Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff

Ming Ye
Pacific Northwest National Laboratory, Richland, Washington, USA

Shlomo P. Neuman
Department of Hydrology and Water Resources, University of Arizona, Tucson, Arizona, USA

Philip D. Meyer
Pacific Northwest National Laboratory, Richland, Washington, USA

Received 5 August 2003; revised 15 January 2004; accepted 23 March 2004; published 25 May 2004.

[1] Hydrologic analyses typically rely on a single conceptual-mathematical model. Yet hydrologic environments are open and complex, rendering them prone to multiple interpretations and mathematical descriptions. Adopting only one of these may lead to statistical bias and underestimation of uncertainty. Bayesian model averaging (BMA) [*Hoeting et al.*, 1999] provides an optimal way to combine the predictions of several competing models and to assess their joint predictive uncertainty. However, it tends to be computationally demanding and relies heavily on prior information about model parameters. *Neuman* [2002, 2003] proposed a maximum likelihood version (MLBMA) of BMA to render it computationally feasible and to allow dealing with cases where reliable prior information is lacking. We apply MLBMA to seven alternative variogram models of log air permeability data from single-hole pneumatic injection tests in six boreholes at the Apache Leap Research Site (ALRS) in central Arizona. Unbiased ML estimates of variogram and drift parameters are obtained using adjoint state maximum likelihood cross validation [*Samper and Neuman*, 1989a] in conjunction with universal kriging and generalized least squares. Standard information criteria provide an ambiguous ranking of the models, which does not justify selecting one of them and discarding all others as is commonly done in practice. Instead, we eliminate some of the models based on their negligibly small posterior probabilities and use the rest to project the measured log permeabilities by kriging onto a rock volume containing the six boreholes. We then average these four projections and associated kriging variances, using the posterior probability of each model as weight. Finally, we cross validate the results by eliminating from consideration all data from one borehole at a time, repeating the above process and comparing the predictive capability of MLBMA with that of each individual model. We find that MLBMA is superior to any individual geostatistical model of log permeability among those we consider at the ALRS. *INDEX TERMS:* 1829 Hydrology: Groundwater hydrology; 1875 Hydrology: Unsaturated zone; 1869 Hydrology: Stochastic processes; 5114 Physical Properties of Rocks: Permeability and porosity; *KEYWORDS:* stochastic continuum, conceptual model uncertainty, predictive uncertainty, cross validation, drift, predictive performance

Citation: Ye, M., S. P. Neuman, and P. D. Meyer (2004), Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff, *Water Resour. Res.*, *40*, W05113, doi:10.1029/2003WR002557.

1. Introduction

[2] Hydrologic analyses are commonly based on a single conceptual-mathematical model. Yet hydrologic environments are open and complex, rendering them prone to multiple interpretations and mathematical descriptions. This is true regardless of the quantity and quality of available hydrologic data. Focusing on only one conceptual-mathematical model may lead to a type I model error, which arises when one rejects (by omission) valid alternative models. It may also result in a type II model error, which arises when one adopts (fails to reject) an invalid conceptual-mathematical framework. Indeed, critiques of hydrologic analyses, and legal challenges to them, typically focus on the validity of the underlying conceptual (and by implication mathematical) model. If severe, these may damage one's professional credibility; result in the loss of a legal contest; and lead to adverse environmental, economic and political impacts [*National Research Council*, 2001; *Neuman and Wierenga*, 2003].

[3] Analyses of model uncertainty based on a single hydrologic concept are prone to statistical bias (by commit-

Copyright 2004 by the American Geophysical Union.
0043-1397/04/2003WR002557$09.00

ting a type II error through reliance on an invalid model) and underestimation of uncertainty (by committing a type I error through under sampling of the relevant model space). *Carrera and Neuman* [1986a, 1986b] and *Samper and Neuman* [1989a, 1989b] have noted that bias and uncertainty resulting from an inadequate model structure (conceptualization) are far more detrimental to the model's predictive ability than is a suboptimal set of model parameters. Yet most hydrologic analyses ignore structural uncertainty and focus exclusively on the optimization of model parameters. This often leads to overconfidence in the predictive capabilities of the model, which the available hydrologic data seldom justify [*National Research Council*, 2001; *Neuman and Wierenga*, 2003].

[4] It is argued by *Beven and Freer* [2001, p. 11] "that, given current levels of understanding and measurement technologies, it may be endemic to mechanistic modeling of complex environmental systems that there are many different model structures and many different parameter sets within a chosen model structure that may be behavioral or acceptable in reproducing the observed behavior of that system." They attribute to *Hornberger and Speer* [1981] the notion that this is not simply a problem of identifying a correct or optimal model given limited data. Instead, this is a generic problem which *Beven* [1993] calls equifinality and attributes to [*Beven*, 2000] limitations of current model structures in representing heterogeneous surface and subsurface flow systems, limitations of measurement techniques and scales in defining system characteristics including initial and boundary conditions for a model, and the uniqueness of individual sites. He points out that to do detailed measurements throughout a site is both impractical and unduly expensive. The unique characteristics of a site are therefore inherently unknowable. All that can be done is to constrain the model representations of the site to those that are acceptably realistic, usually in the sense of being consistent with the data.

[5] To address this issue, *Beven and Binley* [1992] have proposed a strategy to which they refer as GLUE (generalized likelihood uncertainty estimation). The strategy calls for the identification of several alternative structural models and the postulation of a prior probabilistic model of parameter uncertainty for each. Each structural model, coupled with its corresponding parameter uncertainty model, is used to generate Monte Carlo realizations of past hydrologic behaviors and to compare the results with monitored system behavior during the same period. Likelihood measures are defined to gauge the degree of correspondence between each simulated and observed record of system behavior. If a likelihood measure falls below a subjectively defined "rejection criterion," the corresponding combination of model structure and parameter set are discarded. Those combinations which pass this test are retained to provide predictions of system behavior under selected future scenarios. Each prediction is weighted by a corresponding normalized likelihood measure (so as to render the sum of all likelihood measures equal to one), to produce a likelihood-weighted cumulative distribution of all available predictions. For recent discussions of GLUE and its applications the reader is referred to *Beven* [2000] and *Beven and Freer* [2001].

[6] A Bayesian approach to the quantification of errors in a single groundwater model was recently proposed by *Gaganis and Smith* [2001]. Like GLUE, it relies on Monte Carlo simulations without model calibration and on subjective criteria of "model correctness."

[7] *James and Oldenburg* [1997] and *Samper and Molinero* [2000] have calibrated a number of conceptual-mathematical models against available observational data, retained those calibrated models that had reproduced adequately past observations, produced a prediction using each calibrated model, assessed the corresponding predictive uncertainty due to uncertainty in the model parameters, and averaged the predictions as well as their ranges of uncertainty by assigning an equal weight to the results of each model.

[8] Other philosophies of model building under uncertainty are discussed by *Gauch* [1993], *Burnham and Anderson* [2002], and *Christakos* [2000, 2002a, 2002b, 2003, 2004]. A comprehensive strategy for constructing alternative conceptual-mathematical models of subsurface flow and transport, selecting the best among them, and using them jointly to render optimum predictions under uncertainty has recently been proposed by *Neuman and Wierenga* [2003]. The strategy embodies a systematic and comprehensive approach to hydrogeologic conceptualization, model development and predictive uncertainty analysis. It is comprehensive in that it considers all stages of model building and accounts jointly for uncertainties that arise at each of them. These stages include regional and site characterization, hydrogeologic conceptualization, development of conceptual-mathematical model structure, parameter estimation on the basis of monitored system behavior, and assessment of predictive uncertainty. In addition to parameter uncertainty, the strategy concerns itself with uncertainties arising from incomplete definitions of (1) the conceptual framework that determines model structure, (2) spatial and temporal variations in hydrologic variables that are either not fully captured by the available data or not fully resolved by the model, and (3) the scaling behavior of hydrogeologic variables.

[9] *Neuman and Wierenga* [2003] discuss several detailed, real-world examples of situations in which more than one conceptual-mathematical model is supported by available data and how to proceed when this happens. The present paper focuses on a key element of their much broader strategy, which concerns rendering optimum predictions by means of several competing deterministic or stochastic models and assessing their joint predictive uncertainty. It rests on the well-established idea of Bayesian model averaging (BMA) [*Draper*, 1995; *Kass and Raftery*, 1995] (see *Hoeting et al.* [1999] for an excellent tutorial and J. Hoeting (Methodology for Bayesian model averaging: An update, 2004, http://www.stat.colostate.edu/~jah/papers/ibcbma.pdf) for a recent summary of applications) to provide an optimal way of combining the predictions of several competing models and assessing their joint predictive uncertainty. Traditional BMA rests on an exhaustive Monte Carlo simulation of the prior parameter space, which renders it computationally demanding. It also relies heavily on prior information about model parameters. *Neuman* [2002, 2003] suggests obviating the need for such simulations and prior information by adopting a maximum likelihood (ML) version (MLBMA) of BMA, thereby rendering the approach computationally feasible and appli-

cable to a wide range of real-world hydrologic problems. MLBMA utilizes a ML approximation of model posterior probability due to *Kashyap* [1982]. The approach incorporates both site characterization and site monitoring data so as to base the outcome on an optimum combination of prior information (scientific and site knowledge plus data) and model predictions. *Kashyap*'s [1982] expression is closely related to an ML version of the Laplace approximation [e.g., *Draper*, 1995; *Kass and Raftery*, 1995] used successfully in the BMA context by statisticians [*Hoeting et al.*, 1999]. We prefer the former because it conforms more directly to ML-based hydrologic model discrimination and parameter estimation frameworks proposed for deterministic models by *Carrera and Neuman* [1986a, 1986b], for geostatistical models by *Samper and Neuman* [1989a, 1989b], and for stochastic moment models by *Hernandez et al.* [2002, 2003].

[10] In this paper we expand upon the theoretical framework of MLBMA, apply it to seven geostatistical models of air permeability variation at the Apache Leap Research Site (ALRS) in central Arizona, and use cross validation to compare its predictive capabilities with those of each individual model. In the process, we introduce a new way to obtain unbiased ML estimates of variogram parameters and drift coefficients by coupling the adjoint state maximum likelihood cross validation (ASMLCV) method of *Samper and Neuman* [1989a] with universal kriging (UK) and generalized least squares (GLS).

2. Bayesian Model Averaging (BMA)

[11] According to *Hoeting et al.* [1999, p. 382], "standard statistical practice ignores model uncertainty ... leading to over-confident inferences and decisions that are more risky than one thinks they are. ... (BMA) provides a coherent mechanism for accounting for this model uncertainty." They introduce BMA by noting that if Δ is a quantity one wants to predict, then its posterior distribution given a discrete set of data $\mathbf{D}$ is

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

where $\mathbf{M} = (M_1, \ldots, M_K)$ is the set of all models (or hypotheses) considered. In other words, $p(\Delta|\mathbf{D})$ is the average of the posterior distributions $p(\Delta|M_k, \mathbf{D})$ under each model, weighted by their posterior model probabilities $p(M_k|\mathbf{D})$. The posterior probability for model M_k is given by Bayes' rule,

$$p(M_k|\mathbf{D}) = \frac{p(\mathbf{D}|M_k)p(M_k)}{\sum_{l=1}^{K} p(\mathbf{D}|M_l)p(M_l)} \tag{2}$$

where

$$p(\mathbf{D}|M_k) = \int p(\mathbf{D}|\boldsymbol{\theta}_k, M_k) p(\boldsymbol{\theta}_k|M_k) d\boldsymbol{\theta}_k \tag{3}$$

is the integrated likelihood of model M_k, $\boldsymbol{\theta}_k$ is the vector of parameters associated with model M_k, $p(\boldsymbol{\theta}_k|M_k)$ is the prior density of $\boldsymbol{\theta}_k$ under model M_k, $p(\mathbf{D}|\boldsymbol{\theta}_k, M_k)$ is the joint likelihood of model M_k and its parameters $\boldsymbol{\theta}_k$, and $p(M_k)$ is the prior probability that M_k is the correct model. All probabilities are implicitly conditional on $\mathbf{M}$.

[12] The posterior mean and variance of Δ are [*Draper*, 1995]

$$E[\Delta|\mathbf{D}] = \sum_{k=1}^{K} E[\Delta|\mathbf{D}, M_k] p(M_k|\mathbf{D}) \tag{4}$$

$$Var[\Delta|\mathbf{D}] = \sum_{k=1}^{K} Var[\Delta|\mathbf{D}, M_k] p(M_k|\mathbf{D}) + \sum_{k=1}^{K} (E[\Delta|\mathbf{D}, M_k] - E[\Delta|\mathbf{D}])^2 p(M_k|\mathbf{D}). \tag{5}$$

The first term on the right-hand side represents within-model variance; the second term represents between-model variance. Note that the predictive probabilities (1) and leading moments (4) and (5) are weighted by the posterior probabilities of the individual models.

