Multimodel Bayesian analysis of data-worth applied to unsaturated fractured tuffs

Dan Lu a, Ming Ye a,⇑, Shlomo P. Neuman b, Liang Xue b

a Department of Scientific Computing, Florida State University, Tallahassee, FL 32306, United States
b Department of Hydrology and Water Resources, University of Arizona, Tucson, AZ 85721, United States

A R T I C L E   I N F O

Article history:
Received 18 June 2011
Received in revised form 12 October 2011
Accepted 18 October 2011
Available online 3 November 2011

Keywords:
Data worth
Value of information
Data uncertainty
Model uncertainty
Parameter uncertainty
Uncertainty reduction

A B S T R A C T

To manage water resource and environmental systems effectively requires suitable data. The worth of collecting such data depends on their potential benefit and cost, including the expected cost (risk) of failing to take an appropriate decision. Evaluating this risk calls for a probabilistic approach to data-worth assessment. Recently we [39] developed a multimodel approach to optimum value-of-information or data-worth analysis based on model averaging within a maximum likelihood Bayesian framework. Adopting a two-dimensional synthetic example, we implemented our approach using Monte Carlo (MC) simulations with and without lead order approximations, finding that the former approach was almost equally accurate but computationally more efficient. Here we apply our methodology to pneumatic permeability data from vertical and inclined boreholes drilled into unsaturated fractured tuff near Superior, Arizona. In an attempt to improve computational efficiency, we introduce three new approximations that require less computational effort and compare results with those obtained by the original Monte Carlo method. The first approximation disregards uncertainty in model parameter estimates, the second does so for estimates of potential new data, and the third disregards both uncertainties. We find that only the first approximation yields reliable quantitative assessments of reductions in predictive uncertainty brought about by the collection of new data. We conclude that, whereas parameter uncertainty may sometimes be disregarded for purposes of analyzing data worth, the same does not generally apply to uncertainty in estimates of potential new data.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Hydrogeologists face a daunting challenge to help insure that contaminants in the subsurface do not pose unacceptable future risks to humans and the environment [11,12]. To quantify and manage such risks, one must understand their relations to existing pollution and to possible remediation schemes. This in turn requires collecting suitable data (hydrogeological, the focus of this study as well as physiological and other), assessing their uncertainty, incorporating them in subsurface flow and contaminant transport models, and using them to monitor system response to existing and future scenarios. As noted by Back [3], three strategies have traditionally been used to optimize a data collection effort: minimize cost for a specific level of analytic accuracy or precision, minimize analytic uncertainty for a given budget, and/or respond to regulatory demands for a given amount and quality of data. Many today prefer a fourth approach based on value-of-information or data-worth analysis [1,4,9–12,14,16,18,22,24–26,30,33,35,37,40,43,48–50,53,55,56,63]. Here value of information is measured by its cost-effectiveness. A data collection program is cost-effective if its expected benefit exceeds its cost, including the expected cost (risk) of failing to take the right decision. A major benefit of new data would be its potential to help improve one’s understanding of the system, in large part through a reduction in model predictive uncertainty. This benefit would be worth the cost only if it had the potential to impact decisions concerning system management. Corresponding methods of analysis include Bayesian decision-making [24] based on a decision tree [3] and graphic risk-driven approaches [11,12].

In the past, analyses of data-worth have commonly relied on a single model with known parameters. Predictive uncertainty has been attributed at best to estimation uncertainty of model parameters. A more recent trend has been to consider uncertainties in both model parameters and structure. This has been motivated by a growing recognition that environmental systems are open and complex, rendering them prone to multiple conceptualizations and mathematical descriptions, including parameterizations. This is true regardless of the quantity and quality of available and potential data. In the context of multi-model analysis, methods that explore how different sets of conditioning data impact the predictive uncertainty of multiple models include the Generalized Likelihood Uncertainty Estimation method (GLUE [5,6,17]), Bayesian Model Averaging (BMA; [7,15,20,28,31,41,49]), and a combination of GLUE and BMA [47]. The method described in [49] aims to assess the worth of disparate data measured on different scales.

Recently we [39] developed a multimodel approach to optimum value-of-information or data-worth analyses within a Bayesian
Model Averaging (BMA) framework. We focused on a maximum likelihood variant of BMA (MLBMA [38,59]) that (a) is compatible with both deterministic and stochastic models, (b) admits (but does not require) prior information about the parameters, (c) is consistent with modern statistical methods of hydrologic model calibration, (d) allows (but does not require) approximating lead predictive moments of any model by linearization, (e) updates model posterior probabilities as well as parameter estimates on the basis of potential new data both before and after such data become actually available (during so-called pre-posterior and posterior stages of data-worth analysis, respectively), and (f) allows modifying the models and their numbers at both the pre-posterior and posterior stages. We expect the proposed approach to be of help in designing the collection of hydrologic characterization and monitoring data in a cost-effective manner by optimizing the difference between their benefit and cost. Whereas benefits would accrue from optimum gain in information or reduction in predictive uncertainty, costs would include the risk of failing to take an appropriate decision. Evaluating this risk calls for a probabilistic approach to data-worth assessment as we propose.

Implementation of our proposed approach on a synthetic geostatistical problem in two space dimensions demonstrated a need to account for the impact of potential new data on model and parameter uncertainties. Though neither existing nor a potentially augmented set of data were sufficient to identify correctly the underlying geostatistical model (variogram) and its parameters, they nevertheless yielded self-consistent results and allowed identifying quite accurately the impacts of potential new data on the spatial distribution and magnitude of corresponding reductions in predictive variance. Approximating lead predictive moments associated with each model by linearization yielded results comparable to those obtained via Monte Carlo (MC) simulation with a much less expenditure of computational effort. The extent to which such linearization would work in strongly nonlinear situations was deemed an open question.

The purpose of this study is to (1) apply our multimodel Bayesian methodology of data-worth assessment to real data and (2) explore the possibility to improve computational efficiency by minimizing the need for Monte Carlo simulations. We achieve the latter by introducing three approximations that require less computational effort than does our original approach [39]: one approximation that disregards uncertainty in model parameter estimates, another that does so for estimates of potential new data, and a third approximation that disregards both uncertainties. Our test case entails pneumatic permeability data from six vertical and inclined boreholes drilled into unsaturated fractured tuff near Superior, Arizona. We ask two questions: (1) Given data from three boreholes and funds to drill and test one additional borehole at one of two possible locations, at which location should one drill and test so as to reduce predictive uncertainty at yet another target borehole? (2) How does the answer depend on which among the above four MLBMA approaches we adopt? In addition to addressing these questions in the context of real data distributed in three-dimensional space our analysis sheds new light on roles played by parameter and pre-posterior data uncertainties in our approach.

2. Methodology

We start by reviewing key background material of BMA, MLBMA, and the Bayesian method of data-worth analysis in Sections 2.1 and 2.2 summarized from [39]. Section 2.3 presents three new approximation schemes developed here in an attempt, evaluated later, to reduce the computational cost of our original approach.

