] used 50–100 geostatistically generated random fields (i.e., initial seeds = final selected fields) to construct cumulative density functions (CDFs) of travel times using a series of four synthetic problems motivated by the WIPP site to compare and contrast the performance of seven inverse methods through uncertainty analysis. It is noted that the number of parameters (varying for each method) and observed data (total 80 T and head data) in *Zimmerman et al*. [1998] were much smaller than approximately 1300 parameters and 1380 observed T and head data employed in this study. The 100 best selected calibrated fields from 200 multiple starting fields with the robust conceptual model developed at the WIPP site should reflect the uncertainty of the calibrated parameters and travel time prediction.

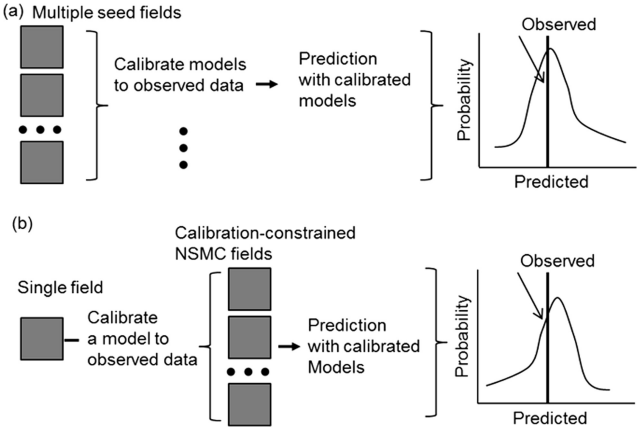

Figure 2. Schematic of (a) MSP and (b) NSMC methods for predictive uncertainty analysis. For NSMC, calibration-constrained NSMC fields require one to two additional recalibration iterations.

3.2.2. Null-Space Monte Carlo

[19] In this work, we examine the consistency between probabilistic ensembles of parameter estimates and predictions using the MSP calibration and the NSMC approaches. As shown in Figure 2, MSP requires intensive model runs and is computationally expensive. A first step of the NSMC method is to calibrate a single model, which is then used to generate calibration-constrained parameter sets very efficiently. For NSMC, 200 sets of parameters were generated from a single calibration, and a total of 100 best fields were selected using the same selection process as for MSP for predictive uncertainty analysis. Here, we briefly describe the NSMC methodology to generate a set of parameter realizations.

[20] From equation (5), parameter error, $\mathbf{p}-\underline{\mathbf{p}}$, can be expressed as follows:

$$\mathbf{p} - \underline{\mathbf{p}} = (\mathbf{I} - \mathbf{R})\mathbf{p} - \mathbf{G}\varepsilon, \tag{6}$$

where $\mathbf{I}$ is the identity matrix. Since $\mathbf{p}$ is unknown, equation (6) cannot be solved. However, statistical properties of the parameter error can be computed from equation (6) as follows:

$$\mathrm{C}\left(\mathbf{p} - \underline{\mathbf{p}}\right) = (\mathbf{I} - \mathbf{R})\mathrm{C}(\mathbf{p})(\mathbf{I} - \mathbf{R})^t + \mathbf{G}\mathrm{C}(\varepsilon)\mathbf{G}^t, \tag{7}$$

where C represents the covariance matrix. The first term of equation (7) arises due to parameter error resulting from inaccessibility of the null-space component of $\mathbf{p}$ to the calibration, and the second term arises due to the measurement noise of the observation $\mathbf{h}$ from which $\underline{\mathbf{p}}$ is estimated. In equation (7), the first term reveals the loss of system details in a calibrated model during calibration process and the C($\mathbf{p}$) matrix informs the null space that consists not only of parameters specified that are not estimated, but also of the world of possible parameters that are not specified and thus not estimated [*Doherty and Hunt*, 2010b]. In contrast to the traditional methods of model parameterization based on well-posed inverse problems, highly parameterized methods with regularization (i.e., SVD in this study) allow us to quantify predictive uncertainty arising from parameter simplification through postcalibration uncertainty analysis with the first term of equation (7). C($\mathbf{p}$) represents the innate variability of parameters that are often informed based on the knowledge of the properties and can be constructed using the prior parameter probability distribution or the postcalibration parameter error covariance [*Doherty et al*., 2010b]. In this work, C($\mathbf{p}$) was constructed with the parameter covariance from the *T* variogram model.

[21] For inversion with TSVD, $\mathbf{I}-\mathbf{R}$ is given by the projection operator onto the null space:

$$\mathbf{I} - \mathbf{R} = \mathbf{V}_2\mathbf{V}_2^{\mathrm{t}}, \tag{8}$$

where the resolution matrix $\mathbf{R}$ represents the projection operator onto the calibration solution subspace ($\mathbf{V}_1\mathbf{V}_1^{\mathrm{t}}$). Equation (8) is used to project differences between stochastic parameter realizations generated using C($\mathbf{p}$) and the calibrated parameter values onto the calibration null space as follows:

$$\left(\underline{\mathbf{p}} - \underline{\mathbf{p}}_{st}\right)' = \mathbf{V}_2\mathbf{V}_2^t\left(\underline{\mathbf{p}} - \underline{\mathbf{p}}_{st}\right), \tag{9}$$

where $\underline{\mathbf{p}}_{st}$ represents the stochastic parameter values, and $(\underline{\mathbf{p}}-\underline{\mathbf{p}}_{st})'$ is the projected parameter difference of $(\underline{\mathbf{p}}-\underline{\mathbf{p}}_{st})$. Finally, the projected differences are added to the calibrated parameters to generate new calibration-constrained parameter sets. The detailed description of equations (6) and (7) is given by *Moore and Doherty* [2005], *Tonkin and Doherty* [2005], and *Tonkin et al*. [2007]. The basis of equations (8) and (9) and development of NSMC method are provided by *Tonkin and Doherty* [2009].

[22] As in calibration, the ordinary kriging system with an exponential variogram model was used to interpolate parameter values from the estimated pilot-point locations over the entire numerical domain onto the model grid. If the model is linear, all NSMC generated parameter sets will calibrate the model to the same level of accuracy as the original calibrated model. However, the model is nonlinear and contains nonzero singular values that are below the TSVD threshold value of 10^{-4}. As a result, the projection of each random realization onto the null space is approximate [*Tonkin and Doherty*, 2009]. Hence, further adjustment of each NSMC parameter set is required. As demonstrated in *Tonkin and Doherty* [2009] and *Herckenrath et al*. [2011], two additional recalibration iterations were employed in this study. For this purpose, SVD is undertaken on the weighted Jacobian matrix, as computed during the initial calibration, to define a set of superparameters. For each NSMC field, the number of superparameters was the same as in the calibration. The first iteration of recalibration is performed by making use of the precalculated Jacobian matrix, dramatically reducing the required model runs. In this study, three to six runs of the model were required to perform the Broyden rank one update for the first recalibration iteration. To further improve the computational efficiency, calibration ceased when the resulting objective function increased.

3.2.3. Multiple Calibration-Constrained NSMC

[23] Pilot-point parameterization employed in this study tends to smooth out the local optima. Like other gradient-based methods, the Marquardt-Levenberg algorithm used in this study cannot avoid finding local optima, and finding the global optimal solution becomes more difficult with increasing complexity of the model and nonlinearity of the problem. It is unclear how well NSMC fields generated from a single calibration can sample a representative range of the solution space spanning more than a local optimum. Hence, M-NSMC and one single NSMC, starting with calibration of the mean seed field, were tested to evaluate the sampling efficiency for finding multiple "equally likely" solutions. M-NSMC has the potential to overcome the local minimum issue by generating random fields constrained by multiple (e.g., one to five) calibrated fields, compared with 100 MSPs. In this study, we develop a practical means of addressing uncertainty without having hundreds of calibrations from multiple starting fields. As described in Table 1, four different strategies of selecting the multiple calibrated fields for input to the NSMC method are evaluated, in addition to the mean field (MF) of 200 stochastic seed (uncalibrated) fields used for the MSP method. Four different cases of starting points for the NSMC analysis include five starting seed fields spanning the range of objective function values based only on a forward model run (OBJ_0), five starting seed fields spanning the range of travel times based

Table 1. Five Approaches for NSMC Analysis[a]

Case[b]	Fields Employed for NSMC Analysis
OBJ_{calib}	Based on objective function values (Φ) of 200 calibrated MSP fields
TT_{calib}	Based on travel times of 200 calibrated MSP fields
OBJ_0	Based on objective function values (Φ) of a forward model run with 200 uncalibrated starting (seed) fields
TT_0	Based on travel times of a forward model run with 200 uncalibrated starting (seed) fields
MF	Mean field of 200 stochastic starting (seed) fields

[a]Uncalibrated MSP fields represent stochastic starting (seed) fields generated by geostatistical simulation conditioned to observations of transmissivity (T).

[b]For objective function value cases, five fields are selected with the lowest, 25, 50, and 75 percentiles, and highest values of the objective function using either uncalibrated (OBJ_0) or calibrated (OBJ_{calib}) fields. For travel time cases, five fields are selected with the fastest, 25, 50, and 75 percentiles, and slowest travel time based on either uncalibrated (TT_0) or calibrated (TT_{calib}) fields.

only on a forward model run (TT_0), five calibrated fields spanning the range of objective function values of previously calibrated fields (OBJ_{calib}), and five calibrated fields spanning the range of travel times of previously calibrated fields (TT_{calib}).

4. Results

4.1. Model Calibration

[24] Predictive uncertainty analysis in this study involves calibration from MSPs. Previous estimation efforts [*Hart et al*., 2009] resulted in 200 calibrated fields using the MSP approach from 200 seed fields created through geostatistical simulation conditional to the T measurements. A large number of parameter values and pilot points used for calibration were possible due to the "SVD-assist" technique implemented in the PEST software that is a computationally efficient and numerically stable truncated SVD method that estimates linear combinations of parameters spanning the calibration solution space.

[25] In addition to satisfying the threshold of the objective function, the resulting calibrated fields also need to meet certain problem-specific criteria. In this study, the selection criteria for the "best" calibrated fields consisted of comparing both the absolute average error of the modeled steady-state heads and the absolute average error of the modeled transient responses (drawdowns) to separate cutoff values. As shown in Figure 3, the cutoff values used were 0.7 and 0.164 m for the steady-state and transient criteria, respectively, which are the mean values of the errors from all 200 MSP fields [*Hart et al*., 2009]. From the ensemble of MSP fields, a set of 102 fields satisfied the cutoff values. After two fields with the largest sum of errors were excluded, the final (i.e., best) 100 fields that were less than the cutoff values remained. In Figure 3, the average errors in both steady-state and transient pumping tests for each field calibrated through the MSP process are shown. The inset of Figure 3 shows that the clustering of the best 100 selected fields (i.e., several distinct solutions are located within the attraction area of the same local minima) is small, highlighting the starting points are perturbed enough to have distinct calibrated solutions.