[13] Given a set of alternative models $\mathbf{M}$, one formally assumes that their prior probabilities sum up to one,

$$\sum_{k=1}^{K} p(M_k) = 1. \tag{6}$$

This implies that all possible models of relevance are included in $\mathbf{M}$, and that all models in $\mathbf{M}$ differ from each other sufficiently to be considered mutually exclusive (the joint probability of two or more models being zero). We interpret prior model probabilities to be subjective values reflecting the analyst's belief about the relative plausibility of each model based on its apparent (qualitative, a priori) consistency with available knowledge and data.

[14] *Hoeting et al.* [1999] point out that (1) the number of potentially feasible models may be exceedingly large, rendering their exhaustive inclusion in $\mathbf{M}$ infeasible and (2) the specification of prior model probabilities $p(M_k)$ remains challenging, having received little attention in the statistical literature. A practical way to eliminate the first difficulty is to adopt the idea of Occam's window [*Jefferys and Berger*, 1992; *Madigan and Raftery*, 1994] according to which one considers only a relatively small set of the most parsimonious models among those which, a priori, appear to be hydrologically most plausible in light of all knowledge and data relevant to the purpose of the model and, a posteriori, explain the data in an acceptable manner [*Neuman and Wierenga*, 2003]. Working with a few plausible models is better than the usual hydrologic practice of adopting a single model, whereas working with many models would render the approach impractical. As demonstrated later by example, the approach can be further simplified by deleting models whose posterior probability turns out to be negligibly small in comparison to that of other models.

[15] When there is insufficient prior reason to prefer one model over another, a "reasonable 'neutral' choice" [*Hoeting et al.*, 1999] is to assume that all models are a priori equally likely. *Draper* [1999] and *George* [1999]

express concern that if two models are near equivalent as regards predictions, treating them as separate equally likely models amounts to giving double weight to a single model of which there are two slightly different versions, thereby "diluting" the predictive power of BMA. One way to minimize this effect is to eliminate at the outset models that are deemed potentially inferior. Another is to retain only models that are structurally distinct and noncollinear. Otherwise, one should consider reducing (diluting) the prior probabilities assigned to models that are deemed closely related. We explore this idea later through an example.

[16] Whereas prior model probabilities must in our view remain subjective, the posterior model probabilities are modifications of these subjective values based on an objective evaluation of each model's consistency with available data. Hence the posterior probabilities are valid only in a comparative, not in an absolute, sense. They are conditional on the choice of models (in addition to being conditional on the data) and may be sensitive to the choice of prior model probabilities (as we demonstrate later by example). This sensitivity is expected to diminish with increased level of conditioning on data.

[17] Given the above, we see no way to assess the uncertainty of hydrologic predictions in an absolute sense as proposed for a single model by *Gaganis and Smith* [2001], only in a relative sense considering several models.

3. Maximum Likelihood Bayesian Model Averaging (MLBMA)

[18] Computing the integral in equation (3) requires exhaustive Monte Carlo simulations of the prior parameter space $\boldsymbol{\theta}_k$ for each model, which may be computationally and hydrologically very demanding. *Neuman* [2002, 2003] proposed obviating the need for such simulations and prior information by adopting a maximum likelihood (ML) version (MLBMA) of BMA. It consists of replacing $\boldsymbol{\theta}_k$ by its maximum likelihood estimate $\hat{\boldsymbol{\theta}}_k$ based on the likelihood $p(\mathbf{D}|\boldsymbol{\theta}_k, M_k)$. *Taplin* [1993] suggested doing so for $p(\Delta|M_k, \mathbf{D})$ in equation (1) by adopting the approximation $p(\Delta|M_k, \hat{\boldsymbol{\theta}}_k, \mathbf{D})$. *Hoeting et al.* [1999] note that this was shown to be useful in the BMA context by *Draper* [1995], *Raftery et al.* [1996], and *Volinsky et al.* [1997].

[19] *Neuman* [2002, 2003] proposed further to evaluate the weights $p(M_k|\mathbf{D})$ in equations (1), (4), and (5) based on a result of *Kashyap* [1982]. We show in Appendix A that Kashyap's expression can be written as

$$p(M_k|\mathbf{D}) = \frac{\exp\left(-\frac{1}{2}\Delta KIC_k\right)p(M_k)}{\sum_{l=1}^{K}\exp\left(-\frac{1}{2}\Delta KIC_l\right)p(M_l)} \tag{7}$$

where

$$\Delta KIC_k = KIC_k - KIC_{\min}, \tag{8}$$

$$KIC_k = NLL_k + N_k \ln\left(\frac{N}{2\pi}\right) + \ln\left|\boldsymbol{F}_k\left(\mathbf{D}|\hat{\boldsymbol{\theta}}_k, M_k\right)\right| \tag{9}$$

KIC_k being the so-called Kashyap information criterion for model M_k, KIC_{min} its minimum value over all candidate models, and $NLL_k = -2\ln p(\mathbf{D}|\hat{\boldsymbol{\theta}}_k, M_k) - 2\ln p(\hat{\boldsymbol{\theta}}_k|M_k)$ the negative log likelihood of M_k evaluated at $\hat{\boldsymbol{\theta}}_k$. Here N_k is the dimension of $\boldsymbol{\theta}_k$ (number of parameters associated with model M_k), N is the dimension of $\mathbf{D}$ (number of discrete data points), and $\mathbf{F}_k$ is the normalized (by N) observed (as opposed to ensemble mean) Fisher information matrix having components

$$F_{k,ij} = -\frac{1}{N}\frac{\partial^2 \ln p(\mathbf{D}|\boldsymbol{\theta}_k, M_k)}{\partial\theta_i\partial\theta_j}\bigg|_{\boldsymbol{\theta}_k=\hat{\boldsymbol{\theta}}_k} \tag{10}$$

In the absence of prior information about the parameters, one simply drops the term $-2\ln p(\hat{\boldsymbol{\theta}}_k|M_k)$ from NLL_k. This reflects common practice in model calibration and is illustrated later by example.

[20] Approximating $p(\mathbf{D}|M_k)$ via equation (9) is closely related to the Laplace approximation [*Kass and Raftery*, 1995] used in BMA [e.g., *Draper*, 1995; *Hoeting et al.*, 1999]. Whereas equation (9) is obtained through expansion of $p(\mathbf{D}|\boldsymbol{\theta}_k, M_k)$ and $p(\boldsymbol{\theta}_k|M_k)$ in Taylor series about $\hat{\boldsymbol{\theta}}_k$, the Laplace approximation follows from an asymptotic expansion of the integral (3). As mentioned in the Introduction, we prefer equation (9) because it conforms more directly to ML-based hydrologic model discrimination and parameter estimation frameworks proposed for deterministic models by *Carrera and Neuman* [1986a, 1986b], for geostatistical models by *Samper and Neuman* [1989a, 1989b], and for stochastic moment models by *Hernandez et al.* [2002, 2003].

[21] Previously, KIC_k has been used [e.g., *Carrera and Neuman*, 1986a, 1998b; *Samper and Neuman*, 1989a, 1989b] as an optimum decision rule for the ranking of competing models. The highest-ranking model is that corresponding to $KIC_{\min}$. Increasing the number of parameters N_k allows $-\ln p(\mathbf{D}|\hat{\boldsymbol{\theta}}_k, M_k)$ to decrease and $N_k \ln N$ to increase. When N_k is large, the rate of decrease does not compensate for the rate of increase and KIC_k grows while $p(M_k|\boldsymbol{D})$ diminishes. This means that a more parsimonious model with fewer parameters is ranked higher and assigned a higher posterior probability. On the other hand, $-\ln p(\boldsymbol{D}|\hat{\boldsymbol{\theta}}_k, M_k)$ diminishes with N at a rate higher than linear so that as the latter grows, there may be an advantage to a more complex model with larger N_k.

[22] The last term in equation (9) gauges the information content of the available data. It thus allows considering models of growing complexity as the data base improves in quantity and quality. As illustrated by *Carrera and Neuman* [1986b], KIC_k recognizes that when the data base is limited and/or of poor quality, one has little justification for selecting an elaborate model with numerous parameters. Instead, one should prefer a simpler model with fewer parameters, which nevertheless reflects adequately the underlying hydrologic structure and regime of the system. Stated otherwise, KIC_k may cause one to prefer a simpler model that leads to a poorer fit with the data over a more complex model that fits the data better.

[23] The information term in equation (9) tends to a constant as N becomes large, so that KIC_k becomes asymptotically equivalent to the Bayes information criterion

$$BIC_k = NLL_k + N_k \ln N \tag{11}$$

derived on the basis of other considerations by *Akaike* [1977], *Rissanen* [1978], and *Schwarz* [1978]. *Raftery* [1993] proposed adopting the asymptotic *BIC* approximation, without the prior information term $-2 \ln p(\hat{\boldsymbol{\theta}}_k|M_k)$, for BMA [see also *Raftery et al.*, 1996; *Volinsky et al.*, 1997; *Hoeting et al.*, 1999]. From equation (11) it follows that equation (7) tends asymptotically to

$$p(M_k|\mathbf{D}) = \frac{\exp\left(-\frac{1}{2}\Delta BIC_k\right)p(M_k)}{\sum_{l=1}^{K}\exp\left(-\frac{1}{2}\Delta BIC_l\right)p(M_l)} \tag{12}$$

where

$$\Delta BIC_k = BIC_k - BIC_{\min} \tag{13}$$

and $BIC_{\min}$ is the smallest value of BIC_k over all candidate models [see also *Burnham and Anderson*, 2002, p. 297].

[24] Since hydrologic models are often data limited, this is less general than the nonasymptotic expression (7), which is at the heart of *Neuman*'s [2002, 2003] MLBMA. Indeed, *Carrera and Neuman* [1986a, 1986b] and *Samper and Neuman* [1989a, 1989b] found KIC_k to provide more reliable rankings of alternative groundwater flow and geostatistical models than do BIC_k or two other commonly used information criteria, $AIC_k = NLL_k + 2N_k$ [*Akaike*, 1974] and $HIC_k = NLL_k + 2N_k \ln(\ln N)$ [*Hannan*, 1980]. For a recent overview of various information criteria the reader is referred to *Burnham and Anderson* [2002, p. 284].

[25] Methods to evaluate $\hat{\boldsymbol{\theta}}_k$ by calibrating a deterministic model M_k against hydrogeologic data $\mathbf{D}$, which may include prior information about the parameters, are described by *Carrera and Neuman* [1986a, 1986b] and *Carrera et al.* [1997]. The same can be done with a stochastic model based on moment equations in a manner similar to that of *Hernandez et al.* [2002, 2003]. The approach yields a negative log likelihood criterion NLL_k that includes two weighted square residual terms: a generalized sum of squared differences between simulated and observed state variables arising from $-2 \ln p(\mathbf{D}|\hat{\boldsymbol{\theta}}_k, M_k)$, and a generalized sum of squared differences between posterior and prior parameter estimates arising from $-2 \ln p(\hat{\boldsymbol{\theta}}_k|M_k)$. The first is weighted by a matrix proportional to the inverse covariance matrix of state observation errors. The second is weighted by a matrix proportional to the inverse covariance matrix of prior parameter estimation errors. Including prior information in the calibration criterion is an option, which allows one to condition the parameter estimates not only on site monitoring (observational) data but also on site characterization data, from which prior parameter estimates are usually derived. When both sets of data are considered to be statistically meaningful, the posterior parameter estimates are compatible with a wider array of measurements than they would be otherwise and are therefore better constrained (potentially rendering the model a better predictor).

[26] Maximum likelihood estimation yields an approximate covariance matrix for the estimation errors of $\hat{\boldsymbol{\theta}}_k$. Upon considering the parameter estimation errors of a calibrated deterministic model M_k to be Gaussian or log Gaussian, one easily determines $p(\Delta|M_k, \hat{\boldsymbol{\theta}}_k, \mathbf{D})$ by Monte Carlo simulation of Δ through random perturbation of the parameters. The simulation also yields corresponding approximations $E[\Delta|M_k, \hat{\boldsymbol{\theta}}_k, \mathbf{D}]$ of $E[\Delta|M_k, \mathbf{D}]$, and $Var[\Delta|M_k, \hat{\boldsymbol{\theta}}_k, \mathbf{D}]$ of $Var[\Delta|M_k, \mathbf{D}]$, in equations (4) and (5). If M_k is a geostatistical (as in our ALRS example below) or stochastic moment (of the kind considered by *Hernandez et al.* [2002, 2003]) model, it yields $E[\Delta|M_k, \hat{\boldsymbol{\theta}}_k, \mathbf{D}]$ and $Var[\Delta|M_k, \hat{\boldsymbol{\theta}}_k, \mathbf{D}]$ directly without Monte Carlo simulation.

[27] As shown in Appendix A, alternative models can have different types and numbers of parameters, but the latter must be estimated and the models compared considering a single data set $\mathbf{D}$. For a comparison of two- and three-dimensional models, data distributed in three-dimensional space may need to be projected onto a two-dimensional plane as done be *Ando et al.* [2003] or averaged in the third dimension as suggested by *Neuman and Wierenga* [2003, Appendix B].

