Optimization and uncertainty assessment of strongly nonlinear groundwater models with high parameter dimensionality

Elizabeth H. Keating,[1] John Doherty,[2,3] Jasper A. Vrugt,[1,4,5] and Qinjun Kang[1]

Received 28 August 2009; revised 12 May 2010; accepted 28 May 2010; published 12 October 2010.

[1] Highly parameterized and CPU-intensive groundwater models are increasingly being used to understand and predict flow and transport through aquifers. Despite their frequent use, these models pose significant challenges for parameter estimation and predictive uncertainty analysis algorithms, particularly global methods which usually require very large numbers of forward runs. Here we present a general methodology for parameter estimation and uncertainty analysis that can be utilized in these situations. Our proposed method includes extraction of a surrogate model that mimics key characteristics of a full process model, followed by testing and implementation of a pragmatic uncertainty analysis technique, called null-space Monte Carlo (NSMC), that merges the strengths of gradient-based search and parameter dimensionality reduction. As part of the surrogate model analysis, the results of NSMC are compared with a formal Bayesian approach using the DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm. Such a comparison has never been accomplished before, especially in the context of high parameter dimensionality. Despite the highly nonlinear nature of the inverse problem, the existence of multiple local minima, and the relatively large parameter dimensionality, both methods performed well and results compare favorably with each other. Experiences gained from the surrogate model analysis are then transferred to calibrate the full highly parameterized and CPU intensive groundwater model and to explore predictive uncertainty of predictions made by that model. The methodology presented here is generally applicable to any highly parameterized and CPU-intensive environmental model, where efficient methods such as NSMC provide the only practical means for conducting predictive uncertainty analysis.

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

1. Introduction

[2] Uncertainty quantification is currently receiving a surge of attention in environmental modeling as researchers try to understand which parts of our models are well resolved, and which parts of our process knowledge and theories are subject to considerable uncertainty, and as decision-makers push to better quantify accuracy and precision of model predictions. In the past decades, a variety of different approaches have appeared in the scientific literature to address parameter, model, and measurement uncertainty, and to provide theories and algorithms for its treatment. See, for example, *Carrera and Neuman* [1986a, 1986b], *Kitanidis* [1997], *Woodbury and Ulrych* [2000], *Jiang and Woodbury* [2006], *Ronayne et al.* [2008], *Fu and Gomez-Hernandez* [2009], *Cooley and Christensen* [2006], *Tiedeman et al.* [2004], *Hill* [1998], *Vrugt et al.* [2005], *Vrugt and Robinson* [2007], and *Tonkin et al.* [2007], to name just a few works on this subject. However, despite the variety of methods available for this purpose, uncertainty analysis is rarely undertaken as an adjunct to model-based decision-making in applied hydrologic analysis [*Pappenberger and Beven*, 2006].

[3] For groundwater modeling in particular, this is an outcome of several practical challenges. First and foremost, many thousands of model evaluations are typically required for calibration, and to search parameter space for other feasible parameter sets that collectively allow appraisal of parameter uncertainty. Yet groundwater models, especially those considering processes such as reactive geochemistry and geomechanical or thermal effects, may require significant CPU time to complete a single forward run. Second, widely available and efficient gradient-based parameter estimation/optimization algorithms may perform very poorly when parameter dimensionality is large, local optima in the objective function surface are numerous, parameter sensitivities are low, and model input/output relationships

[1] Earth and Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico, USA.

[2] National Centre for Groundwater Research and Training, Flinders University, Adelaide, South Australia, Australia.

[3] Watermark Numerical Computing, Brisbane, Queensland, Australia.

[4] Department of Civil and Environmental Engineering, University of California, Irvine, California, USA.

[5] Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Amsterdam, Netherlands.

0043-1397/10/2009WR008584

are highly nonlinear. Third, inverse problems involving groundwater models often violate key assumptions that underpin rigorous statistical analysis of parameter and predictive uncertainty. For example, the assumption that measurement/structural noise is uncorrelated, or even possesses a known and/or stationary covariance matrix, is almost never met [*Cooley*, 2004; *Cooley and Christensen*, 2006; *Doherty and Welter*, 2010]. Unfortunately, model structural error is typically significant and much larger than measurement error [*Vrugt et al.*, 2005, 2008b]; yet there are no clear criteria or inference methods to determine its size, statistical properties, and its effect on the uncertainty associated with values assigned to parameters through the calibration process. Each of these issues can create significant problems in analyzing the uncertainty associated with predictions made by complex groundwater models.

[4] Recent developments in inverse algorithms have produced more robust and efficient tools for uncertainty analysis. As stated above, gradient-based methods are extremely efficient, but have historically been hampered by problems associated with local minima and insensitive parameters. A number of methodologies constituting improvements of traditional gradient based methods have been developed to address these limitations and efficiently compute the uncertainty associated with predictions made by highly parameterized groundwater models. These included the "self-calibration method" (*Hendricks Franssen et al.* [2004], *Gomez-Hernandez et al.* [2003] and dimensionality reduction techniques such as those described by *Tonkin and Doherty* [2005] and *Tonkin et al.* [2007]). On the other hand, global search methods using population-based evolutionary strategies are especially designed to provide robust uncertainty analysis in difficult numerical contexts, and can thus overcome many of the difficulties encountered by gradient-based methods. However, these typically require that a large number of model runs be carried out, and hence cannot be routinely applied in typical groundwater modeling applications. Meanwhile, another line of research has focused on the development of simultaneous, multimethod, global search algorithms whose synergy and interaction of constituent methods significantly increases search efficiency when compared to other methods of this type. The AMALGAM evolutionary search approach developed by *Vrugt and Robinson* [2007], and *Vrugt et al.* [2009a] provides an example of such a method, and has been shown to outperform existing calibration algorithms when confronted with a range of excessively difficult response surfaces. Unfortunately, AMALGAM is especially designed to only provide estimates of optimum parameter values without recourse to estimating parameter and prediction uncertainty. The recently developed DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm [*Vrugt et al.*, 2008a, 2009a, 2009b], which searches for objective function minima while simultaneously providing estimates of parameter uncertainty, overcomes this limitation while maintaining relatively high levels of model run efficiency in the context of Markov chain Monte Carlo (MCMC) simulation. Nevertheless, the number of model runs required for sampling of the posterior probability density function is often prohibitively large. For instance, a simple multivariate normal distribution with dimensionality of about 100, would require at least 300,000–500,000 function evaluations to provide a sample set of the posterior distribution [*Vrugt et al.*, 2009b].

[5] Here, we present a novel strategy for obtaining robust estimates of the uncertainties associated with highly parameterized, field-scale process models. This strategy should be widely applicable for complex model analysis where, due to CPU limitations, standard sequential MCMC approaches that rely on detailed balance are impractical to be used. Our strategy consists of several steps. First, a surrogate model is abstracted from the process model. The surrogate model must have a similar level of parameterization to the process model, and be calibratable to the same data set as the process model. The surrogate model is then used to do the vital work of appropriately posing the inverse problem through which parameters are estimated, including designing the objective function to be minimized through the calibration process, refinement of understanding of environmental processes operative in the study area, evaluation of the magnitude and nature of irreducible model structural error, and so on. This iterative process of model evaluation and calibration/uncertainty analysis continues until a reasonable result is achieved. During this process, a pragmatic methodology for uncertainty analysis, to be later transferred to the process model, can be developed for the surrogate model. Importantly, as is done in the present paper, this method can be tested against more statistically based and theoretically rigorous MCMC methods if this is judged to be necessary. Finally, the efficient method is transferred to predictive uncertainty analysis with the full process model, for which the high computational burden of running the model requires that parameter and predictive analysis be undertaken using as few model runs as possible.

[6] We demonstrate the utility of this strategy by applying it to an ongoing modeling study at Yucca Flat, Nevada Test Site, USA. Although the particulars of the site are unique, this application is representative of many real-world scenarios where management decisions are based, in part, on predictions derived from models, yet where the models are too CPU-intensive to benefit from standard Bayesian uncertainty analysis approaches. The model has 252 parameters, fewer than would normally be accommodated by highly parameterized methods such as those described by (among others) *Gomez Hernandez et al.* [2003], *Woodbury and Ulrych* [2000], and *Kitanidis* [1999] which rely on the use of sophisticated and often model-specific software for computation of numerical derivatives for analysis of parameter and predictive uncertainty, yet many more than are often employed in groundwater modeling applications. The model has long run-times and would therefore be unusable in conjunction with non-gradient uncertainty analysis methods.

[7] In the course of estimating the uncertainty associated with predictions made by this model, we compare two powerful yet very different methodologies using the simpler surrogate model: null-space Monte Carlo (NSMC) analysis and Markov chain Monte Carlo (MCMC) analysis. Because of the recent development of the former of these methods, they have not been compared before. The outcome of this comparison should be of great interest to practitioners who use either of the two methodologies in the face of significant parameter dimensionality. Finally, the understandings gained

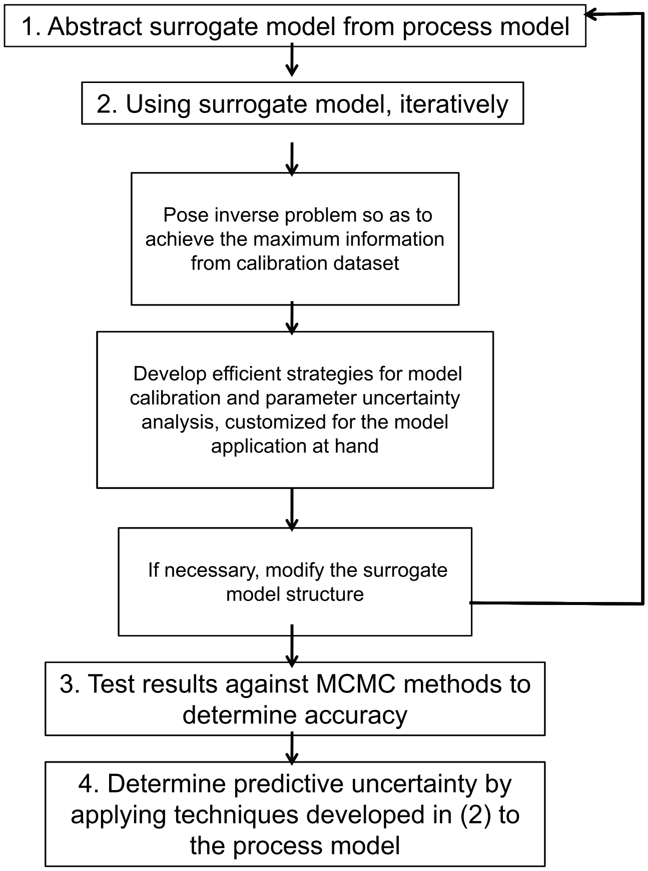
1. Abstract surrogate model from process model
2. Using surrogate model, iteratively
Pose inverse problem so as to achieve the maximum information from calibration dataset
Develop efficient strategies for model calibration and parameter uncertainty analysis, customized for the model application at hand
If necessary, modify the surrogate model structure
3. Test results against MCMC methods to determine accuracy
4. Determine predictive uncertainty by applying techniques developed in (2) to the process model

Figure 1. Outline of proposed methodology.

and refinements made in application of the uncertainty analysis methodology to the surrogate model are applied to the CPU-intensive process model. The end product of this procedure is a population of calibration-constrained parameter sets, together with predictions made by the model on the basis of these parameter sets that are of practical interest to site managers and stakeholder groups.

[8] This paper is organized as follows. First we briefly describe the general methodology, as well as details of the two algorithms used for uncertainty quantification. Then we describe the site-specific practical application of the methodology. This begins with a description of the process model and the surrogate model used as a simple approximation to it. Next, a description is provided of the iterative process of objective function design, calibration, model evaluation, and finally parameter uncertainty estimation using two very different algorithms. Results from the two algorithms are compared. Then we apply the more efficient of the two algorithms to the CPU-intensive process model. The process model is calibrated, and then subjected to post-calibration parameter and prediction uncertainty analysis. The outcomes of this analysis are then discussed and conclusions are drawn.

2. Uncertainty Estimation Methodology

[9] When faced with a practical groundwater application where a process model is developed which is detailed, highly parameterized, and CPU-intensive, it can be very difficult to do the work necessary to evaluate and calibrate the model, and to derive meaningful estimates of parameter and predictive uncertainty. To assist in this process, we propose the sequence of steps shown in Figure 1. The first step is to develop a fast surrogate model. The best strategy for accomplishing this will be highly problem-specific. Ideally, the surrogate model will be characterized by a nearly identical level of parameterization detail as the process model; furthermore the relationships between model parameters and the model-generated counterparts to calibration targets should be similar. However, as will be shown below, the exact mathematical relationship between surrogate and process model parameters need not be exactly known.

