AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE

10.1002/2014WR015503

Multimodel Bayesian analysis of groundwater data worth

Liang Xue[1], Dongxiao Zhang[1], Alberto Guadagnini[2,3], and Shlomo P. Neuman[2]

[1]Department of Energy and Resources Engineering, College of Engineering, Peking University, Beijing, China,
[2]Department of Hydrology and Water Resources, University of Arizona, Tucson, Arizona, USA,
[3]Dipartimento di Ingegneria Civile e Ambientale, Politecnico di Milano, Milano, Italy

Key Points:
- Joint consideration of geostatistical and flow model uncertainties
- Combined assessment of added hydraulic conductivity and head data worth
- Inverse solution of stochastic moment equations combined with MLBMA

Correspondence to:
L. Xue,
xueliang@pku.edu.cn

Citation:
Xue, L., D. Zhang, A. Guadagnini, and S. P. Neuman (2014), Multimodel Bayesian analysis of groundwater data worth, *Water Resour. Res.*, *50*, 8481–8496, doi:10.1002/2014WR015503.

Received 24 FEB 2014
Accepted 4 OCT 2014
Accepted article online 9 OCT 2014
Published online 4 NOV 2014

Abstract We explore the way in which uncertain descriptions of aquifer heterogeneity and groundwater flow impact one's ability to assess the worth of collecting additional data. We do so on the basis of Maximum Likelihood Bayesian Model Averaging (MLBMA) by accounting jointly for uncertainties in geostatistical and flow model structures and parameter (hydraulic conductivity) as well as system state (hydraulic head) estimates, given uncertain measurements of one or both variables. Previous description of our approach was limited to geostatistical models based solely on hydraulic conductivity data. Here we implement the approach on a synthetic example of steady state flow in a two-dimensional random log hydraulic conductivity field with and without recharge by embedding an inverse stochastic moment solution of groundwater flow in MLBMA. A moment-equations-based geostatistical inversion method is utilized to circumvent the need for computationally expensive numerical Monte Carlo simulations. The approach is compatible with either deterministic or stochastic flow models and consistent with modern statistical methods of parameter estimation, admitting but not requiring prior information about the parameters. It allows but does not require approximating lead predictive statistical moments of system states by linearization while updating model posterior probabilities and parameter estimates on the basis of potential new data both before and after such data are actually collected.

1. Introduction

Sustainable development of groundwater resources requires credible analyses of associated subsurface flow regimes and their impact on water quality. Such analyses entail characterizing the spatial distribution of hydraulic parameters and factors controlling groundwater flow, embedding them in groundwater flow models, and using the models to provide predictions (and associated uncertainty) of groundwater flow under various development scenarios. Each such step is affected by uncertainties stemming in part from the complex hydrogeological makeup of the subsurface and lack of precise knowledge about conditions that would control each scenario. A question therefore arises to what extent might the collection of additional information about the system reduce predictive uncertainty, at what cost, and what potential benefits (including risk reduction) might this yield?

A review of literature pertaining to these questions was recently published by *Neuman et al*. [2012]. Most approaches in the literature utilize numerical Monte Carlo simulation. *Trainor-Guitton et al*. [2011, 2013] assess expected data worth of additional aquifer lithology data for decision making in the context of an aquifer vulnerability scenario. *Liu et al*. [2012] quantify the value of information in a groundwater remediation case using a bootstrap filter. *De Barros et al*. [2012] investigate the impact of hydrogeological data on predicting environmental performance metrics such as well drawdown, concentration, or contaminant travel time and human health risk. *Fang et al*. [2014] use data worth analysis to optimize a monitoring network and reduce data redundancy in a CO_2 sequestration scenario. A major limitation of many existing approaches is that they rely on a single conceptual-mathematical model of geologic or watershed makeup and of hydrologic processes therein. Yet hydrologic environments are open and complex, rendering them 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). *Leube et al*. [2012] develop a PreDIA (Preposterior Data Impact Assessor)

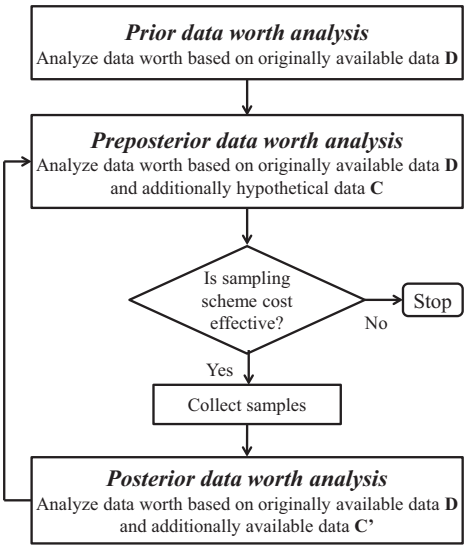

Figure 1. Overview of Bayesian data worth analysis [after *James and Gorelick*, 1994].

method to quantify expected data worth prior to data collection, which takes into account conceptual model uncertainty through Bayesian model averaging (BMA). *Nowak et al.* [2012] use PreDIA to drive optimal data collection design by minimizing the probability of making wrong decisions, i.e., accepting the alternative hypothesis when the null hypothesis is true and vice versa, during Bayesian hypothesis testing. This method relies on Monte Carlo simulations, which tend to be computationally cumbersome. As an alternative to BMA, *Neuman et al.* [2012] propose a multimodel data worth assessment approach based on a Maximum Likelihood version of the Bayesian Model Averaging (MLBMA). The approach is compatible with both deterministic and stochastic models, consistent with modern statistical methods of parameter estimation, admits but does not require prior information on the probability distribution of parameters, allows approximating lead predictive moments of any model by linearization, and updates model posterior probabilities as well as parameter estimates on the basis of potential new data both before and after such data become available.

Previous implementations of MLBMA data-worth assessments were limited to the geostatistical characterization of aquifer heterogeneity in the presence of multiple variogram models (and eventually measured values) of log hydraulic conductivity [*Neuman et al.*, 2012] and air permeability [*Lu et al.*, 2012] in two and three spatial dimensions, respectively. In this study, we explore the ability of their approach to deal simultaneously with system parameters and states, specifically hydraulic conductivities and hydraulic heads. We avoid Monte Carlo simulations by basing our groundwater flow model on stochastic moment equations (ME). Our approach takes advantage of the ME-based geostatistical inverse approach of *Hernandez et al.* [2003, 2006] and *Riva et al.* [2011]. The resulting algorithm is applied to a synthetic example of steady state groundwater flow in a two-dimensional random log hydraulic conductivity field with and without recharge.

2. Multimodel Bayesian Data-Worth Assessment Using MLBMA

Bayesian data-worth assessment is conducted in three stages [*James and Gorelick*, 1994; *Neuman et al.*, 2012]. The first (prior) stage relies entirely on information and data available at the outset. The second (preposterior) stage relies on statistics of potential new data estimated on the basis of prior information and data. The third and last (posterior) stage utilizes joint statistics of prior data and new data made available following the preposterior stage. Posterior statistics serve to assess the quality of their preposterior estimates and, optionally, to furnish statistics for a new prior assessment stage (see flowchart in Figure 1). Key elements of the process, originally described in *Neuman et al.* [2012], are recounted for completeness in subsection 2.1.

2.1. Multimodel Assessment of Data Worth in Groundwater Problems

We limit the description of our approach to a set **M** of K mutually exclusive stochastic moment models, M_k, capable of predicting lead statistics, such as mean and variance/covariance, of steady state groundwater flow in a given domain. The models are uncertain, having prior probabilities $p(M_k)$ that add up to 1. Each flow model considers log hydraulic conductivity to form a spatially correlated random field characterized by a given variogram model, which may differ from one flow model to another. Model M_k predicts the mean (expectation) $E(\mathbf{h}|\mathbf{D}, M_k)$ and the covariance $Cov(\mathbf{h}|\mathbf{D}, M_k)$ of a vector **h** of random hydraulic head values, conditional on a given vector **D** of prior data. The prior data may include measured log hydraulic conductivities as well as hydraulic heads at selected positions. Averaging across all K models renders the following (Bayesian-averaged) lead moments [*Draper*, 1995; *Hoeting et al.*, 1999]

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

$$\begin{aligned} Cov(\mathbf{h}|\mathbf{D}) &= E_{\mathbf{M}|\mathbf{D}} Cov(\mathbf{h}|\mathbf{D},\mathbf{M}) + Cov_{\mathbf{M}|\mathbf{D}} E(\mathbf{h}|\mathbf{D},\mathbf{M}) \\ &= \sum_{k=1}^{K} Cov(\mathbf{h}|\mathbf{D},M_k)p(M_k|\mathbf{D}) \\ &\quad + \sum_{k=1}^{K} [E(\mathbf{h}|\mathbf{D},M_k) - E(\mathbf{h}|\mathbf{D})] \\ &\quad \cdot [E(\mathbf{h}|\mathbf{D},M_k) - E(\mathbf{h}|\mathbf{D})]^T p(M_k|\mathbf{D}) \end{aligned} \tag{2}$$

where T denotes transpose. The conditional covariance, $Cov(\mathbf{h}|\mathbf{D})$, stemming from Bayesian model averaging (BMA), is seen to be the sum of a within-model covariance $E_{\mathbf{M}|\mathbf{D}} Cov(\mathbf{h}|\mathbf{D},\mathbf{M})$ and a between-model covariance $Cov_{\mathbf{M}|\mathbf{D}} E(\mathbf{h}|\mathbf{D},\mathbf{M})$. The posterior model probability, $p(M_k|\mathbf{D})$, weighs the contribution of model M_k to BMA moments. Appendix A shows how $p(M_k|\mathbf{D})$ is computed in BMA or its maximum likelihood version, MLBMA, that we use.