[28] To implement MLBMA one (1) postulates alternative conceptual-mathematical models for a site; (2) assigns a prior probability to each model; (3) optionally assigns prior probabilities to the parameters of each model; (4) obtains posterior ML parameter estimates, and estimation covariance, for each model by inversion (model calibration); (5) calculates a posterior probability for each model; (6) predicts quantities of interest using each model; (7) assesses prediction uncertainty (distribution, variance) for each model using Monte Carlo or stochastic moment methods; (8) weighs predictions and uncertainties by the corresponding posterior model probabilities; and (9) sums the results over all models.

4. Maximum Likelihood Bayesian Averaging of Spatial Variability Models in Unsaturated Fractured Tuff

[29] We apply MLBMA to alternative geostatistical models of log permeability variations in unsaturated fractured tuff at the Apache Leap Research Site (ALRS) in central Arizona. Spatially distributed log air permeability data were obtained by *Guzman et al.* [1994, 1996] based on a steady state interpretation of 184 pneumatic injection tests in 1-m length intervals along 6 vertical and inclined (at 45°) boreholes at the site (Figure 1). Five of the boreholes (V2, W2A, X2, Y2, Z2) are 30-m long and one (Y3) has a length of 45 m; five (W2A, X2, Y2, Y3, Z2) are inclined at 45° and one (V2) is vertical. Figure 2 shows an omnidirectional sample variogram of corresponding $\log_{10}k$ data. *Chen et al.* [2000] fitted three variogram models to these and some 3-m-scale data using an adjoint state maximum likelihood cross validation (ASMLCV) method developed for this purpose by *Samper and Neuman* [1989a, 1989b], coupled with a generalized least squares (GLS) drift removal approach of *Neuman and Jacobson* [1984]. The three models included (1) power (characteristic of a random fractal), (2) exponential with a linear drift, and (3) exponential with a quadratic drift. The data did not support accounting for directional effects by considering the variograms to be anisotropic. The authors found that whereas the exponential variogram model with a quadratic drift provided a best fit to the data (as measured and implied by the smallest negative log likelihood model fit criterion, *NLL*), four model discrimination criteria (*AIC*, *BIC*, *HIC*,

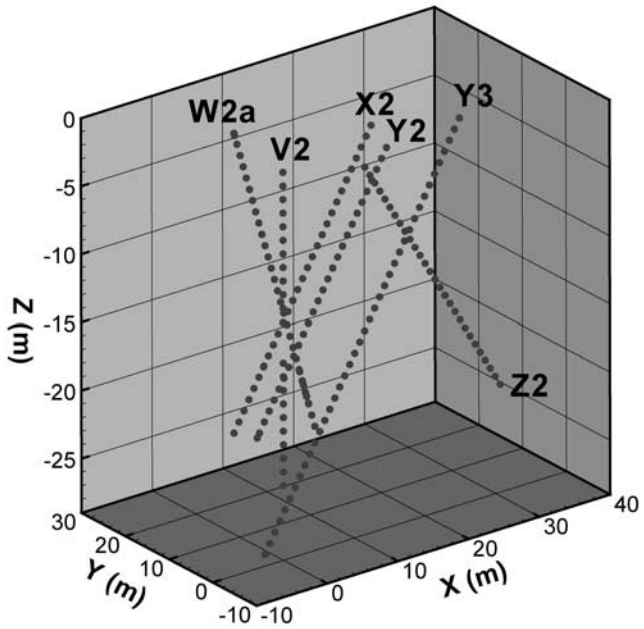

Figure 1. Spatial locations of 184 1-m-scale $\log_{10}k$ data at ALRS. See color version of this figure in the HTML.

KIC) consistently ranked the power model as best, and the former model as least acceptable. The reason was that whereas all three models provided an almost equally good fit to the data, the power model was most parsimonious with only two parameters, and the exponential variogram model with second-order drift was least parsimonious with twelve parameters. They therefore adopted the power model and discarded all other variogram models from further consideration.

[30] For purposes of MLBMA we expand the range of variogram models postulated for 1-m-scale $\log_{10}k$ at the ALRS to seven: (1) Power (*Pow0*), (2) exponential without a drift (*Exp0*), (3) exponential with a linear drift (*Exp1*), (4) exponential with a quadratic drift (*Exp2*), (5) spherical without a drift (*Sph0*), (6) spherical with a linear drift (*Sph1*), and (7) spherical with a quadratic drift (*Sph2*). To estimate the parameter vector $\boldsymbol{\beta}$ of drift-free variogram models (*Pow0*, *Exp0*, *Sph0*) we use ASMLCV as described in Appendix B, implemented in a computer code slightly modified after F. J. Samper (personal communication, 1998). To do the same for models with drift (*Exp1*, *Exp2*, *Sph1*, *Sph2*), we decompose the N-dimensional data vector $\mathbf{D}$ of $\log_{10}k$ measurements into a deterministic drift vector $\boldsymbol{\mu}$ and a random residual vector $\mathbf{R}$,

$$\mathbf{D} = \boldsymbol{\mu} + \mathbf{R} \tag{14}$$

$$\boldsymbol{\mu}(\mathbf{x}) = \sum_{k=0}^{p} g_k(\mathbf{x}) a_k = \mathbf{G}\mathbf{a} \tag{15}$$

where $\mathbf{a} = (a_0, a_1, \ldots, a_p)^T$ is a vector of $p + 1$ drift coefficients and $\mathbf{G}$ is a $N \times (p + 1)$ matrix of linearly independent monomial functions $g_k(\mathbf{x})$ evaluated at the data points $\mathbf{x}_n$, $n = 1, 2, \ldots, N$. Assuming that $\mathbf{D}$ is multivariate Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}_R$ (*Vesselinov* [2000] has shown that the data pass the Kolmogorof-Smirnov test of univariate Gaussianity at a significance level of 0.05), the joint negative log likelihood function of drift and variogram parameters takes the form

$$NLL(\mathbf{a}, \boldsymbol{\beta}|\mathbf{D}) = -2\ln p(\mathbf{D}|\mathbf{a}, \boldsymbol{\beta}) = N\ln 2\pi + \ln|\mathbf{C}_R(\boldsymbol{\beta})| + (\mathbf{D} - \mathbf{G}\mathbf{a})^T \mathbf{C}_R^{-1}(\boldsymbol{\beta})(\mathbf{D} - \mathbf{G}\mathbf{a}). \tag{16}$$

Minimizing equation (16) jointly with respect to $\mathbf{a}$ and $\boldsymbol{\beta}$ yields biased estimates of the variogram parameters, a problem that can be remedied through the use of a restricted ML (RML) approach [*Hoeksema and Kitanidis*, 1985; *Kitanidis and Lane*, 1985; *Cressie*, 1991, p. 92]. We solve the problem differently by formally decoupling the ML estimations of $\mathbf{a}$ and $\boldsymbol{\beta}$. First, we obtain unbiased ML estimates $\hat{\boldsymbol{\beta}}$ of the variogram parameters using ASMLCV in conjunction with universal kriging (ASMLCV-UK (F. J. Samper, personal communication, 1998)), which does not require knowledge of the drift coefficients (Appendix B). Next, we compute corresponding unbiased ML estimates $\hat{\mathbf{a}}$ of the drift coefficients through minimization of

$$NLL\left(\mathbf{a}, \hat{\boldsymbol{\beta}}|\mathbf{D}\right) = N\ln 2\pi + \ln\left|\mathbf{C}_R\left(\hat{\boldsymbol{\beta}}\right)\right| + (\mathbf{D} - \mathbf{G}\mathbf{a})^T \mathbf{C}_R^{-1}\left(\hat{\boldsymbol{\beta}}\right)(\mathbf{D} - \mathbf{G}\mathbf{a}) \tag{17}$$

with respect to $\mathbf{a}$ by generalized least squares, a task we accomplish using PEST-ASP [*Doherty*, 2002]. Our optimum *NLL* is then given by

$$NLL\left(\hat{\mathbf{a}}, \hat{\boldsymbol{\beta}}|\mathbf{D}\right) = N\ln 2\pi + \ln\left|\mathbf{C}_R\left(\hat{\boldsymbol{\beta}}\right)\right| + (\mathbf{D} - \mathbf{G}\hat{\mathbf{a}})^T \mathbf{C}_R^{-1}\left(\hat{\boldsymbol{\beta}}\right)(\mathbf{D} - \mathbf{G}\hat{\mathbf{a}}). \tag{18}$$

[31] Figure 3 depicts profiles of $NLL(\mathbf{a}, \boldsymbol{\beta}|\mathbf{D})$ in equation (16) versus each parameter of model *Exp1* when the remaining parameters are fixed. It clearly demonstrates that $\hat{\boldsymbol{\beta}}$ (the marked values of sill and integral scale [m]) does not

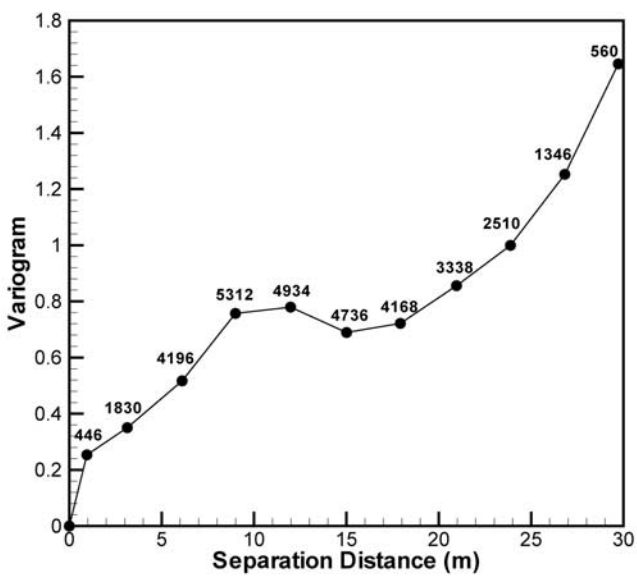

Figure 2. Omnidirectional sample variogram of 1-m-scale $\log_{10}k$ data at the ALRS and numbers of data pairs.

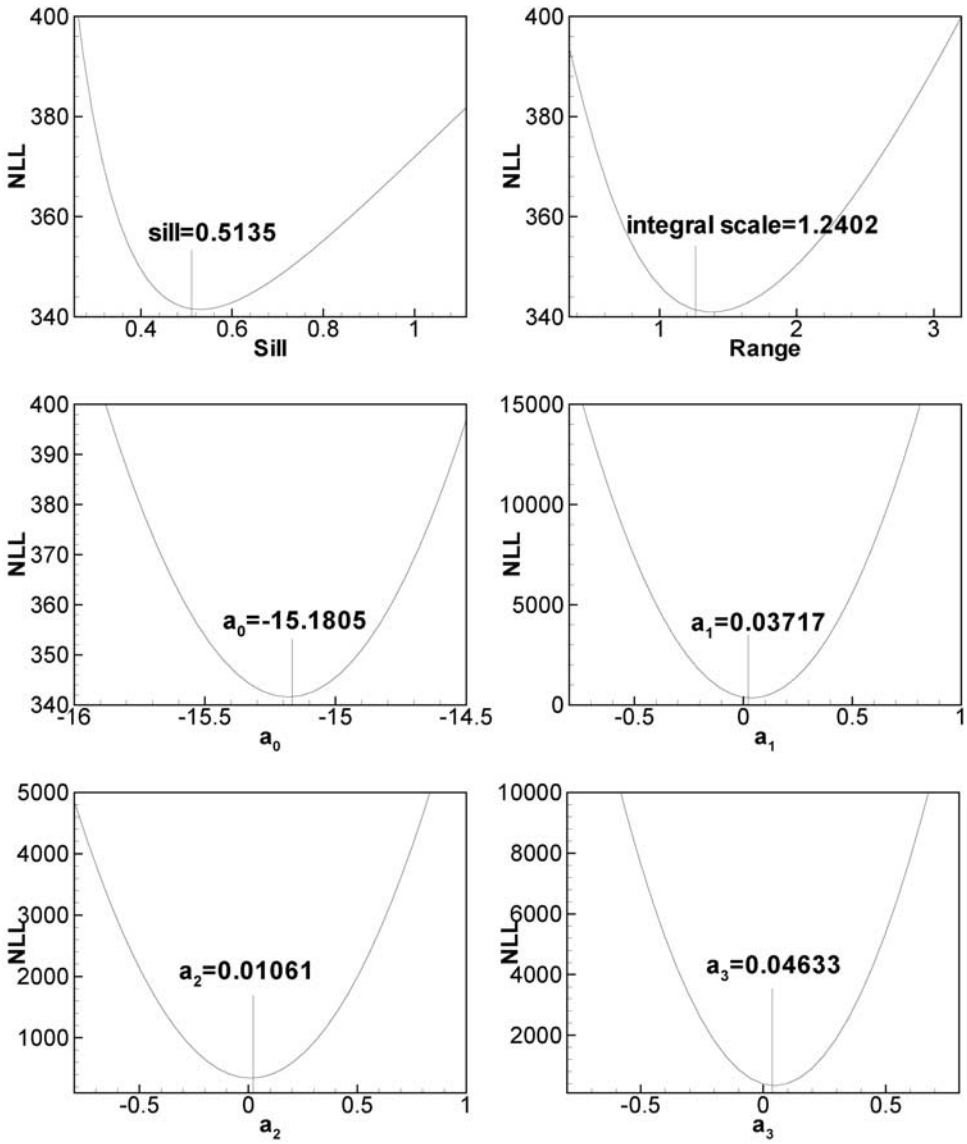

Figure 3. Negative log likelihood functions (*NLL*) as function of each variogram parameter and drift coefficient for exponential model with linear drift (*Exp1*). Vertical lines indicate unbiased ML estimates.

correspond to the minimum of $NLL(\mathbf{a}, \beta|\mathbf{D})$, which would therefore yield biased estimates of variogram parameters.