[10] Step 2 involves an iterative process of model calibration, evaluation, and parameter estimation (with uncertainty). An important task here is to establish a quantitative criterion for model and parameter adequacy. The weighted sum-of-squared-residuals statistic, Φ, is often used for this purpose where Φ is defined as

$$\Phi = \sum_{i=1}^{m} \left(w_i \left(c_i - c_{o,i} \right) \right)^2 \qquad (1)$$

where c_i are model outputs, $c_{o,i}$ are corresponding field or laboratory measurements, and w_i are weighting factors. If measurements are statistically independent and the dominant source of error is measurement error, weights (as defined in the above equation) are normally assigned as inversely proportional to assumed measurement error standard deviation. However, if structural error dominates model-to-measurement misfit, which is usually the case in groundwater model calibration, assignment of weights is less straightforward. Since structural error and its contribution to parameter and prediction error is so difficult to quantify, iterative calibration and model improvement must often be implemented in order to reduce structural error, thus minimizing its contribution to parameter and predictive error.

[11] Seen in this light, the calibration process serves two, somewhat distinct, purposes. First, by identifying the parameter vector $\hat{g}$ corresponding to $\Phi_{\min}$, important information is gained about likely parameter values. This is the standard inverse problem of model calibration. Second, by identifying the magnitude of $\Phi_{\min}$, and the nature and spatial relationships of residuals which constitute it, information is gained about model adequacy. If $\Phi_{\min}$ is unacceptably large, or if posterior parameter estimates are unreasonable, an inference may be made that structural error is too large and that the model should therefore be improved. This iterative process of optimization, model evaluation, and model improvement continues until model structural error is considered to be acceptably low.

[12] In Step 3 an assessment is made of the level of parameter uncertainty associated with the surrogate model, this being a function of prior knowledge of parameter uncertainty and additional constraints on parameters exercised through the calibration process. In the present case we employ two very different methods to assess parameter uncertainty using the surrogate model. The reason for using two methods is that only one of the two, NSMC, is efficient enough to later be applied to the process model (Step 4). However, to gain confidence in the validity of the results

obtained with NSMC analysis we compare its performance to a statistically rigorous MCMC-based approach, DREAM. As we will show, in our case parameter uncertainties computed by the two compare reasonably well. In the general case, however, if significant discrepancies arise, these should be investigated and understood before proceeding.

[13] The fourth and final step is to apply the NSMC methodology, tuned to the surrogate model in Steps 2 and 3, to the process model. This step goes beyond model calibration and parameter uncertainty estimation; ultimately a set of calibration-constrained predictions are provided.

[14] The two uncertainty estimation methodologies employed in this general methodology are now briefly described. The first and most efficient (in terms of its model run requirements) method is embedded in a suite of model calibration and uncertainty analysis tools which comprise the PEST [*Doherty*, 2009] suite. The NSMC methodology implemented by PEST (see *Tonkin and Doherty* [2009] for a description of this methodology) is used to generate a population of parameter sets, all meeting the criteria $\Phi <= \Phi^*$, where Φ^* is a user-specified criterion for "calibration adequacy," given knowledge of measurement noise and the expected degree of model structural error. The method is pragmatic, using a highly efficient subspace approach to generate a collection of calibration-constrained parameter sets. The first step in the process is model calibration, this yielding (if implemented correctly) something approaching a minimum error variance estimate $\hat{g}$ of the possible set of parameters g. Of necessity, this will be a simplified and smoothed parameter set, exhibiting only as much heterogeneity as can be supported by the calibration data set, this being an outcome of the regularization necessary to achieve uniqueness of an ill-posed inverse problem, as *Moore and Doherty* [2006] describe. The PEST suite of tools includes a number of algorithms for achieving this parameter set, including gradient-based methods such as Levenberg-Marquardt and global methods such as the Covariance Matrix Adaptation Evolutionary Scheme (CMAES), the latter being a population-based stochastic global optimizer [*Hansen and Ostermeier*, 2001; *Hansen et al.*, 2003]. Note that appropriate regularization constraints must be employed with these methods to achieve something approaching the minimum error variance parameter set $\hat{g}$ instead of any of the infinite number of other parameters sets that are also capable of minimizing Φ and/or of reducing it to a sufficiently low value.

[15] Through singular value decomposition (SVD) of the linearized model operator, the parameter set $\hat{g}$ is then decomposed into two orthogonal vectors, these being its projection onto the "calibration solution space" (this being comprised of p linear combinations of parameters with are informed by the calibration data set), and "the calibration null space" (comprised of d-p calibration-insensitive parameter combinations, where d is the total number of parameters). Demarcation between these two subspaces is deemed to occur at that singular value at which attempted estimation of orthogonal linear parameter combinations obtained through the SVD process on the basis of the calibration data set results in a gain rather than diminution of post-calibration uncertainty of either that linear combination of parameters, or of a prediction of interest (see *Doherty and Hunt* [2009] for details).

[16] Uncertainty analysis then proceeds with the generation of a set of random parameter realizations, g_m. Each parameter value within each set is drawn from a random distribution (uniform or lognormal) within specified bounds, with bounds and distributions chosen to ensure that all parameter values are physically reasonable. For each g_m, $\hat{g}$ is subtracted; the difference is projected onto the calibration null-space, then $\hat{g}$ is added back in, resulting in a transformed parameter vector g_m'. If the inverse model were perfectly linear, each g_m' would represent a calibrated parameter set. Because of nonlinearity, however, each g_m' must be adjusted through re-calibration. To achieve this, each g_m' is used as a starting point for further optimization. Collectively this results in a set of S parameter sets, $\hat{g}_i$, $i = 1,S$, all of which realize objective functions which are low enough ($\Phi < \Phi^*$) for these parameter sets to be considered to "calibrate" the model. The computational demands of this pragmatic process are relatively light due to the use of the null-space projection operation described above in generating nearly calibrated random parameter fields, and due to the fact that re-calibration of these parameter fields can make use of pre-calculated sensitivities applied only to parameter solution space components. Subsequent adjustment of these solution space components so that the objective function falls below a specified value (that reflects the degree of measurement/structural noise associated with model-to-measurement misfit) ensures that variability of these components is introduced in accordance with the necessity to reflect post-calibration variability that they inherit from measurement/structural noise. It should be noted that, even though methods used to subdivide parameter space into solution and null subspaces are inherently linear, the methodology imposes no constraints on parameter variability and does not presume model linearity. The higher the parameter variability and model nonlinearity, the more model runs be invested in post-null-space-projection model re-calibration for each parameter realization.

[17] A disadvantage of the NSMC method is that, in difficult parameter estimation contexts, parameter sets achieved through application of the method do not necessarily constitute a sample of the posterior probability density function of the parameters in a strictly Bayesian sense. Therefore, for comparison purposes we also apply a second method in Step 3, which is implemented in the DREAM package (see references below). DREAM is a Markov chain Monte Carlo (MCMC) scheme adapted from the Differential Evolution–Markov Chain (DE-MC) method of *ter Braak* [2006]. While not particularly developed for efficiency in highly parameterized contexts, it is capable of producing posterior probability density functions of the parameters that are exactly Bayesian, even in highly parameterized environments characterized by multiple objective function minima and multimodal posterior distributions, where it is able to continually update the orientation and scale of the proposal distribution employed by its underlying Metropolis algorithm while maintaining detailed balance and ergodicity [*Vrugt et al.*, 2008a, 2008b, 2009b]. Contrary to the PEST approach to uncertainty assessment described above, DREAM combines calibration and uncertainty analysis within a single algorithm in which it samples the post-calibration probability density function of the parameters. In summary, DREAM runs N different Markov chains simultaneously in parallel that are typically initialized from a

uniform prior distribution with bounds of the parameters that are chosen based on physical reasoning. The states of the N chains are denoted by the d-dimensional parameter vectors $g_1, \ldots, g_N$,. At each time, the members of the individual chains form a population, stored as an $N \times d$ matrix $\mathbf{G}_t$. DREAM evolves the initial population, ($\mathbf{G}_t$; $t = 0$) into a posterior population using a Metropolis selection rule to decide whether to accept the offspring or not. With this approach, a Markov chain is obtained, the stationary distribution of which is the posterior parameter probability distribution. The proof of this is presented by *ter Braak and Vrugt* [2008] and *Vrugt et al.* [2009b]. After a so-called burn-in period, the convergence of a DREAM run can be monitored using the statistic of *Gelman and Rubin* [1992], which compares the variance within and between the chains. The use of a formal convergence criterion defines a significant difference between the MCMC approach of DREAM and the NSMC technique implemented by PEST. Prior to the present investigation, DREAM had not been tested in highly parameterized contexts; nor had PEST's performance been compared with the theoretically more rigorous, though computationally more demanding, DREAM approach in difficult applications such as that described here. Note that DREAM uses explicit boundary handling to force the parameters to stay within their prior defined bounds. This approach is essential to avoid sampling out of bounds and is especially designed to maintain detailed balance.

3. Example Application

[18] Here we demonstrate the sequence of steps outlined above using a real-world example. The context is the modeling of groundwater flow and transport within saturated volcanic rocks in Yucca Flat, Nevada Test Site, USA (Figure 1), where a total of 659 underground nuclear tests were conducted between 1951 and 1992 [*Fenelon*, 2005]. For the purposes of this paper, we focus on one particular prediction required of the model: the enhanced flux from the volcanic rocks into the lower aquifer due to underground nuclear testing. This enhanced flux has obvious implications for radionuclide transport. Quantification of the uncertainty associated with this prediction is a very important outcome of the analysis. Here we will demonstrate a method for quantifying uncertainty and then test its effectiveness using a post-audit (section 4.5.1).

3.1. Process Model Development

[19] The complex process model simulates groundwater flow in a shallow portion of the saturated zone in Yucca Flat which is characterized by a complex layering of volcanic rocks, dissected by numerous faults, some quite large and hydrologically significant. A unique aspect of the hydrology of this site is the large and abrupt changes in pore pressures and rock properties that occur subsequent to each nuclear test [*Laczniak et al.*, 1996]. A cavity is formed around the working point of each test due to vaporization and melting of rock, which later fills with rubble as the rocks above the cavity collapse, forming a chimney and surface crater. Each of these features has very different hydrologic properties from the pre-testing rocks which form the host material. Upon detonation of a nuclear device, pore pressures are instantaneously increased within tens of meters of the cavity due to compressive stresses associated with the explosion and pore-space compaction. Hydrofracturing occurs in places where stresses reach very high levels. At some locations, elevated pore pressures (and consequently enhanced hydraulic gradients) have persisted for decades [*Fenelon*, 2005].

[20] To appropriately model groundwater flow in this highly dynamic aquifer system, a complex process model was constructed, which sequentially couples a "testing-effects" module (to simulate abrupt head alterations) with a 3-D transient groundwater flow module that simulates long-term groundwater movement in the highly heterogeneous system.

3.1.1. Numerical Details

[21] The first of the two submodels that are coupled to form the process model is a 3-D, transient groundwater flow model developed specifically for the central portion of Yucca Flat. The dimensions of this model are roughly 11 km × 30 km × 1 km; its outline is shown in Figure 2. Several model grids have been developed for different purposes using the grid generation package LaGrit [*Trease et al.*, 1996]; the medium-resolution grid used for the present uncertainty study has 247,464 nodes. The standard transient groundwater flow equation (simplified here, for brevity):

$$S_s \frac{\partial h}{\partial t} = \frac{\rho g}{\mu} \left[\frac{\partial}{\partial x} \left(\kappa_x \frac{\partial h}{\partial x} \right) + \frac{\partial}{\partial y} \left(\kappa_y \frac{\partial h}{\partial y} \right) + \frac{\partial}{\partial z} \left(\kappa_z \frac{\partial h}{\partial z} \right) \right] + Q \quad (2)$$

is solved using the Finite Element Heat and Mass Transfer (FEHM) [*Zyvoloski et al.*, 1997] porous flow simulator. Here, h is hydraulic head, t is time, κ denotes intrinsic permeability (which can be anisotropic), ρ is density, g signifies the gravitational constant, μ is the viscosity, S_s represents the specific storage, and Q is a source or sink term. In the present case the only specified sink or source is a spatially uniform recharge rate applied at the water table; water flows from the bottom of the model domain (into the aquifer below) through specified head nodes. Rock properties (S_s, κ) are spatially distributed according to a deterministic hydrogeologic framework model (HFM) [*Drellack*, 2006] which defines the geometry of 11 units and more than 100 faults within the flow model domain. Faults are lumped into seven categories for the purpose of parameter value assignment.