[26] The modeled steady-state head values and example pumping test results for the 100 selected fields are compared with the observed field values in Figure 4. Comparison of the steady-state head values shows that the 100 selected MSP fields match the observed data well. For the transient pumping test results, the average error of the final 100 fields ranged from 0.12 to 0.164 m across all tests, with an average error of 0.15 m. The maximum error for a single observation well ranged from 0.005 to 2.5 m, with an average of 0.36 m. Overall, calibration reduced the average errors by a factor of 5–7 and objective functions by 2–3 orders of magnitude, compared to the uncalibrated models (seed fields) based on the conceptual model. The details of calibration results compared to all observed data were reported by *Hart et al*. [2009].

[27] Among the four estimated parameters, T fields changed dramatically from initial values and are more heterogeneous, compared with other parameters. The overall storativity values in the confined and transition zones did not change much from the initial values, and standard deviation (SD) of storativity from the 100 best MSP fields was relatively small, less than 0.3 except for a small region in the northwest area. This region has been attributed to the relatively poorer fits of a single pumping test data set [*Hart et al*., 2009]. The final recharge values at three pilot points changed very little compared with other parameters. Since storativity and recharge did not have a strong impact on calibration, we focus on the most sensitive and heterogeneous transmissivity fields. Although the anisotropy was estimated separately, the effective transmissivity fields including the anisotropy are used to describe the resulting T fields. The T_{eff} values are calculated as $\log T + 0.5 \times \log A$, representing the average of $\log T$ in the north-south and east-west directions. Additional details of model calibrations are provided by *Hart et al*. [2009].

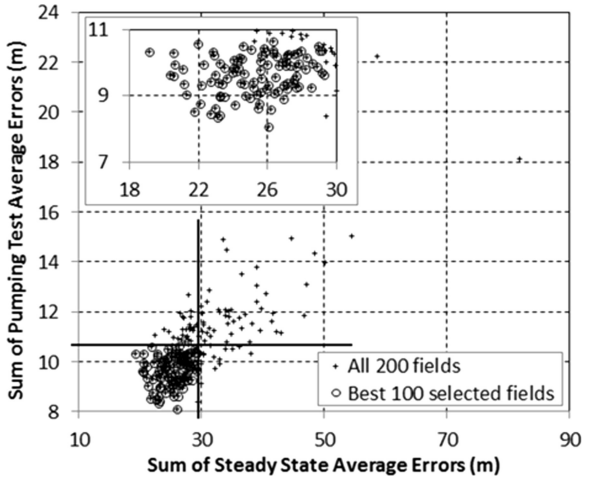

Figure 3. Comparison of the average errors in both steady-state and transient pumping tests for each field calibrated through the MSP process. The best 100 selected fields from all 200 MSP fields are highlighted in circles. Solid lines show the selection criteria. An inset highlights the error distribution of the best 100 selected sets.

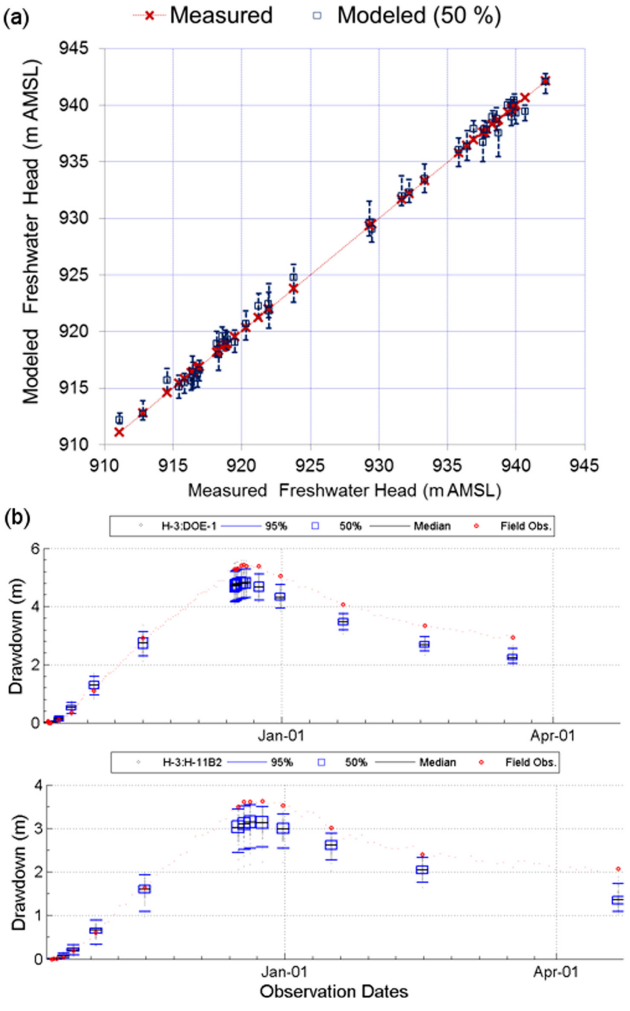

Figure 4. (a) Comparison of measured and modeled steady-state head values and (b) examples of pumping test results at two locations for the 100 best selected MSP fields. Measured steady-state head values at observation wells are presented as "×" marks along a one-to-one axis in (a). The box-and-whisker markers include whiskers at the 2.5 and 97.5 percentile values and a box around the central 50%; the median value is marked with a line in (b).

4.2. Uncertainty Analysis for the Culebra

4.2.1. Single Calibration-Constrained NSMC

[28] The effectiveness of the NSMC method is defined by comparison with the MSP method. As discussed, for the MSP method, the 100 best calibrated fields of 200 seed fields were used to evaluate the predictive uncertainty. The preliminary analysis (results not shown) shows that 50% best (lowest objective function) results from 100 to 200 initial staring fields have similar CDFs of travel times, but 25 best fields from 50 initial seeds have a slightly faster result than the other cases. It also shows that CDFs of travel times for all seed fields are faster than those for 50% best results from 100 to 200 initial seeds, indicating that travel times with lower objective function values tend to be slightly lower. CDFs of travel times for best results of 200 seeds with a different fraction show that 80–150 best fields from 200 initial seed fields are very similar, but using 50 seeds or all 200 seeds has different results. This preliminary analysis confirms that selecting the 100 best calibrated fields of 200 seed fields is appropriate in this work.

[29] The distribution of the calibrated effective transmissivity for all fields selected for the NSMC analysis is shown in Figure 5. For the NSMC method, each calibrated field in Figure 5 is used to generate 200 calibration-constrained Monte Carlo fields, and then the 100 best calibrated fields compared to the criteria were selected. Compared with the MSP approach, the threshold values used for the 100 NSMC field selection were higher by a factor of 1.3–1.6 for most of NSMC fields. Higher selection criteria are attributed to NSMC fields that slightly depart from the calibration solution space (i.e., local minimum). Calibration objectives and travel times for the 100 best NSMC fields generated from each of five calibrated models in Figure 5 are shown in Figure 6. As described earlier, the advective travel time for a particle released at a point to a prescribed boundary within the model domain at the WIPP site was the predictive performance measure [*U.S. Department of Energy*, 1996]. For MSP, the range of objective values is relatively narrow compared with all NSMC fields, and the travel times are distributed over 2 orders of magnitude. For a majority of NSMC fields, objective functions or travel times tend to be biased toward the initial calibration points as highlighted in symbols in Figure 6, indicating that NSMC fields are constrained by their calibrated fields. To further evaluate the efficacy of NSMC approaches, the travel time distribution was compared.

[30] The CDF of travel times from the MSP fields and six single calibration-constrained sets of NSMC fields is shown in Figure 7. The six sets of NSMC fields each include the 100 best results generated from the lowest (OBJ_{calib}-0th and TT_{calib}-0th), median (OBJ_{calib}-50th and TT_{calib}-50th), and highest (OBJ_{calib}-100th and TT_{calib}-100th) fields among the 200 MSP calibrated fields based on the objective function and travel time distribution. For NSMC results, rectangular boxes represent points corresponding to the initial calibrated fields. Figure 7 shows that the NSMC fields constrained by the calibrated model with a travel time close to the median of the MSP case (red symbols) have a travel time CDF similar to that of the MSP, whereas the CDF of the travel time from other NSMC fields is strongly influenced by the travel time of the respective calibrated field. This comparison clearly demonstrates that single calibration-constrained NSMC results are significantly affected by the calibrated solution space and have a limited representation of the range of the solution space.