[32] The estimation covariance matrix of $\hat{\boldsymbol{\theta}} = (\hat{\mathbf{a}}, \hat{\beta})^T$ is generally represented by its asymptotic lower or Cramer-Rao bound, given by the inverse Fisher information matrix [e.g., *Carrera et al.*, 1997]. Components of the observed Fisher information matrix (10) are proportional to those of the Hessian matrix $\mathbf{H}$ which, in turn, can be approximated as [*Kitanidis and Lane*, 1985]

$$H_{k,ij} = -\left.\frac{\partial^2 \ln p(\mathbf{D}|\boldsymbol{\theta}_k, M_k)}{\partial\boldsymbol{\theta}_i \partial\boldsymbol{\theta}_j}\right|_{\boldsymbol{\theta}_k=\hat{\boldsymbol{\theta}}_k} \simeq \frac{1}{2} Tr\left(\mathbf{C}_R^{-1}\frac{\partial \mathbf{C}_R}{\partial\boldsymbol{\theta}_i}\mathbf{C}_R^{-1}\frac{\partial \mathbf{C}_R}{\partial\boldsymbol{\theta}_j}\right) + \left.\frac{\partial \mathbf{R}^T}{\partial\boldsymbol{\theta}_i}\mathbf{C}_R^{-1}\frac{\partial \mathbf{R}}{\partial\boldsymbol{\theta}_j}\right|_{\boldsymbol{\theta}_k=\hat{\boldsymbol{\theta}}_k}. \quad (19)$$

This approximation obviates the need to calculate second-order derivatives of the log likelihood function, which would be computationally more demanding than computing first-order derivatives of $\mathbf{C}_R$ and $\mathbf{R}$. In our case, the latter are easy to obtain analytically as done for exponential and spherical variogram models with drift in Appendix C. An alternative, which in our case yields very similar results, is to compute the observed Fisher information matrix numerically using methods such as the Ridder algorithm [*Press et al.*, 1992, p. 180].

[33] Table 1 confirms that increasing the number of parameters associated with a given class of variogram model (exponential or spherical) brings about an improvement in model fit, as indicated by a reduction in the negative log likelihood criterion *NLL*. Whereas the exponential variogram model with a quadratic drift (*Exp2*) fits the data best (ranks first in terms of fit due to its smallest *NLL* value), it is ranked second by *AIC* and sixth by *BIC* and *KIC*. Whereas the power model (*Pow0*) shows a relatively poor fit with the data (rating fifth), it is ranked highly (first through third) by all three information criteria. The reason is that the difference in fit between the two models is not enough to

Table 1. Quality Criteria, Rankings, and Prior/Posterior Probabilities Associated With Alternative Geostatistical Models of $\log_{10}k$ at the ALRS

Model	*Pow0*	*Exp0*	*Exp1*	*Exp2*	*Sph0*	*Sph1*	*Sph2*
Number of parameters	2	2	6	12	2	6	12
Sill/coefficient	0.286	0.718	0.514	0.501	0.749	0.664	0.662
Correlation/power	0.460	1.840	1.240	1.198	3.184	2.849	2.835
NLL	352.186	361.006	341.565	330.350	379.059	349.596	338.803
NLL Rank	5	6	3	1	7	4	2
AIC	356.186	365.006	353.565	354.350	383.059	361.596	362.803
AIC Rank	3	6	1	2	7	5	4
BIC	362.616	371.436	372.855	392.929	389.489	380.886	401.382
BIC Rank	1	2	3	6	5	4	7
KIC	369.581	370.148	369.454	416.654	390.535	378.072	424.619
KIC Rank	2	3	1	6	5	4	7
$P(M_k)$	1/7	1/7	1/7	1/7	1/7	1/7	1/7
$p(M_k\|\boldsymbol{D})$, %	35.298	26.584	37.612	0	0	0.506	0
$p(M_k)$	1/4	1/4	1/4	-	-	1/4	-
$p(M_k\|\boldsymbol{D})$, %	35.298	26.584	37.612	-	-	0.506	-
$p(M_k)$	1/3	1/9	1/9	1/9	1/9	1/9	1/9
$p(M_k\|\boldsymbol{D})$, %	62.073	15.583	22.047	0	0	0.297	0
$p(M_k)$	1/3	1/6	1/6	-	-	1/3	-
$p(M_k\|\boldsymbol{D})$, %	51.984	19.575	27.696	-	-	0.745	-

compensate for the much more parsimonious nature of *Pow0* (with 2 parameters) than that of *Exp2* (with 12 parameters).

[34] The rankings of the seven models by *AIC*, *BIC* and *KIC* are not entirely consistent. None of these information criteria provide justification for retaining one model while discarding all other models as is commonly done in practice. Nor do they provide clear justification for retaining some models while discarding the rest. We therefore consider all seven models to be valid initial candidates for MLBMA.

[35] Upon assigning an equal prior probability of 1/7 to each model, we find on the basis of *KIC* via equation (7) that the first three models (*Pow0*, *Exp0*, *Exp1*) have much higher posterior probabilities than do the rest. Three of the models (*Exp2*, *Sph0*, *Sph2*) have zero probabilities (to three significant figures) and can therefore be eliminated (considering the low posterior probability of *Sph1*, there is almost equal justification for eliminating it too, but we retain it at this stage for the sake of illustration). Doing so and assigning an equal prior probability of 1/4 to each of the retained models is seen to have no impact on their posterior probabilities. In both cases the posterior probabilities are markedly different from their prior values, reflecting the strong impact of conditioning on data.

[36] To investigate the influence of prior probability selection on the outcome, consider assigning an equal probability of 1/3 to each of the three classes of models (power, exponential and spherical) and also assigning equal probability to models within each class. This results in a prior probability of 1/3 for *Pow0* and of 1/9 for each of the other six models. Though this brings about a marked increase in the posterior probability of *Pow0* and a decrease in those of *Exp0* and *Exp1*, once again the posterior probabilities of *Exp2*, *Sph0* and *Sph2* are zero while that of *Sph1* is very close to zero. Eliminating the three models with zero posterior probability and redistributing the prior probabilities among the remaining models as shown in the eighteenth row of Table 1 brings about a decrease in the posterior probability of *Pow0* and an increase in the posterior probabilities of *Exp0* and *Exp1*. We conclude that posterior model probabilities exhibit some degree of sensitivity to the choice of prior probabilities but expect this sensitivity to diminish with improved conditioning.

[37] We continue our analysis by retaining four (*Pow0*, *Exp0*, *Exp1*, *Sph1*) of the seven models (with the corresponding ML parameter estimates) and assigning to each of them an equal prior probability of 1/4. Using each of these models, we project the available $\log_{10}k$ data by ordinary (in the case of drift-free models) or universal (otherwise) kriging onto a grid of $50 \times 40 \times 30$ 1-m^3 cubes contained within the coordinate ranges $-10 \le x \le 40$ m, $-10 \le y \le 30$ m and $-30 \le z \le 0$ m in Figure 1. If one thinks of Δ as a random value of $\log_{10}k$ in a given grid block then our kriging estimates represent $E[\Delta|M_k, \hat{\boldsymbol{\theta}}_k, \mathbf{D}]$ and their variances stand for $Var[\Delta|M_k, \hat{\boldsymbol{\theta}}_k, \mathbf{D}]$, the ML approximations of $E[\Delta|M_k, \mathbf{D}]$ and $Var[\Delta|M_k, \mathbf{D}]$ in equations (4) and (5), respectively. Figures 4–7 show the kriged estimates and variances of $\log_{10}k$ on a vertical plane $y = 6.5$ m for the four models. Conditioning on borehole data is evident to a lesser degree in the images of $\log_{10}k$ estimates than in those of their variances. Averaging the kriging results across all models using an ML approximation of equations (4) and (5) yields corresponding MLBMA estimates and variances of the kind depicted for $y = 6.5$ m in Figure 8. Figure 9 shows a decomposition of the MLBMA estimation variance in Figure 8b into its within- and between-model components. The largest values of these two components throughout the three-dimensional grid are 1.1 and 0.38, respectively. Whereas the within-model MLBMA variance in Figure 9a reflects conditioning on borehole measurements, it is difficult to discern such conditioning in the image of between-model variance (Figure 9b) due to the faint reflection of such conditioning in the underlying images of $\log_{10}k$ estimates.

[38] Figure 10 shows univariate cumulative distributions of kriging estimates corresponding to each of the four models and MLBMA. The distributions are seen to be sensitive to the choice of model with MLBMA providing

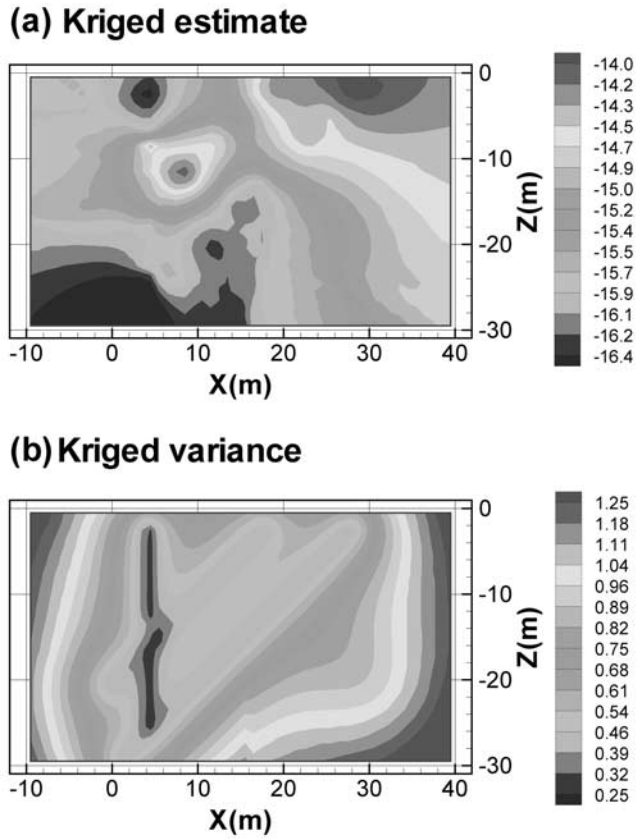

Figure 4. Kriged (a) estimate and (b) variance of $\log_{10}k$ at y = 6.5 m obtained using the power model (*Pow0*). See color version of this figure in the HTML.

a weighted compromise. The same is reflected in the variances of these kriged estimates, listed in Table 2.

5. Assessment of Predictive Performance

[39] To assess the predictive performance of MLBMA, we cross validate the above results by (1) splitting the data **D** into two parts, $\mathbf{D}^A$ and $\mathbf{D}^B$; (2) obtaining ML estimates of model parameters and posterior probabilities conditional on $\mathbf{D}^A$; (3) using these to render MLBMA predictions $\hat{\mathbf{D}}^B$ of $\mathbf{D}^B$; (4) and assessing the quality of the predictions. We do so by eliminating from consideration all $\log_{10}k$ data from one borehole at a time and predicting them with models conditioned on the remaining data. The number and corresponding percentage of data in $\mathbf{D}^A$ for each cross-validation case are listed in Table 3. As *Sph1* has a very small posterior probability in comparison to *Pow0*, *Exp0* and *Exp1* (Table 1), we limit the cross validation to the latter three geostatistical models and recalculate their posterior probabilities by assigning to each of them a prior probability of 1/3.

[40] Figure 11 shows that eliminating data from one borehole at a time may, but need not, have a significant impact on the omnidirectional sample variogram of $\log_{10}k$. The impact that such elimination has on parameter estimates and model quality criteria associated with *Pow0* is indicated in Figure 12. Figure 13 demonstrates that posterior model probability is sensitive to the choice of conditioning data. This sensitivity is greater when posterior probability is computed using *KIC* in equation (7) than *BIC* in equation (12). This illustrates that the nonasymptotic criterion *KIC* is more informative than the asymptotic criterion *BIC*, supporting the choice of the former as the basis for MLBMA [*Neuman*, 2002, 2003].