[22] The second model component is the testing-effects module referred to above, which simulates instantaneous changes in permeability and porosity in the vicinity of a test, as well as increased pore pressures due to elastic and inelastic rock deformation [*Davis*, 1971; *Garber and Wollitz*, 1969; *Halford et al.*, 2005; *Hawkins et al.*, 1989; *Knox et al.*, 1965; *Reed*, 1970]. The model also simulates the creation, immediately following each underground nuclear test, of a high permeability cavity and chimney with radius r_c. Within the region $r < r_c$ hydraulic heads are assumed to be hydrostatic, whereas for larger radial distances $r_c < r < R$ hydraulic head increases instantaneously according to the following relationship, from *Halford et al.* [2005]:

$$dh(r) = H_o(1 - r'/R) \quad (3)$$

where H_o is a test-specific value. In accordance with historical observations that test-effects are often more pronounced laterally than vertically because of rock anisotropy, we introduce a modified distance parameter, $r' = r(1 - \frac{dz}{\gamma})$, where dz is the vertical distance to the working point of the test and γ is an unknown scaling parameter subject to cali-

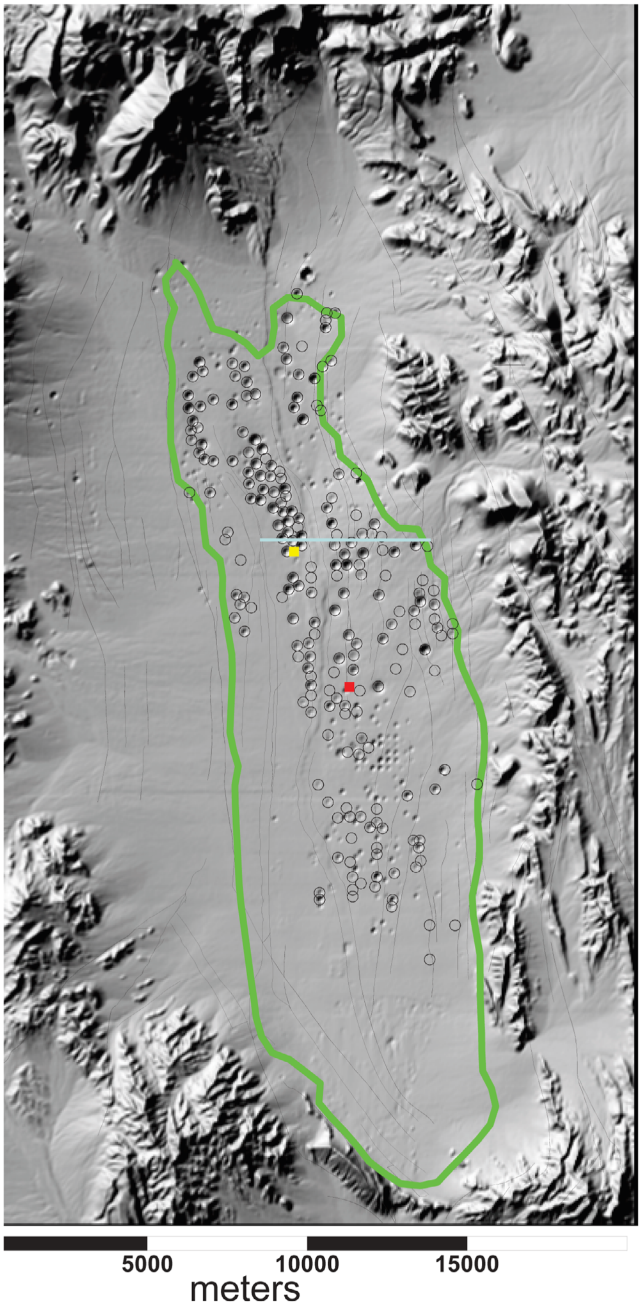

Figure 2. Yucca Flat, Nevada Test Site, Nevada. Circles indicate locations of underground nuclear tests detonated within the model domain. Green line shows outline of numerical model. Black lines show faults. The red square is TW-7; the yellow square is UE-4t. The horizontal light blue line is the location of cross-section indicated in Figure 15.

bration. The smaller the value of γ, the more anisotropic is the effect of an underground nuclear test.

[23] Local rock permeability can be enhanced or reduced by the testing-effects model, this depending on the magnitude of the head increase and on its relationship to lithostatic load. Out to a distance of $\frac{r}{r_c} \leq F$, a permeability reduction factor, ξ, is applied if $dh > H'$. If the pressure at any location exceeds lithostatic pressure by more than a factor λ, the rock breaks and the porosity and permeability are increased. In total, this testing-effects module has k+6 parameters, where k represents the total number of underground tests within the model domain; this is equal to 212 in the present study.

[24] The FEHM and testing-effects models are iteratively coupled to form the overall process model which can then be used to simulate transient and long-term groundwater processes during and after the period of testing (1952–1992). The combined FEHM and testing-effects model is computationally very demanding, requiring between 1 and 3 h to complete a single run on a 3.6 GHz processor. The exact run time depends on the time discretization used to solve the pertinent partial differential equations in FEHM. The combined model has 242 model parameters; these are summarized in Table 1.

3.1.2. Calibration Objectives

[25] A rich and complex data set of hydrographs is available for this site. By reviewing the extensive database developed by *Fenelon* [2005], we were able to filter out unreliable measurements and only use head measurements which are trustworthy and thus unaffected by wellbore and/or drilling affects. A total of 583 head measurements collected at 60 different wells were selected for inclusion in this calibration data set, spanning the time period from 1958 to 2005. Large variations exist in the frequency of data collection in the various wells. One example of a well with head measurements spanning a relatively long time is TW-7 (see Figure 2); measurement data from this well are depicted in Figure 3. Sharp increases in measured head immediately following underground nuclear tests are apparent in Figure 3; each increase is followed by a period of relaxation. An additional complicating factor is that many wells were sampled only once or very few times. Implications of this for the design of the calibration objective function are discussed below.

[26] Due to the spatial and temporal complexities of this data set, it is impossible to determine a priori which of the 242 model parameters may, to any extent, be constrained by model calibration. However, even if the number of parameters constrained by the calibration data set is small, the information pertaining to those parameters acquired through the calibration process could be quite valuable in reducing the uncertainty of key model predictions. Additionally, calibration to the hydrologic data set may reveal important model inadequacies that should be addressed. Meanwhile,

Table 1. Summary of Parameters[a]

Parameter Type	Surrogate	Process	Free
Testing effects			
H'	1	1	1
$f_{\max}$	1	1	1
Vad1, 2, 3	3		
ξ	1	1	1
γ	1	1	1
α	1		
R	1	1	1
H_o	212	212	28
Rock properties			
κ	10	20	9
S_s		2	2
K_z		2	2
Recharge		1	
H_{is}	10		
Total	241	242	

[a] Free parameters are described in section 3.5.

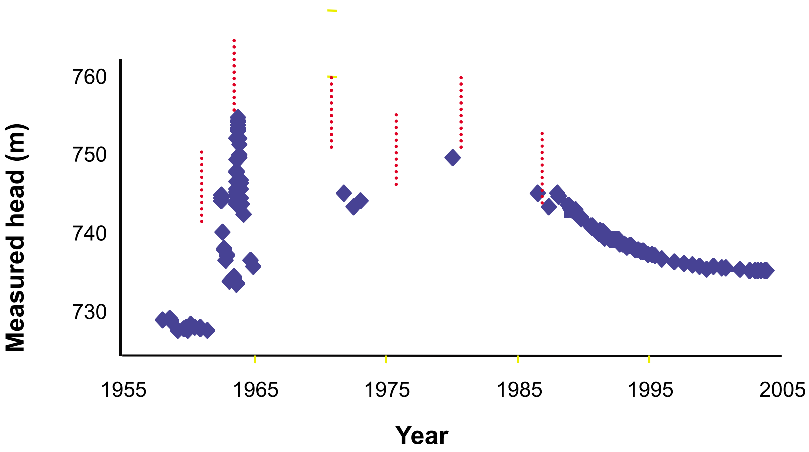

Figure 3. Measured heads at Test Well 7, showing response of well to nearby tests and subsequent relaxation.

our high level of parameterization respects the philosophy that by keeping the level of parameterization commensurate with the complexity of the physical problem, rather than commensurate with the (relatively low) information content of the calibration data set, we will ultimately obtain the best estimates of those parameters that are available to us (provided that appropriate regularization mechanisms are introduced to the solution process of the resulting ill-posed inverse problem). This philosophy also grants us the ability to quantify the uncertainty associated with these parameter estimates and with predictions which are sensitive to them [*Hunt et al.*, 2007].

3.2. Surrogate Model Development and Analysis

[27] To implement the process depicted in Figure 1, we developed a fast-running surrogate model of the complex and CPU-intensive process model. Design of the surrogate model is based on the very simple assumption that for a specific head increase at the location of an explosion, H_o, the immediate head perturbation imposed at a nearby well will be determined by only three factors: 1) distance of the well from the test, 2) time elapsed since the test, and 3) properties of the rock at the point of measurement. The head perturbation is calculated using equation (3). To approximate the transient hydrologic process of depressurization due to groundwater flow away from the pressurized zone, this effect is assumed to decay over time at an exponential rate (κ_s/α), where κ_s is specific to each rock type, s. The principle of superposition is used to calculate the cumulative impact of k tests, where $k = 212$. The resulting equation is as follows:

$$h_i(x,y,z,t) = h_{i(s)} + \sum_{n=1}^{k} dh_n(r) e^{-\frac{\kappa_S(t-t_n)}{\alpha}} \tag{4}$$

where h_i is the head at the ith well and $h_{i(s)}$ is the initial (pre-1950) head in this respective well.

[28] To test the efficacy of this equation in simulating observed groundwater behavior, a number of very simple homogeneous radially symmetric models of groundwater flow based on equation (2) were used to simulate instantaneous pressurization due to a single test (equation (3)) and subsequent relaxation over time. Simulation results were in excellent agreement with equation (4); furthermore a direct relation was evident between parameter κ_s of (equation (4)) and permeability κ featured in equation (2). However, this simple relationship quickly degraded where more complex hydrostratigraphic conditions were represented in these radial simulation models. For this reason, we do not expect there to be a simple relationship between parameters featured in equation (4) and parameters of the full process model which incorporates very complex hydrostratigraphy.

[29] Computation of head using equation (4) does not require an iterative procedure and variable time stepping, and can be used to generate all 60 hydrographs represented in the calibration data set in less than one second on a 3.6 GHz processor. The model has a similar number of parameters to those employed by the process model as summarized in Table 1. One difference is that there are fewer rock-specific values of permeability (κ). This is due to the fact that the surrogate model only considers the rock type at the observation point, whereas the process model considers all rock types within the model domain. This is only one of many simplifications of the surrogate model which prevent an explicit mapping of parameters from the surrogate to process model. The relationship between parameters and outputs of the surrogate model is highly nonlinear, as Figure 4 demonstrates for one example parameter/observation pair; hence, as for the process model, parameters employed by the surrogate model may be difficult to estimate through calibration. Meanwhile, because of its computational efficiency and similarity (to first order) to the process model, use of the surrogate model provides an ideal opportunity for developing strategies for effective calibration and uncertainty estimation that can ultimately be applied to the process model.

