Multimodel Assessment of the Worth of Data under Uncertainty - 11416

Shlomo P. Neuman[*], Liang Xue[*], Ming Ye[**], and Dan Lu[**]
[*]University of Arizona, Tucson, Arizona 85721
[**] Florida State University, Tallahassee, Florida 32306

ABSTRACT

The management of contaminated groundwater systems requires an understanding of their response to alternative remediation strategies. Such understanding requires the collection of suitable data to help characterize the system and monitor its response to existing and future cleanup and/or containment options. It also requires incorporating such data in suitable models of water flow and contaminant transport. As the collection of subsurface characterization and monitoring data is costly, it is important that the design of corresponding data collection schemes be cost-effective, i.e., that the expected benefit of new information exceed its cost. A major benefit of new data is its potential to help improve one's understanding of the system, in large part through a reduction in model predictive uncertainty. Traditionally, value-of-information or data-worth analyses have relied on a single conceptual-mathematical model of site hydrology. Yet there is a growing recognition that analyses and predictions based on a single hydrologic concept are prone to statistical bias and underestimation of uncertainty. This has led to a recent emphasis on conducting hydrologic analyses and rendering corresponding predictions by means of multiple models. We describe a multimodel approach to optimum value-of-information or data-worth analyses based on model averaging within a Bayesian framework. The Bayesian model averaging (BMA) approach is compatible with both deterministic and stochastic models; in its maximum likelihood version (MLBMA) it is additionally consistent with current statistical methods of hydrologic model calibration. Implementation entails either Monte Carlo simulation or linearization. We describe the MLBMA approach and implement it computationally on a synthetic example with and without linearization.

INTRODUCTION

The DOE faces a daunting challenge insuring that contaminants in the subsurface do not pose unacceptable future risks to humans and the environment. To quantify and manage such risks one must understand their relationships to alternative remediation schemes. This in turn requires the collection of suitable data to help characterize the subsurface and monitor its response to existing and future site remediation and management options. It also requires incorporating such data in suitable models of subsurface flow and contaminant transport.

As noted by Back [1], three strategies have traditionally been used to determine the magnitude of a data collection effort: minimizing cost for a specific level of accuracy or precision, minimizing uncertainty for a given budget, or responding to regulatory demands on data quantity and quality. Various combinations of these strategies have also been described such as a fitness-for-purpose approach [2]. Many today prefer a fourth approach based on value-of-information or data-worth analysis. Here the decision to collect additional data, or the design of a data collection program, is based on cost-effectiveness. A program is considered cost-effective if the expected benefit from the new information exceeds its cost. A major benefit of new data is its potential to help improve one's understanding of the system, in large part through a reduction in model predictive uncertainty. This benefit, however, justifies the cost only if it has the potential to impact decisions concerning site remediation and management in a substantive way.

Value-of-information or data-worth analyses incorporating statistical decision theory have been applied to various water-related problems starting in the 1970s and to groundwater problems as of the 1980s. More recent applications to groundwater resource and contamination issues have been reported, among others, in [3, 4]. James and Freeze [5] proposed a Bayesian decision-making framework to evaluate the worth of data in the context of contaminated groundwater that has been widely cited in the subsequent literature. A comprehensive review focusing on health risk assessment can be found in [6]. Additional recent publications of relevance include [7, 8].

A major limitation of many existing approaches is that they rely on a single conceptual-mathematical model of the subsurface and flow/transport processes therein. Yet the subsurface is open and complex, rendering its characteristics and corresponding processes prone to multiple interpretations and mathematical descriptions, including parameterizations. This is true regardless of the quantity and quality of available data. Predictions and analyses of uncertainty based on a single hydrologic concept are prone to statistical bias (by committing a Type II error through reliance on an inadequate model) and underestimation of uncertainty (by committing a Type I error through under sampling of the relevant model space). Analyses of environmental data-worth which explore how different sets of conditioning data impact the predictive uncertainty of multiple models in a Bayesian context include [9, 10]; whereas Freer et al. [9] employ Generalized Likelihood Uncertainty Estimation (GLUE) [11], Rojas et al. [10] combine GLUE with Bayesian Model Averaging (BMA; [12]). Nowak et al. [13] introduce a Bayesian approach to data worth analysis when flow and transport take place in a random log hydraulic conductivity field. Whereas in their analysis flow and transport are described by a single (linearized) model each having known parameters, other than those describing spatial variations in log hydraulic conductivity, the latter is characterized by a single drift model and a continuous family of variogram models having uncertain parameters.

In a similar spirit, we describe in this paper a multimodel approach to optimum value-of-information or data-worth analyses that is based on model averaging within a Bayesian framework. Our approach is general in that it considers multiple models of any kind, all having uncertain parameters; whereas parameterizing models in the manner of [13] is elegant and computationally efficient, it is unfortunately limited to a narrow range of variogram models and does not, generally, apply to other models such as those of flow and transport. We prefer BMA over GLUE because it (a) rests on rigorous statistical theory, (b) is compatible with deterministic as well as stochastic models and, (c) in its maximum likelihood (ML) version (MLBMA), is consistent with current ML methods of hydrologic model calibration. Whereas BMA (like the closely related approach in [13]) relies heavily on prior parameter statistics, MLBMA can do without such statistics or otherwise update them on the basis of potential new data both before and after they are collected. We describe the proposed MLBMA approach and illustrate it on a synthetic example. Our proposed methodology should be of help in designing the collection of hydrologic characterization and monitoring data in a cost-effective manner by maximizing their benefit under given cost constraints. The benefit would accrue from optimum gain in information, or reduction in predictive uncertainty, upon considering jointly not only traditional sources of uncertainty such as those affecting model parameters and the reliability of data but also, most importantly, lack of certainty about the conceptual-mathematical models that underlie the analysis and the scenarios under which the system would operate in the future. The methodology should apply to a broad range of models representing natural processes in ubiquitously open and complex earth and environmental systems.

BACKGROUND

Bayesian Decision Analysis Framework

One way to cast the data-worth issue is within a Bayesian risk-cost-benefit decision framework such as that of Freeze et al. [14, 15]. Suppose without loss of generality that the data are intended to help one decide whether or not a contaminated site should be remediated. This decision problem is illustrated in Fig. 1 [1] by a decision tree in which Φ is the decision objective; Φ_i is an objective function associated with each decision alternative $\left(i=1,2\right)$ defined as

$$\Phi_i = B_i - C_i - \gamma P_i Cf_i \qquad \text{(Eq. 1)}$$

where B_i is the benefit and C_i the investment cost, risk being expressed as the product $\gamma P_i Cf_i$ of a risk aversion factor γ, the probability of failure P_i and the cost of failure Cf_i; C^+ designates a contaminated and C^- an uncontaminated state of the site; costs and benefits occurring at the triangular terminal nodes (only the cost of failure is indicated in the Fig.). Collecting additional information generally causes the risk term to decrease due to a decrease in the probability of failure. The corresponding increase in Φ_i is the expected value (worth) of the new data. The final outcome of the analysis depends on the choice of decision rule one adopts; for example, maximizing Φ_i would result in the largest benefit and lowest cost.