[41] One way to compare the predictive capabilities of alternative models is through their log scores, $-\ln p(\mathbf{D}^B|M_k,\ \mathbf{D}^A)$ [*Good*, 1952; *Volinsky et al.*, 1997]. The lower the predictive log score of model M_k based on data $\mathbf{D}^A$, the smaller the amount of information lost upon eliminating $\mathbf{D}^B$ from the original dataset **D** (i.e., the higher the probability that M_k based on $\mathbf{D}^A$ would reproduce the lost data, $\mathbf{D}^B$). The predictive log score associated with BMA is

$$-\ln p(\mathbf{D}^B|\mathbf{D}^A) = -\ln \sum_{k=1}^{K} p(\mathbf{D}^B|M_k,\mathbf{D}^A)p(M_k|\mathbf{D}^A). \qquad (20)$$

Approximating $p(\mathbf{D}^B|M_k,\ \mathbf{D}^A)$ by $p(\mathbf{D}^B|M_k,\ \hat{\theta}_k,\ \mathbf{D}^A)$, and computing $p(M_k|\mathbf{D}^A)$ via equation (7) after replacing **D** by $\mathbf{D}^A$, yields a corresponding log score for MLBMA.

[42] Let $\hat{\mathbf{D}}^{\boldsymbol{B}}$ be kriged estimates of $\log_{10}k$ data $\mathbf{D}^B$ along a borehole obtained using variogram model M_k with ML parameters $\hat{\theta}_k$ based on $\log_{10}k$ data $\mathbf{D}^A$ in other boreholes.

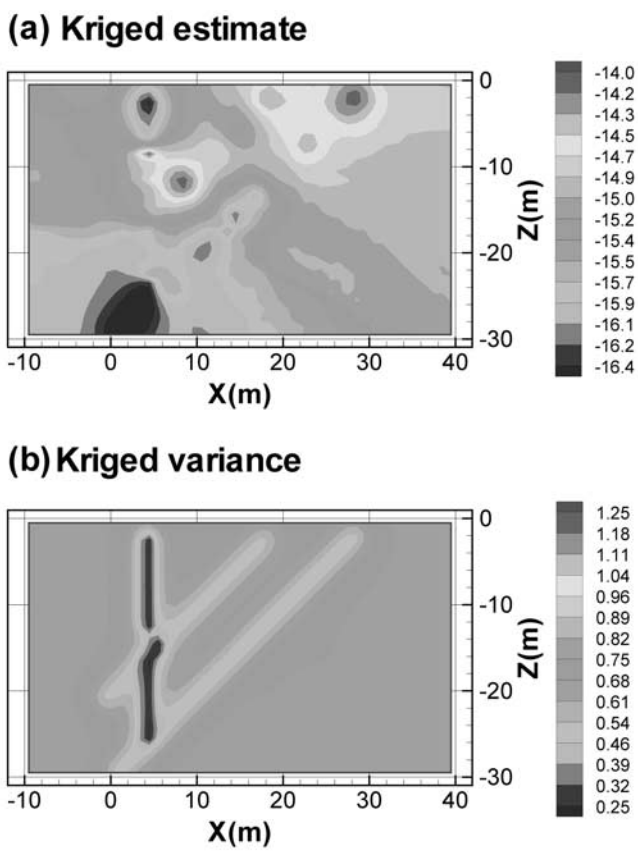

Figure 5. Kriged (a) estimate and (b) variance of $\log_{10}k$ at y = 6.5 m obtained using the exponential model without drift (*Exp0*). See color version of this figure in the HTML.

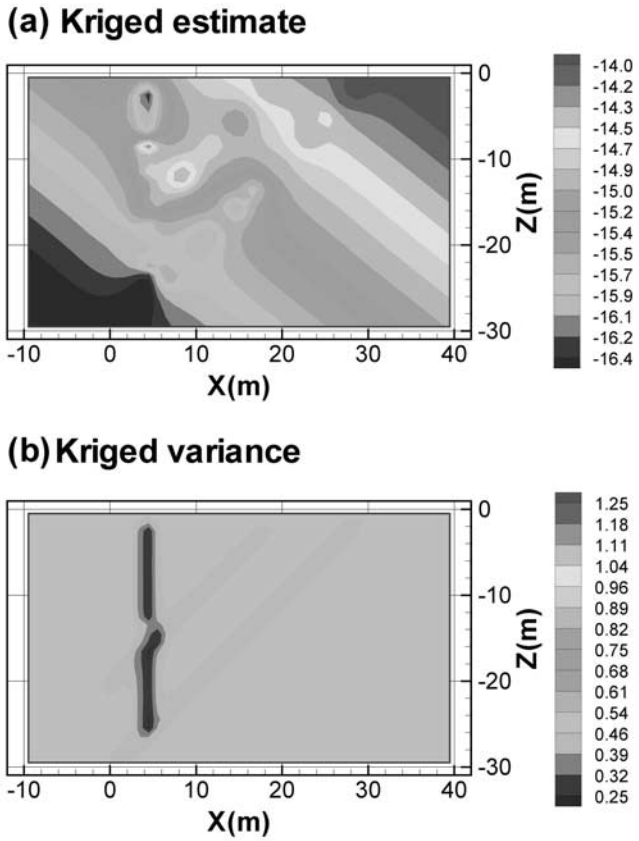

Figure 6. Kriged (a) estimate and (b) variance of $\log_{10}k$ at $y = 6.5$ m obtained using the exponential model with first-order drift (*Exp1*). See color version of this figure in the HTML.

Then in analogy to equation (B4), the ML log score for drift-free models *Pow0* and *Exp0* is

$$-\ln p\left(\mathbf{D}^B|M_k,\hat{\boldsymbol{\theta}}_k,\mathbf{D}^A\right)=\frac{N_d}{2}\ln(2\pi)+\frac{1}{2}\sum_{i=1}^{N_d}\sigma_i^2+\frac{1}{2}\sum_{i=1}^{N_d}\frac{\left(\hat{D}_i^B-D_i^B\right)^2}{\sigma_i^2} \quad (21)$$

where N_d is the dimension of $\mathbf{D}^B$, D_i^B are its components, and σ_i^2 is given by equation (B5). In analogy to equation (17), the ML log score for *Exp1* is

$$-\ln p\left(\mathbf{D}^B|M_k,\hat{\boldsymbol{\theta}}_k,\mathbf{D}^A\right)=\frac{N_d}{2}\ln(2\pi)+\frac{1}{2}\ln\left(|\mathbf{C}_R\left(\hat{\beta}_k\right)|\right)+\frac{1}{2}\left(\mathbf{D}^B-\mathbf{G}_k\hat{\mathbf{a}}_k\right)^T\mathbf{C}_R^{-1}\left(\hat{\beta}\right)\left(\mathbf{D}^B-\mathbf{G}_k\hat{\mathbf{a}}_k\right). \quad (22)$$

[43] Predictive log scores were obtained for each model upon eliminating data from one of six boreholes at a time. Table 4 lists the average of these six scores for each model, as well as the average of corresponding MLBMA scores equation (20). The average predictive log score of MLBMA is seen to be lower than that of any individual model, indicating that MLBMA is a better predictor than any of these models.

[44] Another measure of model performance is its predictive coverage [*Hoeting et al.*, 1999]. This is the percent of measurements D_i^B that fall within a given prediction interval about $\hat{D}_i^B$. In our case, this interval was generated by conducting Monte Carlo simulations of $\log_{10}k$ conditioned on $\mathbf{D}^A$. We used a simulated annealing code [*Deutsch and Journel*, 1998, p. 183] to allow generation of statistically nonhomogeneous random fields characterized by a power variogram. Figures 14a–14c show 90% prediction intervals (dashed) defining the 5% and 95% limits of 500 simulations along borehole X2 using individual models with ML parameter estimates conditioned on measurements in the remaining five boreholes. Figure 14d shows averages of these intervals over the three models, weighted by their posterior probabilities. The percent of measurements (triangles) lying within these and similar intervals, associated with all six boreholes, defines predictive coverage as listed in Table 4. The predictive coverage of MLBMA is larger than that of any individual model, attesting once again to its superior performance.

[45] Figure 15 depicts the cumulative distributions of simulated values at two measurement locations in boreholes V2 and Y3 obtained using individual models and MLBMA, while eliminating data from the corresponding boreholes. The measured values are indicated by vertical lines. In both

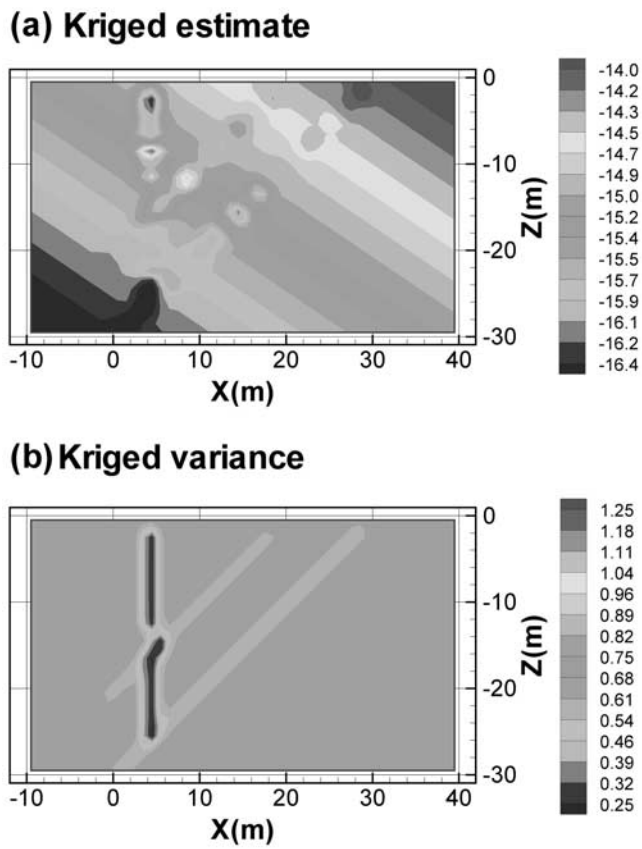

Figure 7. Kriged (a) estimate and (b) variance of $\log_{10}k$ at $y = 6.5$ m obtained using the spherical model with first-order drift (*Sph1*). See color version of this figure in the HTML.

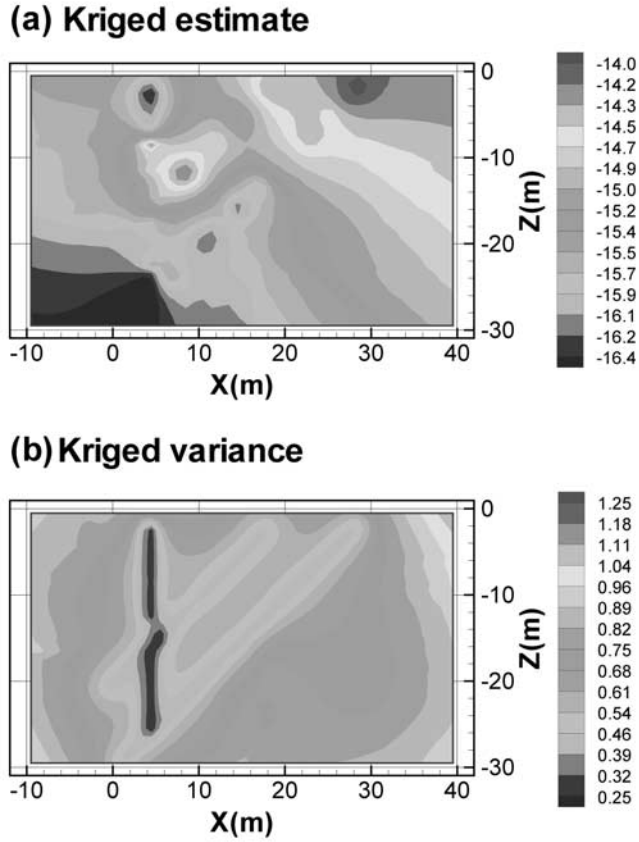

Figure 8. Kriged (a) estimate and (b) variance of $\log_{10}k$ at $y = 6.5$ m obtained using MLBMA. See color version of this figure in the HTML.

cases the MLBMA distribution is strongly influenced by that of *Pow0* and weakly affected by *Exp1*. Figure 16 shows sample predictive variances obtained using individual models and MLBMA at measurement points along each of the two boreholes. Along V2, *Pow0* with a posterior probability of about 83% exerts an overwhelming influence on the predictive variance of MLBMA, which is however lower (closer to those of *Exp0* and *Exp1*). Along Y3, individual models tend to be associated with a somewhat lower predictive variance than MLBMA.

[46] Overall, MLBMA is a more reliable predictor than any individual model, as indicated by its relatively low log score and high predictive coverage.

6. Conclusions

[47] 1. Analyses of model uncertainty based on a single hydrologic concept are prone to statistical bias (by committing a type II error through reliance on an invalid model) and underestimation of uncertainty (by committing a type I error through under sampling of the relevant model space). Bias and uncertainty resulting from an inadequate model structure (conceptualization) are often more detrimental to a model's predictive reliability than are suboptimal model parameters.