Next we ask how might augmenting the original data set, **D**, by an additional set of data, **C′**, affect predictive uncertainty? *Neuman et al*. [2012] propose quantifying this effect at the posterior stage, after **C′** has been collected, by the scalar measure of reference posterior data worth, $Tr[Cov(\mathbf{h}|\mathbf{D})] - Tr[Cov(\mathbf{h}|\mathbf{D},\mathbf{C}')]$, where *Tr* is the trace (sum of diagonal entries) of a matrix. Other measures of uncertainty in the context of optimal sampling design are discussed by *Nowak* [2010]. At the preposterior stage, before **C′** is collected, *Neuman et al*. [2012] generate random estimates **C** of $\mathbf{C}'$ on the basis of available data **D**. One way to generate such random estimates, **C**, is to draw them from a multivariate normal distribution

$$\mathbf{C} \sim N\{E(\mathbf{C}|\mathbf{D}), Cov(\mathbf{C}|\mathbf{D})\} \tag{3}$$

In our case, $E(\mathbf{C}|\mathbf{D})$ and $Cov(\mathbf{C}|\mathbf{D})$ are obtained from (1) and (2), respectively, upon replacing **h** with **C**. By virtue of the law of total covariance, a trace measure of the conditional predictive uncertainty can be decomposed according to

$$Tr[Cov(\mathbf{h}|\mathbf{D})] = Tr\left[E_{\mathbf{C}|\mathbf{D}} Cov(\mathbf{h}|\mathbf{D},\mathbf{C})\right] + Tr\left[Cov_{\mathbf{C}|\mathbf{D}} E(\mathbf{h}|\mathbf{D},\mathbf{C})\right] \tag{4}$$

where $E_{\mathbf{C}|\mathbf{D}} Cov(\mathbf{h}|\mathbf{D},\mathbf{C})$ is the expectation of $Cov(\mathbf{h}|\mathbf{D},\mathbf{C})$ over all **C** vectors generated via (3), and $Cov_{\mathbf{C}|\mathbf{D}} E(\mathbf{h}|\mathbf{D},\mathbf{C})$ is the covariance of $E(\mathbf{h}|\mathbf{D},\mathbf{C})$ over all these **C** vectors. Though the scalar measure of predictive head uncertainty, $Tr[Cov(\mathbf{h}|\mathbf{D})]$, at the preposterior stage is theoretically the same as that at the prior stage, computationally the two may differ somewhat from each other due to the finiteness of the **C** samples obtained through (3). We denote $Cov(\mathbf{h}|\mathbf{D})$ obtained at the prior stage by $Cov(\mathbf{h}|\mathbf{D})_D$ to indicate that it is evaluated on the basis of available data **D**, and that computed at the preposterior stage by $Cov(\mathbf{h}|\mathbf{D})_{DC}$ to indicate that the latter is evaluated on the basis of both available data **D** and generated data **C** in the manner just described. Given that the prior and preposterior predictive uncertainty measures $Tr\left[Cov(\mathbf{h}|\mathbf{D})_D\right]$ and $Tr\left[Cov(\mathbf{h}|\mathbf{D})_{DC}\right]$ coincide in theory, and that $Tr\left[E_{\mathbf{C}|\mathbf{D}} Cov(\mathbf{h}|\mathbf{D},\mathbf{C})\right]$ in (4) is the expected posterior predictive uncertainty measure $Tr[Cov(\mathbf{h}|\mathbf{D},\mathbf{C}')]$, it follows that the scalar measure $Tr\left[Cov_{\mathbf{C}|\mathbf{D}} E(\mathbf{h}|\mathbf{D},\mathbf{C})\right]$ of data worth computed at the preposterior stage is equal to the expected posterior data worth measure $Tr\left[Cov(\mathbf{h}|\mathbf{D})_D\right] - Tr[Cov(\mathbf{h}|\mathbf{D},\mathbf{C}')]$. It is thus clear that our analysis of data worth associated with any sampling scheme entails predictive uncertainty reduction through the collection of additional data. Other factors affecting sampling design include [e.g. *Nowak et al*., 2010, 2012; *De Barros et al*., 2012] its purpose, choice of uncertainty measure(s), type of environmental performance metric considered, and dimensionality of physical system model. Though we do not consider these factors in our work, they could easily be included in our method of assessing data worth.

2.2. ML Estimation of Model Parameters

As noted in Appendix A, MLBMA entails maximum likelihood estimation of model parameters. Here we accomplish this through a stochastic inverse procedure similar to that proposed for steady state groundwater flow by *Hernandez et al*. [2003, 2006] and for transient flow by *Riva et al*. [2009]. Following is a brief synopsis of our approach.

The ensemble mean (expectation), $\langle h(\mathbf{x})\rangle_c$, of hydraulic head $h(\mathbf{x})$ predicted by model M_k conditional (as indicated by the subscript c) on prior data $\mathbf{D}$, equivalent to $E(\mathbf{h}|\mathbf{D}, M_k)$, satisfies (in our case) the conditional mean steady state flow equation [*Guadagnini and Neuman*, 1999]

$$\nabla \cdot [\langle K(\mathbf{x})\rangle_c \nabla \langle h(\mathbf{x})\rangle_c - \mathbf{r}_c(\mathbf{x})] + \langle f(\mathbf{x})\rangle = 0 \tag{5}$$

subject to boundary conditions

$$\langle h(\mathbf{x})\rangle_c = \langle H(\mathbf{x})\rangle \qquad \text{on } \Gamma_D \tag{6}$$

$$[\langle K(\mathbf{x})\rangle_c \nabla \langle h(\mathbf{x})\rangle_c - \mathbf{r}_c(\mathbf{x})] \cdot \mathbf{n}(\mathbf{x}) = \langle Q(\mathbf{x})\rangle \qquad \text{on } \Gamma_N \tag{7}$$

where $\langle K(\mathbf{x})\rangle_c$ is conditional mean hydraulic conductivity, $K(\mathbf{x})$; $K'(\mathbf{x})$ represents zero mean random fluctuations in $K(\mathbf{x})$ about $\langle K(\mathbf{x})\rangle_c$; $\mathbf{r}_c(\mathbf{x}) = -\langle K'(\mathbf{x})\nabla h'(\mathbf{x})\rangle_c$ is residual flux; $h'(\mathbf{x})$ represents zero mean random fluctuations in head about $\langle h(\mathbf{x})\rangle_c$; $\langle f(\mathbf{x})\rangle$ is unconditional mean of a random source term $f(\mathbf{x})$; $\langle H(\mathbf{x})\rangle$ is unconditional mean random head prescribed on Dirichlet boundary segments Γ_D; $\langle Q(\mathbf{x})\rangle$ is unconditional mean random flux prescribed normal to Neumann boundary segments Γ_N; $\mathbf{n}(\mathbf{x})$ is a unit outer vector normal to the boundary $\Gamma = \Gamma_D \cup \Gamma_N$; and $f(\mathbf{x})$, $H(\mathbf{x})$, $Q(\mathbf{x})$ are statistically independent random functions.

The conditional covariance of hydraulic head predictions between locations $\mathbf{x}$ and $\mathbf{y}$, $C_{hc}(\mathbf{x}, \mathbf{y}) = Cov(\mathbf{h}|\mathbf{D}, M_k)$, satisfies the second-moment equation

$$\nabla_x \cdot [\langle K(\mathbf{x})\rangle_c \nabla_x C_{hc}(\mathbf{x}, \mathbf{y}) + \mathbf{p}_c(\mathbf{x}, \mathbf{y}) + C_{hKc}(\mathbf{x}, \mathbf{y}) \nabla_x \langle h(\mathbf{x})\rangle_c] + \langle f'(\mathbf{x}) h'(\mathbf{y})\rangle_c = 0 \tag{8}$$

subject to boundary conditions

$$C_{hc}(\mathbf{x}, \mathbf{y}) = \langle H'(\mathbf{x}) h'(\mathbf{y})\rangle_c \qquad \text{on } \Gamma_D \tag{9}$$

$$[\langle K(\mathbf{x})\rangle_c \nabla_x C_{hc}(\mathbf{x}, \mathbf{y}) + \mathbf{p}_c(\mathbf{x}, \mathbf{y}) + C_{hKc}(\mathbf{x}, \mathbf{y}) \nabla_x \langle h(\mathbf{x})\rangle_c] \cdot \mathbf{n}(\mathbf{x}) = \langle Q'(\mathbf{x}) h'(\mathbf{y})\rangle_c \qquad \text{on } \Gamma_N \tag{10}$$

where $\mathbf{p}_c(\mathbf{x}, \mathbf{y}) = \langle K'(\mathbf{x}) \nabla_x h'(\mathbf{x}) h'(\mathbf{y})\rangle_c$ is a conditional mixed third moment and $C_{hKc}(\mathbf{x}, \mathbf{y})$ is the conditional cross covariance between heads and hydraulic conductivities.

Though the above conditional moment equations are exact, they cannot be solved directly (closed) without high-resolution Monte Carlo simulation through exhaustive sampling of the random parameter space. A finite element approach that allows solving these equations in an approximate manner was developed by *Guadagnini and Neuman* [1999]. The conditional mean $\langle Y(\mathbf{x})\rangle_c$ and covariance $C_{Yc}(\mathbf{x}, \mathbf{y})$ of log hydraulic conductivities, $Y = \ln K$, play the role of model parameters. We estimate them by ML as described in Appendix B. Our method of assessing data worth thus consists of the following steps:

1. Calibrate each postulated model against all available measurements (log hydraulic conductivity and/or head) by minimizing *NLL* in (B4) to obtain estimated log hydraulic conductivity values at measurement and pilot point locations, as described in Appendix B.

2. Project the estimated log hydraulic conductivity values onto the computational grid via kriging to obtain $\langle Y(\mathbf{x})\rangle_c$ and $\langle Y'(\mathbf{x}) Y'(\mathbf{y})\rangle_c$ according to (B1) and (B6), respectively.

3. Solve the conditional moment equations with these parameters for $\langle h(\mathbf{x})\rangle_c$ and $C_{hc}(\mathbf{x}, \mathbf{y})$. Step 1–3 are performed with the Fortran code INME developed by *Riva et al*. [2010].

4. Compute the Kashyap information criterion *KIC* for each model using (B7) and a corresponding posterior weight according to (A3).

5. Compute multimodel conditional statistics of head according to (1)–(2) and analyze the worth of additional data at selected locations in space, as described in subsection 2.1.

3. Illustrative Example

We illustrate our approach to data worth assessment through a synthetic example. In our example, groundwater flows at steady state through a rectangular domain of length 18 and width 10 (all quantities being given in arbitrary consistent units), depicted in Figure 2, similar to the case considered by *Hernandez et al*. [2006]. The domain is discretized into $N_e = 180$ square elements of unit size. Heads are prescribed

deterministically as 10 on the left and 5 on the right boundaries, the top and bottom boundaries being impermeable. An unconditional, zero-mean reference log hydraulic conductivity field (Figure 2a) of point values (having zero measurement or resolution scale) is generated with a sequential Gaussian simulation code developed by *Deutsch and Journel* [1998], modified to accommodate a truncated power variogram model

$$\gamma(s)=\sigma^2(\lambda_u)\left\{1-\exp\left[-\frac{\pi}{4}\left(\frac{s}{\lambda_u}\right)^2\right]+\left[\frac{\pi}{4}\left(\frac{s}{\lambda_u}\right)^2\right]^H\Gamma\left[1-H,\frac{\pi}{4}\left(\frac{s}{\lambda_u}\right)^2\right]\right\};\quad 0<H<1 \tag{11}$$

obtained through superposition of Gaussian modes (TpvG) [*Di Federico and Neuman*, 1997; *Neuman and Di Federico*, 2003; *Neuman et al*., 2008]. Here s is separation distance (lag), λ_u is an upper cutoff scale proportional to domain size, A is a coefficient which assures that the TpvG (11) converges to a power law model when $\lambda_u \to \infty$ [*Neuman and Di Federico*, 2003], H is a Hurst scaling exponent, $\sigma^2(\lambda_u)=A\lambda_u^{2H}/2H$ is variance (sill) and $\Gamma(\cdot,\cdot)$ is the incomplete gamma function. By disregarding the measurement or resolution scale of $Y(\mathbf{x})$ in comparison to domain size, the integral scale of Y becomes equal to $I(\lambda_u)=2H\lambda_u/(1+2H)$. We set the parameter vector characterizing the TpvG model to $\boldsymbol{\theta}=(A,H,\lambda_u)^T=(0.1,0.25,10)^T$. The selected value of H corresponds to an antipersistent log hydraulic conductivity field, consistent with *Neuman* [1994] and *Neuman et al*. [2008]. These TpvG parameters yield $\sigma^2=0.63$ and $I=3.33$, corresponding to a mildly heterogeneous Y field that is well within the range of applicability of the stochastic moment equations we employ. The associated ratio between domain and log conductivity integral scales is consistent with *Neuman et al*. [2008]. Reference hydraulic head contours associated with the above conditions are depicted in Figure 2a.