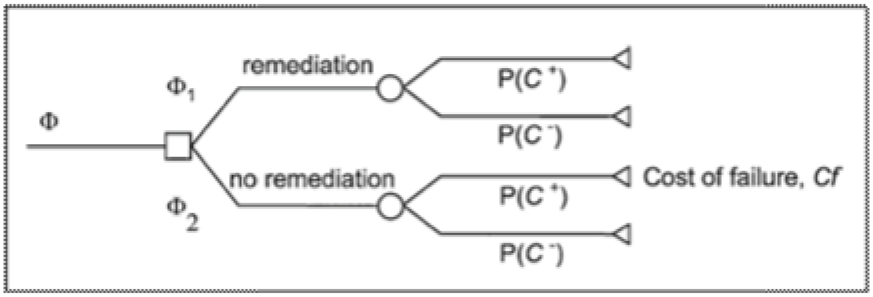

Fig. 1. Decision tree (after [1]).

According to Back [1] the Bayesian approach to data-worth analysis entails five steps: (1) defining one or more data collection (sampling) programs, (2) postulating a prior probability for the state of the site (e.g, contaminated or uncontaminated), (3) using Bayes' theorem to update the prior probability to a posterior probability conditional on the new data, corresponding to each data collection program, (4) estimating corresponding costs and benefits, and (5) computing the worth of data or value of information using a given decision model and using the results to optimize the data collection scheme. Back [1] also considers using linearized estimation of uncertainty to update the prediction variance in step 3.

As noted, collecting additional information generally reduces risk due to a decrease in the probability of failure. A reduction in the probability of failure comes about through a reduction in uncertainty about the expected system state, present or future. The impact of hydrologic data on this expectation and the associated uncertainty are often evaluated by means of a hydrologic model. Commonly, the model is considered to be certain while its parameters (and in some cases its forcing terms such as sources and boundary conditions) are treated as being uncertain due to insufficient and error-prone data. As already noted, we know of only one work that considers the impact of data on model predictive uncertainty within a Bayesian framework by considering the model itself to be uncertain [9] and one other work that parameterizes this uncertainty [13]. Below we provide background about Bayesian model averaging and its maximum likelihood version which we propose to employ for this same purpose.

Bayesian Model Averaging (BMA)

Consider a random vector, $\mathbf{\Delta}$, the multivariate statistics of which are to be predicted with a set **M** of *K* mutually independent models, M_k, each characterized by a vector of parameters $\boldsymbol{\theta}_k$, conditional on a discrete set of data, **D** (the case of correlated models has recently been considered in [16]). In analogy to the case of a scalar Δ [12] we write the joint posterior (conditional) distribution of $\mathbf{\Delta}$ as

$$p\left(\mathbf{\Delta}\middle|\mathbf{D}\right)=\sum_{k=1}^{K}p\left(\mathbf{\Delta}\middle|\mathbf{D},M_k\right)p\left(M_k\middle|\mathbf{D}\right), \qquad \text{(Eq. 2)}$$

i.e., as the average over all models of the joint posterior distributions $p\left(\mathbf{\Delta}\middle|\mathbf{D},M_k\right)$ associated with individual models, weighted by the model posterior probabilities $p\left(M_k\middle|\mathbf{D}\right)$. These weights are given by Bayes' rule in the form

$$p\left(M_k\middle|\mathbf{D}\right)=\frac{p\left(\mathbf{D}\middle|M_k\right)p\left(M_k\right)}{\sum_{l=1}^{K}p\left(\mathbf{D}\middle|M_l\right)p\left(M_l\right)} \qquad \text{(Eq. 3)}$$

where

$$p\left(\mathbf{D}\middle|M_k\right)=\int p\left(\mathbf{D}\middle|M_k,\boldsymbol{\theta}_k\right)p\left(\boldsymbol{\theta}_k\middle|M_k\right)d\boldsymbol{\theta}_k \qquad \text{(Eq. 4)}$$

is the integrated likelihood of model M_k, $p\left(\mathbf{D}\middle|M_k,\boldsymbol{\theta}_k\right)$ being the joint likelihood of this model and its parameters, $p\left(\boldsymbol{\theta}_k\middle|M_k\right)$ the prior density of $\boldsymbol{\theta}_k$ under model M_k, and $p\left(M_k\right)$ the prior probability of M_k. All probabilities are implicitly conditional on the choice of models entering into the set **M**.

The posterior mean and covariance of $\boldsymbol{\Delta}$ are given by

$$E\left(\boldsymbol{\Delta}\middle|\mathbf{D}\right)=\sum_{k=1}^{K}E\left(\boldsymbol{\Delta}\middle|\mathbf{D},M_k\right)p\left(M_k\middle|\mathbf{D}\right), \qquad \text{(Eq. 5)}$$

$$\begin{aligned}Cov\left(\boldsymbol{\Delta}\middle|\mathbf{D}\right)=&\sum_{k=1}^{K}Cov\left(\boldsymbol{\Delta}\middle|\mathbf{D},M_k\right)p\left(M_k\middle|\mathbf{D}\right)\\&+\sum_{k=1}^{K}\left[E\left(\boldsymbol{\Delta}\middle|\mathbf{D},M_k\right)-E\left(\boldsymbol{\Delta}\middle|\mathbf{D}\right)\right]\left[E\left(\boldsymbol{\Delta}\middle|\mathbf{D},M_k\right)-E\left(\boldsymbol{\Delta}\middle|\mathbf{D}\right)\right]^T p\left(M_k\middle|\mathbf{D}\right)\end{aligned} \qquad \text{(Eq. 6)}$$

where the superscript T denotes transpose. (Eq. 6) is a discrete expression of the law of total covariance,

$$Cov\left(\boldsymbol{\Delta}\middle|\mathbf{D}\right)=E_{M_k|\mathbf{D}}Cov\left(\boldsymbol{\Delta}\middle|\mathbf{D},M_k\right)+Cov_{M_k|\mathbf{D}}E\left(\boldsymbol{\Delta}\middle|\mathbf{D},M_k\right), \qquad \text{(Eq. 7)}$$

where $E_{M_k|\mathbf{D}}Cov\left(\boldsymbol{\Delta}\middle|\mathbf{D},M_k\right)$ is the within-model component of $Cov\left(\boldsymbol{\Delta}\middle|\mathbf{D}\right)$ and $Cov_{M_k|\mathbf{D}}E\left(\boldsymbol{\Delta}\middle|\mathbf{D},M_k\right)$ is its between-model component. We also consider the trace