2.1. Background of BMA and MLBMA

Consider a random vector, $\mathbf{A}$, the multivariate statistics of which are to be predicted with a set $\mathbf{M}$ of $K$ mutually independent models (a somewhat ambiguous concept discussed in [60]), $M_k$ each characterized by a vector of parameters $\theta_k$, conditional on a discrete set of data, $\mathbf{D}$ (the case of correlated models has recently been considered in [51]). The joint posterior (conditional) distribution of $\mathbf{A}$ is [15,20]

$$p(\mathbf{A}|\mathbf{D}) = E_{M_k}p(\mathbf{A}|\mathbf{D}, M_k) = \sum_{k=1}^{K} p(\mathbf{A}|\mathbf{D}, M_k)p(M_k|\mathbf{D}),$$

(1)

where $p(\mathbf{A}|\mathbf{D}, M_k)$ is the joint posterior distribution of $\mathbf{A}$ due to model $M_k$ and $P(M_k|\mathbf{D})$ the posterior probability of this model. The latter are given by Bayes’ rule [15,20]

$$P(M_k|\mathbf{D}) = \frac{p(\mathbf{D}|M_k)p(M_k)}{\sum_{i=1}^{K} p(\mathbf{D}|M_i)p(M_i)},$$

(2)

where

$$p(\mathbf{D}|M_k) = \int p(\mathbf{D}|M_k, \theta_k)p(\theta_k|M_k)d\theta_k$$

(3)

is the integrated likelihood of model $M_k$, $p(\mathbf{D}|M_k, \theta_k)$ being the joint likelihood of this model and its parameters, $p(\theta_k|M_k)$ the prior density of $\theta_k$ under model $M_k$, and $P(M_k)$ the prior probability of $M_k$. Whereas the likelihood $p(\mathbf{D}|M_k, \theta_k)$ contains a statistical model of errors associated with $\mathbf{D}$ (due to measurement, stochastic interpolation or both), the prior density $p(\theta_k|M_k)$ may contain a model of parameter measurement errors [8]. All probabilities are implicitly conditioned on the choice of models entering into the set $\mathbf{M}$, which are taken to be mutually independent (the case of correlated models has recently been considered in [51]). The posterior mean and covariance of $\mathbf{A}$ are given by Draper [15] and Hoeting et al. [20]

$$E(\mathbf{A}|\mathbf{D}) = E_{M_k}E(\mathbf{A}|\mathbf{D}, M_k) = \sum_{k=1}^{K} E(\mathbf{A}|\mathbf{D}, M_k)p(M_k|\mathbf{D}),$$

(4)

$$Cov(\mathbf{A}|\mathbf{D}) = E_{M_k}Cov(\mathbf{A}|\mathbf{D}, M_k) + Cov_{M_k}E(\mathbf{A}|\mathbf{D}, M_k)$$

$$= \sum_{k=1}^{K} Cov(\mathbf{A}|\mathbf{D}, M_k)p(M_k|\mathbf{D})$$

$$+ \sum_{k=1}^{K} \left[ E(\mathbf{A}|\mathbf{D}, M_k) - E(\mathbf{A}|\mathbf{D}) \right]\left[ E(\mathbf{A}|\mathbf{D}, M_k) - E(\mathbf{A}|\mathbf{D}) \right]^T p(M_k|\mathbf{D}),$$

(5)

where $T$ denotes transpose, $E_{M_k}Cov(\mathbf{A}|\mathbf{D}, M_k)$ is the within-model component of $Cov(\mathbf{A}|\mathbf{D})$ and $Cov_{M_k}E(\mathbf{A}|\mathbf{D}, M_k)$ is its between-model component. A scalar measure of the posterior variance of $\mathbf{A}$ is given by the trace

$$Tr[Cov(\mathbf{A}|\mathbf{D})] = Tr[EM_kCov(\mathbf{A}|\mathbf{D}, M_k)] + Tr[Cov_{M_k}E(\mathbf{A}|\mathbf{D}, M_k)]$$

(6)

which is of interest because, for $K > 1$, one generally has $Tr[Cov_{M_k}E(\mathbf{A}|\mathbf{D}, M_k)] > 0$ so that $Tr[Cov(\mathbf{A}|\mathbf{D})] > Tr[EM_kCov(\mathbf{A}|\mathbf{D}, M_k)]$. Hence the consideration of multiple models generally results in greater predictive uncertainty, as measured by $Tr[Cov(\mathbf{A}|\mathbf{D})]$, than the uncertainty associated with a single model, as measured by $Tr[EM_kCov(\mathbf{A}|\mathbf{D}, M_k)]$. MLBMA [38,59] is an approximation of BMA obtained through replacement of $\theta_k$ by an estimate, $\hat{\theta}_k$, which maximizes the likelihood $p(\mathbf{D}|M_k, \theta_k)$. Obtaining such ML estimates entails calibrating each model against (conditioning on) the data $\mathbf{D}$ using well-established statistical inverse methods. In MLBMA, the integrated likelihood, $P(M_k|\mathbf{D})$, in (2) is replaced by $P(M_k|\mathbf{D})_{ML}$, where the subscript indicates approximation based on ML estimation of $\theta_k$, using the Laplace method [27,28,38,45,54,59,61,62] via
having components (as opposed to ensemble mean) Fisher information matrix conditioned on the augmented data set \( \{ \Pi \} \), evaluated at \( \theta^j \). Here \( N_k \) is the dimension of \( \theta_k \) (number of adjustable parameters associated with model \( M_k \)). \( N^D \) is the dimension of \( D \) (number of discrete data points, which may include measured parameter values), and \( F_k \) is the normalized (by \( N^D \)) observed (as opposed to ensemble mean) Fisher information matrix having components [27]

\[
F_{k, nm} = -\frac{1}{N^D} \left[ \frac{\partial^2 \ln p(D|\theta_k, M_k)}{\partial \theta_{n} \partial \theta_{m}} \right]_{\theta_k = \hat{\theta}^j}.
\]

The observed Fisher information provides a more reliable assessment of \( KIC \) than does the more commonly used expected Fisher information [29,32]. The ML approximation in MLBMA extends further to approximating \( p(\Delta|D, M_k) \) in (1) by \( p(\Delta|D, M_k)_{\text{ML}} \) [38,57], the accuracy of which is explored numerically in [32,39]. In the limit of large \( N^D/N_k \), \( KIC^D_k \) reduces asymptotically to the Bayesian selection (or information) criterion, \( BIC^D_k \) (e.g. [38,44]). However, a recent study [32] shows that \( KIC^D_k \) yields more accurate approximations of the integrated likelihood and posterior model probability than does \( BIC^D_k \).