3.2.1. Objective Function Definition and Model Assessment

[30] Using the surrogate model, we now implement Step 2 depicted in Figure 1. In implementing this step, our goal was to design the objective function such that the maximum amount of information could be gleaned from the calibration data set through forcing the model, if possible, to reproduce key features in the data set through minimization of this objective function. Beginning with the most simple approach, we used the objective function defined in equation (1), with

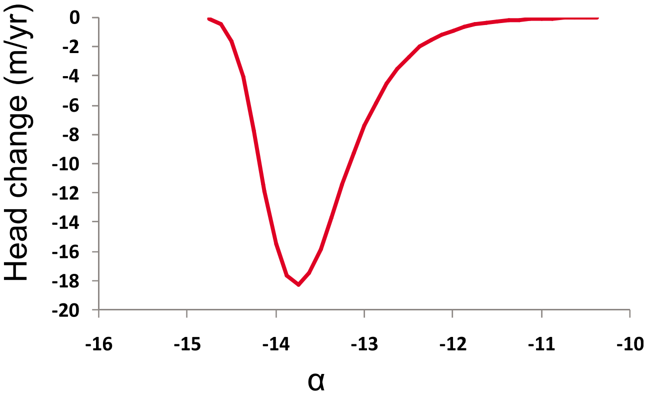

Figure 4. An example of nonlinearity between model parameters and outputs (dimensionless parameter α and calibration target, *dh*/*dt* at well U-3kx).

all 563 head observations given equal weight, this reflecting a presumed uniform measurement error of 0.1 m. We then iterated between parameter estimation, objective function refinement, and model improvement, using the PEST suite of calibration tools to minimize Φ in each case. Early in this process PEST quickly found a solution that explained over 90% of the variance in the head data, (which is remarkable considering the simplicity of the surrogate model and the complexity of the data set), but was unable to further improve the correspondence between measured and modeled heads. Detailed inspection of PEST outcomes revealed that the goodness-of-fit achieved for the mean head at any particular well was strongly controlled by the number of observations made at that well. Similarly, goodness-of-fit within a particular subregion of the model domain was strongly affected by the density of observations in that subregion. Furthermore, temporal trends at any given well were generally poorly reproduced, since in many cases the temporal variance at a given well was much smaller than the overall spatial variance between wells, and so PEST placed much more focus on capturing the latter. This is undesirable, since rock properties of interest, such as permeability, can often be much more reliably deduced from stress-induced head changes (such as a pump test) than from static head measurements over space that are also a function of uncertain recharge and even more uncertain boundary conditions. As *Doherty and Welter* [2010] demonstrate, for maximum efficacy of the calibration process in providing reliable estimates of these parameters, the objective function must be formulated in such a way as to therefore accentuate these head changes induced by known stresses.

[31] We therefore experimented with several modifications to the objective function. Best results were achieved by making the following adjustments. First, weighting was adjusted to accommodate variations in spatial observation density (and, by inference, density of observations in different hydrostratigraphic units), so that approximate equality of weighting normalized by area was obtained. Then hydrographs were filtered so that data representing time periods in which heads changed linearly with time (or not at all) were replaced by two numbers: 1) an initial head for that time period, and 2) *dh*/*dt* for that time period. In some cases, where multiple measurements were taken during a linear-trending time period, this substitution greatly reduced the total number of observations. Finally, weights were adjusted to apportion approximately equal weighting to derivatives (dh/dt) and to absolute heads. The resulting objective function was comprised of 361 observations, constituting a blend of absolute head values and derivatives.

[32] In a similar iterative fashion, we tested and improved conceptual model elements common to both the surrogate and process models. For example, we experimented with allowing the parameter *R* (equation (3)) to be spatially distributed. This change lead to lower values of Φ but produced clearly non-physical simulations of system behavior. As a consequence, we removed spatial dependency in *R*. We also found that the anisotropic test-effects parameter, γ, could be used to great advantage for enhancing model-to-measurement fit over portions of the model domain. Unfortunately, however, a large value was estimated for this parameter (meaning no vertical anisotropy) when attempts were made to fit heads over the entirety of the model domain. It is thus apparent that a simple model of anisotropy has no value; it either must be made much more complex or ignored completely. We adopted the latter option. On other occasions we experimented with various alternative strategies to that of estimating H_o independently for each test. For example we tried to express H_o as a function of maximum announced yield of each test [*Fenelon*, 2005]. We found no advantage in this approach.

[33] Another outcome of this iterative process was that we began to appreciate the sensitivity of the objective function value to the quality of the result, as assessed by visual comparison of measured and modeled hydrographs. It was clear that objective function values ranging from a few hundred to approximately 1,000 gave rise to only subtle visual differences in model-to-measurement misfit. However, misfit noticeably deteriorated above a value of about 1,000. This qualitative relationship between objective function value and parameter field acceptability was employed later in null-space Monte Carlo analysis. It should be noted that use of a subjective likelihood function in this fashion is not uncommon practice in environmental model calibration, where model-to-measurement misfit is dominated by structural noise. It is embraced by the GLUE methodology [*Beven*, 2009] and is given justification by *Beven* [2005], *Beven et al.* [2008], and *Doherty and Welter* [2010].

3.2.2. Parameter Estimation and Uncertainty Analysis

[34] It is at Step 3 of the procedure illustrated in Figure 1 that we applied our two uncertainty analysis methodologies. First, we assigned reasonable a priori bounds to all 242 parameters based on physical arguments and data from previous studies [*Halford et al.*, 2005; *Kwicklis et al.*, 2006; *Stoller-Navarro Joint Venture*, 2006]. All prior parameter probability distributions were assumed to be uniform. We then applied methodologies provided by both the PEST and DREAM packages to parameter estimation and uncertainty analysis of the surrogate model.

[35] Because of the large dimensionality and high degree of nonlinearity of this model, use of PEST was slightly more complicated than use of PEST in simpler calibration contexts. To ensure avoidance of local minima, the calibration process was repeated with 60 random starting values of *g*. In each case several optimization algorithms were employed. These were applied sequentially, each beginning with the result from the previous step. These were 1) CMAES 2) truncated

Table 2a. Summary of Calibration and Uncertainty Analysis Forward Run Requirements Calibration[a]

	Number of Model Runs
CMAES	152542
SVD	206942
AUI	43102
Total	402586

[a] Total for 60 inversions.

singular value decomposition (SVD), and 3) "automatic user intervention." CMAES is a global optimization algorithm described in detail by *Hansen and Ostermeier* [2001] and *Hansen et al.* [2003]; a version of the CMAES algorithm is provided with the PEST suite. Truncated SVD as a mechanism for stable solution of inverse problems is described in many texts, including *Menke* [1984] and *Aster et al.* [2005]. The last of these, "automatic user intervention," is described by *Skahill and Doherty* [2006]; this achieves numerical stability through sequential, temporary, fixing of insensitive parameters while promoting continual parameter movement in difficult objective function terrains through attempting objective function improvement in a number of orthogonal directions. (Note that in many cases this last step was either unnecessary or achieved very little objective function improvement.) The number of forward runs required for each of these steps are listed in Tables 2a and 2b. In total, the entire process required about 400,000 runs and produced 60 parameters sets ($\hat{g}_i$) with Φ values ranging from 203 and 13,000. Of these, those corresponding to the five lowest values of Φ (ranging from 203 to 410) were carried forward into the uncertainty analysis.

[36] For each of the five $\hat{g}_i$ parameter sets the following steps were taken. Calibration solution and null-spaces were defined using singular value decomposition according to the method established by *Doherty and Hunt* [2009]; in all cases the dimensionality of the solution space was calculated to be about 24. Three hundred random parameter sets were then generated and subjected to null-space projection in the manner described previously. Re-calibration of each parameter set was then effected. As already discussed, this re-calibration step was very efficient because parameter sensitivities calculated during the previous calibration process were employed for the first (and often only) iteration of parameter adjustment. If after two iterations an objective function of 1,000 had not been attained, the parameter estimation exercise was aborted and the parameter set rejected. The number of forward runs required by this process are listed in Tables 2a and 2b; in total, this entire process required approximately 200,000 model runs. From one seed ($\hat{g}$, corresponding to $\Phi = 404$), 172 parameter sets ($\hat{g}_i$) were produced, each meeting the criterion $\Phi < 1,000$. We compared these results to those generated using the four other seeds and found that they produced very similar results. Obviously, a larger ensemble of parameter sets could have been produced by combining all sets produced by each of the five $\hat{g}$i seeds and/or by continuing the NSMC process for any one seed. However, we used only the results from one seed in order that our analysis would be comparable to that undertaken using the process model (presented later) where, due to the computational demands of using that model, only one seed could be utilized.

Table 2b. NSMC Analyses Based on Five Calibrated Models[a]

	Calibrated Models				
	1	2	3	4	5
Number of random seeds	300	300	300	300	300
Number of Forward Runs Required					
Definition of null and solution spaces	242	242	242	242	242
Recalibration	37328	42864	45231	32943	41423
Total	37570	43106	45473	33185	41665
Resulting number of parameter sets meeting criterion $\Phi < 1000$	172	162	203	184	151

[a] In all cases 242 runs were undertaken on the basis of the optimized parameters which were an outcome of one of the previous calibration exercises as described in the text. These were required for calculation of a Jacobian matrix on which basis subdivision of parameter space into orthogonal complimentary solution and null-spaces was undertaken.

[37] The DREAM algorithm described in section 3 was not customized for this particular application, other than to specify the value for the number of Markov chains, *N*. Numerical results presented by *Vrugt et al.* [2009b] demonstrate that $N > 1/2d$ is appropriate. For all the calculations presented herein, we set $N = 250$ and used 50 different computational nodes of the LISA cluster at the SARA computing center at the University of Amsterdam, The Netherlands for posterior inference. DREAM required about 50 million model evaluations to generate 100,000 parameter sets that represent samples from the posterior probability density function. As will be shown below, this approach did not yet find the lowest attainable value of the objective function as found by PEST (presumably the global minimum). This demonstrates that this best solution is associated with a small, perhaps negligible probability mass, and that DREAM is inclined to converge to the larger area of attraction encompassing the full range of posterior parameter likelihood.]. Here we will present the results of DREAM with $N = 250$ parallel chains. Increasing the value of *N* did not alter these results.

3.2.3. Comparison of Results From Surrogate Model Analysis

[38] Objective functions attained through use of the two methods are compared in Figure 5. Median values of Φ are similar; however variation in the PEST results was much broader. Interestingly, the PEST method identified the lowest value of Φ (186). Considering the difficulties associated with this problem, it is noteworthy that both methods were able to reduce Φ from values of 1.E6 typically associated with randomly generated parameter sets, to values as low as several hundred.

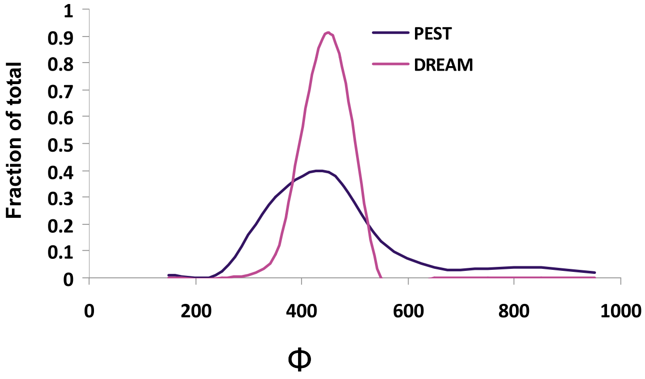

Figure 5. Distribution of Φ values for the two methods.

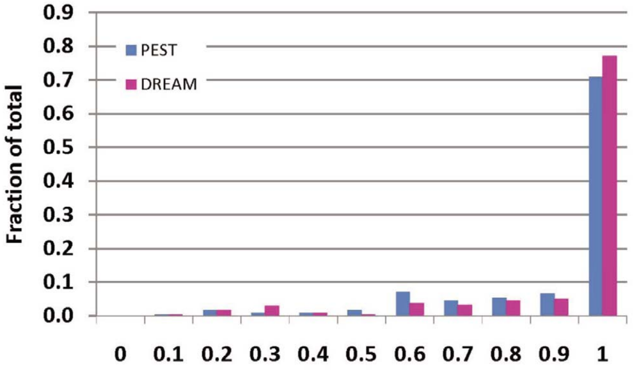

Figure 6. Histogram of the computed ratio of posterior to prior range for all 263 parameters. For over 70% of the parameters, there was no range reduction (ratio = 1.0).