4.2.2. Multiple Calibration-Constrained NSMC

[31] To evaluate the sampling efficiency for finding multiple "equally likely" solutions, M-NSMC fields were constructed by randomly selecting and combining 40 NSMC fields generated from each calibrated model (total 200 fields) shown in Figure 5. The final 100 best fields were selected as previously. The distributions of the 100 best results for five different NSMC sets (Table 1) are compared to the measured steady-state head values used for calibration (Figure 8). Comparison of the steady-state head values shows that the fields selected from the five NSMC cases match the observed data well. The range of errors

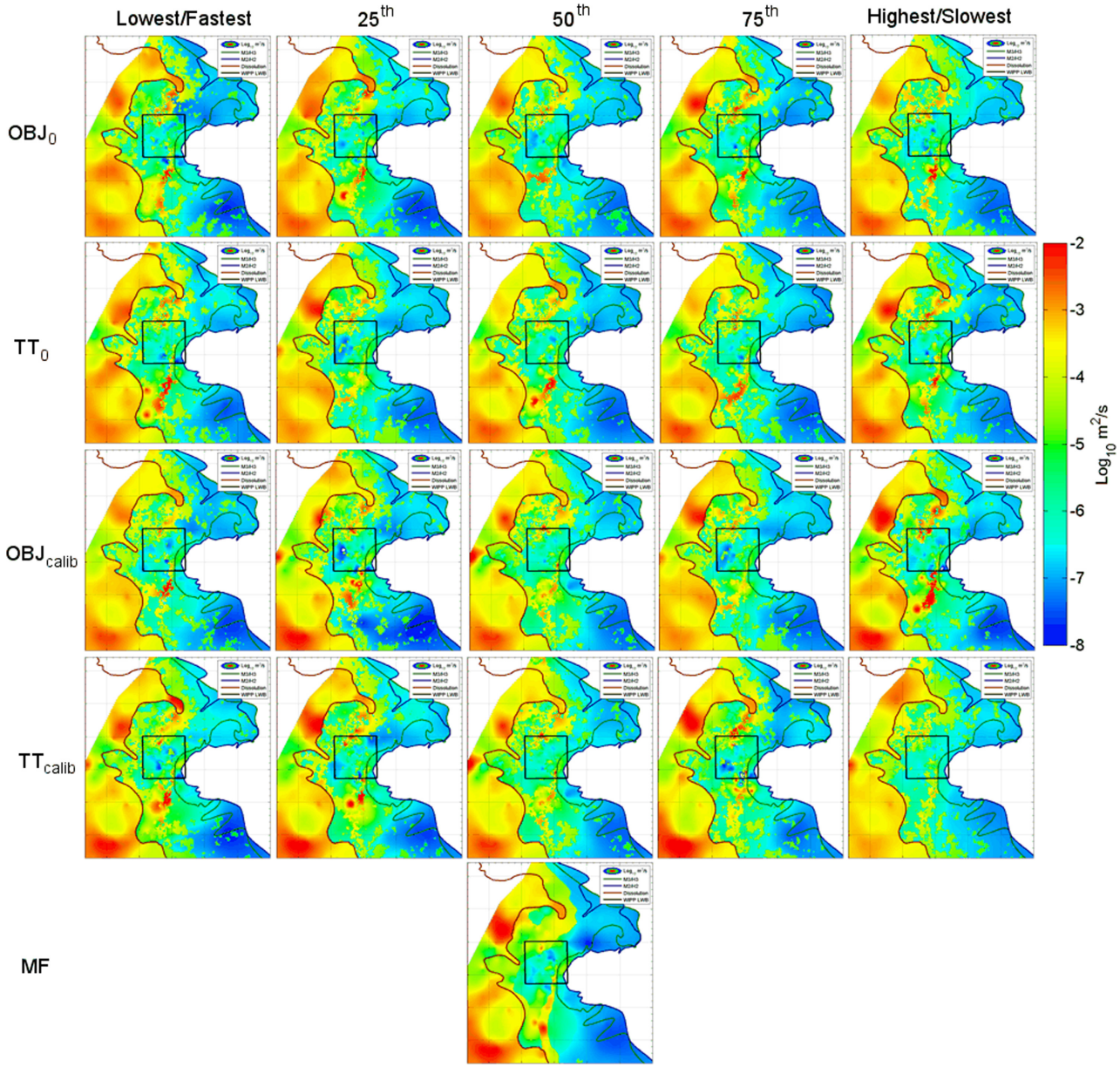

Figure 5. Starting sets of calibrated effective transmissivity (T_{eff}) distribution in zones 0–2 used for M-NSMC analysis. From the left of each case, the calibrated field corresponds to cases with the lowest objective function and fastest travel time to highest objective function and slowest travel time, respectively, through 25th, 50th, and 75th percentiles. MF is the calibrated field with the MF of 200 initial seed fields used for MSP.

(observed–simulated) for five different NSMC sets is larger than that for the MSP fields, but a majority of modeled steady-state head distributions are similar to those from the MSP fields. Comparison of the pumping tests also reveals that the NSMC fields match the observed data relatively well (results not shown), with a majority of the fields (>85%) having a range of errors similar to the MSP fields. This behavior is also shown in the distribution of travel times and calibration objectives in Figure 9. Compared with single calibration-constrained NSMC results, all M-NSMC and MF results show slightly narrower distributions of the objective function and the travel time distributions spanning 2 orders of magnitude similar to MSP results.

[32] To quantitatively evaluate the effectiveness of the NSMC method conditioned on the calibrated state, the CDFs of travel time from M-NSMC results and the MF NSMC results are compared with that from MSP in Figure 10. Comparison of the CDF of travel times shows that OBJ_0 and TT_0 have a better match with the MSP results than OBJ_{calib} and TT_{calib}. The CDF for the calibrated travel time-based case (TT_{calib}) has a greater proportion of faster travel times due to oversampling from the calibrated models with faster travel

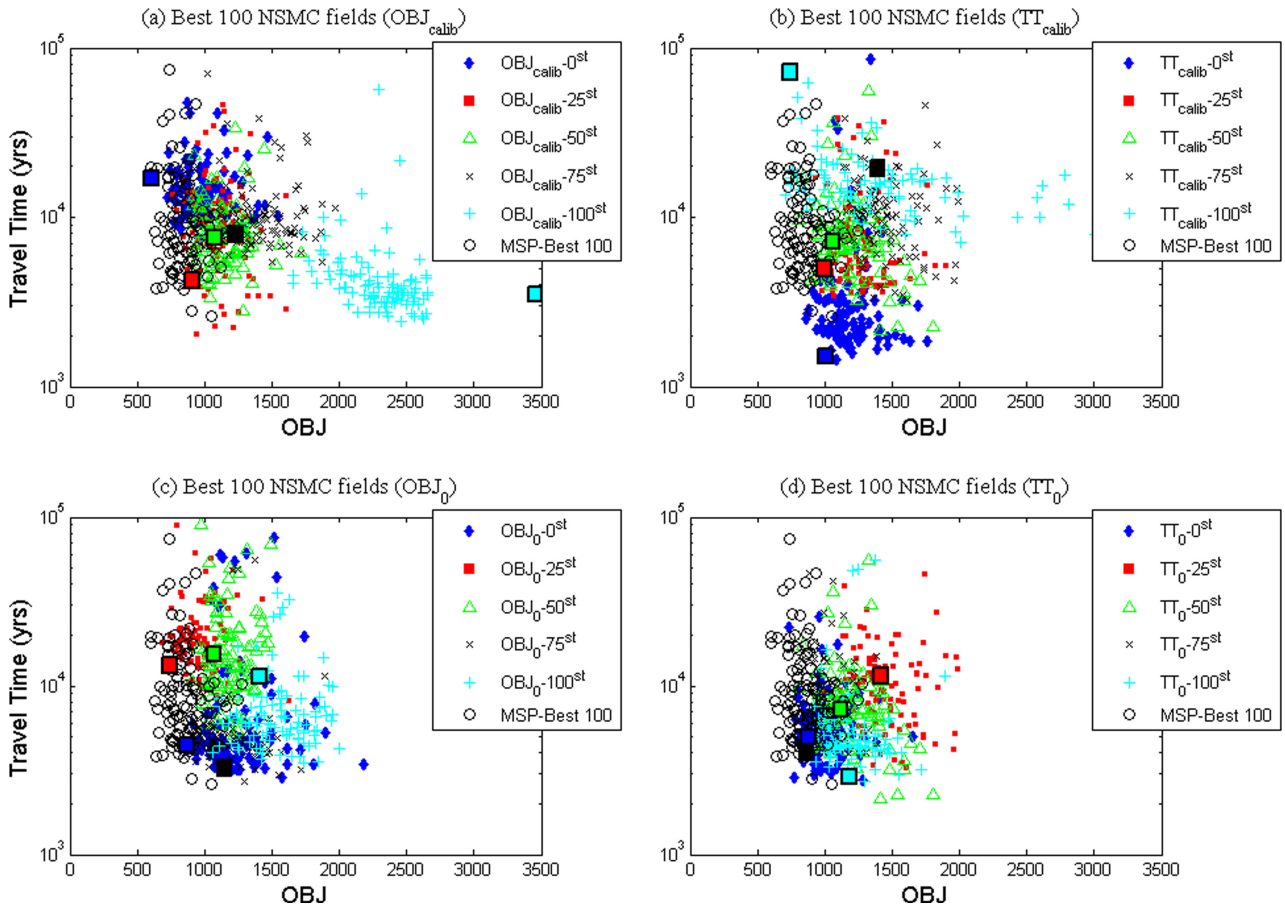

Figure 6. Calibration objectives and travel times for the best 100 NSMC fields generated from each of five calibrated models used for (a) OBJ_{calib}, (b) TT_{calib}, (c) OBJ_0, and (d) TT_0. For NSMC fields, the order of percentiles corresponds to lowest (0th) to highest (100th) objective functions or fastest (0th) to slowest (100th) travel times. The best 100 MSP fields are also shown. Rectangular boxes correspond to the results of calibrated models used for generating NSMC fields.

time as shown in Figure 7b, whereas the CDF for the calibrated OBJ-based case (OBJ_{calib}) has fewer faster travel times due to undersampling from the calibrated models with faster travel time, compared to the MSP results. Although the five uncalibrated fields selected for OBJ_0 and TT_0 include the lowest and highest cases based on uncalibrated starting seed fields, the calibrated models of these fields are all ranked between the 15th and 85th percentiles in terms of both objective function and travel time. This comparison suggests that the NSMC fields selected for OBJ_0 and TT_0 cases tend to cover a wide range of solution space but do not have the influence from the extreme calibrated fields as much as OBJ_{calib} and TT_{calib} sets do.

[33] A two sample Kolmogorov-Smirnov (KS) test was performed between the MSP results and each of the other five data sets to evaluate the CDFs of travel time for all six cases in Figure 10. The null hypothesis (H_0) is that the two data sets come from the same population. The *p* value of the test and the test statistic, the maximum vertical difference between the CDFs, are given in Table 2. Tests are run at a 0.05 significance level. As seen in visual comparison, the two ensembles generated from OBJ_0 and TT_0 are statistically indistinguishable from the MSP results. For the calibrated starting points, TT_{calib} provides a final ensemble that is also indistinguishable from the MSP results, but OBJ_{calib} does not. This indicates that although the resulting rank of the calibrated fields based on initial Φ and travel time can be problem specific, using uncalibrated seed fields as a starting basis is a feasible approach for generating a final ensemble that represents uncertainty in travel times. A faster travel time in the MF results compared with the MSP results may be attributed to a smooth calibrated field (Figure 5), resulting in relatively well-connected high-conductivity pathways. To overcome this smoothness in the MF, subscale heterogeneity can be incorporated into NSMC fields as demonstrated by *Tonkin and Doherty* [2009]. The MF approach with subscale NSMC will be evaluated as an initial starting point for constructing the NSMC fields in future work.