$$Tr\left[Cov\left(\boldsymbol{\Delta}\middle|\mathbf{D}\right)\right]=Tr\left[E_{M_k|\mathbf{D}}Cov\left(\boldsymbol{\Delta}\middle|\mathbf{D},M_k\right)\right]+Tr\left[Cov_{M_k|\mathbf{D}}E\left(\boldsymbol{\Delta}\middle|\mathbf{D},M_k\right)\right] \qquad \text{(Eq. 8)}$$

which provides a scalar measure of the posterior variance of $\boldsymbol{\Delta}$. The latter is of interest because, for $K>1$ (multiple models), one generally has $Tr\left[Cov_{M_k|\mathbf{D}}E\left(\boldsymbol{\Delta}\middle|\mathbf{D},M_k\right)\right]>0$ so that $Tr\left[Cov\left(\boldsymbol{\Delta}\middle|\mathbf{D}\right)\right]>Tr\left[E_{M_k|\mathbf{D}}Cov\left(\boldsymbol{\Delta}\middle|\mathbf{D},M_k\right)\right]$. Hence the consideration of multiple models generally results in greater predictive uncertainty, as measured by $Tr\left[Cov\left(\boldsymbol{\Delta}\middle|\mathbf{D}\right)\right]$, than the uncertainty associated with a single model, as measured by $Tr\left[Cov\left(\boldsymbol{\Delta}\middle|\mathbf{D},M_k\right)\right]$.

Maximum Likelihood Bayesian Model Averaging (MLBMA)

BMA defines the integrated likelihood $p\left(\mathbf{D}\middle|M_k\right)$ of model M_k entirely in terms of the prior parameter density $p\left(\boldsymbol{\theta}_k\middle|M_k\right)$ of model parameters, having thus no provision for the conditioning of model parameters on measurements **D** (i.e., for the estimation of optimum model parameters on the basis of **D** using inverse methods). Instead, it requires computing the integral in (Eq. 4) through exhaustive sampling of the prior parameter space $\boldsymbol{\theta}_k$ for each model followed by numerical integration. One way to resolve both issues is to replace $\boldsymbol{\theta}_k$ by an estimate, $\hat{\boldsymbol{\theta}}_k^D$, which maximizes the likelihood $p\left(\mathbf{D}\middle|M_k,\boldsymbol{\theta}_k\right)$. Obtaining such maximum likelihood (ML) estimates entails calibrating each model against (conditioning on) the data **D** using well-established statistical inverse methods. Approximating $p\left(\boldsymbol{\Delta}\middle|\mathbf{D},M_k\right)$ by $p\left(\boldsymbol{\Delta}\middle|\mathbf{D},M_k\right)_{ML}$, where the subscript indicates ML estimation of $\boldsymbol{\theta}_k$, was shown to be useful (for a scalar Δ) in the statistical literature. Neuman [17] proposed evaluating the weights $p\left(M_k\middle|\mathbf{D}\right)$ based on a result due to Kashyap [18] according to [19]

$$p\left(M_k\middle|\mathbf{D}\right)\approx p\left(M_k\middle|\mathbf{D}\right)_{ML}=\frac{\exp\left(-\frac{1}{2}\delta KIC_k^D\right)p\left(M_k\right)}{\sum_{l=1}^{K}\exp\left(-\frac{1}{2}\delta KIC_l^D\right)p\left(M_l\right)} \qquad \text{(Eq. 9)}$$

where

$$\delta KIC_k^D=KIC_k^D-KIC_{\min}^D, \qquad \text{(Eq. 10)}$$

$$KIC_k^D=-2\ln p\left(\mathbf{D}\middle|M_k,\hat{\boldsymbol{\theta}}_k^D\right)_{ML}-2\ln p\left(\hat{\boldsymbol{\theta}}_k^D\middle|M_k\right)_{ML}+N_k\ln\left(\frac{N^D}{2\pi}\right)+\ln\left|\mathbf{F}_k\left(\mathbf{D}\middle|M_k\right)_{ML}\right|, \qquad \text{(Eq. 11)}$$

KIC_k^D being the so-called Kashyap model selection (or information) criterion for model M_k, $KIC_{\min}^D$ its minimum value over all candidate models, and $-2\ln p\left(\mathbf{D}\middle|M_k,\boldsymbol{\theta}_k\right)_{ML}-2\ln p\left(\boldsymbol{\theta}_k\middle|M_k\right)_{ML}$ a negative log likelihood incorporating prior measurements of the parameters (if available), evaluated at $\hat{\boldsymbol{\theta}}_k^D$. Here N_k is the dimension of $\boldsymbol{\theta}_k$ (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 $\mathbf{F}_k$ is the normalized (by N^D) observed (as opposed to ensemble mean) Fisher information matrix having components

$$F_{k,nm} = -\frac{1}{N^D}\left[\frac{\partial^2 \ln p\left(\mathbf{D}\middle|M_k,\boldsymbol{\theta}_k\right)}{\partial\theta_n\partial\theta_m}\right]_{\boldsymbol{\theta}_k=\hat{\boldsymbol{\theta}}_k^D} . \qquad \text{(Eq. 12)}$$

In the limit of large N^D / N_k, KIC_k^D reduces asymptotically to the so-called Bayesian selection (or information) criterion (e.g. [20])

$$BIC_k^D = -2\ln p\left(\mathbf{D}\middle|M_k,\hat{\boldsymbol{\theta}}_k^D\right)_{ML} + N_k \ln N^D . \qquad \text{(Eq. 13)}$$

EFFECT OF DATA AUGMENTATION ON UNCERTAINTY

We now ask the question how would an augmented data base $\left\{\mathbf{D},\mathbf{C}'\right\}$, where $\mathbf{C}'$ is a potential new dataset not presently available (and thus unknown), impact the above MLBMA uncertainty analysis. We address the question through MLBMA prediction of $\mathbf{C}'$, denoted by **C**, and assessment of the corresponding predictive uncertainty. Our proposed analysis entails the following steps:

1. Postulate a set **M** of *K* mutually independent models, M_k, with parameters $\boldsymbol{\theta}_k$ for the desired output vector, $\boldsymbol{\Delta}$;
2. Obtain ML estimates $\hat{\boldsymbol{\theta}}_k^D$ of $\boldsymbol{\theta}_k$ by calibrating each M_k against available data **D** through minimization of the log likelihood $-2\ln p\left(\mathbf{D}\middle|M_k,\boldsymbol{\theta}_k\right) - 2\ln p\left(\boldsymbol{\theta}_k\middle|M_k\right)$, then compute a corresponding estimation covariance $\boldsymbol{\Gamma}_k^D$ and KIC_k^D;
3. Compute $p\left(M_k\middle|\mathbf{D}\right)_{ML^D} = \exp\left(-\frac{1}{2}\delta KIC_k^D\right)p\left(M_k\right) / \sum_{l=1}^K \exp\left(-\frac{1}{2}\delta KIC_l^D\right)p\left(M_l\right)$ where the subscript ML^D designates the ML estimation process in step 2;
4. For each model M_k estimate $E\left(\boldsymbol{\Delta}\middle|\mathbf{D},M_k\right)_{ML^D}$ and $Cov\left(\boldsymbol{\Delta}\middle|\mathbf{D},M_k\right)_{ML^D}$ either through linearization or via Monte Carlo simulation (both options are explored in our synthetic example below):
    1. Draw random samples (realizations) of $\boldsymbol{\theta}_k$ from a multivariate Gaussian distribution with mean $\hat{\boldsymbol{\theta}}_k^D$ and covariance $\boldsymbol{\Gamma}_k^D$;
    2. Estimate $E\left(\boldsymbol{\Delta}\middle|\mathbf{D},M_k,\boldsymbol{\theta}_k\right)_{ML^D}$ and $Cov\left(\boldsymbol{\Delta}\middle|\mathbf{D},M_k,\boldsymbol{\theta}_k\right)_{ML^D}$ for each realization of $\boldsymbol{\theta}_k$;
    3. Average over all realizations of $\boldsymbol{\theta}_k$ to obtain sample estimates of $E\left(\boldsymbol{\Delta}\middle|\mathbf{D},M_k\right)_{ML^D}$ and $Cov\left(\boldsymbol{\Delta}\middle|\mathbf{D},M_k\right)_{ML^D}$;
5. Compute $E\left(\boldsymbol{\Delta}\middle|\mathbf{D}\right)_{ML^D}$, $Cov\left(\boldsymbol{\Delta}\middle|\mathbf{D}\right)_{ML^D}$ and/or $Tr\left[Cov\left(\boldsymbol{\Delta}\middle|\mathbf{D}\right)_{ML^D}\right]$;
6. Postulate a set **P** of *I* alternative geostatistical, statistical or stochastic models, P_i, with parameters $\boldsymbol{\pi}_i$ for a potential (presently unavailable) data set **C**; 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 examples described below;
7. Predict multivariate statistics of **C**, conditional on **D**, via MLBMA by means of the model set **P** using a procedure paralleling that described for $\boldsymbol{\Delta}$ in steps 2 – 6;
8. Estimate $E\left(\boldsymbol{\Delta}\middle|\mathbf{D},\mathbf{C}\right)_{ML^{D,C}}$ and $Cov\left(\boldsymbol{\Delta}\middle|\mathbf{D},\mathbf{C}\right)_{ML^{D,C}}$, where the subscript $ML^{D,C}$ designates the ML estimation process in step 2 with respect to an augmented data set $\left\{\mathbf{D},\mathbf{C}\right\}$, either through linearization followed by step 10 or via Monte Carlo simulation by using the statistics of **C** from step 7 to generate random realizations of **C** (both options are explored in our synthetic example); for each realization and for each model M_k:
    1. Obtain ML estimates $\hat{\boldsymbol{\theta}}_k^{D,C}$ of $\boldsymbol{\theta}_k$ by minimizing this negative log likelihood with respect to $\boldsymbol{\theta}_k$, then compute the corresponding estimation covariance $\boldsymbol{\Gamma}_k^{D,C}$ and $KIC_k^{D,C}$;

b. Compute $p\left(M_k \left|\mathbf{D},\mathbf{C}\right.\right)_{ML^{D,C}}$;

c. For each model M_k estimate $E\left(\boldsymbol{\Delta}\left|\mathbf{D},\mathbf{C},M_k\right.\right)_{ML^{D,C}}$ and $Cov\left(\boldsymbol{\Delta}\left|\mathbf{D},\mathbf{C},M_k\right.\right)_{ML^{D,C}}$ via Monte Carlo simulation:

   i. Draw random samples (realizations) of $\boldsymbol{\theta}_k$ from a multivariate Gaussian distribution with mean $\hat{\boldsymbol{\theta}}_k^{D,C}$ and covariance $\boldsymbol{\Gamma}_k^{D,C}$;

   ii. Estimate $E\left(\boldsymbol{\Delta}\left|\mathbf{D},\mathbf{C},M_k,\boldsymbol{\theta}_k\right.\right)_{ML^{D,C}}$ and $Cov\left(\boldsymbol{\Delta}\left|\mathbf{D},\mathbf{C},M_k,\boldsymbol{\theta}_k\right.\right)_{ML^{D,C}}$ for each realization of $\boldsymbol{\theta}_k$;

   iii. Average over all realizations of $\boldsymbol{\theta}_k$ to obtain sample estimates of $E\left(\boldsymbol{\Delta}\left|\mathbf{D},\mathbf{C},M_k\right.\right)_{ML^{D,C}}$ and $Cov\left(\boldsymbol{\Delta}\left|\mathbf{D},\mathbf{C},M_k\right.\right)_{ML^{D,C}}$;

d. Compute $E\left(\boldsymbol{\Delta}\left|\mathbf{D},\mathbf{C}\right.\right)_{ML^{D,C}}$, $Cov\left(\boldsymbol{\Delta}\left|\mathbf{D},\mathbf{C}\right.\right)_{ML^{D,C}}$ and/or $Tr\left[Cov\left(\boldsymbol{\Delta}\left|\mathbf{D},\mathbf{C}\right.\right)_{ML^{D,C}}\right]$;

9. Average over all realizations of $\mathbf{C}$ to obtain sample estimates of $E\left(\boldsymbol{\Delta}\left|\mathbf{D}\right.\right)_{ML^{D,C}}$, $Cov\left(\boldsymbol{\Delta}\left|\mathbf{D}\right.\right)_{ML^{D,C}}$ and/or $Tr\left[Cov\left(\boldsymbol{\Delta}\left|\mathbf{D}\right.\right)_{ML^{D,C}}\right]$;

10. Repeat steps 6 - 9 for different sets $\mathbf{C}_1,\mathbf{C}_2,\mathbf{C}_3\ldots$ of potential data and select that set which maximizes the difference $Tr\left[Cov_{\mathbf{C}|\mathbf{D}}E\left(\boldsymbol{\Delta}\left|\mathbf{D},\mathbf{C}\right.\right)_{ML^{D,C}}\right]=Tr\left[Cov\left(\boldsymbol{\Delta}\left|\mathbf{D}\right.\right)_{ML^{D,C}}\right]-Tr\left[E_{\mathbf{C}|\mathbf{D}}Cov\left(\boldsymbol{\Delta}\left|\mathbf{D},\mathbf{C}\right.\right)_{ML^{D,C}}\right]$ between the trace conditional on **D** and the expected trace conditional on **D** and **C** (this step is outside the scope of the present paper).