2.2. Multimodel Bayesian data augmentation

Suppose that the original data set \( D \) is augmented by another hypothetical (pre-posterior) data set, \( C \), which has not yet been collected and is therefore uncertain (random). Assume that the multivariate statistics of \( \Delta \), predicted with model set \( M \), can be conditioned on the augmented data set \( (D, C) \). Assume further that the multivariate statistics of \( C \), conditional on \( D \), can be predicted either via BMA or via MLBMA with a set \( P \) of \( I \) mutually independent statistical models, \( P_i \), having parameters \( \pi_i \). The models \( P_i \) may be independent of \( M_k \), may form extensions of \( M_k \) or may coincide with the latter as in the computational example given later in this paper. For example, whereas \( M_k \) may represent flow and/or transport models, \( P_i \) may be geostatistical models of spatial parameter variability; in our example below the two models coincide. Both sets of models, and their parameters, may change with the collection of new data. In some special cases, such as the Matérn variogram model [34], a range of models may be represented by a single model with one or more structural parameters. In the following models \( P \) are taken to be equivalent to models \( M \). Analogy to (1) and the law of total probability imply

\[
p(\Delta|D) = \text{ECP} p(\Delta|D, C) = \int p(\Delta|D, C)p(C|D) dC,
\]

where

\[
p(\Delta|D, C) = \sum_{k=1}^{K} p(\Delta|D, C, M_k)p(M_k|D, C),
\]

and \( C \) is implicitly conditional on the choice of models \( P \) and on \( D \). Analogy to (4) and by virtue of the law of total expectation lead to

\[
E(\Delta|D) = \text{ECP} E(\Delta|D, C) = \int E(\Delta|D, C)p(C|D) dC.
\]

where

\[
E(\Delta|D, C) = \sum_{k=1}^{K} E(\Delta|D, C, M_k)p(M_k|D, C).
\]

Analogy to (5) and by virtue of the law of total variance lead to

\[
\text{Cov}(\Delta|D, C) = \text{ECP} \text{Cov}(\Delta|D, C) + \text{Cov}_{\text{CP}} E(\Delta|D, C),
\]

where

\[
\text{Cov}(\Delta|D, C) = \sum_{k=1}^{K} \text{Cov}(\Delta|D, C, M_k)p(M_k|D, C) + \sum_{k=1}^{K} E(\Delta|D, C, M_k) - E(\Delta|D, C)^2 p(M_k|D, C).
\]

Analogy to (6) gives

\[
\text{Tr}[\text{Cov}(\Delta|D)] = \text{Tr} [\text{ECP} \text{Cov}(\Delta|D, C)] + \text{Tr} [\text{Cov}_{\text{CP}} E(\Delta|D, C)].
\]

The term \( \text{Tr}[\text{Cov}_{\text{CP}} E(\Delta|D, C)] = \text{Tr}[\text{ECP} \text{Cov}(\Delta|D)] - \text{Tr}[\text{ECP} \text{Cov}(\Delta|D, C)] \) represents the difference between the total trace conditional on \( D \) and the expected trace conditional jointly on \( D \) and \( C \). As this difference is positive, conditioning on \( D \) and \( C \) jointly results in a lower trace than conditioning on \( D \) alone. The difference could be viewed as an extended version of the \( \alpha \)-criterion in optimal design, other measures of uncertainty reduction being possible. Appendix A reproduces the MLBMA procedure developed in [39] for prior conditioning on \( D \) and preposterior conditioning on both \( D \) and \( C \). The procedure accounts for data uncertainty through MC simulations in Step 8 and for parameter uncertainty through MC simulations in Step 8d.

2.3. Proposed new approximations

Implementing the above approach entails MC simulations with respect to both parameter and data uncertainties, rendering it computationally demanding. Here we complement lead-order approximations developed in [39] with three new approximations that reduce computational effort further. We start by rewriting (15) as

\[
\text{Cov}(\Delta|D) = \text{ECP} \text{E}_{\Delta|M_k, D} \text{Cov}(\Delta|M_k, D, C) + \text{ECP} \text{Cov}_{\Delta|M_k} \text{E}(\Delta|M_k, D, C) + \text{Cov}_{\text{CP}} \text{E}_{\Delta|M_k} \text{Cov}_{\Delta|M_k, D} \text{E}(\Delta|M_k, D, C),
\]

where we applied the law of total covariance to \( \text{ECP} \text{Cov}(\Delta|D, C) \) and the law of total expectation to \( \text{Cov}_{\text{CP}} \text{E}(\Delta|D, C) \) in a way that accounts explicitly for model uncertainty. Next we rewrite (18) as

\[
\text{Cov}(\Delta|D) = \text{ECP} \text{E}_{\Delta|M_k, D} \text{E}_{\Delta|M_k} \text{Cov}(\Delta|\theta_k, M_k, D, C) + \text{ECP} \text{Cov}_{\Delta|M_k} \text{E}_{\Delta|M_k} \text{E}(\Delta|\theta_k, M_k, D, C) + \text{Cov}_{\text{CP}} \text{E}_{\Delta|M_k} \text{E}_{\Delta|M_k} \text{E}(\Delta|\theta_k, M_k, D, C),
\]

where we applied the law of total expectation to \( \text{E}(\Delta|M_k, D, C) \) and the law of total covariance to \( \text{Cov}(\Delta|M_k, D, C) \) in a way that accounts explicitly for parameter uncertainty. The covariance decomposition indicates that uncertainties in data \( C \), model \( M \), parameter \( \theta \), and prediction sample \( \Delta \) are quantified in (19). Here \( \text{E}(\Delta|\theta_k, M_k, D, C) \), \( \text{Cov}_{\Delta|M_k} \text{E}(\Delta|\theta_k, M_k, D, C) \) and \( \text{Cov}_{\Delta|M_k, D} \text{E}(\Delta|\theta_k, M_k, D, C) \) represent sampling, parameter and model uncertainties, respectively [45] whereas the last term in (19) represents uncertainty about the pre-posterior data \( C \).
The first approximation eliminates parameter uncertainty from (19). The first term on the right hand side of (19) contains

$$E_{\theta |M_k, D} \text{Cov}(\Lambda(\theta_k, M_k, D, C))$$

which, by virtue of the mean integral value theorem, can be expressed as

$$E_{\theta |M_k, D} \text{Cov}(\Lambda(\theta_k, M_k, D, C)) = \int \text{Cov}(\Lambda(\theta_k, M_k, D, C)d\theta_k = \text{Cov}(\Lambda(\xi_k, M_k, D, C),$$

where $\xi_k$ is an unknown value of $\theta_k$. Replacing $\xi_k$ with $\hat{\theta}_k^{MC}$, the ML estimate of $\theta_k$ based jointly on $(D, C)$, gives $E_{\theta |M_k, D} \text{Cov}$

![Fig. 1. Spatial locations of 184 1-m-scale log data along six boreholes at ALRS.](image)

![Fig. 2. Sample and fitted variograms with number of data pairs per lag based on (a-b) $D \times C_1$, and (c-f) $D \times C_2$.](image)