[48] 2. Bayesian model averaging (BMA) provides an optimal but computationally demanding way of combining the predictions of several competing models and assessing their joint predictive uncertainty. The maximum likelihood (ML) version (MLBMA) of BMA proposed by *Neuman* [2002, 2003], and implemented in this paper, renders the approach computationally feasible and applicable to real-world hydrologic problems. It applies to both deterministic and stochastic models.

[49] 3. Whereas BMA requires specifying a prior distribution for model parameters, MLBMA accepts but does not require such prior information. This is so because, contrary to BMA, MLBMA relies on ML model calibration against observational data.

[50] 4. There appears to be no valid way to assess the uncertainty of hydrologic predictions in an absolute sense for a single model, only in a relative sense for several models conditioned on the choice of models and data.

[51] 5. MLBMA is based on *Kashyap*'s [1982] information criterion, *KIC*, more commonly used as an optimum decision rule for the ranking of competing models. Like *KIC*, MLBMA favors models which, among a given set of alternatives, are least likely to be incorrect. It honors the principle of parsimony by favoring the least complex among models which, otherwise, fit observational data equally well. Among models of equal complexity, MLBMA favors those exhibiting the best fit. It additionally contains an information term which allows one to consider models of growing complexity as the dataset improves in quantity and quality. Stated otherwise, MLBMA recognizes that when

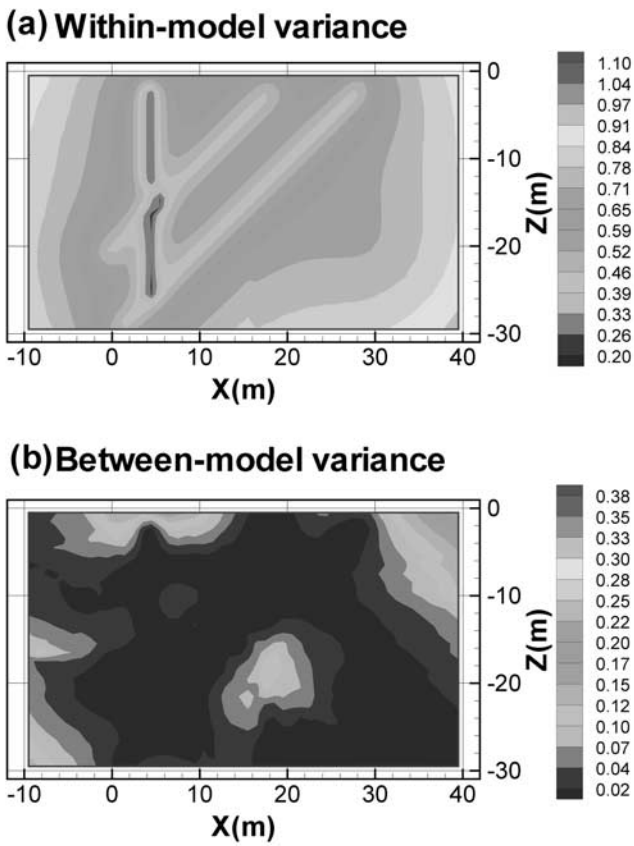

Figure 9. (a) Within- and (b) between-model variance of MLBMA $\log_{10}k$ estimates at $y = 6.5$ m. See color version of this figure in the HTML.

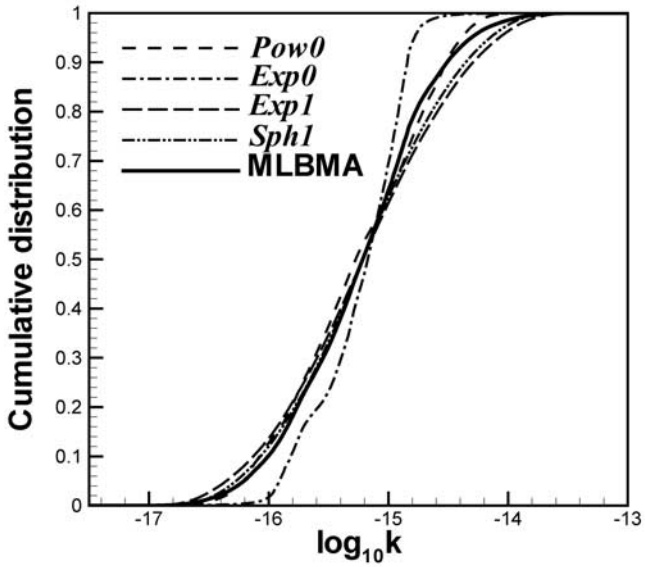

Figure 10. Cumulative distribution of kriged $\log_{10}k$ estimates obtained using various models and MLBMA.

the dataset is limited and/or of poor quality, one should assign relatively low weights to elaborate models with numerous parameters. One should weigh more heavily simpler models with fewer parameters that nevertheless reflect adequately the underlying hydrologic structure and phenomena.

[52] 6. Our example confirms that the nonasymptotic criterion *KIC* is more informative than its asymptotic limit *BIC*, supporting the choice of the former as the basis for MLBMA [*Neuman*, 2002, 2003].

[53] 7. Models considered in MLBMA may have different types and numbers of parameters, but the latter must be estimated and the models weighted based on a single dataset. As an example, to analyze jointly two- and three-dimensional models via MLBMA, a given set of three-dimensional data must be used and either projected onto a two-dimensional plane or averaged in the third dimension for inclusion in the two-dimensional model(s).

[54] 8. Application of MLBMA to alternative geostatistical models of log air permeability variations in unsaturated fractured tuff has shown it to be a better predictor of spatial variability than any individual model.

[55] 9. It is possible to obtain unbiased ML estimates of variogram parameters and drift coefficients by coupling the adjoint state maximum likelihood cross validation (ASMLCV) method of *Samper and Neuman* [1989a] with universal kriging (UK) and generalized least squares (GLS).

Table 2. Variance of Kriged Estimates Across Grid Obtained With Alternative Models and MLBMA

Model	Variance
Pow0	0.334
Exp0	0.134
Exp1	0.467
Sph1	0.404
MLBMA	0.405

Table 3. Number of $\log_{10}k$ Data in $\mathbf{D}^A$ of Each Cross-Validation Case and Their Percentage of the Entire Data Set

Well	Number	Percentage
V2	163	89.1
X2	154	83.7
Y2	156	84.8
Y3	144	78.3
Z2	156	84.8
W2A	147	79.9

Appendix A

[56] *Kashyap* [1982] used asymptotic expansion to show that for linear or nonlinear, Gaussian or non-Gaussian models under some fairly standard conditions,

$$\ln p(M_k|\mathbf{D}) = \ln C_k + \ln p(\mathbf{D}|M_k) \tag{A1}$$

where $C_k = cp(M_k)$, c is a constant determined so as to insure that

$$\sum_{l=1}^{K} p(M_l|\mathbf{D}) = 1, \tag{A2}$$

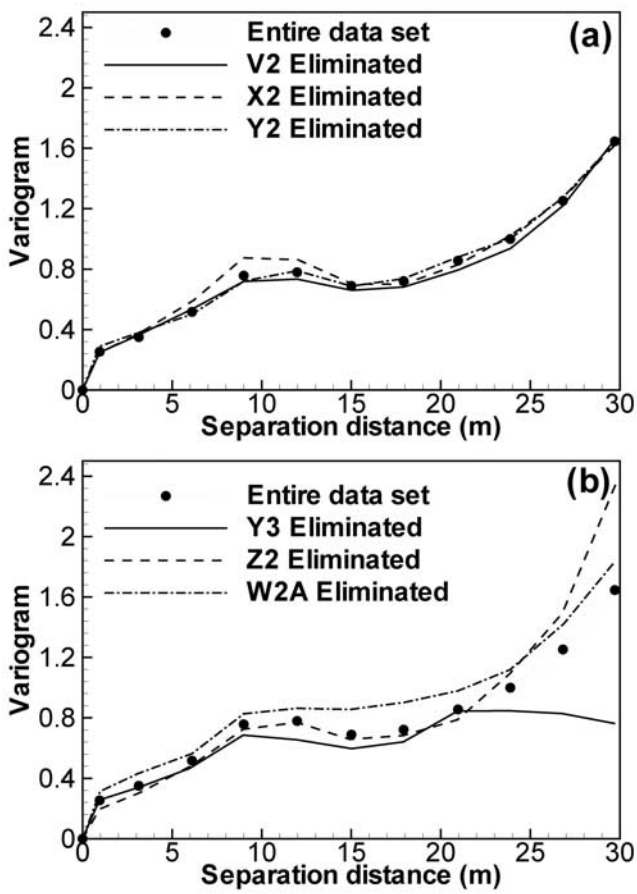

Figure 11. Omnidirectional sample variograms of all data and all but data from boreholes (a) V2, X2, and Y2 and (b) Y3, Z2, and W2A.

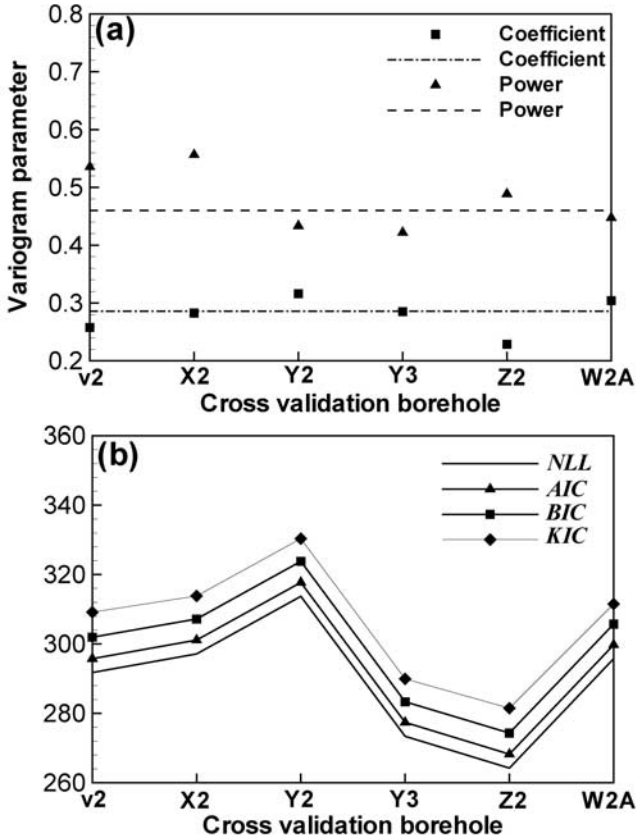

Figure 12. Dependence of power variogram (*Pow0*) (a) parameters and (b) quality criteria on data. In Figure 12a, symbols designate parameter estimates obtained without data from designated borehole; long- and short-dashed line and dashed line designate estimates with all data.

$$\ln p(\mathbf{D}|M_k) = \ln p\left(\mathbf{D}|\hat{\boldsymbol{\theta}}_k, M_k\right) + \ln p\left(\hat{\boldsymbol{\theta}}_k|M_k\right) + \frac{N_k}{2}\ln\left(\frac{2\pi}{N}\right) - \frac{1}{2}\ln\left|\mathbf{F}_k\left(\mathbf{D}\middle|\hat{\boldsymbol{\theta}}_k, M_k\right)\right| + R(N), \tag{A3}$$

and $NR(N)$ tends to a constant almost surely as $N \to \infty$. For a given N, it is possible to write $\ln \alpha = R(N)$ and make α part of the normalizing constant c (see below). Hence equation (A1) can be expressed with the aid of equation (9) as

$$p(M_k|\mathbf{D}) = cp(M_k)\exp\left(-\frac{1}{2}KIC_k\right) \tag{A4}$$

where, by virtue of equation (A2),

$$c = \frac{1}{\sum_{l=1}^{K}\exp\left(-\frac{1}{2}KIC_l\right)p(M_l)}. \tag{A5}$$

To avoid having large arguments in the exponent, we rewrite equations (A4) and (A5) as equation (7) in terms of the difference (equation (8)).