SYNTHETIC GEOSTATISTICAL EXAMPLE

We implement the above procedure of assessing data-worth by considering multiple variogram models (representing alternative assumptions about the spatial structure) of a zero-mean spatially correlated random field (such as the natural logarithm of hydraulic conductivity), $Z\left(\mathbf{x}\right)$, having point support (scale of measurement) in two dimensions, $\mathbf{x}=\left(x_1,x_2\right)^T$. In particular, we use a modified version of the sequential Gaussian simulation code SGSIM [21] to generate an unconditional realization of *Z* on a grid of 50×50 nodes using a truncated power variogram model with Gaussian modes (TpvG) given in [22],

$$\gamma\left(s;\lambda_u\right)=\sigma^2\left(\lambda_u\right)\left\{1-\exp\left[-\frac{\pi}{4}\left(\frac{s}{\lambda_u}\right)^2\right]+\left[\frac{\pi}{4}\left(\frac{s}{\lambda_u}\right)^2\right]^H\Gamma\left[1-H,\frac{\pi}{4}\left(\frac{s}{\lambda_u}\right)^2\right]\right\}\qquad 0<H<1 \qquad \text{(Eq. 14)}$$

where $s=\left\|\mathbf{x}_1-\mathbf{x}_2\right\|$ is separation distance (lag) between *Z* values at any two points $\mathbf{x}_1$ and $\mathbf{x}_2$, λ_u is an upper cutoff scale proportional to domain size, *A* is a coefficient, *H* is a Hurst scaling exponent, $\sigma^2\left(\lambda_u\right)=A\lambda_u^{2H}/2H$ is variance (sill) and $\Gamma\left(\cdot,\cdot\right)$ is the incomplete gamma function. The corresponding integral (spatial correlation) scale is $I\left(\lambda_u\right)=2H\lambda_u/\left(1+2H\right)$. We set the parameters of the TpvG model equal to $\boldsymbol{\theta}=\left(A,H,\lambda_u\right)^T=\left(0.1,0.25,5\right)^T$ which correspond to $\sigma^2=0.45$ and $I=1.67$. We then generate a "true" sample of 2,500 *Z* values at 50×50 nodes of a square grid, spaced a unit distance apart, as shown in Fig. 2. After verifying that a sample variogram based on all the generated values reproduces the original TpvG very closely we select 100 *Z* values at randomly located nodes to comprise a vector $\mathbf{D}$ of “available” data, 20 values to form a vector $\mathbf{C}'$ at other randomly located “potential new” sampling nodes (in real applications these locations will not be random but optimized in step 10 of our proposed procedure), and those at the remaining 2,380 nodes to make up a vector $\boldsymbol{\Delta}$ of "unknown" values that we wish to predict. The latter is based either on **D** or on $\left\{\mathbf{D},\mathbf{C}\right\}$ where **C** are values simulated randomly at the “potential new” sampling nodes conditional on **D**.

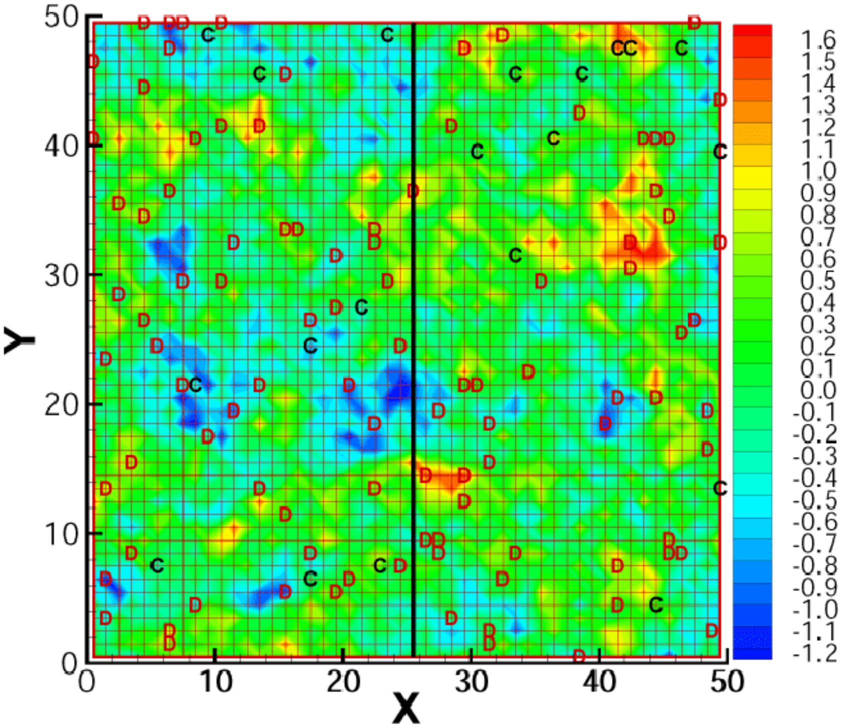

Fig. 2. "True" random field *Z* (colored) generated using TpvG with at 2500 grid nodes (shown). D represent locations of "available" data and C those of "potential new" data.

To predict $\mathbf{\Delta}$ we consider a set $\mathbf{M}$ of $K = 3$ alternative variogram models, M_k, having parameters $\boldsymbol{\theta}_k$ (purposely excluding the generating, or "true," TpvG model): exponential (Exp), Gaussian (Gau) and spherical (Sph) [e.g. 19]. Each model, M_k, is assigned an equal prior probability, $p(M_k) = 1/3$, and is calibrated against $\mathbf{D}$ to yield ML estimates $\hat{\boldsymbol{\theta}}_k^{\mathbf{D}}$ of $\boldsymbol{\theta}_k$ by minimizing the joint negative log likelihood (NLL) $-2\ln p\left(\mathbf{D}\middle|M_k, \boldsymbol{\theta}_k\right)$. The process, denoted by ML^D, also yields corresponding parameter estimation covariance matrices $\boldsymbol{\Gamma}_k^D$, Kashyap criteria $KIC_k^{\mathbf{D}}$ and Bayesian criteria $BIC_k^{\mathbf{D}}$. For comparison we also compute information theoretic criteria

$$AIC_k^D = -2\ln p\left(\mathbf{D}\middle|M_k\right)_{ML} + 2N_k \qquad \text{(Eq. 15)}$$

$$AICc_k^D = -2\ln p\left(\mathbf{D}\middle|M_k\right)_{ML} + 2N_k + \frac{2N_k(N_k+1)}{N_z - N_k - 1} \qquad \text{(Eq. 16)}$$

introduced, respectively, by Akaike [23] and Hurvich and Tsai [24]. $KIC_k^{\mathbf{D}}$ and $BIC_k^{\mathbf{D}}$ are used to compute posterior model probabilities (or, in the case of $AIC_k^{\mathbf{D}}$ and $AICc_k^{\mathbf{D}}$, model weights), $p(M_k \,|\, \mathbf{D})_{ML^D}$, for each model. The results, listed in Table 1, indicate that the sample $\mathbf{D}$ of available data is not large enough to reproduce correctly the TpvG model used to generate it; the corresponding sample and fitted variograms underestimate the true sill and overestimate the true integral scale. All criteria favor the spherical model, assigning a very low posterior probability (or weight) to the Gaussian model. Our model averaged results are based on $KIC_k^{\mathbf{D}}$.