### Table 1

<table>
<thead>
<tr>
<th>Criterion</th>
<th>AIC</th>
<th>KIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>Pow0</td>
<td>Exp0</td>
</tr>
<tr>
<td>$p(M_k</td>
<td>D)$</td>
<td>42.02</td>
</tr>
<tr>
<td>$p(M_k</td>
<td>D, C_1)$</td>
<td>65.80</td>
</tr>
<tr>
<td>$p(M_k</td>
<td>D, C_2)$</td>
<td>62.96</td>
</tr>
<tr>
<td>Case (%)</td>
<td>CV I</td>
<td></td>
</tr>
<tr>
<td>$p(M_k</td>
<td>D)$</td>
<td>75.82</td>
</tr>
<tr>
<td>$p(M_k</td>
<td>D, C_1)$</td>
<td>91.70</td>
</tr>
<tr>
<td>$p(M_k</td>
<td>D, C_2)$</td>
<td>34.03</td>
</tr>
</tbody>
</table>
Fig. 3. Measured $\log k$ (+), predicted (solid) and 95% confidence intervals (dashed) for (a) Pow0, (b) Exp0, (c) Sph0 and (d) MLBMA along borehole V2 based on data $D$ (black), $D + C_1$ (blue), and $D + C_2$ (green) in CV I. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 4. Measured $\log k$ (+), predicted (solid) and 95% confidence intervals (dashed) with (a) Pow0, (b) Exp0, (c) Sph0 and (d) MLBMA along borehole Y3 based on data $D$ (black) and $D + C_1$ (blue) and $D + C_2$ (green) in CV II. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
which eliminates the need for MC simulations in Step 8d of Appendix A. By disregarding parameter uncertainty the approximation introduces an error of order \( O(1/N) \) where \( N \) is the dimension of \( (\mathbf{D}, \mathbf{C}) \), which may be acceptable when model uncertainty exceeds parameter uncertainty [45]. The corresponding equivalent of
\[
\text{Tr}[\text{Cov}_{\text{MC}}(\mathbf{A} | \mathbf{D}, \mathbf{C})] \text{ in (17) is } \text{Tr}[\text{Cov}_{\text{MC}}(\mathbf{A} | \mathbf{D}, \mathbf{C})] \text{ denoted by }
\text{Tr}[\text{Cov}_{\text{MC}}(\mathbf{A} | \mathbf{D}, \mathbf{C})] \text{ in Appendix A.}
\]

The second approximation removes the need for MC simulations over random realizations of the potential new data \( \mathbf{C} \) in step 8 of our procedure (Appendix A) through their replacement by the corresponding conditional mean, \( E(\mathbf{C} | \mathbf{D}) \). This eliminates the second term and the mean operation in the first term at the right hand side of (15). Here the predictive uncertainty reduction cannot be defined as \( \text{Tr}[\text{Cov}_{\text{MC}}(\mathbf{A} | \mathbf{D}, \mathbf{C})] \) because this term now vanishes. Instead, the reduction is defined as the difference between \( \text{Tr}[\text{Cov}(\mathbf{A} | \mathbf{D})] \) and \( \text{Tr}[\text{Cov}(\mathbf{A} | \mathbf{D}, E(\mathbf{C}(\mathbf{D})))] \). Following the derivation of (19) and using the law of total variance with respect to model uncertainty gives
\[
\text{Cov}(\mathbf{A} | \mathbf{D}, E(\mathbf{C}(\mathbf{D}))) = E_{\mathbf{M}_t, \mathbf{D}, E(\mathbf{C})}[\text{Cov}(\mathbf{A} | \mathbf{D}, \mathbf{C})] + \text{Cov}_{\mathbf{M}_t, \mathbf{D}, E(\mathbf{C})}[\text{Cov}(\mathbf{A} | \mathbf{D}, \mathbf{C})].
\]

Applying the law of total variance to \( \text{Cov}(\mathbf{A} | \mathbf{M}_t, \mathbf{D}, E(\mathbf{C}(\mathbf{D}))) \) and the law of total expectation to \( E(\mathbf{A} | \mathbf{M}_t, \mathbf{D}, E(\mathbf{C}(\mathbf{D}))) \), in a way that accounts explicitly for parameter uncertainty yields
\[
\text{Cov}(\mathbf{A} | \mathbf{D}, E(\mathbf{C}(\mathbf{D}))) = E_{\mathbf{M}_t, \mathbf{D}, E(\mathbf{C})}[\text{Cov}(\mathbf{A} | \mathbf{D}, \mathbf{C})] + \text{Cov}_{\mathbf{M}_t, \mathbf{D}, E(\mathbf{C})}[\text{Cov}(\mathbf{A} | \mathbf{D}, \mathbf{C})].
\]

which is the predictive uncertainty given \( \mathbf{D} \) and \( E(\mathbf{C}(\mathbf{D})) \).

The third approximation combines the first two approximations by disregarding both parameter and data uncertainties. Replacing \( \theta_k^M \) in (23) by its ML estimate \( \hat{\theta}_k \) obtained on the basis of \( \mathbf{D} \) and \( E(\mathbf{C}(\mathbf{D})) \), thereby ignoring parameter uncertainty (and thus eliminating the need for MC simulations across random parameter realizations in step 8d of Appendix A), yields a third approximation of (19) which disregard both data and parameter uncertainties,
\[
\text{Cov}(\mathbf{A} | \mathbf{D}) = \text{Cov}(\mathbf{A} | \mathbf{D}, E(\mathbf{C}(\mathbf{D}))) \\
\approx E_{\mathbf{M}_t, \mathbf{D}, E(\mathbf{C})}[\text{Cov}(\mathbf{A} | \mathbf{D}, \mathbf{C})] + \text{Cov}_{\mathbf{M}_t, \mathbf{D}, E(\mathbf{C})}[\text{Cov}(\mathbf{A} | \mathbf{D}, \mathbf{C})].
\]

If \( E(\mathbf{C}(\mathbf{D})) \) and \( \text{Cov}(\mathbf{A} | \mathbf{D}, \mathbf{C}) \) can be estimated directly through geostatistical (as we do below) or stochastic modeling, step 8 in Appendix A, which is computationally the most demanding, becomes redundant. The accuracy of approximations (23) and (24) diminishes with increasing variability of \( \mathbf{C} \) and sensitivities of \( \theta_k \) and \( \mathbf{A} \) to \( \mathbf{C} \). As in the second approximation, uncertainty reduction due to \( E(\mathbf{C}(\mathbf{D})) \) in the third approximation is measured by \( \text{Tr}[\text{Cov}(\mathbf{A} | \mathbf{D})] - \text{Tr}[\text{Cov}(\mathbf{A} | \mathbf{D}, E(\mathbf{C}(\mathbf{D})))]. \)

3. Application to air permeabilities of fractured tuff

We implement our procedure (Appendix A) and its three approximations, based on (21)–(24), on 1-m scale log air permeability (log \( k \)) data from unsaturated fractured tuff at the former Apache Leap Research Site (ALRS) near Superior, Arizona. Spatially distributed log \( k \) data were obtained [19] based on a steady state interpretation of 184 pneumatic injection tests in 1-m-length intervals along six boreholes at the site (Fig. 1). Five of the boreholes (V2, W2a, X2, - Y2, Z2) are 30-m long and one (Y3) has a length of 45 m; five (W2a, X2, Y2, Y3, Z2) are inclined at 45° and one (V2) is vertical. We consider two “cross-validation” cases: CV I, in which log \( k \) measured
Table 2  
Actual and predicted trace variance reductions in CV I and CV II.