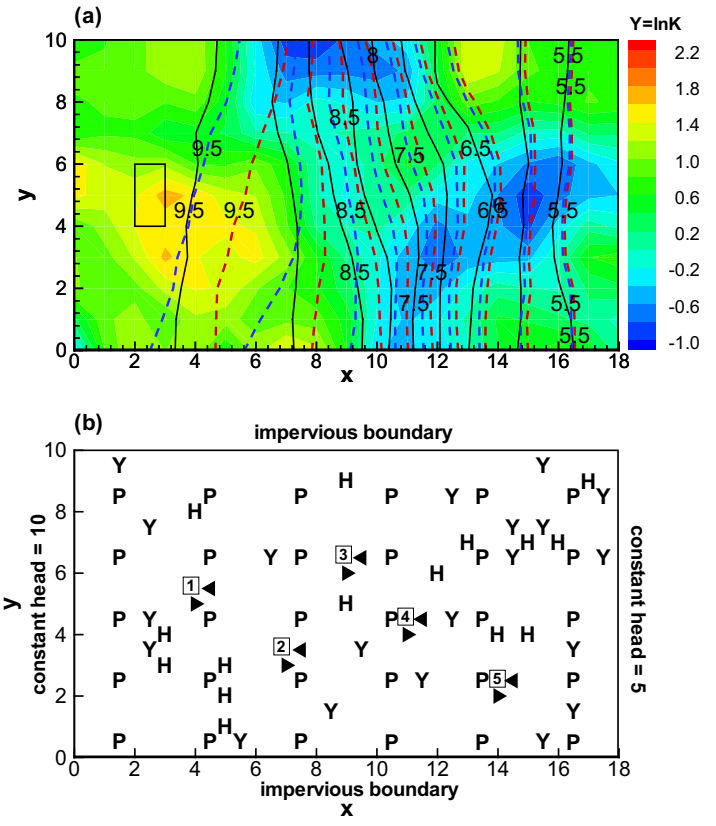

Figure 2. Setup of illustrative example. (a) Reference field of log hydraulic conductivity, Y, and contoured hydraulic head distributions (solid black: no recharge; dashed red: localized recharge; dashed blue: uniform recharge; the rectangle indicates the localized recharge area); (b) Sampling and pilot points locations; Y and h, respectively, represent locations where log conductivities and hydraulic heads are originally available; P indicates pilot points; numbered empty triangles and squares, respectively, indicate locations of 5 additional Y and h samples employed in all test cases except TC3.

We sample log hydraulic conductivities and hydraulic heads at random locations of the corresponding reference fields. We then superimpose zero-mean white Gaussian head and Y measurement errors with variances $\sigma^2_{hE}=0.150$ and $\sigma^2_{YE}=0.001$ on these sampled values.

For the purpose of moment-equations based geostatistical inverse modeling, we distribute 30 pilot points across the domain (Figure 2b); Y values at these points are treated as adjustable parameters.

In addition to the generating TpvG variogram model of Y data, we consider standard two-parameter exponential and spherical variogram models, as illustrated in section 3.1.1. We select three different recharge scenarios based on preliminary simulations: (*a*) the generating scenario of no recharge; (*b*) uniform recharge rate of 0.4 over a localized area of the aquifer (Figure 2a); and (*c*) uniform recharge rate of 0.012 over the entire synthetic aquifer. Reference heads associated with these scenarios are

Table 1. Sets of Models Postulated in Various Test Cases

Test Cases	Model Identifier	*Y* Variogram Model	Rercharge Conditions
All but test case TC2	*Exp_rchall*	Exponential	Uniform recharge
	Exp_rchstp	Exponential	Localized recharge
	Sph_rchall	Spherical	Uniform recharge
	Sph_rchstp	Spherical	Localized recharge
TC2	*Exp_rchno*	Exponential	No recharge
	Exp_rchall	Exponential	Uniform recharge
	Exp_rchstp	Exponential	Localized recharge
	Sph_rchno	Spherical	No recharge
	Sph_rchall	Spherical	Uniform recharge
	Sph_rchstp	Spherical	Localized recharge
	TpvG_rchno	TpvG	No recharge
	TpvG_rchall	TpvG	Uniform recharge
	TpvG_rchstp	TpvG	Localized recharge

contoured and juxtaposed in Figure 2a. Whereas contours representing diverse recharge scenarios differ from each other, each is plausible in light of the available head data; though one cannot rule out any of the scenarios a priori, our approach will help rank them a posteriori. Table 1 lists two sets of models we shall consider jointly in our analysis: whereas one set excludes models used to generate our reference conditions, the other set includes them.

We examine a set of test cases, listed in Table 2. The table also lists the range of variability of the coefficient of variation (expressed as ratio between standard deviations of measurement errors and sampled reference values) for original h (i.e., $\varepsilon_{h,data}^{TCi}$, $i = 1, 2, \ldots, 8$) and Y (i.e., $\varepsilon_{Y,data}^{TCi}$) data, together with corresponding ranges of the coefficients of variation ε_h^{TCi} and ε_Y^{TCi} ($i = 1, 2, \ldots, 8$) associated with additional h and Y measurements, respectively, made available in some of the test cases.

Test cases TC1 and TC2 are designed to assess the reliability of our proposed data worth methodology. Both rely on a set of 20 Y and 15 h measurements distributed randomly throughout the flow domain and placed originally at the analyst's disposal. Five additional measurement points of Y and of h, planned to be collected at other random locations in these two (and some other) test cases, are depicted in Figure 2b. The combined data worth of these additional measurements is quantified on the basis of the two model sets listed in Table 1 for test cases TC1 and TC2, respectively.

Test cases other than TC1 and TC2 are patterned after these two and are designed to investigate the influence of diverse factors on the performance of the methodology with TC1 as the base case. Test case TC3 differs from TC1 in the (randomly selected) locations of the additional Y and h measurements; its purpose is to test our ability to discriminate between diverse sampling schemes.

Test cases TC4 and TC5 compare situations in which additional samples of only Y or only h, respectively, are made available. Their purpose is to compare the relative worths of these two data types.

Test cases TC6, TC7, and TC8 investigate the effects of diverse prior data sets on the worth of additional samples. The prior data set in each of these three test cases forms a subset of those employed in test cases TC1–TC5.

Table 2. Main Characteristics of Test Cases Examined; Ranges of Variability of Coefficient of Variation (Ratios Between Standard Deviations of Measurement Errors and Sampled Reference Values) for Original Y (i.e., $\varepsilon_{Y,data}^{TCi}$; $i = 1, 2, \ldots, 8$) and h (i.e., $\varepsilon_{h,data}^{TCi}$) Are Also Listed, Together With Corresponding Ranges of Coefficients of Variation ε_h^{TCi} and ε_Y^{TCi}, Respectively, for Additional Y and h Measurement Samples Selected From Reference Field

Test Case	No. of Original Y	No. of Original h	No. of Additional Y	No. of Additional h	$\varepsilon_{Y,data}^{TCi}$, ε_Y^{TCi} ($i = 1, 2, \ldots, 8$)	$\varepsilon_{h,data}^{TCi}$, ε_h^{TCi} z($i = 1, 2, \ldots, 8$)	Description
TC1	20	15	5	5	$-0.63 \le \varepsilon_{Y,data}^{TC1} \le 0.18$ $-0.078 \le \varepsilon_Y^{TC1} \le 0.15$	$0.040 \le \varepsilon_{h,data}^{TC1} \le 0.073$ $0.041 \le \varepsilon_h^{TC1} \le 0.062$	Method validation, base case
TC2	20	15	5	5	$-0.63 \le \varepsilon_{Y,data}^{TC2} \le 0.18$ $-0.078 \le \varepsilon_Y^{TC2} \le 0.15$	$0.040 \le \varepsilon_{h,data}^{TC2} \le 0.073$ $0.041 \le \varepsilon_h^{TC2} \le 0.062$	Method validation
TC3	20	15	5	5	$-0.63 \le \varepsilon_{Y,data}^{TC3} \le 0.18$ $0.020 \le \varepsilon_Y^{TC3} \le 0.083$	$0.040 \le \varepsilon_{h,data}^{TC3} \le 0.073$ $0.039 \le \varepsilon_h^{TC3} \le 0.042$	Alternative sampling scheme for additional data
TC4	20	15	5	0	$-0.63 \le \varepsilon_{Y,data}^{TC4} \le 0.18$ $-0.078 \le \varepsilon_Y^{TC4} \le 0.15$	$0.040 \le \varepsilon_{h,data}^{TC4} \le 0.073$ -	Influence of additional data type
TC5	20	15	0	5	$-0.63 \le \varepsilon_{Y,data}^{TC5} \le 0.18$ -	$0.040 \le \varepsilon_{h,data}^{TC5} \le 0.073$ $0.041 \le \varepsilon_h^{TC5} \le 0.062$	Influence of additional data type
TC6	5	0	5	5	$0.020 \le \varepsilon_{Y,data}^{TC6} \le 0.11$ $-0.078 \le \varepsilon_Y^{TC6} \le 0.15$	- $0.041 \le \varepsilon_h^{TC6} \le 0.062$	Effect of prior data content
TC7	10	5	5	5	$-0.63 \le \varepsilon_{Y,data}^{TC7} \le 0.11$ $-0.078 \le \varepsilon_Y^{TC7} \le 0.15$	$0.041 \le \varepsilon_{h,data}^{TC7} \le 0.073$ $0.041 \le \varepsilon_h^{TC7} \le 0.062$	Effect of prior data content
TC8	15	10	5	5	$-0.63 \le \varepsilon_{Y,data}^{TC8} \le 0.18$ $-0.078 \le \varepsilon_Y^{TC8} \le 0.15$	$0.040 \le \varepsilon_{h,data}^{TC8} \le 0.073$ $0.041 \le \varepsilon_h^{TC8} \le 0.062$	Effect of prior data content

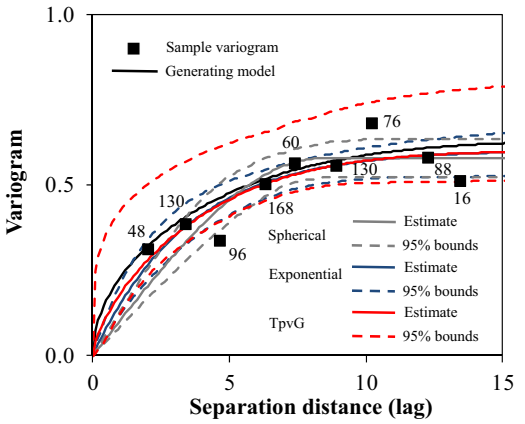

Figure 3. Generating, sample, and estimated variograms. Dashed curves indicate 95% estimation confidence intervals. The sample variogram is calculated on the basis of the available *Y* data in TC1. The number of pairs associated with each lag is reported.