Augmenting the sample to include $\left\{\mathbf{D}, \mathbf{C}'\right\}$ yields results listed in Table 2. They show that such data augmentation is still not enough to reproduce correctly the TpvG model; the sample and fitted variograms underestimate the true sill to a greater extent than was the case with **D** alone but overestimate the true integral scale to a lesser extent. The preference of all criteria for the spherical model is now more pronounced (less ambiguous) than it was in the case of **D**. These results indicate a need to account for the effect of potential new data on parameter estimation and model weighting, as we do next.

Table 1. Parameter estimates; negative log likelihoods *NLL*; model selection criteria *AIC*, *AICc*, *BIC* and *KIC*; prior and posterior model probabilities; and rankings of variogram models based on **D**.

Model	**Exp**	**Gau**	**Sph**
Sill Estimate	0.404	0.401	0.399
std* of Sill	0.226	0.233	0.222
Integral Scale Estimate	3.686	1.979	2.681
std* of Integral scale	0.574	0.187	0.283
NLL	-145.14	-136.13	-147.23
NLL Rank	2	3	1
$p(M_k)$	1/3	1/3	1/3
AIC	-139.14	-130.13	-141.23
AIC Rank	2	3	1
$p\left(M_k \middle\| \mathbf{D}\right)_{AIC}$	25.98%	0.29%	73.74%
AICc	-138.91	-129.90	-141.00
AICc Rank	2	3	1
$p\left(M_k \middle\| \mathbf{D}\right)_{AICc}$	25.98%	0.29%	73.74%
BIC	-131.04	-122.03	-133.12
BIC Rank	2	3	1
$p\left(M_k \middle\| \mathbf{D}\right)_{BIC}$	25.98%	0.29%	73.74%
KIC	-146.89	-135.94	-147.29
KIC Rank	2	3	1
$p\left(M_k \middle\| \mathbf{D}\right)_{KIC}$	44.92%	0.19%	54.90%
* std represents standard deviation			

Predictions of $\mathbf{\Delta}$ are uncertain due to random spatial fluctuations in *Z* as well as uncertainty about the variogram model, M_k, and its parameters, $\boldsymbol{\theta}_k$. As $\mathbf{\Delta}$ is generally nonlinear in $\boldsymbol{\theta}_k$, one can estimate its lead moments either through linearization or via Monte Carlo simulation. We start with the latter option by drawing $R_\theta = 2{,}000$ random realizations, $\boldsymbol{\theta}_k^{r_\theta}$, of $\boldsymbol{\theta}_k$ from a multivariate normal distribution $\boldsymbol{\theta}_k \sim \mathcal{N}(\hat{\boldsymbol{\theta}}_k^D, \Gamma_k^D)$ where $r_\theta = 1, 2, \ldots R_\theta$ for each M_k; obtaining kriging (minimum variance unbiased linear) estimates $E\left(\Delta_p \middle| \mathbf{D}, M_k, \boldsymbol{\theta}_k^{r\theta}\right)_{ML^D}$ and kriging (estimation) variances $Var\left(\Delta_p \middle| \mathbf{D}, M_k, \boldsymbol{\theta}_k^{r\theta}\right)_{ML^D}$ and covariances $Cov\left(\Delta_p \Delta_q \middle| \mathbf{D}, M_k, \boldsymbol{\theta}_k^{r\theta}\right)_{ML^D}$ for all components Δ_p and Δ_q of $\mathbf{\Delta}$; averaging these over all R_θ realizations to obtain $E\left(\Delta_p \middle| \mathbf{D}, M_k\right)_{ML^D}$, $Cov\left(\Delta_p \Delta_q \middle| \mathbf{D}, M_k\right)_{ML^D}$, $Cov\left(\Delta_p \Delta_q \middle| \mathbf{D}, M_k, \boldsymbol{\theta}_k^{r\theta}\right)_{ML^D}$; verifying that $R_\theta = 2{,}000$ is large enough for $Tr\left[Cov\left(\mathbf{\Delta} \middle| \mathbf{D}, M_k\right)_{ML^D}\right]$ to stabilize for each model; and computing the model-averaged quantities $E\left(\Delta_p \middle| \mathbf{D}\right)_{ML^D}$ and $Cov\left(\Delta_p \Delta_q \middle| \mathbf{D}\right)_{ML^D}$.

A similar procedure is used to obtain $E\left(\Delta_p \middle| \mathbf{D}, \mathbf{C}'\right)_{ML^{D,C'}}$ and $Cov\left(\Delta_p \Delta_q \middle| \mathbf{D}, \mathbf{C}'\right)_{ML^{D,C'}}$ where, we recall, $\mathbf{C}'$ are additional data yet to be collected (known *a priori* in our example). Fig. 3 shows how the difference $Var\left(\Delta_p \middle| \mathbf{D}\right)_{ML^D} - Var\left(\Delta_p \middle| \mathbf{D}, \mathbf{C}'\right)_{ML^{D,C'}}$ varies across the grid. It is seen that enlarging the data base from 100 (the dimension of **D**) to 120 (the dimension of $\left\{\mathbf{D}, \mathbf{C}'\right\}$) reduces the prediction variance across the grid, most noticeably in its upper right quadrant.

Table 2. Parameter estimates; negative log likelihoods *NLL*; model selection criteria *AIC*, *AICc*, *BIC* and *KIC*; prior and posterior model probabilities; and rankings of variogram models based on $\{\mathbf{D}, \mathbf{C}'\}$.

Model	**Exp**	**Gau**	**Sph**
Sill Estimate	0.371	0.372	0.373
std* of Sill	0.214	0.220	0.210
Integral Scale Estimate	2.409	1.907	2.672
std* of Integral scale	0.403	0.151	0.311
NLL	-183.86	-175.59	-188.36
NLL Rank	2	3	1
$p(M_k)$	1/3	1/3	1/3
AIC	-177.86	-169.59	-182.36
AIC Rank	2	3	1
$p\left(M_k \middle\| \mathbf{D}\right)_{AIC}$	9.52%	0.15%	90.33%
AICc	-177.67	-169.40	-182.17
AICc Rank	2	3	1
$p\left(M_k \middle\| \mathbf{D}\right)_{AICc}$	9.52%	0.15%	90.33%
BIC	-169.25	-160.99	-173.76
BIC Rank	2	3	1
$p\left(M_k \middle\| \mathbf{D}\right)_{BIC}$	9.52%	0.15%	90.33%
KIC	-184.80	-174.87	-188.51
KIC Rank	2	3	1
$p\left(M_k \middle\| \mathbf{D}\right)_{KIC}$	13.51%	0.09%	86.39%
* std represents standard deviation			