[34] The computational efficiency of M-NSMC (OBJ_0 and TT_0) is compared with MSP in terms of the required number of forward model runs (Table 3). As described earlier, initial forward runs are used to compute objective function values and travel time to select five seed fields for

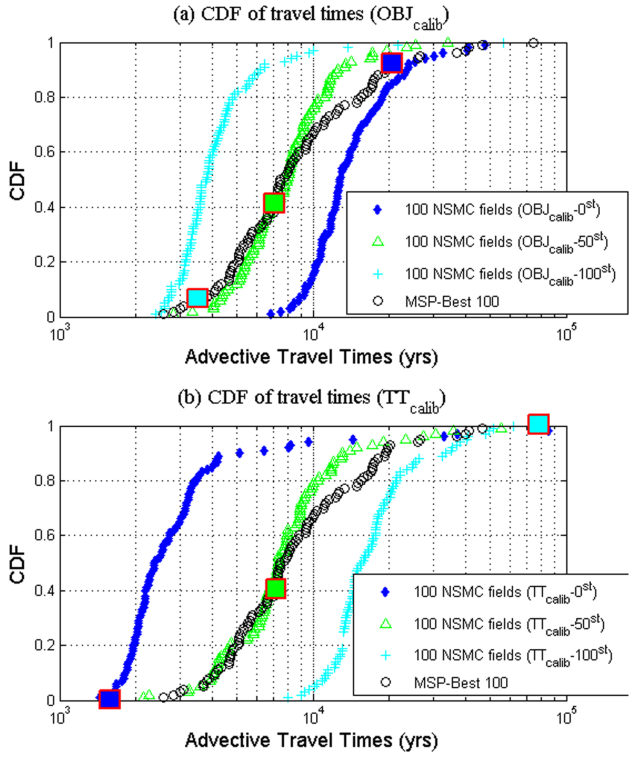

Figure 7. CDF of travel times from the MSP fields and NSMC fields generated from single calibrated models based on (a) the calibrated objective function (OBJ_{calib}-0th, OBJ_{calib}-50th, and OBJ_{calib}-100th percentiles) and (b) the travel time (TT_{calib}-0th, TT_{calib}-50th, and TT_{calib}-100th percentiles). The best 100 MSP fields are also shown. Rectangular boxes on each plot correspond to travel times calculated from calibrated models used for generating NSMC fields.

OBJ_0 and TT_0. All fields require construction of the Jacobian matrix prior to factorization with TSVD. For all fields, calibration solution and null spaces are determined using TSVD with a threshold value of 10^{-4}. During recalibration of NSMC fields, the cases with increasing Φ after the first iteration were abandoned, since the central difference scheme to construct the Jacobian matrix doubles the number of model runs. Overall, the forward model runs required for OBJ_0 and TT_0 are approximately 4% of those required for MSP. This result demonstrates that M-NSMC is computationally efficient, the performance measure of advective travel time between M-NSMC and MSP is statistically identical, and the method is model independent. It is evident that other selection methods such as maximum likelihood Bayesian model averaging (MLBMA) with information criteria [e.g., *Neuman*, 2003; *Singh et al.*, 2010], GLUE [*Beven and Binley*, 1992; *Singh et al.*, 2010] or clustering methods [*Skahill et al.*, 2009] could be easily employed for all field selection procedures used in this study. It is also evident that we can test the effect of different calibration fields on M-NSMC performance. However, here, we used simple criteria (i.e., errors in steady-state heads and pumping tests) for model selections to demonstrate the effectiveness of M-NSMC. Model selection methods will be the topic to be thoroughly addressed in the future.

4.2.3. Effective *T* Fields

[35] The effective T (T_{eff}) distributions from the final 100 selected fields for five NSMC (four M-NSMC and MF) and MSP methods are compared to evaluate spatial patterns of T fields. The mean and SD of T_{eff} distributions in zones 0–2 are compared in Figure 11. For qualitative comparison, the mapping accuracy, which is the proportion of properties (mean and SD of T_{eff}) matching between MSP and NSMC cases, is also computed. For the mapping accuracy of the high T_{eff} defined as those greater than -5.41 $\log_{10}$ m^2/s (i.e., the mean value of the 200 MSP seed fields in zones 0–2), all five NSMC sets are able to match the high and low T_{eff} of MSP with greater than 91% accuracy. Although the characteristics of the mean T_{eff} distribution are similar for all cases in Figure 11, the SD distribution of T_{eff} fields is very different between the MSP and NSMC results. For MSP, the SD values are higher along the high K paths in

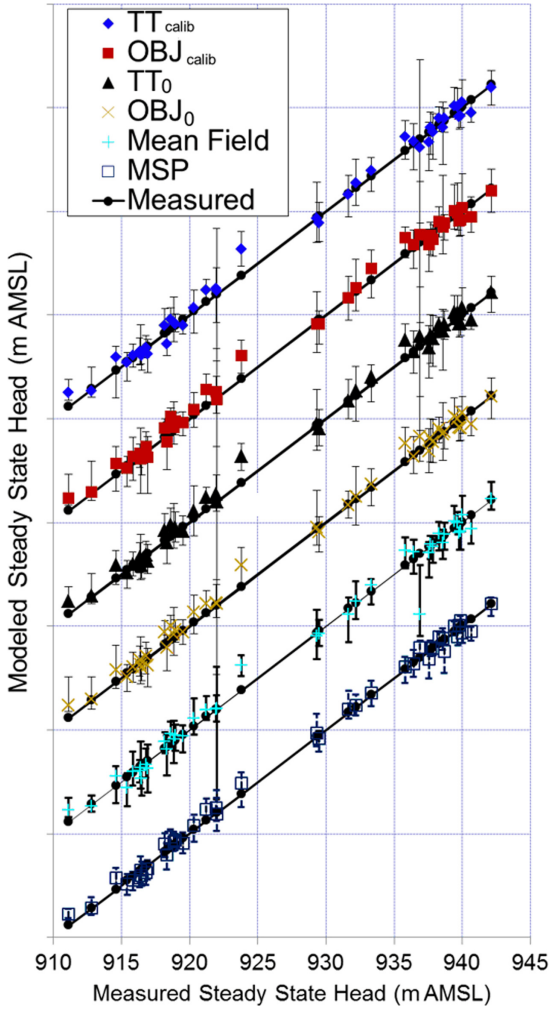

Figure 8. Comparison of measured and modeled steady-state head values for five different NSMC and MSP approaches. Measured values at observation wells are presented as solid circles along a one-to-one axis. For the purpose of comparison, modeled head values are separated by 10 m.

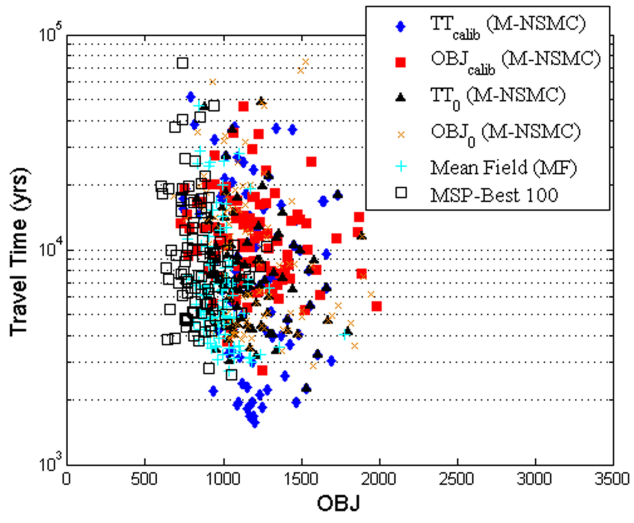

Figure 9. Calibration objective functions and travel times for the best 100 NSMC fields generated from four M-NSMC and MF approaches. The best 100 MSP fields are also shown.

the south of the model domain where a majority of particle paths lie. For all four M-NSMC sets, the SD distribution is more heterogeneous compared with that of MSP in zones 0 and 1, but the SD distribution is qualitatively similar. The mapping accuracy of the high SD, greater than 0.7, for all four M-NSMC was greater than 60%. However, the mapping accuracy for MF was approximately 45%, and the SD distribution for the MF results is relatively lower around the center and higher in the lower domain of zones 0 and 1. This difference is mainly due to generating NSMC fields from a single calibration-constrained MF where the variability of the T_{eff} field is small along the pilot points belonging to the calibration solution space.

[36] As shown in Figure 1a, a majority of pumping and monitoring wells (i.e., observed data and fixed points) are located within and around the center of the model domain. Hence, pilot points located within and around the center

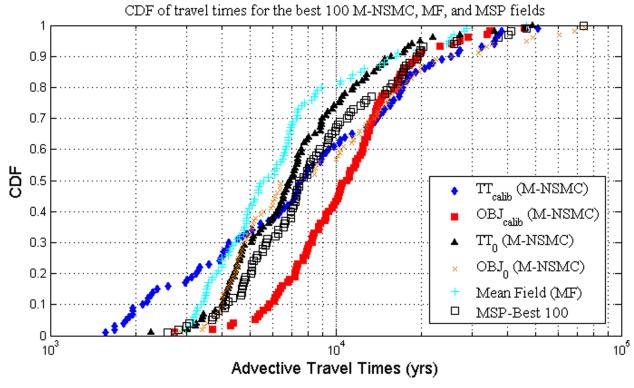

Figure 10. CDF of travel times from the 100 best selected fields from four M-NSMC, one MF and one MSP approaches.

Table 2. Summary of KS Test[a]

Data Set	Reject H_0[b]	p Value	Test Statistic[c]
OBJ_0	0	0.1930	0.150
TT_0	0	0.2778	0.137
OBJ_{calib}	1	0.0010	0.270
TT_{calib}	0	0.0505	0.188
MF	1	0.0003	0.290
OBJ_{calib}[d]	1	0.0003	0.290
TT_{calib}[d]	1	0.0010	0.270

[a]The threshold value for TSVD is 10^{-4}, otherwise indicated separately.
[b]The null hypothesis (H_0) is that the CDF of travel times of MSP and a NSMC data set is from the same distribution, which is accepted with p value greater than α (= 0.05) (0), otherwise rejected (1) at the 5% significance level.
[c]Test statistic is the supremum of the set of distances between two samples.
[d]The threshold value for TSVD is 10^{-5}.

were calibrated well, and a majority of parameters at these pilot points belonged to the calibration solution space. Since parameters at pilot points in the area where observed data are scarce and pilot points are loosely located are likely to lie within the calibration null space, the variability of the null-space projected parameters in this area is high (see north and south sides of SD maps in Figure 11). Since these parameters are highly likely to belong to the null space during recalibration, the final variability in their values will be close to that which was used to generate the random fields. This result is also shown in the T_{eff} distribution of 100 selected fields for all NSMC and MSP cases along the transect from south to north mostly in zones 0 and 1 (Figure 12). Compared with the MSP results, the MF case shows strong evidence of the calibration-constrained T_{eff} field, and four M-NSMC cases show more heterogeneous distributions, in particular, along the locations away from the pilot points belonging to the solution space.