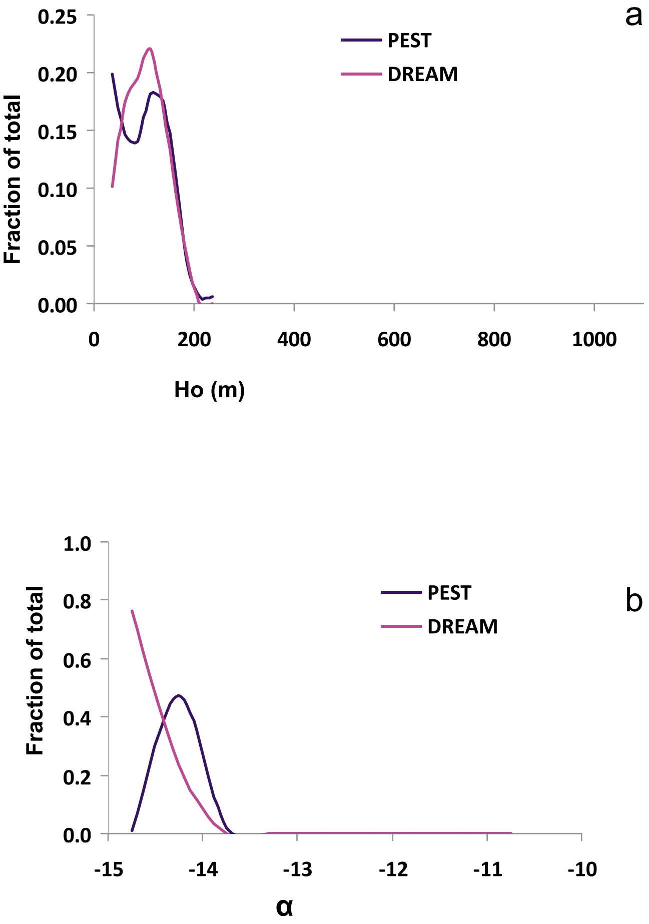

Figure 7. Comparison of posterior distributions for (a) α and (b) H_o (Aleman test). The X range reflected the prior range.

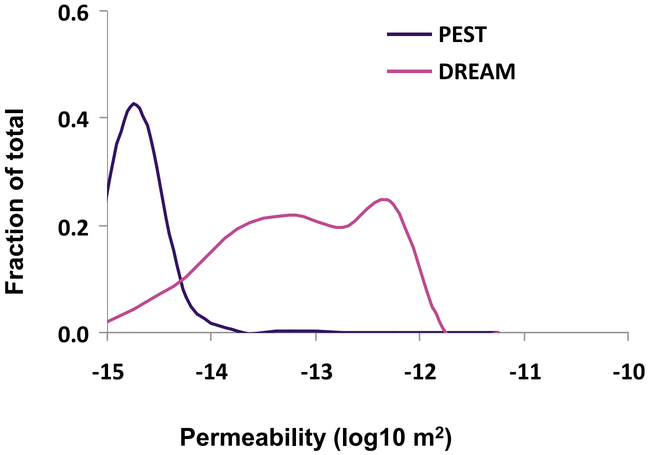

Figure 8. Comparison of posterior distributions for κ, Timber Mountain Volcanic Tuff Aquifer. The x axis ranges are defined by prior distributions.

[39] It was expected that the calibration process would constrain only a small fraction of model parameters; by comparing the posterior-to-prior ranges of all parameters obtained using both methods we see confirmation of this. As shown in Figure 6, both methods demonstrate that for 70% of the parameters, the posterior range is not significantly smaller than the prior range. Detailed examination of posterior distributions generated by each method revealed both similarities and differences, as shown in Figure 7, which illustrates frequency distributions for two parameters. For each parameter shown here, both methods yield a posterior parameter range that is significantly narrower than its prior range (the range of the x axis in each of these plots reflects prior parameter bounds), and in Figure 7a it is evident that the characters of the two posterior distributions are remarkably similar. However, as shown in Figure 7b, for the α parameter the lower end of the distribution is marked by the parameter bound in the case of the DREAM distribution, and by a drop in likelihood in the case of the PEST distribution; the reason for this is unknown. In Figure 8 an example is provided where the two results differ significantly. In this case the DREAM posterior distribution is commensurate with the prior. In contrast, PEST considers this parameter to be strongly skewed toward its lower bound; once again, reasons for these differences are unknown. However, it should be noted that because the prior parameter bounds were constructed so as to ensure that any estimated parameter would be physically reasonable, no physical arguments can be made in favor of either PEST or DREAM producing more “realistic” parameter estimates.

[40] The median value of the posterior distributions generated by the two methods was compared for each parameter. To facilitate comparisons among parameters with very different magnitudes, the posterior values were first normalized with respect to the prior parameter ranges. As a measure of the degree to which the posterior median value is constrained by the calibration process, the ratio (σ^*) of the posterior standard deviation to the prior was also calculated. Ratios approaching 1.0 indicate parameters poorly constrained by calibration. In Figure 9 a comparison of the median values forthcoming from the two methods is shown, with symbols colored according to the σ^* calculated by PEST (Figure 9a) and DREAM (Figure 9b). The red and black symbols indicate values with significantly reduced standard deviations. Most symbols cluster near the center of Figure 9, and have light blue color. These are parameters that the method did not constrain, implying therefore that the prior and posterior standard deviations are similar, and that posterior medians therefore coincide with prior medians in the center of the

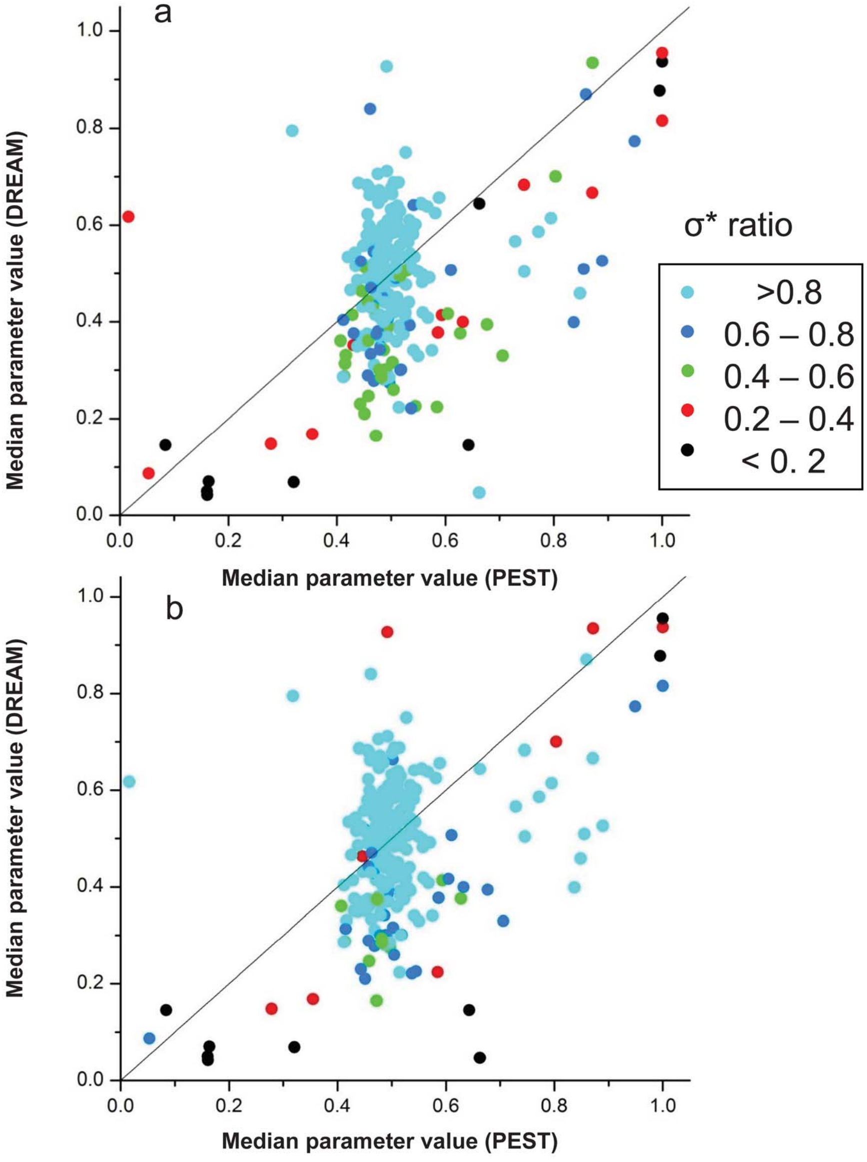

Figure 9. A comparison of median normalized parameter values estimated by PEST and DREAM. Color indicates the ratio of posterior to prior variance according to (a) PEST and (b) DREAM.

prior range. There are some light blue symbols, however, with normalized median values significantly different from 0.5. The reasons for this are not completely understood. One plausible explanation is that the sampling of parameter space is insufficiently dense (due to the large number of parameters) and that if the sampling density had been much larger, these poorly constrained parameters would possess normalized median values of 0.5. Clearly, in a high-dimensional problem with numerous local minima such as this, the normalized median values estimated by these methods may not be reliable for poorly constrained parameters. This may be particularly the case for the DREAM method, which has a tendency to yield greater spread in the normalized median of poorly constrained parameters.

[41] Most of the significant departures from the 1:1 line depicted in Figure 9 are for parameters that are not considered well-constrained by the calibration data set by one or both methods; for the reasons described above in these cases the discrepancy is not important. Table 3 lists those parameters considered well-constrained by calibration ($\sigma^* < 0.5$) by both methods. As shown in Figure 10, the median values compare moderately well; however there is a tendency for DREAM medians to be lower than those of PEST. We expect that those parameters listed in Table 3 which are shared by the process model will be important in the process model calibration and uncertainty analysis as well. This type of parameter estimability information is a very useful outcome of analysis based on the surrogate model.

[42] Overall, there are some interesting differences in the degree to which each method was able to reduce parameter standard deviations. Figure 11 compares PEST and DREAM results in this regard. The standard deviation ratio for both methods clusters near 1.0 (i.e., no reduction). For parameters with $\sigma^* < 1.0$ according to one or both methods, in most cases DREAM reduced the standard deviation less than PEST. One reason for this may be that the PEST

Table 3. Parameters That Both Methods Reduced Standard Deviation by at Least 50%.

	Normalized Median		Standard Deviation Reduction		Ratio of Posterior to Prior Range	
	PEST	DREAM	PEST	DREAM	PEST	DREAM
Test effects	0.64	0.15	0.10	0.06	0.14	0.07
Permeability	0.08	0.15	0.10	0.06	0.14	0.07
Test effects	0.16	0.04	0.01	0.15	0.03	0.23
Permeability	0.16	0.05	0.12	0.17	0.30	0.20
Permeability	1.00	0.88	0.06	0.17	0.12	0.21
H_o	0.32	0.07	0.18	0.14	0.24	0.17
Permeability	0.16	0.07	0.16	0.19	0.42	0.24
$H_{i(s)}$	1.00	0.94	0.03	0.20	0.12	0.28
Permeability	0.28	0.15	0.25	0.23	0.80	0.31
H_o	0.35	0.17	0.22	0.31	0.46	0.36
$H_{i(s)}$	1.00	0.95	0.34	0.19	1.00	0.26
H_o	0.45	0.46	0.44	0.21	0.53	0.31
H_o	0.87	0.93	0.46	0.32	0.66	1.00
H_o	0.48	0.29	0.44	0.47	0.52	0.60
H_o	0.59	0.41	0.39	0.49	0.53	0.57
$H_{i(s)}$	0.47	0.16	0.43	0.50	1.00	0.78

methodology is explicit in exploring the "null space," this being that portion of the parameter space which is not constrained, or is only partially constrained, by the calibration data set. In contrast, MCMC methods only exploit the unconstrained variability offered by the null-space in the natural course of sampling full parameter space. Where there are many insensitive parameters and/or insensitive parameter combinations, an extremely large number of model runs is required for exhaustive sampling of the null-space in this way. MCMC methods, which are not specifically focused on doing so, may thereby overestimate the uncertainty associated with parameters whose projections onto the null-space are large.

3.2.4. Discussion

[43] The surrogate model allowed facile iteration between objective function definition, model evaluation, and calibration. It also allowed us to apply and compare two techniques for estimating parameter uncertainty. Generally, the parameter estimation and uncertainty analysis results forthcoming from these two very different methods are remarkably similar. Nevertheless, there are some notable differences between the two. For this highly parameterized, highly nonlinear, ill-posed problem, neither method can provide an unequivocal guarantee that it has sampled the true posterior parameter distribution, or has identified the global minimum of Φ. This exercise has, however, demonstrated that each method is capable of providing consistent estimates of parameter uncertainty and of providing samples from posterior parameter distributions that, in the main, are consistent with each other. As expected, the MCMC method encapsulated in DREAM was much less efficient, requiring several million forward runs to converge to a limiting posterior distribution. The method was not customized for this particular problem, and perhaps could have been made more efficient if it had been. For instance, significant efficiency improvements could have been made by generating candidate points in each individual chain using an archive of past states. This would require far fewer chains to be run in parallel ($N \ll 250$), which significantly speeds up convergence to a limiting distribution. Furthermore, for high parameter dimensions, it would certainly be desirable to use dimensionality reduction sampling in DREAM. This, in combination with multiple-try Metropolis search and Metropolis adjusted Langevin sampling, which uses the gradient of the likelihood function to sample preferentially in direction of largest objective function improvement (violating detailed balance), should significantly increase computationally efficiency of DREAM for highly parameterized models. Conceptually, such an approach could be implemented in a way that is automatic and therefore requires no user intervention, in harmony with the DREAM design philosophy. Such a method is presently under development; results will be reported in due course.