Since in real applications $\mathbf{C}'$ would not be available, the next step is to compute the statistics of potential data $\mathbf{C}$ conditional on **D**. As in our case $\mathbf{\Delta}$ and $\mathbf{C}$ describe the same attribute at different locations, the statistics of $\mathbf{C}$ are obtained simply upon replacing $\mathbf{\Delta}$ in the above procedure with $\mathbf{C}$. We use these statistics to generate $R_c = 200$ random realizations, $\mathbf{C}^{r_C}$, $r_c = 1, 2, \ldots R_c$, of $\mathbf{C}$ by considering it to be multivariate normal, $\mathbf{C} \sim \mathcal{N}\left[E\left(\mathbf{C}\middle|\mathbf{D}\right)_{ML^D}, Cov\left(\mathbf{C} \mid \mathbf{D}\right)_{ML^D}\right]$. For every realization $\mathbf{C}^{r_C}$ we predict $\mathbf{\Delta}$ the same way as before but now conditional on an expanded data base $\left\{\mathbf{D}, \mathbf{C}^{r_C}\right\}$. This yields $E\left(\Delta_p \middle| \mathbf{D}, \mathbf{C}^{r_C}\right)_{ML^{D,C}}$ and $Cov\left(\Delta_p \Delta_q \middle| \mathbf{D}, \mathbf{C}^{r_C}\right)_{ML^{D,C}}$ where the subscript $ML^{D,C}$ denotes ML estimation based on $\left\{\mathbf{D}, \mathbf{C}^{r_C}\right\}$. The latter are then averaged over all realizations of $\mathbf{C}^{r_C}$ to obtain $E\left(\Delta_p \middle| \mathbf{D}\right)_{ML^{D,C}}$, $Cov\left(\Delta_p \Delta_q \middle| \mathbf{D}\right)_{ML^{D,C}}$ and $Tr\left[Cov\left(\mathbf{\Delta}\middle|\mathbf{D}\right)_{ML^{D,C}}\right]$. The final step entails computing $Tr\left[Var_{\mathbf{C}|\mathbf{D}} E\left(\mathbf{\Delta}\middle|\mathbf{D}, \mathbf{C}\right)_{ML^{D,C}}\right]$ which, we recall, represents $Tr\left[Var\left(\mathbf{\Delta}\middle|\mathbf{D}\right)_{ML^{D,C}}\right] - Tr\left[E_{\mathbf{C}|\mathbf{D}} Var\left(\mathbf{\Delta}\middle|\mathbf{D}, \mathbf{C}\right)_{ML^{D,C}}\right]$; we verify that sample estimates of all these three terms stabilize after 200 realizations. Variation of $Var_{\mathbf{C}|\mathbf{D}} E\left(\mathbf{\Delta}\middle|\mathbf{D}, \mathbf{C}\right)_{ML^{D,C}}$ across the grid is shown in Fig. 4. A comparison of Fig.s 3 and 4 confirms that the estimated variance reduction $Var_{\mathbf{C}|\mathbf{D}} E\left(\mathbf{\Delta}\middle|\mathbf{D}, \mathbf{C}\right)_{ML^{D,C}}$ varies in a manner similar to that of the true variance reduction $Var\left(\mathbf{\Delta}\middle|\mathbf{D}\right)_{ML^D}$ -

$Var\left(\mathbf{\Delta}\middle|\mathbf{D},\mathbf{C}'\right)_{ML^{D,C'}}$. Correspondingly $Tr\left[Var_{C|D}E(\mathbf{\Delta}\mid\mathbf{D},\mathbf{C})_{ML^{D,C}}\right]$ = 54.61 approximates closely the true trace reduction $Tr\left[Var\left(\mathbf{\Delta}\middle|\mathbf{D}\right)_{ML^{D}}\right]-Tr\left[Var\left(\mathbf{\Delta}\middle|\mathbf{D},\mathbf{C'}\right)_{ML^{D,C'}}\right]$ = 58.95.

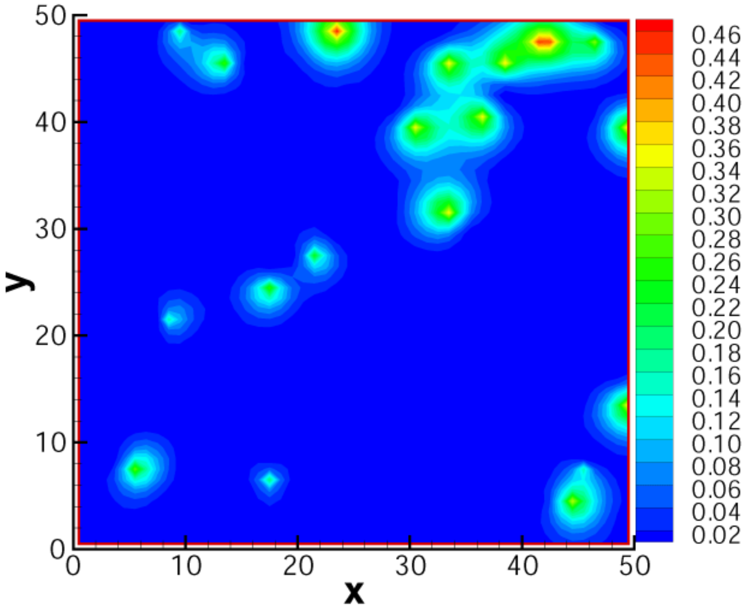

Fig. 3. Variation of $Var\left(\Delta_p\middle|\mathbf{D}\right)_{ML^{D}}-Var\left(\Delta_p\middle|\mathbf{D},\mathbf{C}'\right)_{ML^{D,C'}}$ across the grid.

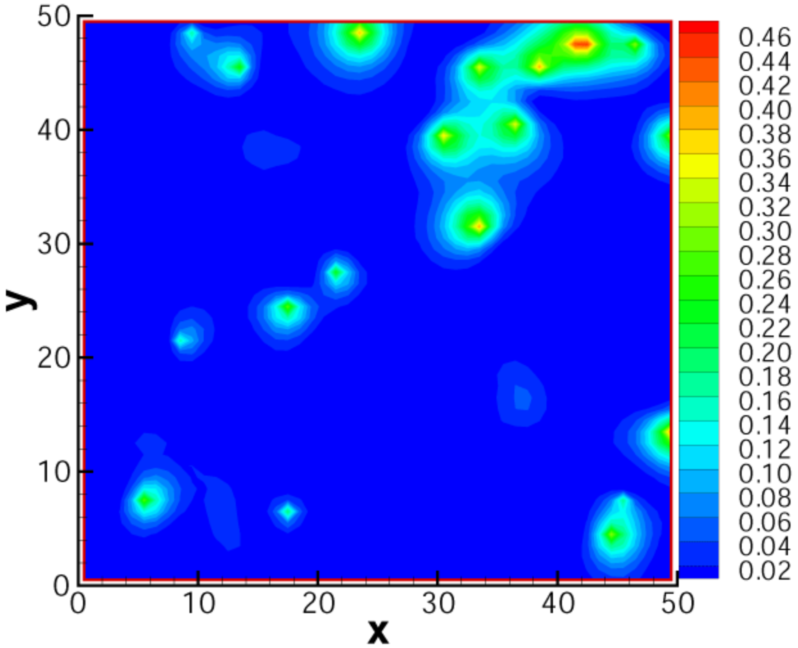

Fig. 4. Variation of $Var_{\mathbf{C}|\mathbf{D}}E\left(\mathbf{\Delta}\middle|\mathbf{D},\mathbf{C}\right)_{ML^{D,C}}$ across the grid.