[57] For the purpose of dimensional analysis we combine and rewrite equations (A1) and (A3) as

$$p(M_k|\mathbf{D}) = cp(M_k)p\left(\mathbf{D}|\hat{\boldsymbol{\theta}}_k, M_k\right)p\left(\hat{\boldsymbol{\theta}}_k|M_k\right)\left(\frac{2\pi}{N}\right)^{\frac{N_k}{2}} \cdot \left|\mathbf{F}_k\left(\mathbf{D}|\hat{\boldsymbol{\theta}}_k, M_k\right)\right|^{-1/2}\alpha. \tag{A6}$$

We note that $p(M_k|\mathbf{D})$ is the posterior (discrete, dimensionless) probability of model M_k; $p(M_k)$ is the prior (discrete, dimensionless) probability of M_k; $p(\mathbf{D}|\hat{\boldsymbol{\theta}}_k, M_k)$ is the probability density function (continuous, having inverse dimensions of $\mathbf{D}$, i.e., $(d_1d_2d_3\ldots d_N)^{-1}$ where d_i is the dimension of D_i) of the data vector $\mathbf{D}$ under model M_k with parameters $\hat{\boldsymbol{\theta}}_k$; $p(\boldsymbol{\theta}_k|M_k)$ is the prior probability density of $\boldsymbol{\theta}_k$ under model M_k (continuous, having inverse dimensions of $\boldsymbol{\theta}_k$, i.e., $(t_1t_2t_3\ldots t_{N_k})^{-1}$ where t_i is the dimension of θ_i)); $|\mathbf{F}_k(\mathbf{D}|\hat{\boldsymbol{\theta}}_k, M_k)|^{-1/2}$, by virtue of equation (10), is continuous with dimensions of $\boldsymbol{\theta}_k$, i.e., $(t_1t_2t_3\ldots t_{N_k})$); α is dimensionless; and hence the normalizing constant c (whether or not one absorbs α into it) has dimensions of $\mathbf{D}$, i.e., $(d_1d_2d_3\ldots d_N)$. As the dimensions of $p(\boldsymbol{\theta}_k|M_k)$ and $|\mathbf{F}_k(\mathbf{D}|\hat{\boldsymbol{\theta}}_k, M_k)|^{-1/2}$ cancel, it is legitimate to add $p(M_k|\mathbf{D})$ corresponding to models M_k having different types and

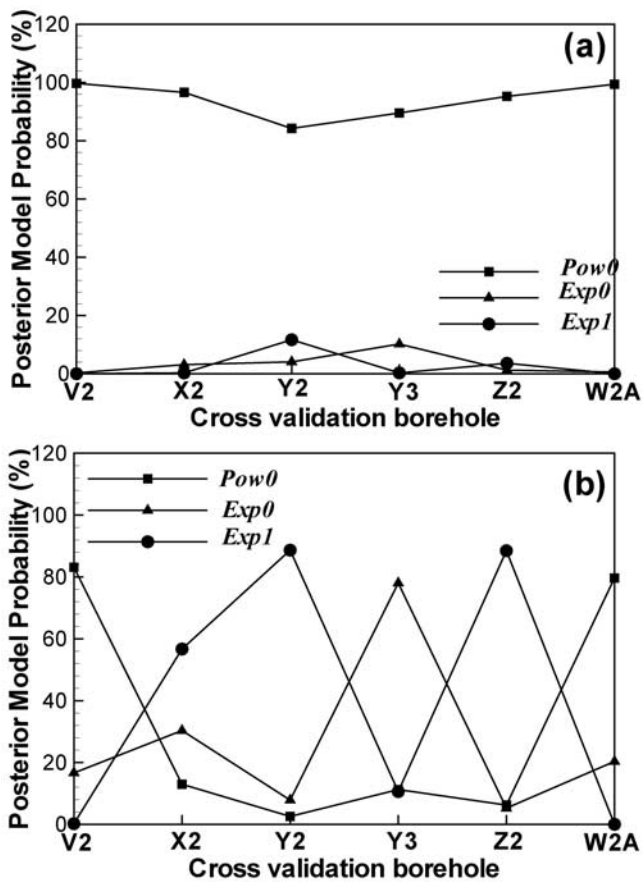

Figure 13. Posterior model probabilities based on (a) *BIC* and (b) *KIC* upon eliminating data from designated borehole.

Table 4. Average Predictive Log Score and Predictive Coverage of Individual Models and MLBMA

Model	*Pow0*	*Exp0*	*Exp1*	MLBMA
Predictive log score	34.1	35.2	34.0	31.4
Predictive coverage, %	86.5	80.8	83.7	87.5

numbers of parameters $\boldsymbol{\theta}_k$. On the other hand, $p(M_k|\mathbf{D})$ must contain the same data $\mathbf{D}$ for this addition to be valid across all models.

Appendix B

[58] Following *Samper and Neuman* [1989a], let $\mathbf{D} = (D_1, D_2, \ldots, D_N)^T$ be a vector of measurements at N points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$. A kriged estimate, $\hat{D}_i$, of D_i is given by

$$\hat{D}_i = \sum_{m \in N_i} \lambda_{im} D_m \tag{B1}$$

where N_i is the number of measurements included in the kriging neighborhood of $\mathbf{x}_i$ and λ_{im} are kriging coefficients. Assuming that the vector $\mathbf{e} = (e_1, e_2, \ldots, e_M)^T$ of M cross-validation errors

$$e_i = \hat{D}_i - D_i \tag{B2}$$

is Gaussian with zero mean and covariance matrix $\mathbf{C}$, the negative log likelihood of variogram parameters β given $\mathbf{D}$ is

$$NLL(\beta|\mathbf{D}) = -2 \ln p(\mathbf{D}|\beta) = M \ln 2\pi + \ln |\mathbf{C}| + \mathbf{e}^T \mathbf{C}^{-1} \mathbf{e}. \tag{B3}$$

In practice, it is convenient to replace $\mathbf{C}$ by a diagonal matrix with terms $C_{ij} = \delta_{ij} \sigma_i^2$ where δ_{ij} is the Kronecker delta and σ_i^2 the kriging variance, so that equation (B3) simplifies to

$$NLL(\beta|\mathbf{D}) = -2 \ln p(\mathbf{D}|\beta) = M \ln 2\pi + \sum_{i=1}^{M} \ln \sigma_i^2 + \sum_{i=1}^{M} \frac{e_i^2}{\sigma_i^2}. \tag{B4}$$

The corresponding kriging variance is given by

$$\sigma_i^2 = \sum_{m \in N_i} \lambda_{im} \gamma_{mi} - v_i \tag{B5}$$

in the case of ordinary kriging (drift-free models) and by

$$\sigma_i^2 = \sum_{m \in N_i} \lambda_{im} \gamma_{mi} - \sum_{k=0}^{p} v_k g_{ki} \tag{B6}$$

in the case of universal kriging with polynomial drift equation (15) where γ_{mi} is the variogram of D_m and D_i, v_k

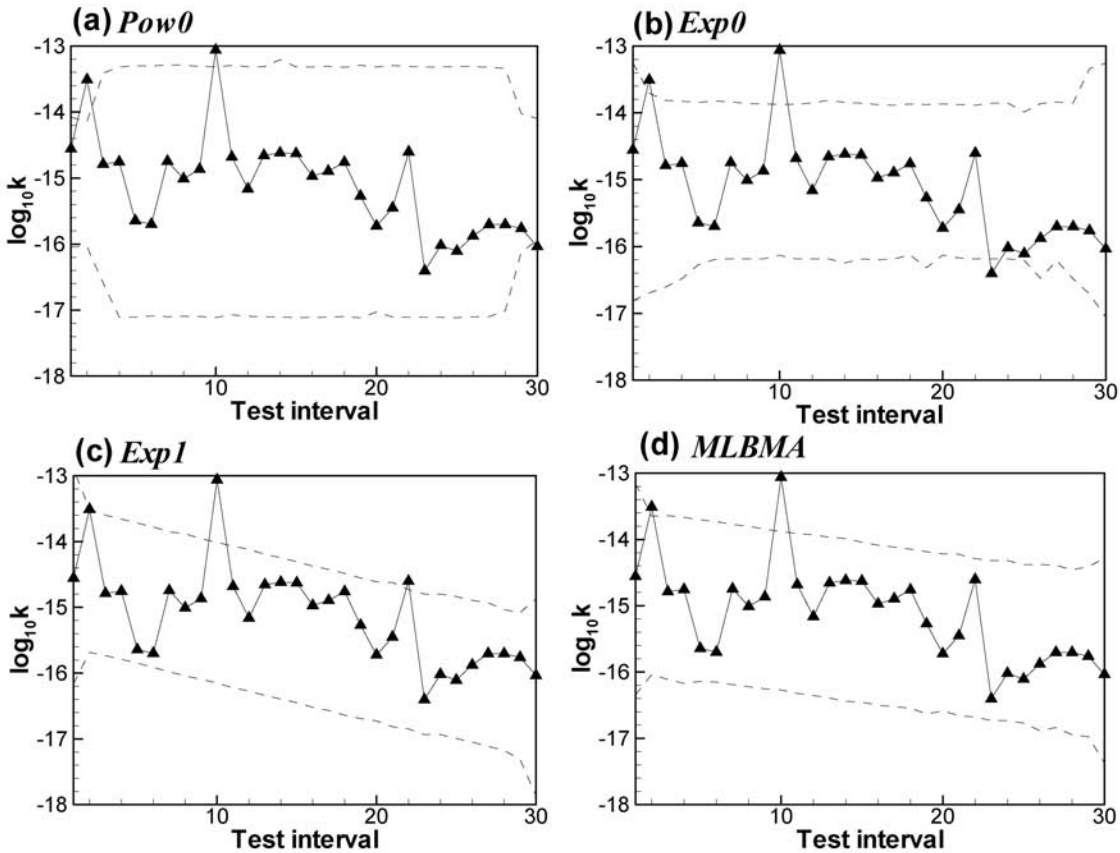

Figure 14. The 5% (bottom dashed line) and 95% (top dashed line) limits of simulated prediction interval of $\log_{10}k$ along borehole X2. Triangles designate measured values.

are Lagrange multipliers and $g_{ki} = g_k(\mathbf{x}_i)$. The kriging coefficients $\boldsymbol{\lambda}$ and Lagrange multipliers $\mathbf{v}$ are obtained by solving a linear system of algebraic universal kriging equations which require knowing the functional form of the drift function $\mathbf{G}$ but not its coefficients $\mathbf{a}$ [e.g., *Cressie*, 1991, p. 153; *Deutsch and Journel*, 1998, p. 67].

Appendix C

[59] Let $\boldsymbol{\beta} = (\sigma^2, \lambda)^T$ where σ^2 is the sill and λ the integral scale or range of a variogram. Given a separation distance s_{ij} between two points $\mathbf{x}_i$ and $\mathbf{x}_j$, $\partial \mathbf{C}_R/\partial\beta_i$ for a corresponding exponential covariance

$$C_R(h_{ij}) = \sigma^2 \exp\left(-\frac{h_{ij}}{\lambda}\right) \tag{C1}$$

is given by

$$\frac{\partial C_{R,ij}}{\partial \sigma^2} = \exp\left(-\frac{h_{ij}}{\lambda}\right)$$
$$\frac{\partial C_{R,ij}}{\partial \lambda} = \frac{\sigma^2 h_{ij}}{\lambda^2}\exp\left(-\frac{h_{ij}}{\lambda}\right) \tag{C2}$$

and for a spherical covariance

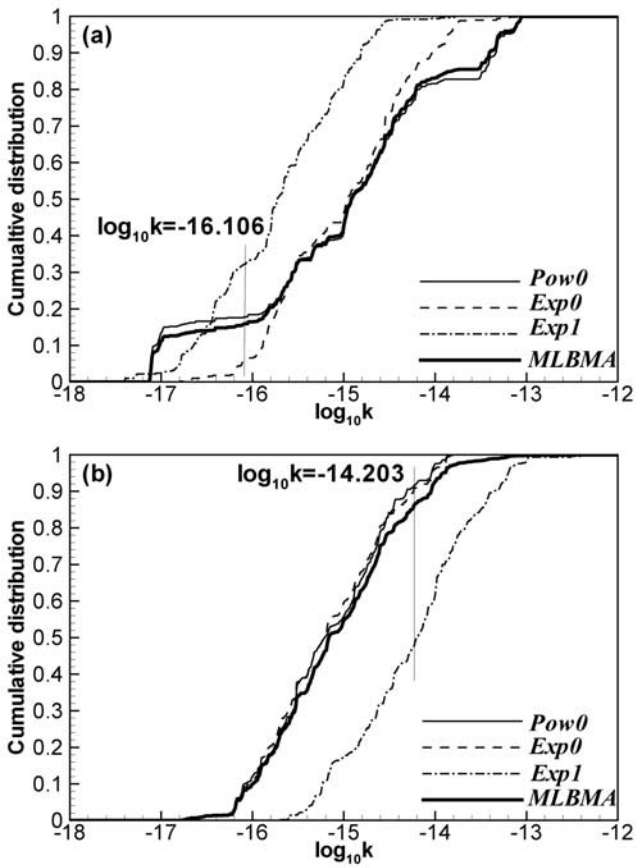

Figure 15. Cumulative distribution of simulated $\log_{10}k$ values at a measurement location in boreholes (a) V2 and (b) Y3. Vertical line indicates measured value.