4.2.4. Variograms

[37] Due to a limited number of observed data and smoothing characteristics of the inverse estimation process,

Table 3. Comparison of MSP and M-NSMC Computational Performance[a]

	MSP	OBJ_0	TT_0	MF
Initial forward runs[b]		200	200	
Determination of null and solution spaces[c]	233,865	5,866	5,826	982
Model calibration	1,431,571	39,126	40,043	9,739
Recalibration for NSMC fields[d]		20,496	21,545	24,288
Total	1,665,436	65,688	67,614	35,009

[a]For MSP, model runs are counted for 200 MSP fields and for OBJ_0 and TT_0; model runs are counted for 5 fields for model calibration and 40 NSMC fields per each calibrated field (i.e., total 200 NSMC fields) for recalibrations up to two iterations.
[b]Initial forward runs to compute objective function values and travel times for 200 multiple starting (seed) fields.
[c]Calibration solution and null spaces are determined using TSVD. The number of forward runs is the same as that of adjustable parameters (i.e., no fixed parameters).
[d]For the first iteration of recalibration, parameter sensitivity constructed during the model calibration process is used; hence, only three to six forward runs are required. For the second iteration, the number of model runs is the same as that of adjustable parameters.

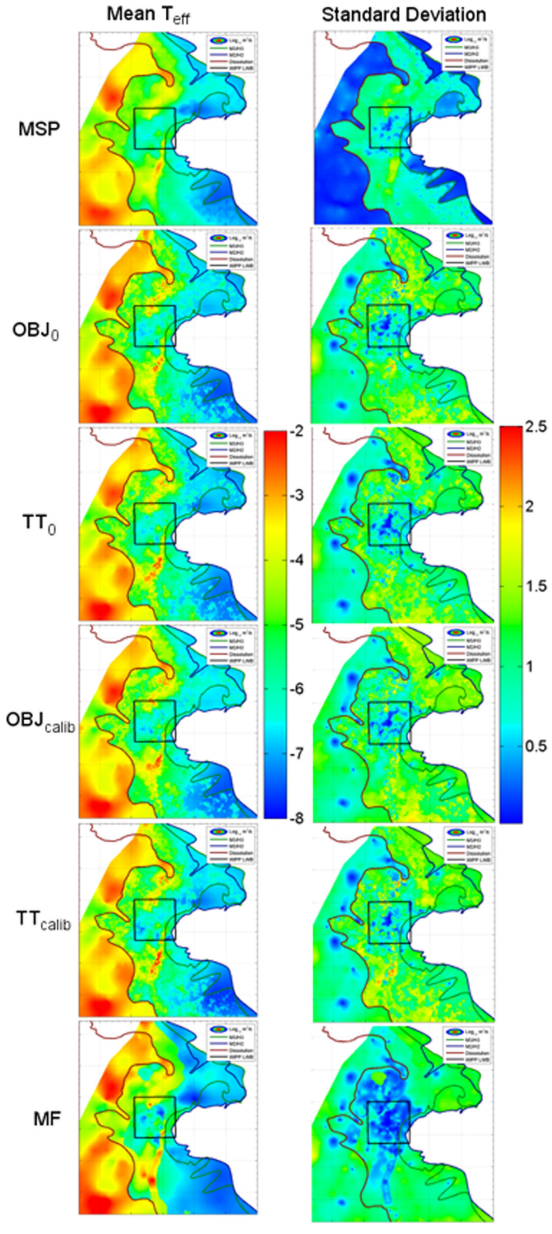

Figure 11. Comparison of the distributions of the mean effective T (T_{eff}, $\log_{10}$ m^2/s) and SD of T_{eff} in zones 0–2 for the 100 best selected fields from (top) MSP, four M-NSMC, and one MF approaches.

in many studies, it is observed that calibrated parameter fields do not reflect the heterogeneity of the observed hydraulic properties [*Moore and Doherty*, 2005; *Herckenrath et al*., 2011]. In this study, the variogram of the 100 best selected T_{eff} fields for the NSMC and MSP methods is compared with the model variogram in Figure 13. The MSP fields match the experimental variogram well, and the variation of sill and range values is narrow. However, NSMC fields have a higher variability (i.e., higher sill) than the model variogram as well as the MSP fields. This is not consistent with the observation in *Herckenrath et al*. [2011], who show that the calibrated model used for the NSMC method had much lower variability (i.e., a lower sill value) than the model variogram, and it was necessary to add heterogeneity at the numerical grid scale to match the model variogram. In this study, uncalibrated stochastic fields generated for the MSP have initial parameter values assigned to each pilot point based on site-specific conditions (e.g., stochastic zonal distribution) used for the seed T fields [*Hart et al*., 2008, 2009]. Hence, initial parameter fields reflect prior information regarding the characteristics of the site-specific hydrological features.

[38] During calibration, approximately 80%–85% of all parameters (approximately 1000 parameters) lie in the calibration null space, and the null-space components remain close to the initially assigned parameter values, since they are insensitive to the data set used for calibration. Unlike problems with uniformly distributed parameter values as initial conditions [e.g., *Herckenrath et al*., 2011], null-space parameter values in this work can be significantly perturbed by the null-space projection during the NSMC process as shown in the T_{eff} distribution along the transects (Figure 12). As a result, the null-space variability is larger than the solution-space variability. The newly introduced null-space variability will be lowered to a minimum degree during recalibration, resulting in higher variability than the model variogram. It is important to note that each null-space field includes a degree of heterogeneity compatible to that in the real world through the covariance matrix of innate parameter variability C(**p**). Comparison of model calibration and postcalibration uncertainty analysis between a regularized inversion method used in this study and a traditional method with relatively simplified models is discussed by *Doherty and Hunt* [2009, 2010b] and *Hill* [2010].

4.2.5. Effect of Truncated SVD Thresholds

[39] The effect of the dimensionality of the solution space on calibration and predictive uncertainty is evaluated by changing the SVD threshold value above which the calibration solution space is obtained by the truncated SVD method. In this study, 10 fields based on the calibrated objective function and travel time (OBJ_{calib} and TT_{calib}) with an SVD threshold of 10^{-4} were recalibrated with a lower threshold value of 10^{-5}. For both SVD threshold values, initial conditions were identical. Calibration with a higher threshold value (i.e., 10^{-4}) used in the previous section requires fewer model runs than the lower threshold value due to the lower solution-space dimension (92–133 superparameters) as listed in Table 4. As expected with larger numbers of superparameters, Φ values in 8 of 10 calibrations were lower with the threshold value of 10^{-5} than 10^{-4}. It should be noted that the range of Φ and travel time with the threshold value of 10^{-5} was much narrower, indicating that calibrated models with higher calibration solution dimensions are more constrained to the observed data used for calibration.

[40] Overall, the NSMC fields generated from the calibrated model with a lower threshold value (i.e., higher dimension of calibration solution space) tend to have less departure from the calibrated solution space due to a lower null-space dimension. Hence, the predictive uncertainty in terms of travel time will have a narrower range with a higher number of superparameters. The selection criteria (i.e., steady-state and pumping test errors) for the 100 best

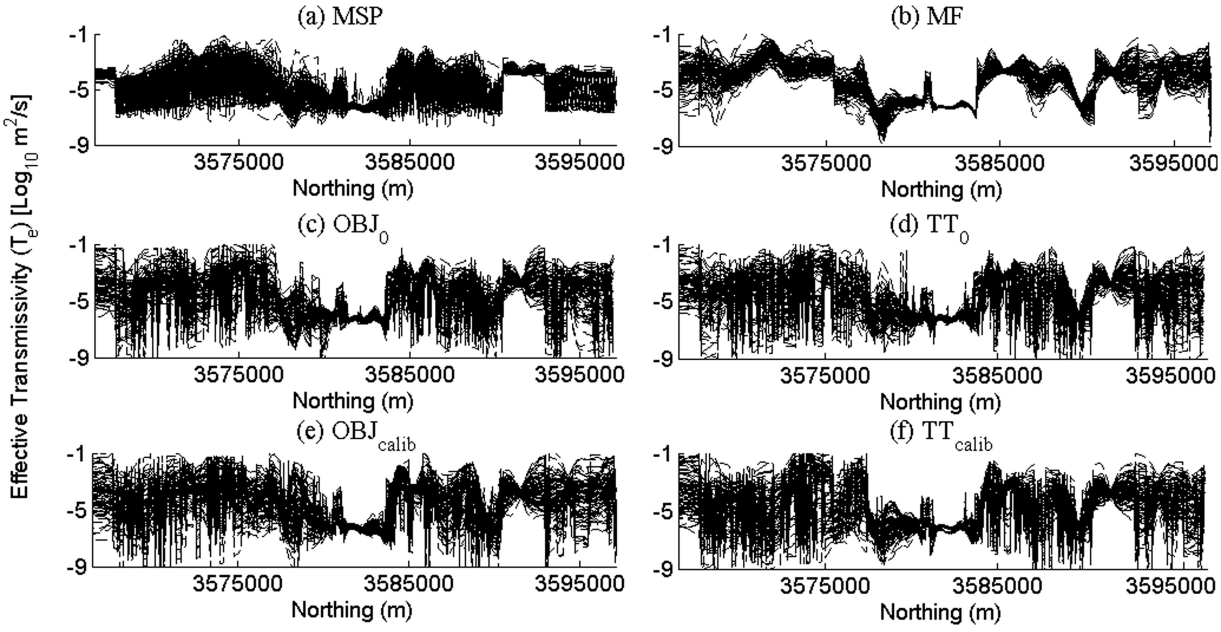

Figure 12. The effective T distribution of the 100 best selected fields for all NSMC and MSP approaches along the transect in the south-north direction ($x = 615{,}000$ m) shown in Figure 1: (a) MSP, (b) MF, (c) OBJ_0, (d) TT_0, (e) OBJ_{calib}, and (f) TT_{calib}.