The above results are based on Monte Carlo evaluation of all moments. Estimating the lead moments of $\mathbf{\Delta}$ through linearization brings about a ten-fold reduction in central processor time without any serious effect on accuracy.

CONCLUSIONS

Our paper leads to the following major conclusions:

1. A multimodel approach to optimum value-of-information or data-worth analyses has been proposed based on a Bayesian model averaging (BMA) framework. We have focused on a maximum likelihood (MLBMA) variant of BMA 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 approximating lead predictive moments of any model by linearization, and (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.
2. The proposed approach should be of help to the DOE in designing the collection of characterization and monitoring data at contaminated sites in a cost-effective manner by maximizing their benefit under given cost constraints. Benefits would accrue from optimum gain in information, or reduction in predictive uncertainty (and risk), upon considering jointly not only traditional sources of uncertainty such as those affecting model parameters and the reliability of data but also lack of certainty about the underlying models.
3. Implementation of the proposed approach on a synthetic geostatistical problem in two space dimensions demonstrates 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 are sufficient to identify correctly the underlying geostatistical model (variogram) and its parameters, they nevertheless yield self-consistent results and allow identifying quite accurately the impacts of potential new data on the spatial distribution and magnitude of corresponding reductions in predictive variance.
4. Approximating lead predictive moments associated with each model by linearization has yielded results comparable to those obtained via Monte Carlo simulation with a much lesser expenditure of computational time and effort.

ACKNOWLEDGEMENT

This research was 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 U.S. Department of Energy. The third and fourth authors were supported in part by NSF-EAR grant 0911074 and DOE-ERSP grant DE-SC0002687.

REFERENCES

1. P.-E. BACK, A model for estimating the value of sampling programs and the optimal number of samples for contaminated soil, Environ. Geol., doi:10.1007/s00254-0488-6, 52, 573-585 (2007).
2. M. THOMPSON and T. FEARN, What exactly is fitness for purpose in analytical measurements? Analyst, 121(March), 275-278 (1996).
3. K.T. RUSSELL and A.J. RABIDEAU, Decision analysis for pump-and-treat design, Groundwater Monitoring and Remediation, Summer, 159-268 (2000).
4. A.M. DAUSMAN, J. DOHERTY, C.D. LANGEVIN, and M.C. SUKOP (2010), Quantifying data worth toward reducing predictive uncertainty, Ground Water, 48(5), 729-740.
5. B.R. JAMES and R.A. FREEZE, The worth of data in predicting aquitard continuity in hydrogeological design, Water Resour. Res., 29(7), 2049-2065 (1993).
6. F. YOKOTA and K. THOMPSON, Value of information literature analysis: A review of applications in health risk assessment, Med. Decis. Mak., 24, 287 (2004).
7. L. FEYEN and S.M. GORELICK, Framework to evaluate the worth of hydraulic conductivity data for optimal groundwater resources management in ecologically sensitive areas, Water Resour. Res., 41, W03019, doi:10.1029/2003WR002901 (2005).
8. T. NORBERG and L. ROSEN, Calculating the optimal number of contaminant samples by means of data worth analysis, Environmetrics, 17, 705-719 (2006).

9. J. FREER, K. BEVEN, and B. AMBROISE, Bayesian estimation of uncertainty in runoff predictioin and the value of data: An application of the GLUE approach, Water Resour. Res., 32(7), 2161-2173 (1996).
10. R. ROJAS, L. FEYEN, O. BATELAAN, and A. DASSARGUES, On the value of conditioning data to reduce conceptual model uncertainty in groundwater modeling, Water Resour. Res., doi:10.1029/2009WR008822, in press (2010).
11. K.J. BEVEN and J. FREER, Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. Jour. Hydrology, 249: 11–29 (2001).
12. J.A. HOETING, D. MADIGAN, A.E. RAFTERY, and C.T. VOLINSKY, Bayesian model averaging: A tutorial, Statist. Sci., 14(4), 382-417 (1999).
13. W. NOWAK, F.P.J. DE BARROS, and Y. RUBIN, Bayesian geostatistical design: Task-driven site investigation when the geostatistical model is uncertain, Water Resour. Res., 46, W035535, doi:10.1029/2009WR008312 (2010).
14. R.A. FREEZE, J. MASSMAN, L. SMITH, T. SPERLING, and B. JAMES, Hydrogeological decision analysis: 1. A framework, Ground Water, 28(5), 738-266 (1990).
15. R.A. FREEZE, J. BRUDE, J. MASSMAN, T. SPERLING, and L. SMITH, Hydrogeological decision analysis: 4. The concept of data worth and its use in the development of site investigation strategies, Ground Water, 30(4), 574-288 (1992).
16. S.R. SAIN and R. FURRER, Combining climate model output via model correlations, Stochastic Environmental Research and Risk Assessment, 24(6), 821-829, DOI: 10.1007/s00477-010-0380-5 (2010).
17. S.P. NEUMAN, Maximum likelihood Bayesian averaging of alternative conceptual-mathematical models, Stochastic Environmental Research and Risk Assessment, 17(5), 291-305, DOI: 10.1007/s00477-003-0151-7 (2003).
18. R.L. KASHYAP, Optimal choice of AR and MA parts in autoregressive moving average models, IEEE Transactions on Pattern Analysis and Machine Intelligence, 4(2), 99-104 (1982).
19. M. YE, S.P. NEUMAN, and P.D. MEYER, Maximum Likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff, Water Resour. Res., 40, W05113, doi:10.1029/2003WR002557 (2004).
20. M. YE, P.D. MEYER and S.P. NEUMAN, On model selection criteria in multimodel analysis, *Water Resour. Res.*, 44, W03428, doi:10.1029/2008WR006803 (2008).
21. C.V. DEUTSCH, and A.G. JOURNEL, GSLIB: Geostatistical Software Library and User's Guide, 2nd ed., Oxford Univ. Press, New York (1998).
22. V. DI FEDERICO and S.P. NEUMAN, scaling of random fields by means of truncated power variograms and associated spectra. Water Resouces Research, 33(5), 1075-1085 (1997).
23. H. AKAIKE, A new look at statistical model identification, IEEE Transactions on Automatic Control, AC-19, 716-722 (1974).
24. C.M. HURVICH, and C.-L. TSAI, Regression and time series model selection in small sample, Biometrika, 76(2), 99-104 (1989).