3.1. Results and Discussions

3.1.1. Variogram Analysis

In test cases TC1 and TC2, parameters of alternative variogram models selected to describe the spatial covariance structure of *Y* are initially estimated by least squares on the basis of prior *Y* data, collected at locations identified in Figure 2b. Figure 3 depicts sample variograms of *Y* together with the number of data pairs associated with each separation distance (lag). Exponential, spherical, and TpvG variogram models fitted to the sample variograms by least squares, together with their associated 95% confidence intervals and the generating TpvG model, are also shown. Confidence intervals were constructed through Monte Carlo generation of 500 normally distributed parameter sets around their least squares estimates, considering the associated estimation covariance matrices. Parameter estimates and their standard estimation errors are listed in Table 3. The estimated TpvG variogram model does not differ significantly from the estimated exponential variogram model, the two being virtually indistinguishable from each other at lags greater than 2. The TpvG model, however, has a much wider 95% confidence interval than do the other variogram models due in part to its larger number of parameters (3 versus 2). The similarity we observe between exponential and TpvG variogram models is consistent with that noted by *Neuman et al*. [2008]. It reinforces the conclusion of these authors that fitting standard exponential models to hierarchical data may mask the multiscale nature of the underlying random field. As there is nothing in these results to support, a priori, preference for one variogram model over the rest, we retain all of them for further consideration. Though for simplicity we ignore variogram parameter uncertainty in all subsequent test cases, one could account for it in principle as described by *Neuman et al*. [2012]. Augmenting the prior *Y* data with five additional measurements, either generated at the preposterior stage or measured at the posterior stage, has virtually no effect on variogram parameter estimates and we therefore continue relying on the prior estimates in Table 3 throughout our preposterior and posterior analyses.

3.1.2. Data Worth Assessment Excluding Generating Models (Test Case TC1)

We start with test case TC1 that excludes from consideration variogram and flow models used to generate our synthetic data. Each of the remaining models is assigned equal prior probability and is calibrated against the conditioning data set within the Maximum Likelihood framework summarized in Appendix B.

At the prior stage of the analysis, the prior data vector, **D**, consists of 20 and 15 noisy *Y* and *h* measurements, respectively. Geostatistical inversion of the groundwater flow moment equations yields negative log likelihoods *NLL* (B4), Kashyap discrimination critera *KIC* (B7), and posterior model weights $p(M_k|\mathbf{D})$ listed in the upper part of Table 4. *KIC* is seen to prefer the *Exp_rchall* variogram and flow model combination over all other combinations, assigning to it a posterior weight of 39.95%. This is likely due to similarity between the fitted exponential and TpvG variogram models noted earlier, and greater similarity between the uniform and no recharge cases than between the latter and the case of localized recharge. The scalar measure of prior predictive head uncertainty, $Tr\left[Cov(\mathbf{h}|\mathbf{D})_D\right]$, is 3.24. Figure 4a depicts the spatial distribution of prior predictive head variance, $Var(\mathbf{h}|\mathbf{D})_D$. The latter tends to be largest in the central part of the domain farthest from the deterministically prescribed head boundaries and slightly elevated near the localized recharge zone. Heads in the

Table 3. Variogram Parameter Estimates and Associated Standard Deviations

Model	Parameter	Estimated Value	Standard Deviation
Exponential	Sill	0.60	0.047
	Integral scale	3.45	0.87
Spherical	Sill	0.58	0.034
	Range	8.37	1.25
TpvG	*A*	0.10	0.014
	H	0.33	0.23
	λ_u	8.15	5.44

Table 4. Values of *NLL*, *KIC*, Posterior Weights and Model Rank in Prior and Posterior Analyses of TC1

Prior				
Model	*NLL*	*KIC*	$p(M_k\|\mathbf{D})$	Rank
Exp_rchall	−94.83	204.62	39.95%	1
Exp_rchstp	−94.27	205.25	29.04%	2
Sph_rchall	−103.84	207.19	11.00%	4
Sph_rchstp	−105.12	206.00	20.01%	3
Posterior				
Model	*NLL*	*KIC*	$p(M_k\|\mathbf{D},\mathbf{C}')$	Rank
Exp_rchall	−128.50	226.22	48.11%	1
Exp_rchstp	−125.92	228.79	13.29%	3
Sph_rchall	−139.24	227.42	26.41%	2
Sph_rchstp	−137.73	228.96	12.19%	4

vicinity of this zone are sufficiently different from those under uniform recharge to cause head predictive uncertainty near this zone to increase.

At the preposterior stage, the vector **C**′, consisting of five as yet unsampled potential new *Y* and *h* values at locations shown in Figure 2b, is estimated on the basis of prior data **D**. Random estimates **C** of **C**′ were generated by means of (3); 200 such estimates were found sufficient to yield stable mean and variance values. Figure 5 compares mean values and 95% confidence intervals of these preposterior estimates with corresponding posterior (reference) *Y* and *h* values at these locations. The figure shows that mean values are close to the reference values, all of which lie within 95% confidence intervals of the estimates. The scalar measure of preposterior predictive head uncertainty, $Tr\left[Cov(\mathbf{h}|\mathbf{D})_{DC}\right]=3.14$, is remarkably close to its prior value, $Tr\left[Cov(\mathbf{h}|\mathbf{D})_{D}\right]=3.24$. Likewise, the spatial distribution of preposterior predictive head variance $Var(\mathbf{h}|\mathbf{D})_{DC}$

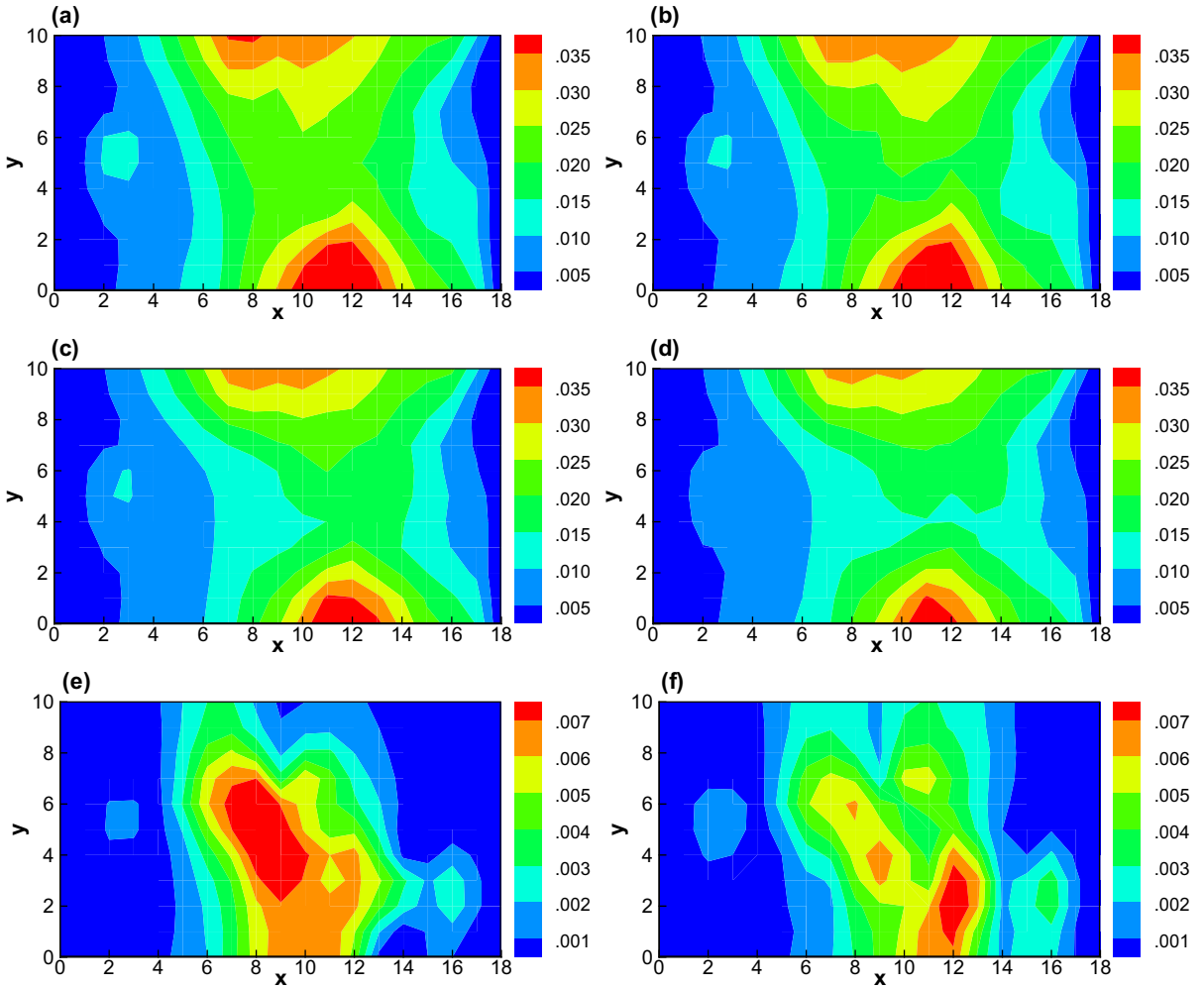

Figure 4. Spatial distributions of predictive head variances and data worth for TC1. (a) prior predictive head variance, (b) preposterior predictive head variance, (c) posterior predictive head variance, (d) expected posterior predictive head variance, (e) reference posterior data worth, and (f) expected preposterior data worth.