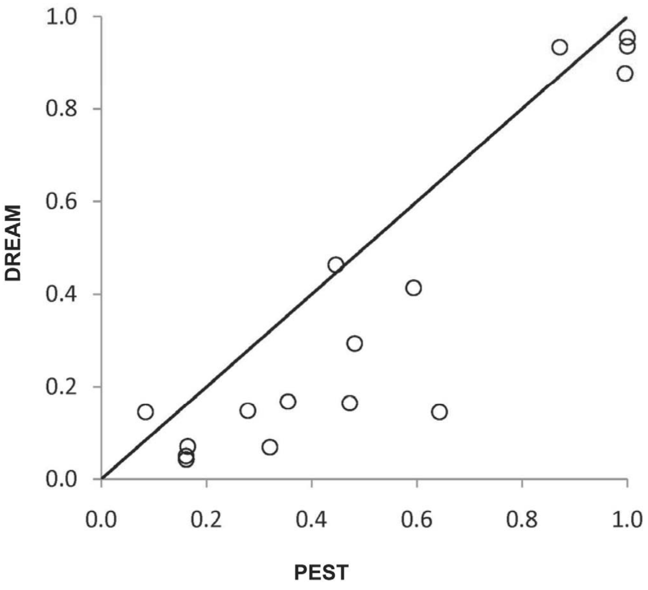

Figure 10. Comparison a normalized median estimates for parameters that both methods reduced standard deviation by at least 50% (Tables 2a and 2b).

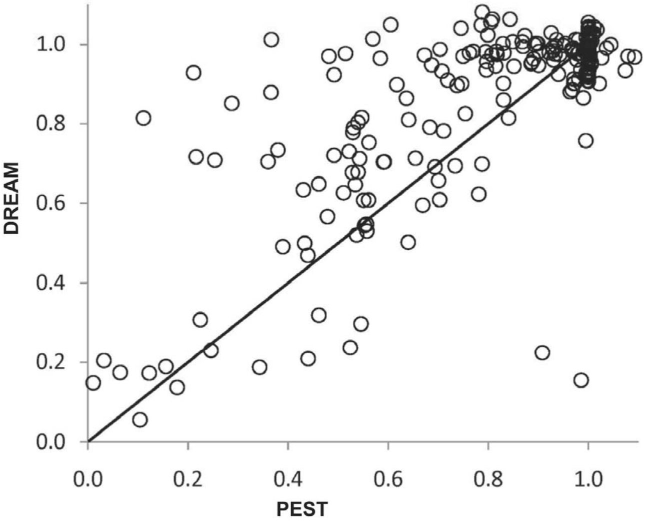

Figure 11. Comparison standard deviation ratio for each parameter.

[44] PEST-based analysis was relatively run-efficient. There are several reasons for this. Where objective function surfaces are reasonably concave and model outputs are differentiable with respect to its parameters, gradient-based

methods provide fast attainment of an objective function minimum than global methods. A problem, of course, is that they may all too easily find a local rather than global minimum. In the present case use of the CMAES method as a "starter" for each SVD-based parameter estimation process increased PEST's ability to locate minima with low objective function values, this contributing to overall optimization efficiency. Model run efficiency of the ensuing uncertainty analysis was attained by focusing null-space Monte Carlo analysis on exploring variability associated with nonuniqueness as it exists around each of several minima located in previous calibration exercises. This stepwise approach (first calibrate, then explore uncertainty), appears to have brought significant computational efficiency to the process.

[45] Some of the differences in the results between the two methods may be an artifact of the relatively small number of calibrated parameter sets generated by PEST (172) through use of its null-space Monte Carlo methodology compared to those generated by DREAM (100,000). Theoretically, further null-space Monte Carlo generation of new, and widely different, calibration-constrained parameter sets could have been undertaken by simply continuing the null-space Monte Carlo process described above; the cost would have been about 150 model runs per additional calibrated parameter field. It was for reasons of computational expediency that the process was halted at 172. The similarity of the statistical measures obtained from 172 parameter sets to those obtained using the more rigorous (in a Bayesian sense) DREAM results based on 100,000 parameter sets engendered optimism that use of PEST in conjunction with the CPU-intensive process model would yield results which are representative of the parameter uncertainty associated with that model. We also note some differences between the results of NSMC and DREAM, yet argue that those differences appear rather marginal within the context of the current modeling approach.

3.3. Prediction Uncertainty Using the Process Model

[46] The ultimate purpose of the analysis is to generate useful predictions, together with the uncertainty associated with those predictions (Step 4). The high parameter dimensionality and long run-times associated with the process model demands that the NSMC, rather than MCMC, method be used for this purpose. Most of the methodological details developed in Steps 2 and 3 can be directly applied to the process model. We applied the design of the objective function as attained and refined through the surrogate model analysis directly to estimation of parameters for the process model. Likewise, as will be described below, we were able to directly apply the NSMC approach used with the surrogate model to the process model. For the initial calibration step, however, it was necessary to deviate from the method used on the surrogate model. We do not expect this adjustment to affect the validity of our results since the final NSMC analysis should not be sensitive to the $\hat{g}$ obtained in the initial calibration. This follows from the observation that the nature of the NSMC methodology is such that it is relatively independent of the particular method used to achieve the initial calibration which then provides the basis for post-calibration exploration of calibration-constrained parameter space.

[47] Specifically, in contrast to the surrogate model calibration, in calibration of the process model, CPU requirements prevented us from beginning the calibration process with 60 random starting points, or from applying truncated SVD in conjunction with a full adjustable parameter set. This is an outcome of the large computational expense of calculating a full Jacobian matrix at every iteration of the inversion process. In principle, the computational burden could be lessened by taking advantage of model-computed adjoints. However, FEHM does not provide adjoints, and perhaps more importantly, the sequential coupling of the testing-effects and groundwater flow models would make computing adjoints very difficult. Instead, we used physical arguments to limit the number of model parameters varied in the calibration step (but not the uncertainty analysis step). Only 46 parameters were allowed to vary, as indicated in Table 1 (column labeled "free"). These included 18 of the 29 material property parameters, emphasizing rock units in which calibration wells were completed, as well as a subset of 28 H_o values, specifically for those tests located within approximately 1000 m of calibration wells. As an additional efficiency measure for calibration, we used the "SVD-Assist" technique described by *Tonkin and Doherty* [2005] to minimize the objective function; this gains efficiency through directly estimating linear combinations of parameters spanning the calibration solution space, thereby eliminating the need to compute sensitivities with respect to all adjustable parameters. We found that by using knowledge of the spatial relation between the calibration data set and model parameters to reduce the parameter dimensionality a priori, followed by the use of the "SVD-assist" methodology to estimate values for the reduced parameter set we were able to achieve a reasonable calibration outcome (an objective function of 682) using a small fraction of the model runs that were necessary to calibrate the surrogate model using the full parameter space. The calibrated model explains over 98% of measured variance in head (see Figure 12), compared to less than 5% based on prior information alone. The range of residuals (simulated –

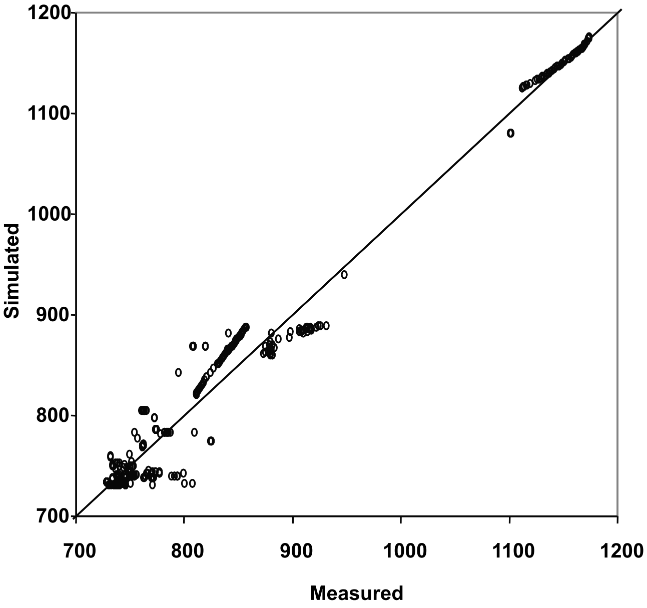

Figure 12. Comparison of measured and simulated heads (m), 583 observations.

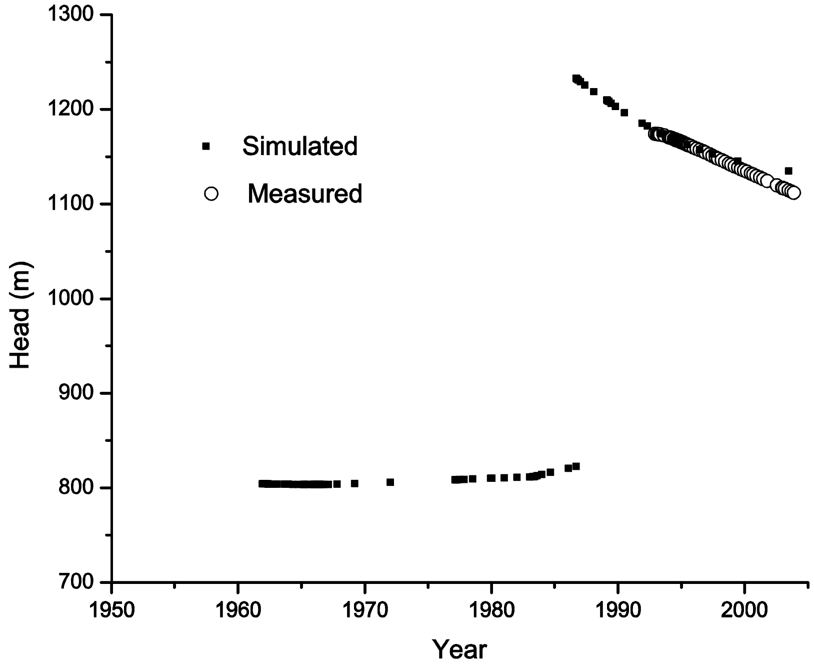

Figure 13. Comparison of measured and simulated heads at well Ue-4t. The abrupt rise in simulated heads around 1987 is due to a nearby underground nuclear test.

measured) is rather large (77 to −59 m), however, most (~60%) are less than 10 m.

[48] Subsequent NSMC analysis used the same methodology as for the surrogate model. First a Jacobian matrix was calculated for all 242 parameters, in this case at great computational expense. This was only done once, at the $\hat{g}$ described above. We do not expect that our results would change if we had repeated this procedure multiple times at estimates of $\hat{g}$ corresponding to other possible local objective function minima, since as described earlier, we found the results of NSMC analysis to be fairly insensitive to starting $\hat{g}$ when using the surrogate model. The NSMC methodology was then employed to generate a suite of parameter sets that all met the criterion ($\Phi < 1100$), for this was heuristically considered to be the objective function value below which the model could be deemed to be "calibrated." (Note: a slightly larger cutoff value was used for the process model than for the surrogate model due to the difficulty of obtaining lower values given the relatively small number of forward runs at our disposal). Because of the computationally demanding nature of this process, we were able to generate only 62 such calibration-constrained random parameter sets. An example hydrograph generated from one of these parameter sets is shown in Figure 13. With enormous additional computational effort, a larger set of calibration-constrained parameter sets could have been obtained. Although this sample size is too small to derive "true" posterior distributions characterizing parameter and predictive uncertainty, it can nevertheless be used to provide some indication of the magnitude of these. The distributions of objective function and prediction (enhanced flux) values, β, arising from this process are shown in Figure 14.