NSMC fields with the threshold value of 10^{-5} were lower than those with the threshold value of 10^{-4} and even slightly lower than those for the MSP fields calibrated with the threshold value of 10^{-4}. The CDFs of travel times for the 100 best NSMC fields for OBJ_{calib} and TT_{calib} with the threshold value of 10^{-5} are shown in Figure 14. The CDFs of travel times from the NSMC results with the threshold value of 10^{-5} are narrower than those from the MSP results and NSMC results with the threshold value of 10^{-4}. For the NSMC results with the threshold value of 10^{-5}, the fastest travel time (approximately 5200 years) was longer than other cases, whereas the tailing of the travel time (>15,000 years) was relatively similar to other cases. This indicates that a better fit to the observed data with the lower threshold value led to a lower variability of the T fields along the high-permeability path over the southern boundary around the central region. This result is also supported by the fact that the NSMC results (OBJ_{calib} and TT_{calib}) with the threshold value of 10^{-5} have similar travel time CDF profiles.

[41] This comparison shows that the case with a higher threshold value (i.e., larger null-space components) creates a wider CDF including both fast and slow travel times. Without any vigorous estimation of the proper range of the calibration solution space, this suggests that using the smaller solution space with the higher threshold value that provides acceptable fits to the observed data is the preferred approach for a robust estimation of predictive uncertainty. In PEST, the threshold value between 10^{-5} and 10^{-7} with a typical start value of 5×10^{-7} was recommended [*Doherty and Hunt*, 2010a]. Comparison of two SVD thresholds suggests that the threshold value of 10^{-4} much higher than the recommended range can be also used for stable inversion and acceptable fits to the observed data. It should be noted that the robust conceptual model developed with geological information based on more than 3000 oil and gas well logs, and numerous pumping and tracer tests results may allow us to use a larger threshold than that might be possible in some other studies. Overall, the idea of using a higher threshold (e.g., 10^{-4}) but still only selecting the fields that reasonably fit the observed data will generate a larger uncertainty space, which will more likely capture the true, but unknown, underlying systems.

4.3. Implications for Practical Applications

[42] For predictive uncertainty estimation under environmental management applications, such as at the WIPP site, small-scale heterogeneity at length scales smaller than distances between boreholes (approximately 100 m) has a significant impact on transport processes. In general, the pumping tests do not possess the characteristics of small-scale variability, so the information of small-scale heterogeneity in this study is likely to lie within the null-space parameter and at subgrid scales not resolved here. The travel times predicted from the calibrated T fields with small-scale variability will tend to be longer than those without small-scale variability (e.g., tailing of breakthrough curves) [*Neuman*, 1990; *Alcolea et al*., 2008; *Singh et al*., 2010]. In addition, highly connected high-conductivity fields can result in fast arrival times for contaminants leaving the domain of concern. This is particularly important in water quality issues with emerging contaminants such as pathogens, pharmaceuticals, hormones, and other organic wastewater contaminants whose concentrations in groundwater resources are typically very low [*U.S. Environmental Protection Agency*, 2006; *Barnes et al*., 2008]. In the assessment of sewer source contamination of drinking water wells, *Hunt et al*. [2010a] pointed out that the transport of

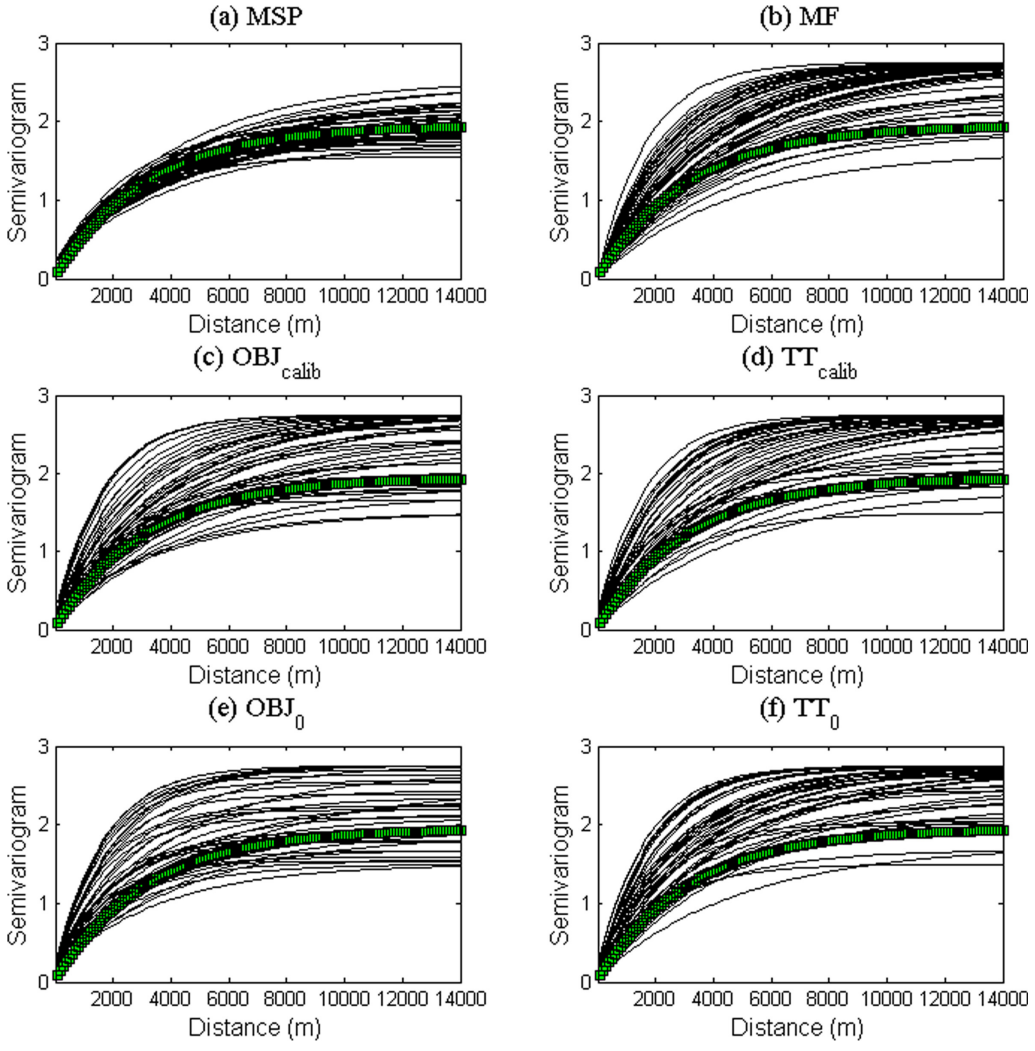

Figure 13. Comparison of the model variogram (green squares) that was used to generate the 200 MSP fields with variograms (lines) of the 100 best selected fields from (a) MSP, (b) MF, (c) OBJ_{calib}, (d) TT_{calib}, (e) OBJ_0, and (f) TT_0.

infectious viruses can be induced by high-capacity pumping, and vulnerability assessments in drinking water wells require both low water yield-fast transport information (e.g., virus transport) and high water yield-slower transport information (e.g., traditional average flow/plume migration). They highlighted that our current prediction ability to characterize such detailed information is not sufficient to do model prediction without undue uncertainty.

[43] *Moore et al*. [2010] recently investigated uncertainty in travel time prediction using a Pareto front approach in conjunction with global search optimizers, demonstrating that the correct travel time is encompassed only when the model calibration has a very high objective function that most decision makers would reject as being unrepresentative of the system. They also highlighted that system (aquifer) details for predicting the first arrival travel time belong to the characteristics of small-scale variability, so are beyond our ability to predict without much larger uncertainty than our uncertainty methods encompass. *Yoon and McKenna* [2012] demonstrated that even if very dense pilot points with highly parameterized models are used at every 1 cm^3 in a well-controlled three-dimensional sandbox experiment that is the most favorable condition for parameter estimation, local-heterogeneity features below the representative elementary volume (REV) scale are still difficult to incorporate into parameterized models. Overall, these recent studies highlight that model prediction of first arrival travel time needs to make use of high-level uncertainty methods to address small-scale variability (i.e., uncaptured variability) in modeling practice.

5. Conclusions

[44] Quantification of uncertainty associated with estimated model parameters and its impact on predictive uncertainty is becoming increasingly important as a fundamental tool to support decision making. The computationally demanding nature of complex and highly nonlinear models often limits the usage of model-based decision making. Due to the complexity (and subjectivity) of model

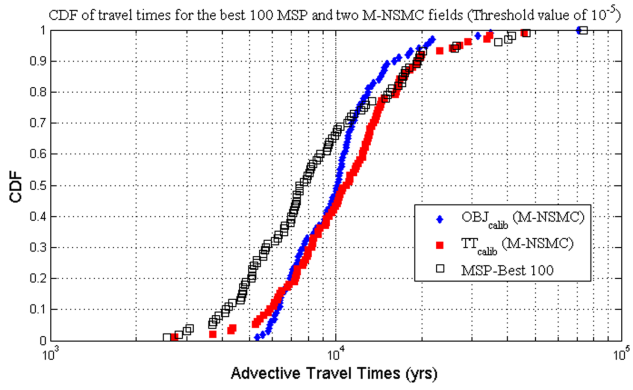

Figure 14. CDF of travel times from the MSP fields and two different M-NSMC methods (OBJ_{calib} and TT_{calib}) with the threshold value of 10^{-5} for TSVD.

constraints and our imperfect knowledge of problem-specific properties, the question arises as to how uncertainty defined by one approach compares with that of another. In particular, can computationally efficient approaches remain consistent with the results of more computationally intensive approaches? Here, we compare the recently developed and computationally efficient NSMC approach to the MSP method, which has been the accepted means of quantifying uncertainty in inverse parameter estimation problems with pilot-point parameterization over the past 20 years. Both methods were applied to a highly parameterized model of the Culebra dolomite previously developed for the WIPP project where the stochastic conceptual model of the heterogeneity creates multiple, equally probable, starting points for the inverse estimation. A total of 100 estimated fields are retained from the MSP approach, and the ensemble of results defining the model fit to the data, prediction of an advective travel time, and the reproduction of the variogram model is compared with the same results obtained using NSMC.