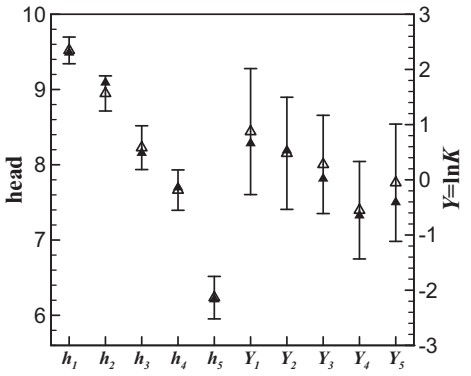

Figure 5. Reference and MLBMA-generated additional *Y* and *h* values in TC1. The solid and empty triangles, respectively, represent reference and mean values; bars indicating the width of the 95% confidence interval around mean values are reported.

in Figure 4b closely resembles that of the prior variance, $Var(\mathbf{h}|\mathbf{D})_D$, in Figure 4a. The scalar measure $Tr\left[E_{C|D}Cov(\mathbf{h}|\mathbf{D},\mathbf{C})\right]$ of expected posterior predictive head uncertainty is 2.70, the corresponding spatial distribution of expected posterior predictive head variance, $E_{C|D}Var(\mathbf{h}|\mathbf{D},\mathbf{C})$, being depicted in Figure 4d. Augmenting the original data base is seen to have reduced the expected posterior head variance across much of the flow domain. The scalar measure $Tr\left[Cov_{C|D}E(\mathbf{h}|\mathbf{D},\mathbf{C})\right]$ of expected preposterior data worth is 0.44, the corresponding spatial distribution of expected preposterior data worth, $Var_{C|D}E(\mathbf{h}|\mathbf{D},\mathbf{C})$, being depicted in Figure 4f.

After $\mathbf{C}'$ becomes available at the posterior stage, calibrating the flow moment equations against the augmented data set $\{\mathbf{D},\mathbf{C}'\}$ yields *NLL*, *KIC*, and posterior model weights listed in the bottom part of Table 4. For reasons noted earlier, *KIC* again prefers model combination *Exp_rchall* with an even larger posterior weight of 48.11%. The reference posterior predictive uncertainty, $Tr[Cov(\mathbf{h}|\mathbf{D},\mathbf{C}')]=2.78$, corresponds closely to the expected posterior predictive head uncertainty of 2.70. Spatial distribution of the posterior predictive head variance $Var(\mathbf{h}|\mathbf{D},\mathbf{C}')$ in Figure 4c resembles closely that of its expected posterior counterpart in Figure 4d. The scalar measure $Tr\left[Cov(\mathbf{h}|\mathbf{D})_D\right]-Tr[Cov(\mathbf{h}|\mathbf{D},\mathbf{C}')]=0.46$ of posterior data worth is very close to its preposterior estimate $Tr\left[Cov_{C|D}E(\mathbf{h}|\mathbf{D},\mathbf{C})\right]=0.44$ (see summary of scalar measures for all test cases in Table 6). The spatial distribution of posterior data worth in Figure 4e resembles closely its preposterior estimate in Figure 4f.

3.1.3. Effect of Including Generating Models (Test Case TC2)

Prior and posterior values of *NLL*, *KIC*, and posterior model weights obtained in test case TC2, which differs from TC1 in its inclusion of the generating models, are listed in Table 5. Here the generating model combination *TpvG_rchno* is ranked best at both stages, with weights equal, respectively, to 28.09% and 27.75%. In other words, our methodology consistently identifies the generating model as best when this combination is included in the postulated model set. The combination *Exp_rchno* is consistently ranked second best with posterior probabilities between 16 and 17 percent, most likely due to the similarity noted earlier between exponential and TpvG variogram models. Once again we see how difficult it is to diagnose the multiscale (hierarchical) structure of a random field on the basis of its variogram alone, without resorting to more advanced methods of geostatistical analysis such as that employed by, e.g., *Guadagnini et al.* [2014].

Table 5. Values of *NLL*, *KIC*, Posterior Weights and Model Rank in Prior and Posterior Analyses of TC2

Prior				
Model	*NLL*	*KIC*	$p(M_k\|\mathbf{D})$	Rank
Exp_rchall	−103.33	196.08	9.06%	5
Exp_rchno	−104.77	194.93	16.11%	2
Exp_rchstp	−103.18	196.32	8.04%	6
Sph_rchall	−110.88	200.13	1.20%	9
Sph_rchno	−114.48	196.87	6.11%	7
Sph_rchstp	−112.65	198.45	2.76%	8
Tpvg_rchall	−100.56	194.96	15.89%	3
Tpvg_rchno	−101.96	193.82	28.09%	1
Tpvg_rchstp	−100.19	195.40	12.74%	4
Posterior				
Model	*NLL*	*KIC*	$p(M_k\|\mathbf{D},\mathbf{C}')$	Rank
Exp_rchall	−139.35	216.66	7.72%	7
Exp_rchno	−141.16	215.10	16.85%	2
Exp_rchstp	−139.43	216.56	8.11%	6
Sph_rchall	−148.29	219.71	1.69%	9
Sph_rchno	−151.90	216.42	8.70%	5
Sph_rchstp	−150.41	217.61	4.81%	8
Tpvg_rchall	−135.91	215.59	13.17%	3
Tpvg_rchno	−137.63	214.10	27.75%	1
Tpvg_rchstp	−135.56	215.92	11.20%	4

Figure 6 compares the spatial distributions of expected preposterior and reference posterior data worth in case TC2. Table 6 shows that the scalar measure of preposterior predictive head uncertainty, $Tr\left[Cov(\mathbf{h}|\mathbf{D})_{DC}\right]=3.53$, is again close to its prior value, $Tr\left[Cov(\mathbf{h}|\mathbf{D})_D\right]=3.47$, and the scalar measure $Tr\left[E_{C|D}Cov(\mathbf{h}|\mathbf{D},\mathbf{C})\right]=2.98$ of expected posterior predictive head uncertainty is slightly smaller than the reference posterior predictive uncertainty, $Tr[Cov(\mathbf{h}|\mathbf{D},\mathbf{C}')]=3.08$. All four measures are slightly larger than their values in test case TC1 due to the wider range of models included in TC2. On the other hand, the scalar measure *Tr*

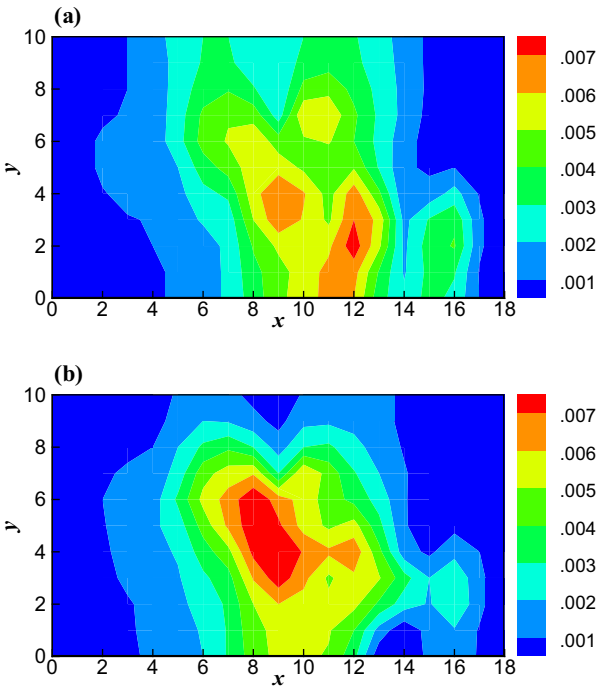

Figure 6. Spatial distributions of the (a) expected preposterior and (b) reference posterior data worth for TC2.

$[Cov(\mathbf{h}|\mathbf{D})_D] - Tr[Cov(\mathbf{h}|\mathbf{D},\mathbf{C}')] = 0.45$ of posterior data worth is very close to its TC1 value of 0.46 and to its preposterior estimate $Tr[Cov_{C|D}E(\mathbf{h}|\mathbf{D},\mathbf{C})] = 0.48$ as well as the latter's TC1 value of 0.44. The spatial distribution of preposterior data worth in Figure 6a resembles closely its posterior counterpart in Figure 6b as well as their TC1 counterparts in Figures 4f and 4e, respectively.

Considering that inclusion of the generating models in TC2 led to results similar to those obtained upon excluding these models from TC1, and that in nature generating models are usually unknown, we exclude these models from all subsequent test cases.

3.1.4. Effect of Altering Sampling Locations (Test Case TC3)

Test case TC3 is identical to TC1 except that additional sampling is performed not at locations indicated in Figure 2b but at new random locations identified in Figure 7. Also shown are spatial distributions of corresponding expected preposterior and reference posterior data worth. Comparison with corresponding distributions in Figures 4 and 6 reveals a shift of elevated values to the left due to a similar shift in new sampling locations. Expected preposterior data worth patterns in Figure 7a anticipate quite closely their reference posterior patterns in Figure 7b. Scalar measures of predictive head uncertainty in the first four rows of Table 6 do not differ markedly from those corresponding to test cases TC1 and TC2. On the other hand, scalar measures of data worth in the last two rows of Table 6 are now significantly smaller due, we believe, to (a) greater proximity of the new sampling locations to the leftmost deterministic head boundary and (b) their more pronounced clustering relative to each other and to original sampling locations. Effect (b) is in line with recent observations [*Hendricks Franssen et al.*, 2009; *Riva et al.*, 2010] according to which the quality of inverse groundwater flow modeling results depends on the density of available measurement location relative to the correlation scale of the underlying log-conductivity field. The new additional sampling locations in Figure 7 are thus revealed to be less advantageous those in Figures 2 and 6. We therefore consider the latter in all remaining test cases.