<table>
<thead>
<tr>
<th>Case</th>
<th>CV I</th>
<th>CV II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consider both parameter and data uncertainties</td>
<td>( C_1 (X2) )</td>
<td>( C_1 (Y2) )</td>
</tr>
<tr>
<td>Data estimates (C)</td>
<td>2.31</td>
<td>3.11</td>
</tr>
<tr>
<td>Approximation (21) disregarding parameter uncertainty (Eq. (21))</td>
<td>0.59</td>
<td>1.79</td>
</tr>
<tr>
<td>Data estimates (C)</td>
<td>0.42</td>
<td>0.84</td>
</tr>
<tr>
<td>Approximation (22) disregarding data uncertainty</td>
<td>0.69</td>
<td>1.94</td>
</tr>
<tr>
<td>Data estimates (C)</td>
<td>0.53</td>
<td>0.84</td>
</tr>
<tr>
<td>Approximation (23) disregarding both parameter and data uncertainties</td>
<td>3.66</td>
<td>10.15</td>
</tr>
<tr>
<td>Data estimates (C)</td>
<td>3.64</td>
<td>8.96</td>
</tr>
</tbody>
</table>

Fig. 7. C (red), \( E(C | D)_{\text{post}} \) (solid blue), 200 realizations of predicted C (gray) and 95% confidence interval of all 200 realizations (dashed blue) for (a) \( C_1 \) along borehole X2 and (b) \( C_2 \) along borehole Y2 in CV I. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 8. Variation of \( \text{Var}(J \mid D)_{\text{post}} \), \( E_{C_{j}} \text{Var}(\Delta D \mid C, C_{j})_{\text{post}} \), and \( V_{C_{j}} E_{\Delta D} (\Delta D, C_{j})_{\text{post}} \) along borehole V2 based on (a) \( D \times C_{i} \) and (b) \( D \times C_{j} \) in CV I.

Fig. 9. Variation of \( \text{Var}(J \mid D)_{\text{post}} \), \( E_{C_{j}} \text{Var}(\Delta D \mid C, C_{j})_{\text{post}} \), and \( V_{C_{j}} E_{\Delta D} (\Delta D, C_{j})_{\text{post}} \) along borehole Y3 based on (a) \( D \times C_{i} \) and (b) \( D \times C_{j} \) in CV II.

in W2a, Y3 and Z2 play the role of existing data D, boreholes X2 and Y2 are the sites of potential new data \( C_1 \) and \( C_2 \), the goal being to predict \( \log k(A) \) along V2; CV II, in which \( \log k \) measured in W2a, X2 and Y2 play the role of D, boreholes V2 and Z2 are the sites of potential new data \( C_1 \) and \( C_2 \), the goal being to predict \( \log k(A) \) along Y3 (true posterior data are denoted by \( C \) and their pre-posterior estimates by \( \hat{C} \)). Given that in each case one has funds to measure \( \log k \) in only one borehole, along which among boreholes X2 and Y2 should one conduct such measurements in case CV I, and along which among V2 and Z2 in case CV II?

The issue is compounded by uncertainty about the correct geostatistical model and parameters to be employed in each case. To address it, the data (D and C) are viewed as a Gaussian random field with covariance described by three alternative variogram models: exponential (Exp0), spherical (Sph0) and power (Pow0) each having two parameters (sill and range in the case of Exp0 and Sph0, variance coefficient and power in the case of Pow0); our abbreviations are consistent with those in [59] where the 0 indicates no (zero-order) drift. In each case, variogram parameters are unknown and are estimated from the data using maximum likelihood. Uncertainty of the estimated parameters, \( \theta \), is quantified by a covariance matrix, \( \Gamma \). The parameters are assumed to have a multivariate normal distribution, \( N(\theta, \Gamma) \), about their estimates, from which random samples are drawn by a Monte Carlo method (Appendix A, Step 4a). Details about these models,
nature of our study, no new models emerge in the pre-posterior estimation of model probabilities are given in [59–61]. Because of the high probability of model Exp0 obtained on the basis of data augmentation; Fig. 3b and d are similar to each other due to confidence intervals, which are seen to narrow down slightly with confidence intervals obtained with each geostatistical model and with MLBMA based on D, {D, C} and {D, C} in CV I and CV II. Anisotropy and multi-modality are not exhibited in the sample variograms. The corresponding posterior probabilities based on the information theoretic model discrimination criterion AIC [2] and on KIC are listed in Table 1; those based on the information theoretic criterion AICc [21] and on BIC are similar to those based on AIC and are therefore not listed. Whereas KIC favors model Exp0 in all 6 cases listed in Table 1, AIC prefers Exp0 in 2 cases and Pow0 in 4 cases. Considering the Bayesian nature of our methodology we present below only results associated with KIC.

3.1. Effect of augmentation with actual data

To evaluate the predictive uncertainty of \( \Delta \) associated with data sets D, \( \{D, C_1\} \) and \( \{D, C_2\} \) where primes indicate actual (posterior, measured) values we employ MC according to steps 1–5 in Appendix A. We generate \( R_0 = 1000 \) realizations of \( \theta_0 \) according to step 4a while insuring that their mean and variance have stabilized for each variogram model. Values of \( \Delta \) are predicted by using the KT3D package of [13], revised to produce a predictive (kriging) covariance according to [52]. Fig. 3 compares measured log K values along borehole V2 with predicted values and their 95% confidence intervals obtained with each geostatistical model and with MLBMA based on D, \( \{D, C_1\} \) and \( \{D, C_2\} \) in CV I and Fig. 4 does so for CV II. All measurements in Fig. 3 are contained within 95% confidence intervals, which are seen to narrow down slightly with data augmentation; Fig. 3b and d are similar to each other due to the high probability of model Exp0 obtained on the basis of KIC. Table 1. While similar patterns are observed in Fig. 4, some measurements lie outside the 95% confidence intervals and predictions

![Fig. 10](image-url) Variation of \( \text{Var}(\Delta |D)_{\text{MC}} \), \( \text{Exp} \text{Var}(\Delta |D, C)_{\text{MC}} \) and \( \text{Var}_{\text{Exp}} \text(E(\Delta |D, C)_{\text{MC}}) \) along predicted borehole V2 based on (a) \( D + C_1 \) and (b) \( D + C_2 \) disregarding parametric uncertainty in CV I.