[45] We demonstrated that the NSMC fields based on a single calibrated model can be significantly constrained by the calibrated solution space, and the resulting distribution of advective travel times was biased toward the travel time from the single calibrated field. To overcome this, a proposed strategy to employ M-NSMC approach was evaluated. Five different means of choosing the starting fields that are the basis of the NSMC sampling were evaluated. Starting fields drawn from both previously calibrated and uncalibrated seed fields were examined. Among them, five seed fields spanning the range of Φ (OBJ_0) and the range of travel times (TT_0) based on a single forward model run predicted the CDFs of travel times that were most similar to that of the MSP results. Comparison of the reproduction of the variogram model and parameter distribution shows that M-NSMC can provide a set of solutions that span the uncertainty space for predictive uncertainty analysis in highly nonlinear and complex groundwater models without having hundreds of calibrations from multiple starting fields.

[46] To our knowledge, this work is the first comparison of the NSMC method against the MSP method on a large, highly parameterized, real world problem. The NSMC method combines the calibration solution-space parameters with the ensemble of null-space parameters, creating sets of calibration-constrained parameters. The issue of selecting multiple calibration fields is important in the M-NSMC approach by aiming to avoid multiple local minima within the influence region of each calibrated solution. This is critical to improving the computational efficiency in the NSMC method. Within the framework demonstrated here, other more sophisticated strategies of selecting starting seed fields such as clustering [*Skahill et al*., 2009], multiobjective criteria [*Singh et al*., 2010], and image ranking [e.g., *T* field distribution] can be used. The other remaining question would be how many starting seed fields are necessary for generating NSMC fields. From the computational perspective, the fewer calibration models (i.e., single calibration-constrained NSMC), the more computational efficiency can be achieved. This work suggests that M-NSMC with three to five calibrated models can provide a set of solutions that span the uncertainty space for predictive uncertainty analysis in highly nonlinear and complex groundwater models without having hundreds of calibrations from multiple starting fields. In the present work, M-NSMC improves the computational efficiency by a factor of approximately 25 compared with the MSP method. In addition to the simple strategy used in this study, other methods will help develop some guidelines regarding this aspect and possibly further improve the computational efficiency.

[47] In this work, a comprehensive geologic conceptual model developed with a large set of hydraulic data was used. Since the model is developed in a highly heterogeneous and complex field, the "true" *T* field is unknown. For synthetic test cases with a limited number of pilot points [e.g., *Herckenrath et al*., 2011], calibrated models tend to underestimate the heterogeneity primarily due to limited information of the calibration data set. In this study, the large number of pilot points placed in the stochastic zones through the center of the model domain and the 100 selected fields from both MSP and NSMC methods matched the model variogram well. In particular, the degree of heterogeneity from M-NSMC fields exceeds that of the model variogram (larger sill) without incorporating any randomization of the fine-scale (i.e., grid-scale) heterogeneity in contrast to the results of *Herckenrath et al*.

Table 4. Comparison of the Number of Superparameters (SP), Objective Function Values (OBJ), and Travel Time (TT) for the Calibrated Models Based on the Calibrated OBJ and TT Values With Two Different SVD Threshold Values

	Threshold Value $= 10^{-4}$			Threshold Value $= 10^{-5}$[a]		
	SP	OBJ	TT	SP	OBJ	TT
OBJ_{calib}-0th	118	600.8	17,299.1	201	517.5	9096.3
OBJ_{calib}-25th	108	834.4	5709.3	186	906.5	7716.6
OBJ_{calib}-50th	133	1019.6	6989.8	213	658	6032.5
OBJ_{calib}-75th	107	1247.9	7221.8	187	801.2	11,636.6
OBJ_{calib}-100th	109	4421	3475.7	192	664.6	4937.8
TT_{calib}-0th	116	1023.2	1608.1	197	800.1	17,763.6
TT_{calib}-25th	101	1018.5	4921.6	172	1030.1	8473.8
TT_{calib}-50th	126	1136.9	7082.0	209	807.7	14,001.5
TT_{calib}-75th	92	1394.4	11,046.2	193	845.1	12,171.9
TT_{calib}-100th	111	759	73,911.7	191	684	6050.1

[a]Ten fields based on the calibrated objective function and travel time (OBJ_{calib} and TT_{calib}) with an SVD threshold of 10^{-4} were recalibrated with a lower threshold value of 10^{-5}.

[2011]. An important problem in parameter and predictive uncertainty analysis is the lack of information on the covariance of the parameters (and measurement errors). In the NSMC context, the **C**(**p**) matrix in equation (7) is the estimate of parameter variability supplied by the user, and its uncertainty is site specific. Although the impact of **C**(**p**) on predictive uncertainty analysis is not considered here, the impact, in particular, with real study problems needs to be addressed.

[48] The computational efficiency of NSMC comes from the fact that sets of calibration-constrained parameters can be recalibrated very efficiently by adjusting only solution-space parameters. The choice of the threshold value of SVD to determine the solution-space components (and the number of superparameters) is often subjective. Selection of a higher threshold value (or higher objective function level) is often preferred, because the resulting higher null-space dimensions can retain the parameter uncertainty (or error) dominated by the null-space term, in addition to increased computational efficiency. In this work, the higher threshold value (10^{-4}), versus the lower one (10^{-5}), provides the preferred approach for a robust estimation of predictive uncertainty spanning the fast and slow travel times with acceptable fits to the observed data. This suggests that the idea of using a higher threshold (e.g., 10^{-4}) with selected calibrated fields that are reasonably fit to the observed data can generate a larger uncertainty space that encompasses the set of possible parameters not specified and thus not estimated. For the Culebra application, calibration to steady-state water levels and transient pumping drawdowns results in the solution space being dominated by parameter combinations that respond to the second-order diffusive process of pressure migration. These are larger-scale features, whereas the small-scale heterogeneity of hydraulic material properties that controls first-order solute transport processes is not resolved by calibration to water levels and drawdowns and remains mainly within the null space. In the NSMC framework, small-scale heterogeneity can be incorporated to account for solute transport processes using a subscale NSMC method at model grid scale [*Tonkin and Doherty*, 2009] and nested variogram models with small-scale variability [*Alcolea et al*., 2008]. Inclusion of small-scale heterogeneity in NSMC will enhance our ability to provide a practical tool of utilizing model predictive uncertainty methods in management practices for many decision policies.

[49] **Acknowledgments.** This material is based upon work supported as part of the Center for Frontiers of Subsurface Energy Security, an Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under award DE-SC0001114. Sandia National Laboratories is a multiprogram laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. We also acknowledge the effort of Randall Hunt and two anonymous reviewers for their careful and constructive reviews, which led to significant improvement of our manuscript.

References

Alcolea, A., J. Carrera, and M. Medina (2006), Inversion of heterogeneous parabolic-type equations using the pilot-points method, *Int. J. Numer. Methods Fluids*, *51*, 963–980, doi:10.1002/fld.1213.

Alcolea, A., J. Carrera, and A. Medina (2008), Regularized pilot points method for reproducing the effect of small scale variability: Application to simulations of contaminant transport, *J. Hydrol.*, *355*(1–4), 76–90, doi:10.1016/j.jhydrol.2008.03.004.

Barnes, K. K., D. W. Kolpin, E. T. Furlong, S. D. Zaugg, M. T. Meyer, and L. B. Barber (2008), A national reconnaissance of pharmaceuticals and other organic wastewater contaminants in the United States—I. Groundwater, *Sci. Total Environ.*, *402*(2–3), 192–200, doi:10.1016/j.scitotenv.2008.04.028.

Beven, K., and A. Binley (1992), The future of distributed models—Model calibration and uncertainty prediction, *Hydrol. Processes*, *6*(3), 279–298, doi:10.1002/hyp.3360060305.

Beven, K. J. (2006), A manifesto for the equifinality thesis, *J. Hydrol.*, *320*(1–2), 18–36.

Capilla, J. E., J. J. Gómez-Hernández, and A. Sahuquillo (1998), Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric head data. 3. Application to the Culebra formation at the Waste Isolation Pilot Plan (WIPP), New Mexico, USA, *J. Hydrol.*, *207*, 254–269.

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, doi:10.1029/WR022i002p00199.

de Marsily, G., C. Lavedan, M. Boucher, and G. Fasanion (1984), Interpretation of interference tests in a well field using geostatistical techniques to fit the permeability distribution in a reservoir model, in *Geostatistics for Natural Resources Characterization. Part 2*, edited by G. Verly et al., pp. 831–849, D. Reidel, Dordetcht, Netherlands.

de Marsily, G., J. P. Delhomme, F. Delay, and A. Buoro (1999), 40 years of inverse problems in hydrogeology, *Earth Planet. Sci.*, *329*(2), 73–87.

Doherty, J. (2003), Ground water model calibration using pilot points and regularization, *Ground Water*, *41*(2), 170–177.

Doherty, J. (2010), PEST: Model independent parameter estimation, Watermark Numer. Comput., Corinda, Queensland, Australia. [Available at http://www.pesthomepage.org.].

Doherty, J., and R. J. Hunt (2009), Two statistics for evaluating parameter identifiability and error reduction, *J. Hydrol.*, *366*, 119–127, doi:10.1016/j.jhydrol.2008.12.018.

Doherty, J., and R. J. Hunt (2010a), Approaches to highly parameterized inversion: A guide to using PEST for groundwater-model calibration, *U.S. Geol. Surv. Sci. Invest. Rep. 2010-5169*, U.S. Geol. Surv., Middleton, Wis.

Doherty, J., and R. J. Hunt (2010b), Response to comment on ''Two statistics for evaluating parameter identifiability and error reduction'', *J. Hydrol.*, *380*, 489–496, doi:10.1016/j.jhydrol.2009.10.012.

Doherty, J., M. N. Fienen, and R. J. Hunt (2010a), Approaches to highly parameterized inversion: Pilot-point theory, guidelines, and research directions: *U.S. Geol. Surv. Sci. Invest. Rep. 2010-5168*, 36 pp., U.S. Geol. Surv., Middleton, Wis.

Doherty, J., R. J. Hunt, and M. J. Tonkin (2010b), Approaches to highly parameterized inversion: A guide to using PEST for model-parameter and predictive-uncertainty analysis: *U.S. Geol. Surv. Sci. Invest. Rep. 2010-5211*, 71 pp., U.S. Geol. Surv., Middleton, Wis.

Duan, Q., S. Sorooshian, and V. K. Gupta (1992), Effective and efficient global optimization for conceptual rainfall-runoff models, *Water Resour. Res.*, *28*(4), 1015–1031.