3.1.5. Effect of Altering Type of Sampled Data (Test Cases TC4 and TC5)

Test case TC4 differs from TC1 in that no additional h values are sampled, and TC5 differs from TC1 in that no additional Y values are made available. Figure 8 depicts spatial patterns of expected preposterior and reference posterior data worth for TC4 and TC5. The former pattern approximates the latter quite closely in both cases. Scalar measures of data worth in the last two rows of Table 6 are, in both cases, smaller than those in test case TC1. The fact that these measures are significantly smaller in TC5 than in TC4 implies that, in our steady state flow example, sampling additional Y data provides a greater benefit than measuring an

Table 6. Scalar Measures in Prior, Preposterior, and Posterior Analyses

Scalar Measures	TC1	TC2	TC3	TC4	TC5	TC6	TC7	TC8
Prior predictive uncertainty	3.24	3.53	3.24	3.24	3.24	15.29	6.60	4.29
Preposterior predictive uncertainty	3.14	3.47	3.24	3.27	3.18	16.76	6.55	4.19
Reference Posterior predictive uncertainty	2.78	3.08	2.94	2.92	2.99	5.02	4.04	3.16
Expected posterior predictive uncertainty	2.70	2.98	2.90	2.87	2.97	4.97	4.18	3.27
Reference posterior data worth	0.46	0.45	0.30	0.32	0.25	10.27	2.56	1.14
Expected preposterior data worth	0.44	0.48	0.34	0.40	0.21	11.79	2.36	0.92

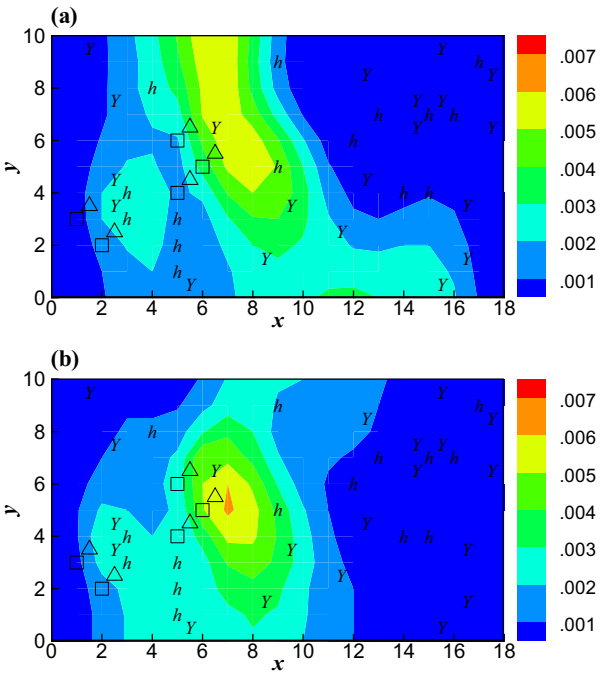

Figure 7. Spatial distributions of (a) expected preposterior and (b) reference posterior data worth for TC3. *Y* and *h*, respectively, represent locations where log-conductivities and hydraulic heads are originally available; empty triangles and squares, respectively, indicate locations of 5 additional *Y* and *h* samples.

equal number of corresponding additional *h* values; the same may not necessarily be true under other, say transient, flow regimes. For example, *Panzeri et al.* [2013] and others, referenced by them, found that increasing the number of early head observations improves parameter estimates of transient stochastic flow models to a greater degree than does observing additional heads under pseudo-steady state conditions at later time.

3.1.6. Effect of Reduced Prior Data Sets (Test Cases TC6, TC7, TC8)

We end our analysis by altering the number and type of prior data entering into test case TC1. Test case TC6 includes 5 instead of 20 prior *Y* measurements and 0 instead of 15 prior head measurements. The number of prior *Y* and *h* data, respectively, entering into TC7 is 10 and 5 and into TC8 is 15 and 10. Figure 9 shows spatial patterns of expected preposterior and reference posterior data worth for TC6, TC7, and TC8. Expected preposterior patterns approximate the posterior patterns reasonably well in all three cases, though not as well as in previous cases. Scalar measures of data worth in the last two rows of Table 6 are, in all three cases, much larger than those in previous test cases. The smaller is the number of prior data, the greater is the benefit of

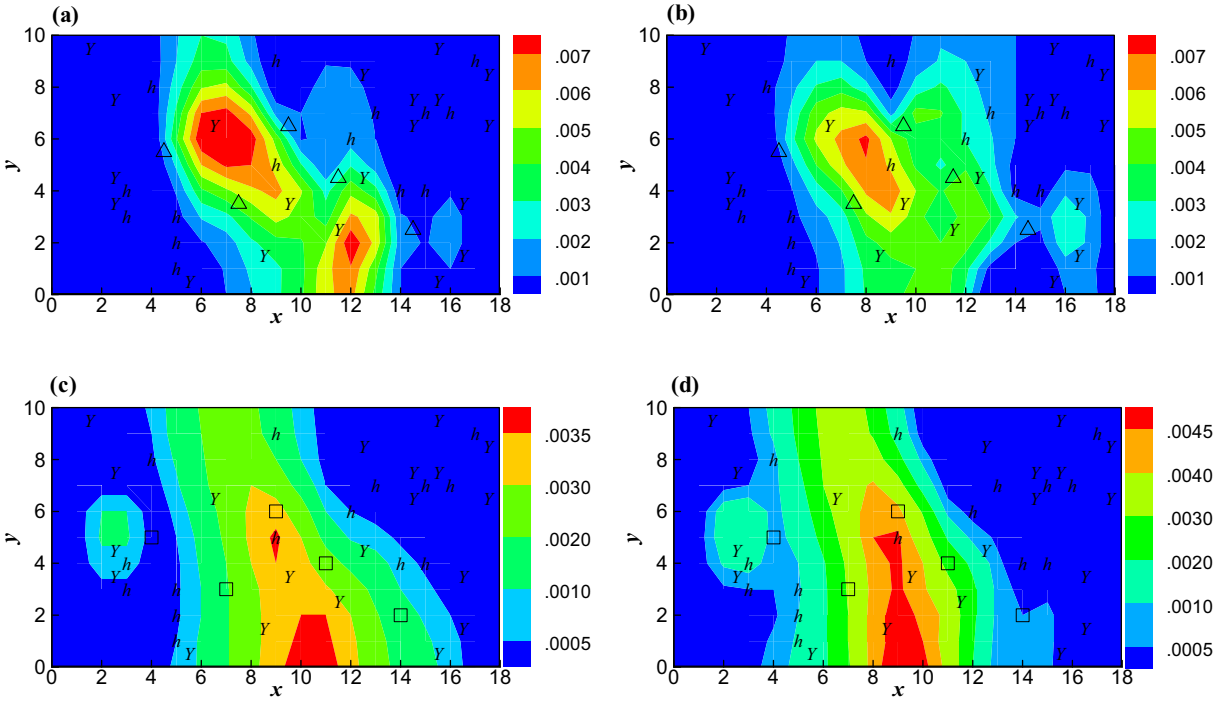

Figure 8. Spatial distributions of (a, c) expected preposterior and (b, d) reference posterior data worth for (a, b) TC4 and (c, d) TC5 (symbols are illustrated in Figure 7).

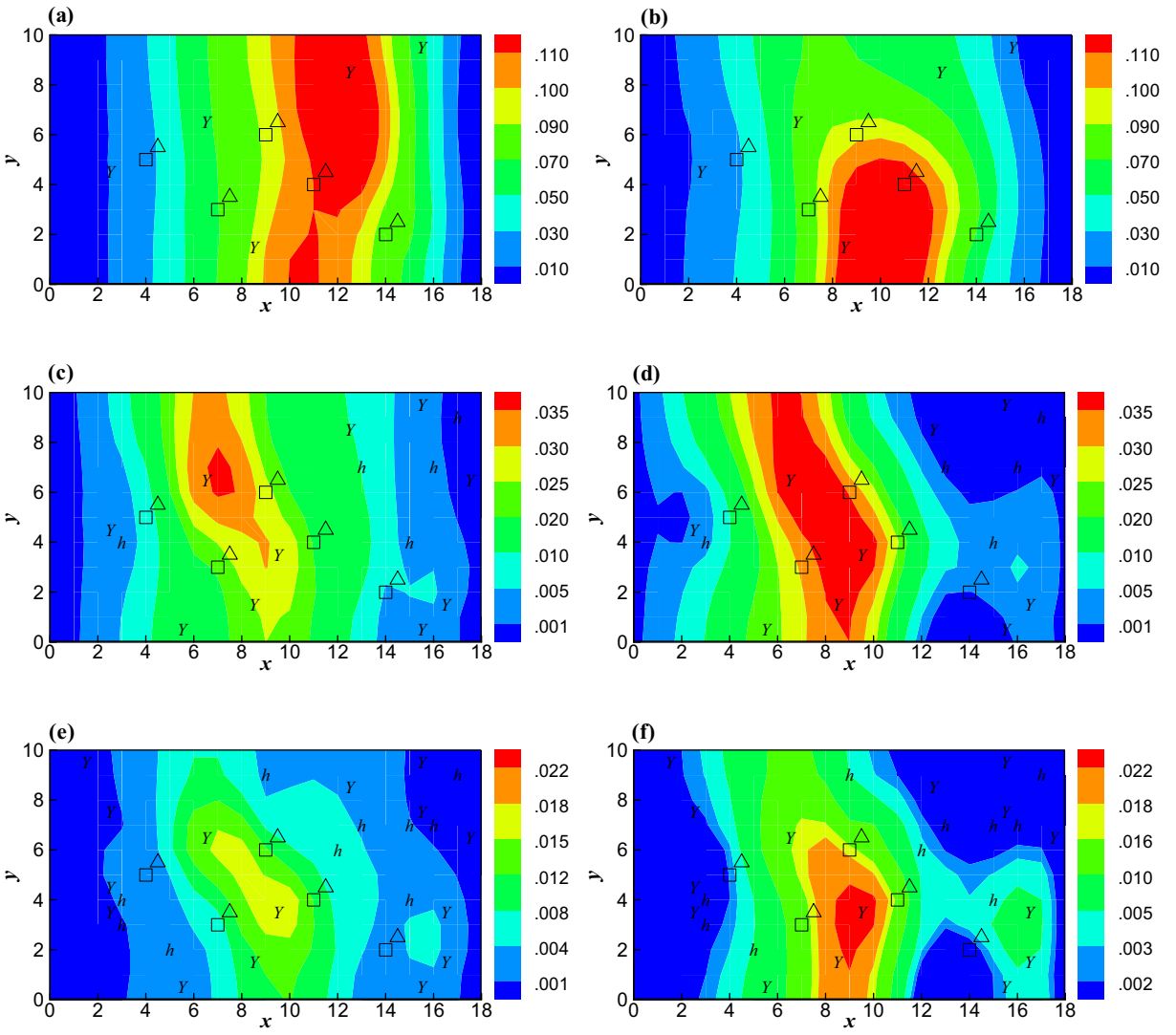

Figure 9. Spatial distributions of the (a, c, e) expected preposterior and (b, d, f) reference posterior data worth for (a, b) TC6, (c, d) TC7, and (e, f) TC8 (symbols are illustrated in Figure 7).

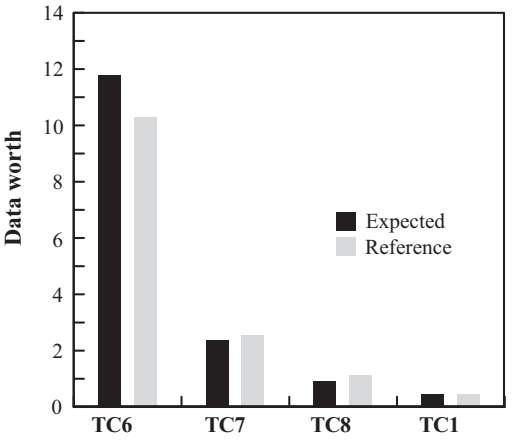

Figure 10. Variation of the expected preposterior and reference posterior data worth with the size of the prior data in our tests.

collecting additional data. The scalar worth of additional data is much greater in case TC6 than in any other case, dropping sharply as the number of prior data increases from TC6 through TC7 and TC8 to TC1 (Figure 10).