[49] The range of prediction values arising from the NSMC analysis described above is remarkably small, varying by less than one order of magnitude, and does not include 0. In contrast, an ensemble of predictions calculated on the basis of 62 randomly generated parameter sets ranges from 0 to 1462 m^3 X 1E6. It is possible, of course, that the range of the calibration-constrained posterior prediction as calculated using NSMC would increase (even to include 0) if more calibration-constrained parameter sets were generated. But if the present result were to hold with a larger number of parameter sets, it would imply that there is no question that testing did create some degree of enhanced flux to the lower aquifer.

3.4. Post Audit

[50] It is unfortunately often the case in applied groundwater modeling that predictions must be made in the absence of any practical means for testing their veracity. In the present case, however, the availability of a data set of different type to any discussed previously affords us an opportunity to evaluate the ability of the process model to make predictions, and to quantify the uncertainty associated with those predictions. This data set is comprised of measurements of ground surface subsidence (made using airborne synthetic aperture radar) caused by post-testing depressurization of the aquifer [*Vincent et al.*, 2003]). Subsidence was measured over three time periods during the 1990s at a spatial resolution of 30 m. Accuracy of the method was estimated to be of the order of about 1 cm. As testing-induced over–pressurization relaxes over time, pore spaces compress due to release of water from elastic storage; as a consequence the ground surface subsides. Simulations based on the same parameters used to make the flux predictions can be used to predict surface subsidence over time. If the simplifying assumption is made that all volume change is accommodated in the vertical direction, the porosity changes due to changes in elastic storage simulated in each column of cells at each time step can be directly

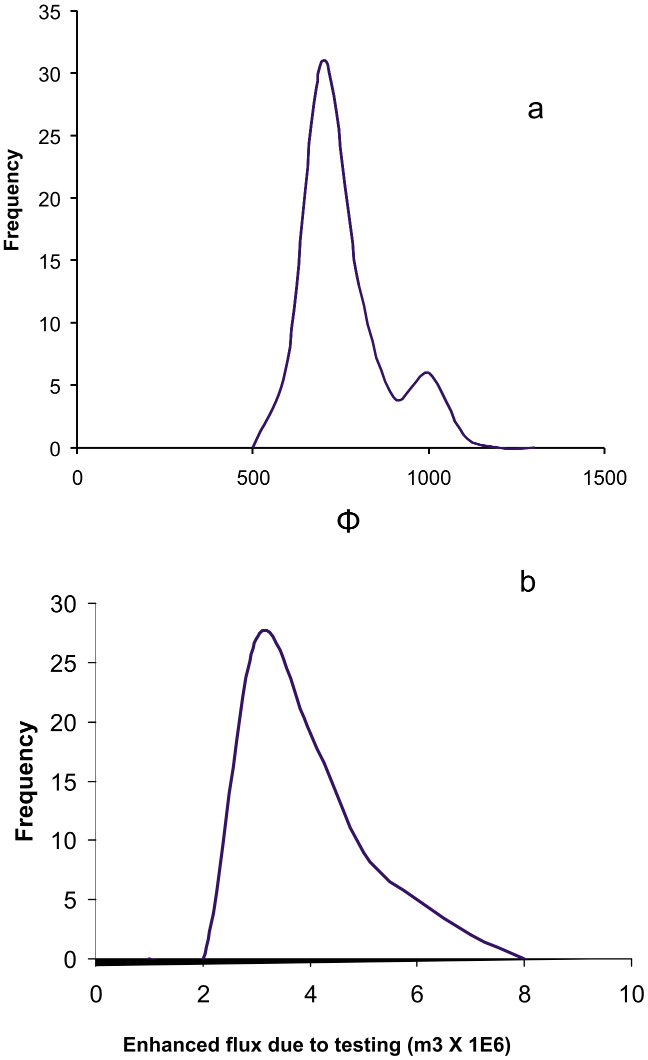

Figure 14. Null-space Monte Carlo results: (a) objective function distribution and (b) prediction distribution.

converted to a 2-D map of ground surface elevation changes over time.

[51] We compared simulated to observed subsidence for the time period 1992–1993 at each point in the model domain that overlaps with the subsidence data (this comprising the northern half of the model domain). Within this overlapping area, 10,000 comparisons between simulated and observed subsidence could be made. Of these, 85.6% of observations (±1 cm) fell between the minimum and maximum predictions as defined by the model runs based on the 64 sets of NSMC-generated parameters. In Figure 15 a comparison between observed subsidence and model-generated uncertainty bounds of predicted subsidence is shown for a representative east–west cross-section (located on Figure 1). The comparison, though not perfect, shows that predictive uncertainty bounds computed using the process model are able to span a range that largely includes observations that were not used in the model calibration process. This is quite remarkable and inspires confidence that the process model has been calibrated appropriately, and provides estimates of uncertainty that appear reasonable when making the prediction of most interest to the present study (Figure 13), as this prediction is dependent on similar aspects of the system as those which lead to ground subsidence.

[52] It should be noted that both the enhanced flux and the ground surface subsidence predictions discussed here are primarily sensitive to relatively large-scale structural variations. Other types of model predictions might be more sensitive to local features, which may not be accurately represented in this model. The potential impact of unresolved local features would have to be considered in any uncertainty analysis of model predictions sensitive to small-scale, localized effects.

4. Conclusions

[53] The methodology we propose here is of practical value to assist in the development and analysis of complex, CPU-intensive groundwater models. Our example application has characteristics that are typical of many practical problems. Because the model is developed for a complex field setting rather than for a synthetic test problem, “true” model structure, parameters, and predictions are unknown. Parameters must be estimated based on limited prior knowledge and using a calibration data set of variable quality whose spatial and temporal sampling is far from optimal. Furthermore, the ability of the model to extract information from that data set is compromised by the fact that the model, like all models, is an imperfect simulator of system behavior. The statistical character of the resulting structural noise, including its spatial and temporal correlation structure, are unknown. As a result of this, special attention needed to be given to formulation of an objection function whose minimization allowed maximum receptivity of information within the calibration data set by the model. An added problem is the computationally demanding nature of the model, and the complex nature of the processes that it simulates and of the hydraulic properties of the host environment in which those processes operate.

[54] As is often the case, in the process of model development, calibration, and predictive uncertainty analysis, many qualitative decisions needed to be made. These were, of necessity, based on expert knowledge of the site, as well as on the strengths and weaknesses of the model components and calibration data set. These decisions included 1) choice of optimal formulation of the objective function, 2) choice of a weighting scheme to use in conjunction with this objective function, 3) value of the objective function for which model-to-measurement fit is deemed to be satisfactory, and 4) choice of the number of calibration-constrained parameter fields that is considered “adequate” for sufficient characterization of predictive uncertainty. It is the authors’ opinion that careful model-based hydrogeologic analyses will always require subjective inputs such as these. We also argue that subjectivity does not detract from the value of information yielded by analyses such as those documented herein. In fact, as we have tried to demonstrate, without the proper use of informed subjectivity, based in part on experimentation allowed by use of a simple surrogate model, the use of a complex model as a basis for environmental decision-making would be seriously compromised.

[55] To our knowledge, the study documented herein represents the first occasion on which the NSMC and MCMC methods have been formally compared. It also constitutes the first demonstration of the use of the DREAM

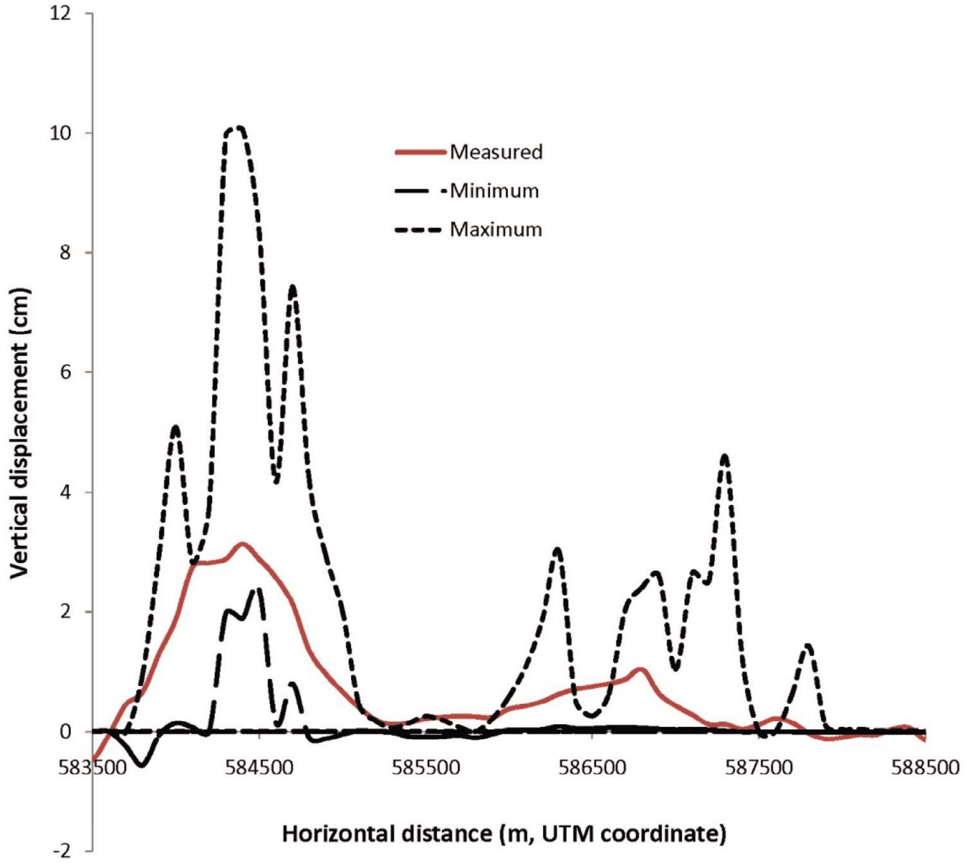

Figure 15. Measured ground-surface subsidence for an EW cross-section through northern Yucca Flat, 1992–1993. Range of simulated values for the 64 cases generated by the NSMC analysis are shown for comparison.

package on a high dimension parameter estimation problem, with low sensitivities for most parameters and a high dimensional null-space. The high dimensionality and nonlinearity of the inverse problem and the low sensitivity of most parameters provided a challenge to which both the NSMC and MCMC algorithms were able to rise. Given the fact that these two uncertainty analysis techniques are based on such different theoretical foundations, the degree of similarity between the outcomes of their analyses was both pleasing and remarkable. Given the sound theoretical basis of DREAM on the one hand, and the computational efficiency of NSMC on the other hand, this is an important finding, for it suggests that widespread use of the NSMC method in conjunction with the large complex models that often form the basis for environmental decision-making will be of widespread worth. Furthermore, our work suggests that it would be advantageous to combine the strengths of the NSMC and MCMC algorithms to enhance the efficiency of MCMC methods in highly parameterized contexts while providing more theoretical rigor to NSMC analysis. Development of a Metropolis MCMC sampling algorithm that incorporates a proposal density function that accommodates the null-space may provide the means through which this can be achieved. This study has stimulated the development of a hybrid parallel MCMC scheme that combines multiple-try Metropolis search, adjusted Langevin sampling, and sampling from a past archive of states to speed up finding the posterior distribution for highly parameterized problems.

[56] In the study documented herein, not all of the information gained, and techniques developed, for optimal calibration and uncertainty analysis using the surrogate model were directly transferable to the process model. For example, the method used to initially calibrate the surrogate model had to be modified to reduce the number of forward model runs required for objective function minimization to take place. Such differences in methodologies through which pre-uncertainty analysis calibration is achieved should not impair the ability of the NSMC methodology to obtain a wide range of parameters sets which satisfy calibration constraints, and hence should not affect the ability of the methodology to properly represent parameter and predictive uncertainty. This is demonstrated in the present study where our final NSMC analysis achieved a suite of widely different parameter fields which could be used to make any kind of prediction required of the model. Our post-audit analysis based on predictions of recovery from subsidence, demonstrated that the range of uncertainty established through use of the NSMC methodology was indeed able to encapsulate the true behavior of the system.

[57] We finish this paper with a summary of our conclusions in point form.

[58] 1. Use of an appropriately simplified surrogate model in conjunction with a highly complex CPU-intensive process model, provides a user with considerable freedom in testing different model concepts and data processing alternatives. This, in turn, promulgates optimal use of the pro-

cess model in estimation of system properties, in simulating future system behavior, and in quantifying the uncertainty associated with predictions of this behavior.