Duan, Q., V.K. Gupta, and S. Sorooshian (1993), A shuffled complex evolution approach for effective and efficient global minimization, *J. Optim. Theory Appl.*, *76*(3), 501–521.

Gómez-Hernández, J. J., H. J. Hendricks Franssen, and A. Sahuquillo (2003), Stochastic conditional inverse modeling of subsurface mass transport: A brief review of the self-calibrating method, *Stochastic Environ. Res. Risk Assess.*, *17*, 319–328, doi:10.1007/s00477-003-0153-5.

Guadagnini, A., and S. P. Neuman (1999), Nonlocal and localized analyses of conditional mean steady state flow in bounded, randomly nonuniform domains: 1. Theory and computational approach, *Water Resour. Res.*, *35*(10), 2999–3018.

Hansen, N., and A. Ostermeier (2001), Completely derandomized self-adaptation in evolution strategies, *Evol. Comput.*, *9*, 159–195.

Hansen, N., S. D. Muller, and P. Koumoutsakos (2003), Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES), *Evol. Comput.*, *11*(1), 1–18.

Harbaugh, A. W., E. R. Banta, M. C. Hill, and M. G. McDonald (2000), MODFLOW-2000, the U.S. Geological Survey modular ground-water

model—User guide to modularization concepts and the ground-water flow process, *U.S. Geol. Surv. Open File Rep. 00-92*, 121 pp., Reston, Va.

Hart, D. B., R. M. Holt, and S. A. McKenna (2008), Analysis report for Task 5 of AP-114: Generation of revised base transmissivity fields, *ERMS# 541153*, Sandia Natl. Lab., WIPP Rec. Cent., Carlsbad, N. M.

Hart, D. B., R. L. Beauheim, and S. A. McKenna (2009), Analysis report for Task 7 of AP-114: Calibration of Culebra transmissivity fields, *WIPP:1.4.1.1:TD:QA-L:RECERT:541153*, *Sandia Natl. Lab*., WIPP Rec. Cent., Carlsbad, N. M.

Harvey, C. F., and S. M. Gorelick (1995), Mapping hydraulic conductivity: Sequential conditioning with measurements of solute arrival time, hydraulic-head, and local conductivity, *Water Resour. Res.*, *31*(7), 1615–1626, doi:10.1029/95WR00547.

Herckenrath, D., C. D. Langevin, and J. Doherty (2011), Predictive uncertainty analysis of a saltwater intrusion model using null-space Monte Carlo, *Water Resour. Res.*, *47*, W05504, doi:10.1029/2010WR009342.

Hernandez, A. F., S. P. Neuman, A. Guadagnini, and J. Carrera (2006), Inverse stochastic moment analysis of steady state flow in randomly heterogeneous media, *Water Resour. Res.*, *42*, W05425, doi:10.1029/2005WR004449.

Hill, M. C. (2010), Comment on ''Two statistics for evaluating parameter identifiability and error reduction'' by John Doherty and Randall J. Hunt, *J. Hydrol.*, *380*, 481–488, doi:10.1016/j.jhydrol.2009.10.011.

Hu, L.Y. (2000), Gradual deformation and iterative calibration of Gaussian-related stochastic models, *Math. Geol.*, *32* (1), 87–108.

Hunt, R. J., M. A. Borchardt, K. D. Richards, and S. K. Spencer (2010a), Assessment of sewer source contamination of drinking water wells using tracers and human enteric viruses, *Environ. Sci. Technol.*, *44*(20), 7956–7963, doi:10.1021/es100698m.

Hunt, R. J., J. Luchette, W. A. Shreuder, J. Rumbaugh, J. Doherty, M. J. Tonkin, and D. Rumbaugh (2010b), Using the cloud to replenish parched groundwater modeling efforts, *Ground Water*, *48*(3), 360–365, doi:10.1111/j.1745-6584.2010.00699.x.

Keating, E. H., J. Doherty, J. A. Vrugt, and Q. Kang (2010), Optimization and uncertainty assessment of strongly non-linear groundwater models with high parameter dimensionality, *Water Resour. Res.*, *46*, W10517, doi:10.1029/2009WR008584.

Kitanidis, P. K. (1996), On the geostatistical approach to the inverse problem, *Adv. Water Resour.*, *19*(6), 333–342, doi:10.1016/0309-1708(96)00005-x.

Larrson, A., K. Pers, K. Skagius, and B. Dverstorp (1996), The international INTRAVAL project, Phase II, *Working Group 2 Rep*., Finnsjön, Stripa and WIPP2, OECD/NEA Publ.

LaVenue, A. M., B. S. RamaRao, G. de Marsily, and M. G. Marietta (1995), Pilot-point methodology for automated calibration of an ensemble of conditionally simulated transmissivity fields: 2. Application, *Water Resour. Res.*, *31*(3), 495–516, doi:10.1029/94WR02259.

Lavenue, M., and G. de Marsily (2001), Three-dimensional interference test interpretation in a fractured aquifer using the pilot point inverse method, *Water Resour. Res.*, *37*(11), 2659–2675.

McKenna, S. A., J. Doherty, and D. B. Hart (2003), Non-uniqueness of inverse transmissivity field calibration and predictive transport modeling, *J. Hydrol.*, *281*, pp. 265–280.

Moore, C., and J. Doherty (2005), Role of the calibration process in reducing model predictive error, *Water Resour. Res.*, *41*, W05020, doi:10.1029/2004WR003501.

Moore, C., T. Wöhling, and J. Doherty (2010), Efficient regularization and uncertainty analysis using a global optimization methodology, *Water Resour. Res.*, *46*, W08527, doi:10.1029/2009WR008627.

Neuman, S. P. (1990), Universal scaling of hydraulic conductivities and dispersivities in geological media, *Water Resour. Res.*, *26*, 1749–1758.

Neuman, S. P. (2003), Maximum likelihood Bayesian averaging of alternative conceptual-mathematical models, *Stochastic Environ. Res. Risk Assess.*, *17*(5), 291–305, doi:10.1007/s00477-003-0151-7.

Pappenberger, F., and K. J. Beven (2006), Ignorance is bliss: Or seven reasons not to use uncertainty analysis, *Water Resour. Res.*, *42*, W05302, doi:10.1029/2005WR004820.

RamaRao, B. S., A. M. LaVenue, G. de Marsily, and M. G. Marietta (1995), Pilot point methodology for automated calibration of an ensemble of conditionally simulated transmissivity fields: 1. Theory and computational experiments, *Water Resour. Res.*, *31*(3), 475–493.

Renard, B., D. Kavetski, G. Kuczera, M. A. Thyer, and S. W. Franks (2010), Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors, *Water Resour. Res.*, *46*, W05521, doi:10.1029/2009WR008328.

Rubin, Y., and G. Dagan (1987), Stochastic identification of transmissivity and effective recharge in steady state groundwater flow: 1. Theory, *Water Resour. Res.*, *23*(7), 1192–1200.

Rubin, Y., and K. Seong (1994), Investigation of flow and transport in certain cases of non-stationary conductivity fields, *Water Resour. Res.*, *30*(11), 2901–2911.

Rudeen, D. K (2003), User's Manual for DTRKMF version 1.00, *ERMS# 523246*, Sandia Natl. Lab., WIPP Rec. Cent., Carlsbad, N. M.

Seber, G. A. F., and C. J. Wild (1989), *Nonlinear Regression*, 768 pp., Wiley, New York.

Seong, K., and Y. Rubin, 1999, Field investigation of the waste isolation pilot plant (WIPP) site (New Mexico) using a non-stationary stochastic model with a trending hydraulic conductivity field, *Water Resour. Res.*, *35*(4), pp. 1011–1018.

Skahill, B. E., J. S. Baggett, S. Frankenstein, and C.W. Downer (2009), More efficient PEST compatible model independent model calibration, *Environ. Model. Software*, *24*, 517–529.

Singh, A., D. D. Walker, B. S. Minsker, and A. J. Valocchi (2010), Incorporating subjective and stochastic uncertainty in an interactive multi-objective groundwater calibration framework, *Stochastic Environ. Res. Risk Assess.*, *24*(6), 881–898, doi:10.1007/s00477-010-0384-1.

Tolson, B. A., and C. A. Shoemaker (2007), Dynamically dimensioned search algorithm for computationally efficient watershed model calibration, *Water Resour. Res.*, *43*, W01413, doi:10.1029/2005WR004723.

Tonkin, M., and J. Doherty (2009), Calibration-constrained Monte Carlo analysis of highly parameterized models using subspace techniques, *Water Resour. Res.*, *45*, W00B10, doi:10.1029/2007WR006678.

Tonkin, M. J., and J. Doherty (2005), A hybrid regularized inversion methodology for highly parameterized environmental models, *Water Resour. Res.*, *41*, W10412, doi:10.1029/2005WR003995.

Tonkin, M. J., J. Doherty, and C. Moore (2007), Efficient non-linear predictive error variance for highly parameterized models, *Water Resour. Res.*, *43*, W07429, doi:10.1029/2006WR005348.

U.S. Department of Energy (1996), Title 40 CFR Part 191 Compliance Certification Application for the Waste Isolation Pilot Plant, *DOE/CAO-1996-2184*, U.S. DOE, Carlsbad Area Off., Carlsbad, N. M.

U.S. Department of Energy (2009), Title 40 CFR Part 191 Compliance Certification Application for the Waste Isolation Pilot Plant, *DOE/WIPP-09-3424*, U.S. DOE, Carlsbad Field Off., Carlsbad, N. M.

U.S. Environmental Protection Agency (2006), National primary drinking water regulations: Ground water rule; Final rule, *Fed. Regist.*, *71*, 67,427–65,660.

Woodbury, A. D., and T. J. Ulrych (2000), A full-Bayesian approach to the groundwater inverse problem for steady state flow, *Water Resour. Res.*, *36*(8), 2081–2093, doi:10.1029/2000WR900086.

Yoon, H., and S. A. McKenna (2012), Highly parameterized inverse estimation of hydraulic conductivity and porosity in a three-dimensional, heterogeneous transport experiment, *Water Resour. Res.*, *48*, W10536, doi:10.1029/2012WR012149.

Zimmerman, D. A., et al. (1998), A comparison of seven geostatistically based inverse approaches to estimate transmissivities for modeling advective transport by groundwater flow, *Water Resour. Res.*, *34*(6), 1373–1413, doi:10.1029/98WR00003.