4. Conclusions

Our analysis leads to the following major conclusions:

1. The multimodel Maximum Likelihood Bayesian method of assessing data worth proposed by *Neuman et al.* [2012] and tested by them on geostatistical models has been shown here to

work well on synthetic data generated by a combination of geostatistical and steady state groundwater flow models in two spatial dimensions. Embedding the moment equations (ME) of groundwater flow in MLBMA in the context of the geostatistical inverse modeling method of *Hernandez et al*. [2003, 2006] and *Riva et al*. [2011] allows circumventing the need for computationally expensive numerical Monte Carlo simulations.

2. We have demonstrated by means of examples based on such synthetic data that our methodology is able to identify the models that have generated the data when these models are included in the set of potential alternatives. In real-world situations, the generating models are seldom if ever known. For this reason, all but one of eight test cases we present excludes them from consideration in our multimodel assessment of data worth.
3. Our synthetic log hydraulic conductivity data were generated randomly on the basis of a truncated power variogram (TPV) characteristic of truncated fractional Brownian motion. Correspondingly, the underlying random field has a hierarchical structure. When TPV is excluded from the set of models entering into our analysis, the latter favors an exponential variogram model over a spherical model. This is so because, as shown by *Neuman et al*. [2008], TPV and exponential variogram models may appear to be quite similar. Our finding reinforces a conclusion by these authors that fitting standard exponential models to hierarchical data may mask the multiscale nature of the underlying random field.
4. In all eight test cases we consider, our Bayesian method of analysis discriminates quite sharply between the worth of alternative sampling schemes and the relative worth of various data types (in our case log hydraulic conductivities and hydraulic heads).
5. In our steady state flow examples, log hydraulic conductivity data are found to be worth more than an equal number of corresponding head measurements. We do not expect this to be necessarily the case under different, e.g. transient, flow regimes.
6. Our examples confirm that the smaller is the set of prior data, the greater is the value of supplementing them with additional measurements. When prior data are scarce, the scalar worth of a few additional measurements may be very large but decreases sharply as the number of additional measurements grows.

Appendix A

In the standard BMA framework, the posterior model weight is given, according to Bayes' rule, by

$$p(M_k|\mathbf{D}) = \frac{p(\mathbf{D}|M_k)p(M_k)}{\sum_{l=1}^{K} p(\mathbf{D}|M_l)p(M_l)} \tag{A1}$$

where

$$p(\mathbf{D}|M_k) = \int p(\mathbf{D}|M_k, \boldsymbol{\theta}_k)p(\boldsymbol{\theta}_k|M_k)d\boldsymbol{\theta}_k \tag{A2}$$

is the marginal likelihood of model M_k, $p(\mathbf{D}|M_k, \boldsymbol{\theta}_k)$ is the joint likelihood of model M_k and its parameter vector $\boldsymbol{\theta}_k$, $p(\boldsymbol{\theta}_k|M_k)$ is the prior probability of parameters associated with model M_k, and $p(M_k)$ is the prior model probability. All probabilities in equation (A1) are implicitly conditional on the choice of models included in the set $\mathcal{M}$.

In MLBMA, the parameter probability is made conditional on prior data **D** through maximization of the likelihood $p(\mathbf{D}|M_k, \boldsymbol{\theta}_k)$. In accord with *Draper* [1995], $p(\mathbf{D}|M_k, \boldsymbol{\theta}_k)$ in (A2) is approximated by its maximum likelihood value $p(\mathbf{D}|M_k, \hat{\boldsymbol{\theta}}^D{}_k)$ where $\boldsymbol{\theta}_k$ has been replaced by its maximum likelihood estimate $\hat{\boldsymbol{\theta}}_k^D$. *Neuman* [2003] proposed evaluating the posterior model weights $p(M_k|\mathbf{D})$ based on *Kashyap's* [1982] information criterion *KIC* (for further details on *KIC*, refer to *Ye et al*. [2008]) according to

$$p(M_k|\mathbf{D}) \approx p(M_k|\mathbf{D})_{ML} = \frac{\exp\left(-\frac{1}{2}\Delta KIC_k^D\right)p(M_k)}{\sum_{l=1}^{K} \exp\left(-\frac{1}{2}\Delta KIC_l^D\right)p(M_l)} \tag{A3}$$

where

$$\Delta KIC_k^D = KIC_k^D - KIC_{\min}^D \tag{A4}$$

$$KIC_k^D = -2\ln p(\mathbf{D}|M_k, \hat{\boldsymbol{\theta}}_k^D)_{ML} - 2\ln p(\hat{\boldsymbol{\theta}}_k^D|M_k)_{ML} + N_k \ln\left(\frac{N_D}{2\pi}\right) + \ln|\mathbf{F}_k^D| \tag{A5}$$

KIC_k^D being the Kashyap model selection criterion for model M_k based on data vector $\mathbf{D}$ of dimension N_D, $KIC_{\min}^D$ is the smallest KIC value among all K models, $-2\ln p(\mathbf{D}|M_k, \hat{\boldsymbol{\theta}}_k^D)_{ML} - 2\ln p(\hat{\boldsymbol{\theta}}_k^D|M_k)_{ML}$ is negative log likelihood incorporating prior probability of the parameters evaluated at $\hat{\boldsymbol{\theta}}_k^D$, N_k is the dimension of $\hat{\boldsymbol{\theta}}_k^D$, and $\mathbf{F}_k^D$ is a normalized (by N_D) observed (as opposed to ensemble mean) Fisher information matrix having components

$$F_{k,nm}^D = -\frac{1}{N_D}\left[\frac{\partial^2 \ln\, p(\mathbf{D}|M_k, \theta_k)}{\partial\theta_{kn}\partial\theta_{km}}\right]_{\theta_k = \hat{\theta}_k^D} \tag{A6}$$

KIC is chosen due to its unique discriminatory power in the context of our inverse approach [*Riva et al.*, 2011] and consistently reliable indication of model quality [*Lu et al.*, 2011].

Appendix B

In a manner similar to *Hernandez et al.* [2003, 2006], we parameterize $\langle Y(\mathbf{x})\rangle_c$ as

$$\langle Y(\mathbf{x})\rangle_c = \sum_{i=1}^{N_M} \omega_i(\mathbf{x}) Y_{Mi} + \sum_{j=1}^{N_P} \omega_j(\mathbf{x}) Y_{Pj} = \sum_{k=1}^{N_Y} \omega_k(\mathbf{x}) Y_{Lk} \tag{B1}$$

Here, Y_{Mi} and Y_{Pj}, respectively, are log hydraulic conductivities at N_M measurement points and at N_P pilot points $\mathbf{x}_p$ [*De Marsily et al.*, 1984], $\mathbf{Y}_L = (\mathbf{Y}_M, \mathbf{Y}_P)^T$, $N_Y = N_M + N_P$ is the dimension of $\mathbf{Y}_L$, and ω_i, ω_j and ω_k are kriging weights. We characterize the spatial structure of $Y(\mathbf{x})$ by a variogram model $\gamma(\mathbf{s};\vartheta)$ in which $\mathbf{s}$ is separation distance vector or lag and ϑ is a vector of variogram parameters such as nugget effect (being ignored in this study), sill and integral scale. In this work, we estimate ϑ a priori on the basis of available log hydraulic measurements, $\mathbf{Y}_M$. Though we do not do so here, it is possible to improve these estimates by conditioning them additionally on measured head values as proposed by *Riva et al.* [2011].

Let Y_{Mi}^* represents measured values of Y_{Mi} and Y_{Pp}^* prior kriged estimates of Y_{Pp} obtained through

$$Y_{Pp}^* = \sum_{i=1}^{N_M} \varpi_i(\mathbf{x}_p) Y_{Mi}^* \qquad p = 1, 2, \cdots, N_p \tag{B2}$$

where $\varpi_i(\mathbf{x}_p)$ are kriging weights. If one uses ordinary kriging, then the covariance of the corresponding estimation (kriging) errors $\varepsilon_{Yp}^* = Y_{Pp}^* - Y_{Pp}$ is given by

$$\langle \varepsilon_{Yp}^* \varepsilon_{Yq}^* \rangle = -\gamma(\mathbf{x}_p - \mathbf{x}_q) + \sum_{i=1}^{N_M} \varpi_i(\mathbf{x}_p)\gamma(\mathbf{x}_i - \mathbf{x}_q) + \mu(\mathbf{x}_p) \qquad p, q = 1, 2, \cdots, N_P \tag{B3}$$

where $\mu(\mathbf{x}_p)$ are Lagrange multipliers.

Obtaining ML estimates of $\langle Y(\mathbf{x})\rangle_c$ is thus equivalent to obtaining the estimates of model parameter $\mathbf{Y}_L$ in the ML estimation process. We do so by minimizing the negative log likelihood (NLL) criterion [*Carrera and Neuman*, 1986]

$$NLL = \frac{F_h}{\sigma_{hE}^2} + \frac{F_Y}{\sigma_{hE}^2} + \ln|\mathbf{V}_Y| + \ln|\mathbf{V}_h| + N_h \ln\sigma_{hE}^2 + N_Y \ln\sigma_{YE}^2 + N_Z \ln 2\pi \tag{B4}$$

with respect to model parameters such as log conductivities at measurement and pilot point locations and hydraulic heads at measurement locations, in equation (B4) $F_h = (\mathbf{h}^* - \langle\mathbf{h}\rangle_{Mc})^T \mathbf{V}_h^{-1} (\mathbf{h}^* - \langle\mathbf{h}\rangle_{Mc})$ is a weighted sum of squared head residuals, $\mathbf{h}^*$ representing measured head values and $\langle\mathbf{h}\rangle_{Mc}$ conditional mean heads at corresponding measurement locations, $\mathbf{C}_h = \sigma_{hE}^2 \mathbf{V}_h$ being the covariance matrix of head measurement errors, σ_{hE}^2 acting as a scaling factor; $F_Y = (\mathbf{Y}^* - \mathbf{Y}_L)^T \mathbf{V}_Y^{-1} (\mathbf{Y}^* - \mathbf{Y}_L)$ is a penalty or regularization function

consisting of the weighted sum of squared log conductivity residuals, the vector $\mathbf{Y}^* = (\mathbf{Y}_M^*, \mathbf{Y}_P^*)$ including log conductivity measurements and prior log conductivity estimates at pilot points, $\mathbf{C}_Y = \sigma_{YE}^2 \mathbf{V}_Y$ being the covariance matrix of the corresponding measurement and estimation errors, σ_{YE}^2 acting as a scaling factor; N_h is the number of head measurements; and $N_z = N_h + N_Y$ is the total number of head and hydraulic conductivity measurements. In this paper, we treat measurement errors as being spatially uncorrelated, taking $\mathbf{C}_h = \sigma_{hE}^2 \mathbf{V}_h$ and $\mathbf{C}_{YM} = \sigma_{YE}^2 \mathbf{V}_{YM}$ to be diagonal with known σ_{hE}^2 and σ_{YE}^2 values; it is possible to treat these matrices as nondiagonal and the scaling factors as unknown parameters in the manner of *Riva et al.* [2011]. The covariance matrix $\mathbf{C}_Y$ consists of two blocks

$$\mathbf{C}_Y = \begin{bmatrix} \mathbf{C}_{YM} & 0 \\ 0 & \mathbf{C}_{YP} \end{bmatrix} \tag{B5}$$

The covariance $\mathbf{C}_{YP}$ of pilot point estimates is generally nondiagonal, its components being given by (B3). Specifying all components of $\mathbf{C}_h$, $\mathbf{C}_{YM}$, and $\mathbf{C}_{YP}$ as we do here reduces the ML problem of minimizing *NLL* in (B4) to a relatively simple problem of minimizing a sum of squared residuals criterion $(\sigma_{hE}^2/\sigma_{hE}^2)F_h + F_Y$.