![Fig. 11](image-url) \( E_{\text{Exp}} \text{E}(\Delta |\theta_0, M_0, D, C, \{D, C\} \text{ red) and } E(\Delta |\theta_0, M_0, D, C) \text{ (blue corresponding to predicted data index 13 (Fig. 10) in borehole V2 obtained for 200 realizations of } C_1 \text{ in borehole X2, CV I, using model (a) Pow0, (b) Exp0 and (c) Sph0; } E_{\text{Exp}} \text{E}(\Delta |\theta_0, M_0, D, C) \text{ (red) and } E_{\text{MLBMA}} \text{E}(\Delta |\theta_0, M_0, D, C) \text{ (blue obtained using (d) MLBMA. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)} \)
deviate from measurements to a greater extent than they do in Fig. 3. This confirms [61] that it is easier to predict \( \log k \) along borehole V2 than along borehole Y3 on the basis of other site data. Below we focus exclusively on results of MLBMA analyses.

Figs. 5 and 6 plot variations of \( \text{Var}(\Delta \mathbf{D})_{\text{ab}} \), \( \text{Var}(\Delta \mathbf{D}, \mathbf{C})_{\text{abc}} \) and their difference, \( \text{Var}(\Delta \mathbf{D})_{\text{ab}} - \text{Var}(\Delta \mathbf{D}, \mathbf{C})_{\text{abc}} \), along boreholes V2 and Y3 in CV I and CV II, respectively. The figures show that augmenting the samples reduces predictive variances along both boreholes. Table 2 lists actual trace variance reductions, \( \text{Tr} \left[ \text{Var}(\Delta \mathbf{D}, \mathbf{C})_{\text{abc}} \right] \), and their predicted counterparts, \( \text{Tr} \left[ \text{Var}_{\text{MLD}}(\Delta \mathbf{D}, \mathbf{C})_{\text{abc}} \right] \), for the two cross-validation cases. In CV I, collecting new data along borehole X2 is seen to result in greater uncertainty reduction than doing the same along borehole Y2 regardless of whether one relies on actual data or their predictions; in CV II collecting new data along borehole V2 is more advantageous than doing so along borehole Z2 for actual data and their predictions.

### 3.2. Effect of augmentation with randomly generated data

Since in reality the actual data \( \mathbf{C} \) are unknown, we generate \( R_s = 200 \) pre-posterior realizations \( \mathbf{C} \) of the same using available data \( \mathbf{D} \) in accord with steps 6 and 7 of our procedure (Appendix A). Fig. 7 shows \( \mathbf{E}(\mathbf{C})_{\text{abc}} \), the 200 realizations and their 95% confidence intervals in CV I. \( \mathbf{E}(\mathbf{C})_{\text{abc}} \) is seen to represent a smoothed version of \( \mathbf{C} \), suggesting a slight bias in the generated \( \mathbf{C} \) values. Some values of \( \mathbf{C} \) lie outside the confidence intervals, indicating that the generated \( \mathbf{C} \) values do not reflect all information contained in \( \mathbf{C} \). The same is true for CV II (result not shown).

Following Steps 8 and 9 allows us to compute and plot variations in \( \text{Var}(\Delta \mathbf{D})_{\text{abc}} \), \( \text{E}_{\text{MLD}} \text{Var}(\Delta \mathbf{D}, \mathbf{C})_{\text{abc}} \) and \( \text{Var}_{\text{MLD}}(\Delta \mathbf{D}, \mathbf{C})_{\text{abc}} \) along borehole V2 in Fig. 8 and along borehole Y3 in Fig. 9. We verified that 200 realizations are enough for sample estimates of all three terms to stabilize. A comparison of Figs. 8 and 5 reveals that, although \( \text{Tr} \left[ \text{Var}(\Delta \mathbf{D})_{\text{abc}} \right] \) is smaller than its true counterpart, \( \text{Tr} \left[ \text{Var}_{\text{MLD}}(\Delta \mathbf{D}, \mathbf{C})_{\text{abc}} \right] \), the spatial patterns in the two figures are similar. Likewise, though \( \text{E}_{\text{MLD}} \text{Var}(\Delta \mathbf{D}, \mathbf{C})_{\text{abc}} \) and \( \text{Var}_{\text{MLD}}(\Delta \mathbf{D}, \mathbf{C})_{\text{abc}} \) in Fig. 8 tends to exceed \( \text{Var}(\Delta \mathbf{D}, \mathbf{C})_{\text{abc}} \) in Fig. 5, these too have near identical spatial patterns. One therefore expects the estimated variance reduction \( \text{Var}_{\text{MLD}}(\Delta \mathbf{D}, \mathbf{C})_{\text{abc}} \) to exhibit a pattern similar to that of the true variance reduction \( \text{Var}(\Delta \mathbf{D})_{\text{abc}} - \text{Var}(\Delta \mathbf{D}, \mathbf{C})_{\text{abc}} \). The predicted variance reduction measure \( \text{Tr} \left[ \text{Var}_{\text{MLD}}(\Delta \mathbf{D}, \mathbf{C})_{\text{abc}} \right] = 0.59 \) Table 2 is smaller than its true (posterior) counterpart \( \text{Tr} \left[ \text{Var}(\Delta \mathbf{D})_{\text{abc}} \right] - \text{Tr} \left[ \text{Var}_{\text{MLD}}(\Delta \mathbf{D}, \mathbf{C})_{\text{abc}} \right] = 2.31 \) while \( \text{Tr} \left[ \text{Var}_{\text{MLD}}(\Delta \mathbf{D}, \mathbf{C})_{\text{abc}} \right] = 0.42 \) is smaller than \( \text{Tr} \left[ \text{Var}(\Delta \mathbf{D}, \mathbf{C})_{\text{abc}} \right] - \text{Tr} \left[ \text{Var}_{\text{MLD}}(\Delta \mathbf{D}, \mathbf{C})_{\text{abc}} \right] = 1.01 \); comparing Figs. 9 and 6 reveals similar relationships in the CV II case Table 2.
As already noted, our generated C values do not contain all information about the actual data C. This explains, for example, why actual data along boreholes V2 (C1) and Z2 (C2) help reduce predictive uncertainty along the entire borehole Y3 (Fig. 6) but generated values along boreholes V2 and Z2 help reduce uncertainty only along the bottom (Fig. 9a) and top (Fig. 9b) parts of borehole Y3. This makes intuitive sense considering that the bottom part of borehole Y3 is closer to V2 than to Z2 and its upper part is closer to Z2 than to V2.

Even though our generated C values underestimate the potential of such data to help reduce predictive uncertainty, Table 2 makes clear that they lead to correct choices of boreholes to sample (X2 in CV I and V2 in CV II).

4. Accuracies of the three approximations

Upon disregarding parameter uncertainty approximation (21) eliminates the need to simulate θ k in Step 8d. Fig. 10 plots variations in Var(ΔD)MC, ECEBVar(ΔD,C)MC, and VarCEB E(ΔD,C)MC along borehole V2 in CV I computed in this manner. A comparison of Figs. 10 and 8 reveals that patterns of uncertainty reduction in CV I are not affected by the approximation; though we do not show it, the same holds true for CV II. Table 2 indicates that the approximation has minimal effect on computed measures of uncertainty reduction, causing them to increase on average by about 15% in the two cross validation cases and to remain unaffected along borehole Z2 in CV II. Table 2 suggests selecting boreholes X2 and V2 in CV I and CV II, respectively, as targets for collecting additional data regardless of whether or not parameter uncertainty is considered.