[59] 2. The sophisticated version of the MCMC methodology that is encapsulated in DREAM is able to provide robust posterior parameter distributions even in highly nonlinear, highly parameterized contexts characterized by many local optima, this being a context that often presents challenges to more traditional MCMC software.

[60] 3. Post-calibration parameter and predictive probability distributions can also be obtained in these same contexts at much smaller computational cost using the NSMC methodology encapsulated in PEST. Though lacking the theoretical rigor of MCMC methods in general, and the DREAM method in particular, comparison of results achieved by the two methods suggests that NSMC-generated probability distributions have integrity.

[61] 4. The uncertainty associated with predictions made by a large and complex process model can be explored using the NSMC methodology, notwithstanding the high run times that these model may possess.

[62] The last point is of considerable importance. Model-based decision making cannot take place with integrity unless the uncertainty associated with our ability to predict future system behavior, after all site-relevant data has been processed, is quantified. Models find their proper place in the decision-making process when their ability to absorb the information available in site data through the calibration process is optimized (thereby reducing the uncertainty associated with critical predictions by the largest possible amount), and when the remaining uncertainty is made plainly visible to the model user.

[63] **Acknowledgments.** This work was supported by the U.S. Department of Energy Underground Test Area Project. The third author is supported by a J. Robert Oppenheimer Fellowship from the Los Alamos National Laboratory Postdoctoral Program. The source code of DREAM is written in MATLAB and sequential and parallel implementations can be obtained from the third author upon request (jasper@uci.edu). PEST and ancillary software can be downloaded from http://www.pesthomepage.org. We appreciate the many useful comments and suggestions of the anonymous reviewers and the Associate Editor, and the thoughtful internal review by Andy Wolfsberg. We thank R. J. Laczniak and Keith Halford of the U.S. Geological Survey for providing the InSAR data sets.

References

Aster, R. C., B. Borchers, and C. H. Thurber (2005), *Parameter Estimation and Inverse Problems*, 301 pp., Elsevier, Amsterdam.

Beven, K. (2005), On the concept of model structural error, *Water Sci. Technol.*, *52*(6), 167–175.

Beven, K. (2009), *Environmental Modeling: An Uncertain Future*, 301 pp., Routledge, London.

Beven, K. J., P. J. Smith, and J. E. Freer (2008), So why would a modeller choose to be incoherent?, *J. Hydrol.*, *354*, 15–32, doi:10.1016/j.jhydrol.2008.02.007.

Carrera, J., and S. P. Neuman (1986a), 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.

Carrera, J., and S. P. Neuman (1986b), Estimation of aquifer parameters under transient and steady state conditions: 2. Uniqueness, stability, and solution algorithms, *Water Resour. Res.*, *22*(2), 211–227, doi:10.1029/WR022i002p00211.

Cooley, R. L. (2004), A theory for modeling ground-water flow in heterogeneous media, *U.S. Geol. Surv. Prof. Pap., 1679*, 220 pp.

Cooley, R. L., and S. Christensen (2006), Bias and uncertainty in regression-calibrated models of groundwater flow in heterogeneous media, Adv. *Water Resour.*, *29*(5), 639–656, doi:10.1016/j.advwatres.2005.07.012.

Davis, S. N. (1971), Discussion of paper by Walter E. Reed, "Transport of water away from a buried heat source with special reference to hydrologic phenomena observed at aardvark nuclear detonation," *J. Geophys. Res.*, *76*, 630–632, doi:10.1029/JB076i002p00630.

Doherty, J. (2009), 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*(1–4), 119–127, doi:10.1016/j.jhydrol.2008.12.018.

Doherty, J., and D. Welter (2010), A short exploration of structural noise, *Water Resour. Res.*, *46*, W05525, doi:10.1029/2009WR008377.

Drellack, S., Jr. (2006), A hydrostratigraphic model and alternatives for the groundwater flow and contaminant transport model of corrective action unit 97: Yucca Flat-Climax Mine, Lincoln and Nye Counties, Nevada, Geotech. Sci. Group, Las Vegas, Nev.

Fenelon, J. M. (2005), Analysis of ground-water levels and associated trends in Yucca Flat, Nevada Test Site, Nye County, Nevada, 1951–2003, 97 pp., *U.S. Geol. Surv. Sci. Invest. Rep., 2005-5175*.

Fu, J., and J. J. Gomez-Hernandez (2009), Uncertainty assessment and data worth in groundwater flow and mass transport modeling using a blocking Markov chain Monte Carlo method, *J. Hydrol.*, *364*(3–4), 328–341, doi:10.1016/j.jhydrol.2008.11.014.

Garber, M. S., and L. E. Wollitz (1969), Measuring underground explosion effects on water levels in surrounding aquifers, *Ground Water*, *7*(4), 3–7, doi:10.1111/j.1745-6584.1969.tb01283.x.

Gelman, A., and D. B. Rubin (1992), Inference from iterative simulation using multiple sequences, *Stat. Sci.*, *7*, 457–472, doi:10.1214/ss/1177011136.

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

Halford, K. J., R. J. Laczniak, and D. L. Galloway (2005), Hydraulic characterization of overpressured tuffs in Central Yucca Flat, Nevada Test Site, Nye County, Nevada, *U.S. Geol. Surv. Sci. Invest. Rep., 2005-5211*.

Hansen, N., and A. Ostermeier (2001), Completely derandomized self-adaptation in evolution strategies, *Evol. Comput.*, *9*(2), 159–195, doi:10.1162/106365601750190398.

Hansen, N., S. D. Muller, and P. Koumoutasakos (2003), Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES), *Evol. Comput.*, *11*(1), 1–18, doi:10.1162/106365603321828970.

Hawkins, W. L., D. A. Trudeau, and T. M. Mihevc (1989), Hydrologic testing in exploratory drill hole UE4t Yucca Flat, the Nevada Test Site, in *Proceedings of the 5th Symposium on Containment of Underground Nuclear Detonations, Rep. LLNL CONF 89-09163*, pp, 387–398, Los Alamos Natl. Lab., Los Alamos, N. M.

Hendricks Franssen, H.-J., F. Stauffer, and W. Kinzelbach (2004), Joint estimation of transmissivities and recharges—Application: Stochastic characterization of well capture zones, *J. Hydrol.*, *294*(1–3), 87–102, doi:10.1016/j.jhydrol.2003.10.021.

Hill, M. (1998), Methods and guidelines for effective model calibration, *U.S. Geol. Surv. Sci. Invest. Rep., 98-4005*, 90 pp.

Hunt, R. J., J. Doherty, and M. J. Tonkin (2007), Are models too simple? Arguments for increased parameterization, *Ground Water*, *45*(3), 254–262, doi:10.1111/j.1745-6584.2007.00316.x.

Jiang, Y., and A. D. Woodbury (2006), A full-Bayesian approach to the inverse problem for steady-state groundwater flow and heat transport, *Geophys. J. Int.*, *167*(3), 1501–1512, doi:10.1111/j.1365-246X.2006.03145.x.

Kitanidis, P. K. (1997), The minimum structure solution to the inverse problem, *Water Resour. Res.*, *33*(10), 2263–2272, doi:10.1029/97WR01619.

Kitanidis, P. K. (1999), Generalized covariance functions associated with the Laplace equation and their use in interpolation and inverse problems, *Water Resour. Res.*, *35*(5), 1361–1367, doi:10.1029/1999WR900026.

Knox, J. B., et al. (1965), Analysis of groundwater anomaly created by an underground nuclear explosion, *J. Geophys. Res.*, *70*(4), 823–835, doi:10.1029/JZ070i004p00823.

Kwicklis, E. A., et al. (2006), Multiphase, multicomponent parameter estimation for liquid and vapor fluxes in deep arid systems using hydrologic data and natural environmental tracers, *Vadose Zone J.*, *5*, 934–950, doi:10.2136/vzj2006.0021.

Laczniak, R. L., J. C. Cole, D. A. Sawyer, and D. A. Trudeau (1996), Summary of hydrogeologic controls on the movement of groundwater flow at the Nevada test site, Nye County, Nevada, *U.S. Geol. Surv. Sci. Invest. Rep., 96-4109*.

Menke, W. (1984), *Geophysical Data Analysis: Discrete Inverse Theory*, Academic, Orlando, Fla.

Moore, C., and J. Doherty (2006), The cost of uniqueness in groundwater model calibration, *Adv. Water Resour.*, *29*(4), 605–623, doi:10.1016/j.advwatres.2005.07.003.

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

Reed, W. E. (1970), Transport of water away from a buried heat source with special reference to hydrologic phenomena observed at aardvark nuclear detonation, *J. Geophys. Res.*, *75*(2), 415–430, doi:10.1029/JB075i002p00415.

Ronayne, M. J., S. M. Gorelic, and J. Caers (2008), Identifying discrete geologic structures that produce anomalous hydraulic response: An inverse modeling approach, *Water Resour. Res.*, *44*, W08426, doi:10.1029/2007WR006635.

Skahill, B. E., and J. Doherty (2006), Efficient accommodation of local minima in watershed model calibration, *J. Hydrol.*, *329*(1–2), 122–139, doi:10.1016/j.jhydrol.2006.02.005.

Stoller-Navarro Joint Venture (2006), Phase I hydrologic data for the groundwater flow and contaminant transport model of corrective action unit 97: Yucca Flat/Climax Mine, Nevada Test Site, Nye County, Nevada, Las Vegas, Nev.

ter Braak, C. J. F. (2006), A Markov Chain Monte Carlo version of the genetic algorithm differential evolution: Easy Bayesian computing for real parameter space, *Stat. Comput.*, *16*(3), 239–249, doi:10.1007/s11222-006-8769-1.

ter Braak, C. J. F., and J. A. Vrugt (2008), Differential evolution Markov chain with snooker updater and fewer chains, *Stat. Comput.*, *18*(4), 435–446, doi:10.1007/s11222-008-9104-9.

Tiedeman, C. R., D. M. Ely, M. C. Hill, and G. M. O'Brien (2004), A method for evaluating the importance of system state observations to model predictions, *Water Resour. Res.*, *40*, W12411, doi:10.1029/2004WR003313.

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

Trease, H., et al. (1996), *The X3D grid generation system, Proceedings of Numerical grid generation system in computational fluid dynamics and related fields*, 1129 pp., Miss. State Univ., Mississippi State.

Vincent, P., S. Larsen, D. Galloway, R. J. Laczniak, W. R. Walter, W. Foxall, and J. J. Zucca (2003), New signatures of underground nuclear tests revealed by satellite radar interferometry, *Geophys. Res. Lett.*, *30*(22), 2141, doi:10.1029/2003GL018179.

Vrugt, J. A., and B. A. Robinson (2007), Improved evolutionary optimization from genetically adaptive multimethod search, *Proc. Natl. Acad. Sci. U. S. A.*, *104*(3), 708–711, doi:10.1073/pnas.0610471104.

Vrugt, J. A., C. G. H. Diks, W. Bouten, H. V. Gupta, and J. M. Verstraten (2005), Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation, *Water Resour. Res.*, *41*, W01017, doi:10.1029/2004WR003059.

Vrugt, J. A., P. H. Stauffer, T. Wohling, B. A. Robinson, and V. V. Vesselinov (2008a), Inverse modeling of subsurface flow and transport properties: A review with new developments, *Vadose Zone J.*, *7*(2), 843–864, doi:10.2136/vzj2007.0078.

Vrugt, J. A., C. J. F. ter Braak, M. P. Clark, J. M. Hyman, and B. A. Robinson (2008b), Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, *Water Resour. Res.*, *44*, W00B09, doi:10.1029/2007WR006720.

Vrugt, J. A., B. A. Robinson, and J. M. Hyman (2009a), Self-adaptive multimethod search for global optimization in real-parameter spaces, *IEEE Trans. Evol. Comput.*, *13*(2), 243–259, doi:10.1109/TEVC.2008.924428.

Vrugt, J. A., C. J. F. ter Braak, C. G. H. Diks, B. A. Robinson, J. M. Hyman, and D. Higdon (2009b), Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling, *Int. J. Nonlinear Sci. Numer. Simul.*, *10*(3), 273–290.

Woodbury, A., and T. 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.

Zyvoloski, G. A., et al. (1997), Summary of the models and methods for the FEHM application—A finite-element heat- and mass-transfer code, 72 pp., Los Alamos Natl. Lab., Los Alamos, N. M.

J. Doherty, National Centre for Groundwater Research and Training, Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia.

Q. Kang, E. H. Keating, and J. A. Vrugt, Earth and Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. (ekeating@lanl.gov)