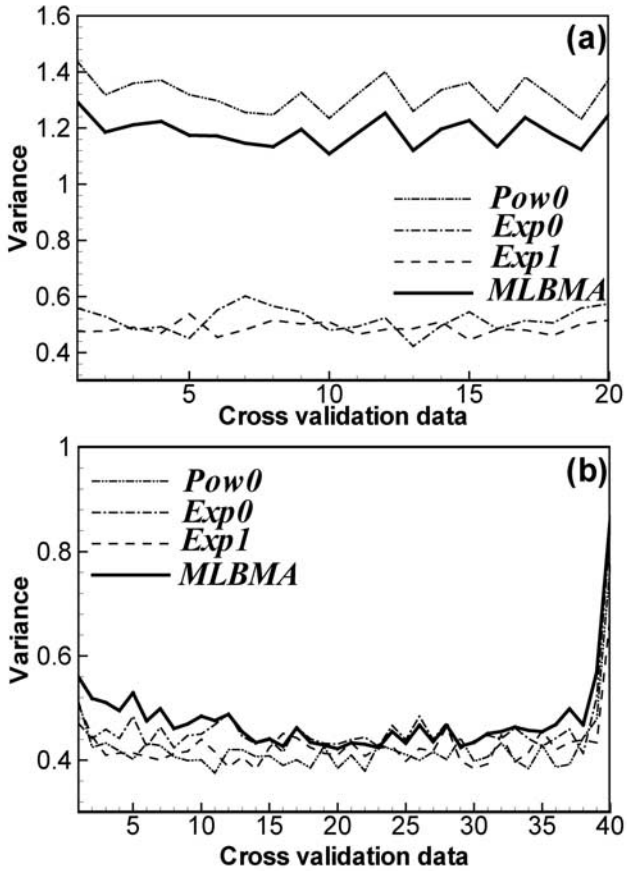

Figure 16. Sample variances of $\log_{10}k$ values simulated using various models and MLBMA along boreholes (a) V2 and (b) Y3 while eliminating the corresponding data.

$$C_R(h_{ij}) = \begin{cases} \sigma^2 - \sigma^2\left[\dfrac{h_{ij}}{\lambda} - 0.5\left(\dfrac{h_{ij}}{\lambda}\right)^3\right] & h_{ij} \le \lambda \\ 0 & h_{ij} > \lambda \end{cases} \tag{C3}$$

by

$$\frac{\partial C_{R,ij}}{\partial \sigma^2} = \begin{cases} 1 - 1.5\dfrac{h_{ij}}{\lambda} + 0.5\left(\dfrac{h_{ij}}{\lambda}\right)^3 & h_{ij} \le \lambda \\ 0 & h_{ij} > \lambda \end{cases}$$
$$\frac{\partial C_{R,ij}}{\partial \lambda} = \begin{cases} 1.5\dfrac{\sigma^2 h_{ij}}{\lambda^2}\left[1 - \left(\dfrac{h_{ij}}{\lambda}\right)^2\right] & h_{ij} \le \lambda \\ 0 & h_{ij} > \lambda \end{cases} \tag{C4}$$

By virtue of equations (14) and (15) the derivatives of residuals with respect to drift coefficients are

$$\frac{\partial R_i}{\partial a_k} = -g_{ki} \tag{C5}$$

where $g_{ki} = g_k(\mathbf{x}_i)$.

[60] **Acknowledgments.** This research was supported by the U.S. Nuclear Regulatory Commission, Office of Nuclear Regulatory Research, under contract JCN Y6465 with Pacific Northwest National Laboratory.

The development of MLBMA was supported by the same Office under contract NRC-04-95-038 with the University of Arizona.

References

Akaike, H. (1974), A new look at statistical model identification, *IEEE Trans. Autom. Control*, *AC-19*, 716–722.

Akaike, H. (1977), On entropy maximization principle, in *Applications of Statistics*, edited by P. R. Krishnaiah, pp. 27–41, North-Holland, New York.

Ando, K., A. Kostner, and S. P. Neuman (2003), Stochastic continuum modeling of flow and transport in a crystalline rock mass: Fanay-Augères, France, revisited, *Hydrogeol. J.*, *11*(5), 521–535.

Beven, K. J. (1993), Prophecy, reality and uncertainty in distributed hydrological modeling, *Adv. Water Resour.*, *16*, 41–51.

Beven, K. (2000), Uniqueness of place and non-uniqueness of models in assessing predictive uncertainty, in *Computational Methods in Water Resources XIII*, edited by L. R. Bentley et al., pp. 1085–1091, A. A. Balkema, Brookfield, Vt.

Beven, K. J., and A. M. Binley (1992), The future of distributed models: Model calibration and uncertainty prediction, *Hydrol. Processes*, *6*, 279–298.

Beven, K. J., and J. Freer (2001), Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, *J. Hydrol.*, *249*, 11–29.

Burnham, K. P., and A. R. Anderson (2002), *Model Selection and Multiple Model Inference: A Practical Information-Theoretical Approach*, 2nd ed., Springer-Verlag, New York.

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

Carrera, J., and S. P. Neuman (1986b), Estimation of aquifer parameters under transient and steady state conditions: 3. Application to synthetic and field data, *Water Resour. Res.*, *22*, 228–242.

Carrera, J., A. Medina, C. Axness, and T. Zimmerman (1997), Formulations and computational issues of the inversion of random fields, in *Subsurface Flow and Transport: A Stochastic Approach*, edited by G. Dagan and S. P. Neuman, pp. 62–79, Cambridge Univ. Press, New York.

Chen, G., W. A. Illman, D. L. Thompson, V. V. Vesselinov, and S. P. Neuman (2000), Geostatistical, type curve and inverse analyses of pneumatic injection tests in unsaturated fractured tuffs at the Apache Leap Research Site near Superior, Arizona, in *Dynamics of Fluids in Fractured Rock*, *Geophys. Monogr. Ser.*, vol. 122, edited by B. Faybishenko et al., pp. 73–98, AGU, Washington, D. C.

Christakos, G. (2000), *Modern Spatiotemporal Geostatistics*, Oxford Univ. Press, New York.

Christakos, G. (2002a), On the assimilation of uncertain physical knowledge bases: Bayesian and non-Bayesian techniques, *Adv. Water Resour.*, *25*(8–12), 1257–1274.

Christakos, G. (2002b), The role of conceptual frameworks in hydrologic research and development, in *Calibration and Reliability in Groundwater Modelling: A Few Steps Closer to Reality*, edited by K. Kovar and Z. Hrkal, *IAHS Publ.*, *19*, 277–285.

Christakos, G. (2003), Critical conceptualism in environmental modeling and prediction, *Environ. Sci. Technol.*, *37*(20), 4685–4693.

Christakos, G. (2004), *Conceptual Blending and Formal Tools for Multi-Disciplinary Systems in Uncertain Environments*, Springer-Verlag, New York, in press.

Cressie, N. (1991), *Statistics of Spatial Data*, John Wiley, Hoboken, N. J.

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

Doherty, J. (2002), PEST model—Independent parameter estimation, software manual, S. S. Papadopulos & Assoc., Bethesda, Md.

Draper, D. (1995), Assessment and propagation of model uncertainty, *J. R. Stat. Soc., Ser. B*, *57*(1), 45–97.

Draper, D. (1999), Comment to "Bayesian model averaging: A tutorial," *Stat. Sci.*, *14*(4), 405–409.

Gaganis, P., and L. Smith (2001), A Bayesian approach to the quantification of the effect of model error on the predictions of groundwater models, *Water Resour. Res.*, *37*, 2309–2322.

Gauch, H. G. Jr. (1993), Prediction, parsimony and noise, *Am. Sci.*, *81*, 468–478.

George, E. I. (1999), Comment, *Stat. Sci*, *14*(4), 409–412.

Good, I. J. (1952), Rational decisions, *J. R. Stat. Soc., Ser. B*, *57*(1), 107–114.

Guzman, A. G., S. P. Neuman, C. Lohrstorfer, and R. Bassett (1994), Chapter 4, in *Validation Studies for Assessing Unsaturated Flow and Transport Through Fractured Rock*, edited by R. L. Bassett et al., pp. 4-1–4-58, *Rep. NUREG/CR-6203*, U.S. Nucl. Regul. Comm., Washington, D. C.

Guzman, A. G., A. M. Geddis, M. J. Henrich, C. Lohrstorfer, and S. P. Neuman (1996), Summary of air permeability data from single-hole injection tests in unsaturated fractured tuffs at the Apache Leap Research Site: Results of steady-state test interpretation, *Rep. NUREG/CR-6360*, U.S. Nucl. Regul. Comm., Washington, D. C.

Hannan, E. S. (1980), The estimation of the order of ARMA process, *Ann. Stat.*, 1791–1801.

Hernandez, A. F., S. P. Neuman, A. Guadagnini, and J. Carrera-Ramirez (2002), Conditioning steady state mean stochastic flow equations on head and hydraulic conductivity measurements, in *Proceedings of 4th International Conference on Calibration and Reliability in Groundwater Modelling (ModelCARE 2002)*, edited by K. Kovar and Z. Hrkal, pp. 158–162, Charles Univ., Prague, Czech Republic.

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, *Stochastic Environ. Res. Risk Assess.*, *17*(5), 329–338, doi:10.1007/s00477-003-0154-4.

Hoeksema, R. J., and P. K. Kitanidis (1985), Analysis of the spatial structure of properties of selected aquifers, *Water Resour. Res.*, *21*, 563–572.

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

Hornberger, G. M., and R. C. Speer (1981), An approach to the preliminary analysis of environmental systems, *J. Environ. Manage.*, *12*, 7–18.

James, A. L., and C. M. Oldenburg (1997), Linear and Monte Carlo uncertainty analysis for subsurface contaminant transport simulation, *Water Resour. Res.*, *33*, 2495–2508.

Jefferys, W., and J. Berger (1992), Ockham's razor and Bayesian analysis, *Am. Sci.*, *80*, 64–72.

Kashyap, R. L. (1982), Optimal choice of AR and MA parts in autoregressive moving average models, *IEEE Trans. Pattern Anal. Mach. Intel.*, *4*(2), 99–104.

Kass, R. E., and A. E. Raftery (1995), Bayesian factor, *J. Am. Stat. Assoc.*, *90*, 773–795.

Kitanidis, P. K., and R. W. Lane (1985), Maximum likelihood parameter estimation of hydrologic spatial processes by the Gaussian-Newton method, *J. Hydrol.*, *79*, 53–71.

Madigan, D., and A. E. Raftery (1994), Model selection and accounting for model uncertainty in graphical models using Occam's window, *J. Am. Stat. Assoc.*, *89*, 1535–1546.

National Research Council (2001), *Conceptual Models of Flow and Transport in the Fractured Vadose Zone*, Natl. Acad. Press, Washington, D. C.

Neuman, S. P. (2002), Accounting for conceptual model uncertainty via maximum likelihood model averaging, in *Proceedings of 4th International Conference on Calibration and Reliability in Groundwater Modelling (ModelCARE 2002)*, edited by K. Kovar and Z. Hrkal, pp. 529–534, Charles Univ., Prague, Czech Republic.

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

Neuman, S. P., and E. A. Jacobson (1984), Analysis of nonintrinsic spatial variability by residual kriging with application to regional groundwater levels, *Math. Geol.*, *16*, 491–521.

Neuman, S. P., and P. J. Wierenga (2003), A comprehensive strategy of hydrogeologic modeling and uncertainty analysis for nuclear facilities and sites, *Rep. NUREG/CR-6805*, U.S. Nucl. Regul. Comm., Washington, D. C.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992), *Numerical Recipe in Fortran 77*, 2nd ed., Cambridge Univ. Press, New York.

Raftery, A. E. (1993), Bayesian model selection in structural equation models, in *Testing Structural Equation Models*, edited by K. Bollen and J. Long, pp. 163–180, Sage, Newbury Park, Calif.

Raftery, A. E., D. Madigan, and C. T. Volinsky (1996), Accounting for model uncertainty in survival analysis improves predictive performance, in *Bayesian Statistics*, edited by J. Bernardo et al., pp. 323–349, Oxford Univ. Press, New York.

Rissanen, J. (1978), Modeling by shortest data description, *Automatica*, *14*, 465–471.

Samper, J., and J. Molinero (2000), Predictive uncertainty of numerical models of groundwater flow and solute transport, *Eos. Trans. AGU*, *81*(48), Fall Meet. Suppl., Abstract H11G-04.

Samper, F. J., and S. P. Neuman (1989a), Estimation of spatial covariance structures by adjoint state maximum likelihood cross- validation: 1. Theory, *Water Resour. Res.*, *25*, 351–362.

Samper, F. J., and S. P. Neuman (1989b), Estimation of spatial covariance structures by adjoint state maximum likelihood cross-validation: 2. Synthetic experiments, *Water Resour. Res.*, *25*, 363–371.

Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Stat.*, *6*(2), 461–464.

Taplin, R. H. (1993), Robust likelihood calculation for time series, *J. R. Stat. Soc. Ser. B*, *55*, 829–836.

Vesselinov, V. V. (2000), Numerical inverse interpretation of pneumatic tests in unsaturated fractured tuffs at the Apache Leap Research Site, Ph. D. dissertation, Univ. of Ariz., Tucson.

Volinsky, C. T., D. Madigan, A. E. Raftery, and R. A. Kronmal (1997), Bayesian model averaging in proportional hazard models: Assessing the risk of a stroke, *J. R. Stat. Soc., Ser. C*, *46*, 433–448.

P. D. Meyer and M. Ye, Pacific Northwest National Laboratory, P.O. Box 999, Richland, WA 99352, USA. (philip.meyer@pnl.gov; ming.ye@pnl.gov)

S. P. Neuman, Department of Hydrology and Water Resources, University of Arizona, Tucson, AZ 85721, USA. (neuman@hwr.arizona.edu)