To gain insight into the question why parameter uncertainty has little impact on our results, we compare the predictive uncertainty reduction measure Tr[CovCEB E(Δθ k, M k, ΔC)] in (19) or equivalently by Tr[CovCEB E(ΔD,C)] in (17), with its approximation Tr[CovCEB E(Δθ k, M k, ΔC)] in (21). Fig. 11a–c plot E(Δθ k, M k, ΔC) and E(ΔD,C) corresponding to predicted data index 13 (Fig. 10) in borehole V2 (for which the difference between these measures is the largest)
the sample variance of obtained for 200 realizations of \( C \) in borehole X2 (CV I) using models Pow0, Exp0 and Sph0; Fig. 11d plots \( E_{\theta \theta}^{\lambda} E_{M_{\lambda}} D_{C} E(\theta_{\lambda}, M_{\lambda}, D, C) \) (red) and \( E_{\theta \theta}^{\lambda} E_{M_{\lambda}} D_{C} E(\theta_{\lambda}^{2}, M_{\lambda}, D, C) \) (blue) obtained using MLBMA.

The exact and approximate measures are similar in all realizations, the sample variance of \( E_{\theta \theta}^{\lambda} E_{M_{\lambda}} D_{C} E(\theta_{\lambda}, M_{\lambda}, D, C) \) being 0.042 and that of \( E_{\theta \theta}^{\lambda} E_{M_{\lambda}} D_{C} E(\theta_{\lambda}^{2}, M_{\lambda}, D, C) \) 0.048. Fig. 12 plots coefficients of variation (CV) of 1000 realizations of the two parameters of models Pow0, Exp0, and Sph0 for each of the 200 realizations in Fig. 11. Fig. 13 plots coefficients of variation of \( E(\theta_{\lambda}, M_{\lambda}, D, C) \) predicted on the basis of these 1000 parameter realizations for each of the 200 data realizations along X2. The coefficients of parameter variation in Fig. 12 are seen to be much larger than those of corresponding variations in \( E(\theta_{\lambda}, M_{\lambda}, D, C) \), indicating that the latter is not very sensitive to the former. This explains why our first approximation works.

Upon disregarding data uncertainty approximation (23) eliminates the need to simulate \( C \) in Steps 7–9. Here uncertainty reduction is no longer measured by \( Tr[CoV_{CSE}(\lambda, D, C)_{MBC}] \) but by \( Tr[CoV_{CSE}(\lambda, D_{C})_{MBC}] \). Fig. 14 plots variations of \( Var(\lambda_{D}C_{SE}), Var(\lambda_{D}E(\lambda_{C}D))_{MBC} \) and their difference along borehole V2 in CV I. Comparing Fig. 14 with Figs. 8 and 10 shows a pattern of uncertainty reduction similar to those obtained earlier while taking data uncertainty into account; the same happens in CV II (results not shown). However, disregarding data uncertainty is seen to cause the predicted uncertainty reduction to be significantly exaggerated as evidenced further by corresponding measures of uncertainty reduction in Table 2. These measures nevertheless show a preference for boreholes X2 and V2 as new data collection targets in CV I and CV II, respectively, consistent with previous results. However, the estimated variance reductions for boreholes X2 and V2 in CV I are almost identical, implying that borehole selection is tenuous in this case.

To gain insight into the question why data uncertainty has major impact on our results, we compare the first three terms on the right hand side of (19), representing \( E_{CSE} Cov(\lambda, D, C) \), with the three terms on the right hand side of (23), representing \( Cov(\lambda, D, E(\lambda_{C}D)) \) in our second approximation. As shown in Fig. 15, all three terms, but particularly the first one, are seriously underestimated upon disregarding data uncertainty. The pattern and magnitude of the differences between \( E_{CSE} E_{\lambda_{\lambda}} E_{\lambda_{\lambda}} D_{C} E(\lambda_{\lambda}, M_{\lambda}, D, C) \) and \( E_{CSE} E_{\lambda_{\lambda}} E_{\lambda_{\lambda}} D_{C} E(\lambda_{\lambda}, M_{\lambda}, D, E(\lambda_{C}D)) \) shown in Fig. 15a are similar to those between \( Var(\lambda_{D}C_{SE}) \) and \( Var(\lambda_{D}E(\lambda_{C}D))_{MBC} \) in Fig. 14a. This is so because kriging variance conditioned on \( C \) is significantly larger than that conditioned on \( E(\lambda_{C}D) \) due to the smooth nature of the latter, as seen in Fig. 7. Correspondingly, disregarding data uncertainty results in underestimation of predictive uncertainty and overestimation of its reduction. The smoother is \( E(\lambda_{C}D) \) in comparison to \( C \), the more pronounced this effect would be.

Disregarding both data and parameter uncertainties according to (24) eliminates the need to simulate both \( \theta \) and \( C \) in Steps 7–9, rendering this approach most efficient computationally. Here again uncertainty reduction is measured by \( Tr[CoV_{CSE}(\lambda, D, C)_{MBC}] \). Fig. 16 plots variations of \( Var(\lambda_{D}C_{SE}), Var(\lambda_{D}E(\lambda_{C}D))_{MBC} \) and the difference between them along borehole V2 in CV I. The corresponding pattern of uncertainty reduction is practically identical to that in Fig. 14; the same happens in CV II (results not shown). Corresponding trace variance reductions in Table 2 indicate a preference for boreholes X2 and V2 as new data collection targets in CV I and CV II, respectively, consistent with previous results. However, the estimated uncertainty reductions in CV I do not support an unambiguous preference for borehole X2.
We thus see that whereas disregarding parameter uncertainty has minimal effect on computed measures of uncertainty reduction, disregarding data uncertainty exaggerates these measures significantly regardless of whether or not parameter uncertainty is considered. This is so because disregarding data uncertainty renders the predictive variance in our geostatistical analysis equivalent to the kriging variance, which in turn is proportional to the coefficient of the power model and to the sill of the exponential and spherical models. As shown in Table 3, estimates of these parameters become suppressed when data uncertainty is ignored. This causes predictive uncertainty, $Var(\Delta D, E(C(D))_{ML|obs}|v)$, to be suppressed and the variance reduction measure $Tr(Cov(\Delta D, E(C(D))_{ML|obs}|v))$ to become inflated.

5. Conclusions and discussion

1. Our multimodel Bayesian approach to data-worth analysis [39] appears to work well when applied to log air permeability data from vertical and inclined boreholes drilled into unsaturated fractured tuff. In particular, the selection of targets for additional data collection during the pre-posterior stage is validated against actual data collected during the posterior stage of the analysis.