At each iteration of the nonlinear optimization process, we compute a conditional covariance of Y across the entire computational grid, according to

$$\begin{aligned} \langle Y'(\mathbf{x})Y'(\mathbf{y})\rangle_c &= \left\langle \left[Y(\mathbf{x}) - Y(\mathbf{x})_c\right]\left[Y(\mathbf{y}) - Y(\mathbf{y})_c\right] \right\rangle_c \\ &= -\gamma(\mathbf{x}-\mathbf{y}) - \sum_{k=1}^{N_Y}\lambda_k(\mathbf{x})\sum_{i=1}^{N_Y}\lambda_i(\mathbf{y})[\gamma(\mathbf{x}_k - \mathbf{x}_i) - Q_{ki}] \\ &\quad + \sum_{i=1}^{N_Y}\lambda_i(\mathbf{y})\gamma(\mathbf{x}-\mathbf{x}_i) + \sum_{i=1}^{N_Y}\lambda_i(\mathbf{y})\gamma(\mathbf{y}-\mathbf{x}_i) \end{aligned} \tag{B6}$$

where λ_k are kriging coefficients and Q_{ki} are components of the parameter estimation covariance matrix $\mathbf{Q} \equiv \langle(\mathbf{Y} - \langle\mathbf{Y}_L\rangle_c)(\mathbf{Y} - \langle\mathbf{Y}_L\rangle_c)^T\rangle$. The Kashyap information criterion (A5) for model k then becomes

$$KIC = NLL + N_Y \ln\left(\frac{1}{2\pi}\right) - \ln|\mathbf{Q}| \tag{B7}$$

Acknowledgments

This work is partially funded by the National Science and Technology Major Project of China through grants 2011ZX05009-006 and 2011ZX05052, the National Key Technology R&D Program of China (Grant 2012BAC24B02), National Science Foundation for Young Scientists of China (Grant 41402199) and by the China Postdoctoral Science Foundation (Grant 2012M520118). Funding from MIUR (Italian ministry of Education, Universities and Research-PRIN2010-11; project: "Innovative methods for water resources under hydro-climatic uncertainty scenarios") is acknowledged by the third author. 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), funded by the U.S. Department of Energy.

References

Carrera, J., and S. P. Neuman (1986), Estimation of aquifer parameters under transient and steady state conditions: 1. Maximum likelihood method incorporating prior information, *Water Resour. Res.*, *22*(2), 199–210, doi:10.1029/WR022i002p00199.

De Barros, F. P. J., S. Ezzedine, and Y. Rubin (2012), Impact of hydrogeological data on measures of uncertainty, site characterization and environmental performance metrics, *Adv. Water Resour.*, *36*, 51–63, doi: 10.1016/j.advwatres.2011.05.004.

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

Deutsch, C. V., and A. G. Journel (1998), *GSLIB: Geostatistical Software Library and User's Guide*, 2nd ed., Oxford Univ. Press, N. Y.

Di Federico, V., and S. P. Neuman (1997), Scaling of random fields by means of truncated power variograms and associated spectra, *Water Resour. Res.*, *33*(5), 1075–1085, doi:10.1029/97WR00299.

Draper, D. (1995), Assessment and propagation of model uncertainty, *J. R. Stat. Soc. Ser. B*, *57*(1), 45–90.

Fang, Z., Z. Hou, G. Lin, D. Engel, Y. Fang, and P. Eslinger (2014), Exploring the effects of data quality, data worth, and redundancy of CO2 gas pressure and saturation data on reservoir characterization through PEST inversion, *Environ. Earth Sci.*, *71*(7), 3025–3037, doi: 10.1007/s12665-013-2680-9.

Guadagnini, A., and S. P. Neuman (1999), Nonlocal and localized analyses of conditional mean steady state flow in bounded, randomly nonuniform domains: 1. Theory and computational approach, *Water Resour. Res.*, *35*(10), 2999–3018, doi:10.1029/1999WR900160.

Guadagnini, A., S. P. Neuman, M. G. Schaap, and M. Riva (2014), Frequency distributions and scaling of soil texture and hydraulic properties in a stratified deep vadose zone near Maricopa, Arizona, in *Mathematics of Planet Earth*, pp. 189–192, Springer, Berlin.

Hendricks Franssen, H. J., A. Alcolea, M. Riva, M. Bakr, N. van der Wiel, F. Stauffer, and A. Guadagnini (2009), A comparison of seven methods for the inverse modelling of groundwater flow. Application to the characterisation of well catchments, *Adv. Water Resour.*, *32*, 851–872, doi:10.1016/j.advwatres.2009.02.011.

Hernandez, A.F., S. P. Neuman, A. Guadagnini, and J. Carrera (2003), Conditioning mean steady state flow on hydraulic head and conductivity through geostatistical inversion, *Stoch. Environ. Res. Risk Assess.*, *17*(5), 329–338.

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

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999), Bayesian model averaging: A tutorial, *Stat. Sci.*, *14*(4), 382–417.

James, B. R., and S. M. Gorelick (1994), When enough is enough: The worth of monitoring data in aquifer remediation design, *Water Resour. Res.*, *30*(12), 3499–3513, doi:10.1029/94WR01972.

Kashyap, R. L. (1982), Optimal choice of AR and MA parts in autoregressive moving average models, *IEEE Trans. Pattern Anal. Mach. Intell.*, *4*(2), 99–104.
Leube, P. C., A. Geiges, and W. Nowak (2012), Bayesian assessment of the expected data impact on prediction confidence in optimal sampling design, *Water Resour. Res.*, *48*, W02501, doi:10.1029/2010WR010137.
Liu, X., J. Lee, P. K. Kitanidis, J. Parker, and U. Kim (2012), Value of information as a context-specific measure of uncertainty in groundwater remediation, *Water Resour. Manage.*, *26*(6), 1513–1535.
Lu, D., M. Ye, and S. P. Neuman (2011), Dependence of Bayesian model selection criteria and Fisher information matrix on sample size, *Math. Geosci.*, *43*(8), 971–993.
Lu, D., M. Ye, S. P. Neuman, and L. Xue (2012), Multimodel Bayesian analysis of data-worth applied to unsaturated fractured tuffs, *Adv. Water Resour.*, *35*, 69–82.
Neuman, S. P. (1994), Generalized scaling of permeabilities: Validation and effect of support scale, *Geophys. Res. Lett.*, *21*(5), 349–352, doi:10.1029/94GL00308.
Neuman, S. P. (2003), Maximum likelihood Bayesian averaging of uncertain model predictions, *Stoch. Environ. Res. Risk Assess.*, *17*(5), 291–305.
Neuman, S. P., and V. Di Federico (2003), Multifaceted nature of hydrogeologic scaling and its interpretation, *Rev. Geophys.*, *41*, 1014, doi:10.1029/2003RG000130.
Neuman, S. P., M. Riva, and A. Guadagnini (2008), On the geostatistical characterization of hierarchical media, *Water Resour. Res.*, *44*, W02403, doi:10.1029/2007WR006228.
Neuman, S. P., L. Xue, M. Ye, and Lu, D. (2012), Bayesian analysis of data-worth considering model and parameter uncertainties, *Adv. Water Resour.*, *36*, 75–85.
Nowak, W. (2010), Measures of parameter uncertainty in geostatistical estimation and geostatistical optimal design, *Math. Geosci.*, *42*(2), 199–221, doi:10.1007/s11004-009-9245-1.
Nowak, W., F. P. J. de Barros, and Y. Rubin (2010), Bayesian geostatistical design: Task-driven optimal site investigation when the geostatistical model is uncertain, *Water Resour. Res.*, *46*, W03535, doi:10.1029/2009WR008312.
Nowak, W., Y. Rubin, and F. P. J. de Barros (2012), A hypothesis-driven approach to optimize field campaigns, *Water Resour. Res.*, *48*, W06509, doi:10.1029/2011WR011016.
Panzeri, M., M. Riva, A. Guadagnini, and S. P. Neuman (2013), Data assimilation and parameter estimation via ensemble Kalman filter coupled with stochastic moment equations of transient groundwater flow, *Water Resour. Res.*, *49*, 1334–1344, doi:10.1002/wrcr.20113.
Riva, M., A. Guadagnini, S. P. Neuman, E. Bianchi Janetti, B. Malama (2009), Inverse analysis of stochastic moment equations for transient flow in randomly heterogeneous media, *Adv. Water Res.*, *32*, 1495–1507, doi:10.1016/j.advwatres.2009.07.003.
Riva, M., A. Guadagnini, F. De Gaspari, and A. Alcolea (2010), Exact sensitivity matrix and influence of the number of pilot points in the geostatistical inversion of moment equations of groundwater flow, *Water Resour. Res.*, *46*, W11513, doi:10.1029/2009WR008476.
Riva, M., M. Panzeri, A. Guadagnini, and S. P. Neuman (2011), Role of model selection criteria in geostatistical inverse estimation of statistical data- and model-parameters, *Water Resour. Res.*, *47*, W07502, doi:10.1029/2011WR010480.
Trainor-Guitton, W. J., J. K. Caers, and T. Mukerji (2011), A methodology for establishing a data reliability measure for value of spatial information problems, *Math. Geosci.*, *43*(8), 929–949, doi:10.1007/s11004-011-9367-0.
Trainor-Guitton, W. J., T. Mukerji, and R. Knight (2013), A methodology for quantifying the value of spatial information for dynamic Earth problems, *Stoch. Environ. Res. Risk Assess.*, *27*(4), 969–983, doi:10.1007/s00477-012-0619-4.
Ye, M., P. D. Meyer, and S. P. Neuman (2008), On model selection criteria in multimodel analysis, *Water Resour. Res.*, *44*, W03428, doi:10.1029/2008WR006803.