2. In our case study, partial linearization of the posterior variance equations achieved by disregarding parameter uncertainty brings about an improvement in computational efficiency without affecting the results in any significant way. Though parameter uncertainty is significant, our results are only marginally sensitive to this uncertainty.

3. Partial or complete linearization of our posterior variance equations, achieved by disregarding data uncertainty while either considering or disregarding parameter uncertainty, brings about a significant reduction in computational effort. Whereas in our case study such linearization has only a minor effect on predicted pattern of uncertainty reduction due to the potential collection of new data, it significantly overestimates the magnitude of this reduction. The overestimation is due to replacement of variable data by their smooth estimate without taking account of its uncertainty. Correspondingly, selection among alternative data collection targets becomes tenuous. It follows that whereas parameter uncertainty may sometimes be disregarded for purposes of analyzing the worth of data, the same does not generally apply to uncertainty in estimates of potential new data.

4. Our study was limited to predetermined borehole locations so as to allow cross-validating our results against known data. In practice one may want to employ an optimization algorithm to select among alternative candidate locations.

5. Though demonstration of our Bayesian data-worth analysis has been limited to geostatistical models, the approach is general enough to apply to any models including those of groundwater flow and contaminant transport. Considering that the choice of variogram model may affect flow and transport modeling [23,42,46], a study such as ours could be considered either a prelude to groundwater flow and transport modeling or an integral part thereof.

### Acknowledgement

The first and second authors were supported in part by NSF-EAR Grant 0911074 and DOE-ERSP Grant DE-SC0002687. The second author was also supported by the COFRS project of the Florida State University. The third and fourth authors were supported in part through a contract between the University of Arizona and Vanderbilt University under the Consortium for Risk Evaluation with Stakeholder Participation (CRESP) III, funded by the US Department of Energy.

### Appendix A. Computational implementation of MLBMA framework

To assess the impact of data augmentation within the above MLBMA framework computationally we propose the following approach:

1. Postulate a set $M$ of $K$ mutually independent geostatistical, statistical or stochastic models, $M_k$, with parameters $\theta_k$ for the desired output vector, $\Delta$.

2. Obtain ML estimates $\theta_k^M$ of $\theta_k$ by calibrating each $M_k$ against available data $D$ through minimization of the log likelihood $-2\ln p(D|M_k, \theta_k)$, then compute the corresponding estimation covariance $\Gamma^M_k$.

3. Compute $p(M_k|D)_{ML} = \exp \left( -\frac{1}{2} \Gamma^M_k \theta_k \right)$, where the subscript $ML$ designates the ML estimation process in step 2.

4. For each model $M_k$ estimate $\hat{E}(\Delta|D, M_k)_{ML}$ and $\hat{Cov}(\Delta|D, M_k)_{ML}$ either through second-order approximations of [36] or via Monte Carlo simulation (used in this study):

   a. Draw random samples (realizations) of $\theta_k$ from a multivariate Gaussian distribution with mean $\theta_k^M$ and covariance $\Gamma^M_k$.

5. The first and second authors were supported in part by NSF-EAR Grant 0911074 and DOE-ERSP Grant DE-SC0002687. The second author was also supported by the COFRS project of the Florida State University. The third and fourth authors were supported in part through a contract between the University of Arizona and Vanderbilt University under the Consortium for Risk Evaluation with Stakeholder Participation (CRESP) III, funded by the US Department of Energy.

### Table 3

<table>
<thead>
<tr>
<th>Model</th>
<th>Pow0</th>
<th>Exp0</th>
<th>Sph0</th>
<th>Pow0</th>
<th>Exp0</th>
<th>Sph0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case</td>
<td>CV I</td>
<td>CV II</td>
<td>CV I</td>
<td>CV II</td>
<td>CV I</td>
<td>CV II</td>
</tr>
<tr>
<td>$C_1$</td>
<td>0.28</td>
<td>0.68</td>
<td>1.20</td>
<td>0.18</td>
<td>0.67</td>
<td>0.86</td>
</tr>
<tr>
<td>$E(C_1</td>
<td>D)$</td>
<td>0.23</td>
<td>0.61</td>
<td>0.97</td>
<td>0.14</td>
<td>0.61</td>
</tr>
<tr>
<td>$C_2$</td>
<td>0.26</td>
<td>0.78</td>
<td>1.07</td>
<td>0.26</td>
<td>0.59</td>
<td>0.82</td>
</tr>
<tr>
<td>$E(C_2</td>
<td>D)$</td>
<td>0.23</td>
<td>0.62</td>
<td>0.98</td>
<td>0.13</td>
<td>0.36</td>
</tr>
</tbody>
</table>

Fig. 16. Variation of $Var(\Delta D|_{ML}, Var(\Delta D, E(C(D))_{ML|obs}|v)$ and $Var(\Delta D|_{ML}) - Var(\Delta D, E(C(D))_{ML|obs}|v)$ along predicted borehole V2 based on (a) $D \times E(C_i|D)$ and (b) $D \times E(C_0|D)$ disregarding data and parametric uncertainties in CV I.
b. Estimate $E(\Delta D, M_k, \theta_{MB})$ and $Cov(\Delta D, M_k, \theta_{MB})$ for each realization of $\theta_k$;

c. Average over all realizations of $\theta_k$ to obtain sample estimates of $E(\Delta D, M_k, \theta_{MB}) = E_k E(\Delta D, M_k, \theta_{MB})$ and $Cov(\Delta D, M_k, \theta_{MB}) = E_k Cov(\Delta D, M_k, \theta_{MB})$;

d. Compute $E(\Delta D, M_k, \theta_{MB}) = \sum_k E(\Delta D, M_k, \theta_{MB})$ and $Cov(\Delta D, M_k, \theta_{MB}) = \sum_k Cov(\Delta D, M_k, \theta_{MB})$;

e. Postulate a set $P$ of $L$ alternative geostatistical, statistical or stochastic models, $P_k$, with parameters $\pi_k$ for a potential data set $C$; the models $P_k$ may be independent of $M_k$, may form extensions of $M_k$ or may coincide with the latter as in the computational examples given in this paper;

9. Average over all realizations of $\theta_k$ to obtain sample estimates of $E(\Delta D) = E_k E(\Delta D, M_k, \theta_{MB})$ and $Cov(\Delta D) = E_k Cov(\Delta D, M_k, \theta_{MB})$ and/or $Tr Cov(\Delta D)$ obtained at this step would generally differ, though ideally not much, from $E(\Delta D) = E_k E(\Delta D)$ and $Tr Cov(\Delta D)$ obtained at step 5;

10. Repeat steps 6–9 for different sets $C_1, C_2, C_3, \ldots$ of potential data and select that set which maximizes the difference $Tr Cov(\Delta D, C_{MB}) = Tr Cov(\Delta D, C_{MB}) - Tr Cov(\Delta D)$ between the trace conditional on $D$ and the expected trace conditional on $D$ and $C$.

References


