Sensitivity analysis and assessment of prior model probabilities in MLBMA with application to unsaturated fractured tuff

Ming Ye,[1] Shlomo P. Neuman,[2] Philip D. Meyer,[3] and Karl Pohlmann[1]

Received 16 May 2005; revised 3 September 2005; accepted 3 October 2005; published 24 December 2005.

[1] Previous application of maximum likelihood Bayesian model averaging (MLBMA, Neuman (2002, 2003)) to alternative variogram models of log air permeability data in fractured tuff has demonstrated its effectiveness in quantifying conceptual model uncertainty and enhancing predictive capability (Ye et al., 2004). A question remained how best to ascribe prior probabilities to competing models. In this paper we examine the extent to which lead statistics of posterior log permeability predictions are sensitive to prior probabilities of seven corresponding variogram models. We then explore the feasibility of quantifying prior model probabilities by (1) maximizing Shannon's entropy H (Shannon, 1948) subject to constraints reflecting a single analyst's (or a group of analysts') prior perception about how plausible each alternative model (or a group of models) is relative to others, and (2) selecting a posteriori the most likely among such maxima corresponding to alternative prior perceptions of various analysts or groups of analysts. Another way to select among alternative prior model probability sets, which, however, is not guaranteed to yield optimum predictive performance (though it did so in our example) and would therefore not be our preferred option, is a minimum-maximum approach according to which one selects a priori the set corresponding to the smallest value of maximum entropy. Whereas maximizing H subject to the prior perception of a single analyst (or group) maximizes the potential for further information gain through conditioning, selecting the smallest among such maxima gives preference to the most informed prior perception among those of several analysts (or groups). We use the same variogram models and log permeability data as Ye et al. (2004) to demonstrate that our proposed approach yields the least amount of posterior entropy (residual uncertainty after conditioning) and enhances predictive model performance as compared to (1) the noninformative neutral case in which all prior model probabilities are set equal to each other and (2) an informed case that nevertheless violates the principle of parsimony.

Citation: Ye, M., S. P. Neuman, P. D. Meyer, and K. Pohlmann (2005), Sensitivity analysis and assessment of prior model probabilities in MLBMA with application to unsaturated fractured tuff, *Water Resour. Res.*, *41*, W12429, doi:10.1029/2005WR004260.

1. Introduction

[2] Conceptualization is the foundation of hydrologic modeling. As hydrologic systems are open and complex, they are open to multiple interpretations and mathematical descriptions (postulation of alternative conceptual-mathematical models). Hydrologic analyses based on a single conceptual-mathematical model 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) [*Neuman and Wierenga*, 2003]. Ignoring conceptual (structural) model uncertainty and focusing solely on the optimization of model parameters may lead to overconfidence in model predictive capabilities [*National Research Council*, 2001]. *Neuman and Wierenga* [2003] proposed rendering optimum hydrologic predictions by means of several competing deterministic or stochastic models and assessing their joint predictive uncertainty using maximum likelihood Bayesian model averaging (MLBMA) [*Neuman*, 2002, 2003]. The latter is an approximate version of Bayesian model averaging (BMA) [e.g., *Draper*, 1995; *Hoeting et al.*, 1999; J. A. Hoeting, Methodology for Bayesian model averaging: An update, unpublished paper, 2002, available at http://www.stat.colostate.edu/~jah/papers/ibcbma.pdf], which offers two advantages over BMA: It avoids the need for exhaustive Monte Carlo simulations and obviates the need for (though it can incorporate) prior information about model parameters, which is often difficult to obtain.

[3] *Ye et al.* [2004] expanded upon the theoretical framework of MLBMA and applied it 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 in central Arizona. To obtain maximum

[1]Desert Research Institute, Las Vegas, Nevada, USA.
[2]Department of Hydrology and Water Resources, University of Arizona, Tucson, Arizona, USA.
[3]Pacific Northwest National Laboratory, Richland, Washington, USA.

0043-1397/05/2005WR004260$09.00

likelihood (ML) estimates of variogram and drift parameters, they used adjoint state maximum likelihood cross validation [*Samper and Neuman*, 1989a] in conjunction with universal Kriging and generalized least squares. Standard information criteria had been found to provide an ambiguous ranking of the models, which did not justify selecting one of them and discarding all others (as is commonly done in hydrologic research and practice). Instead, the authors eliminated some of the models based on their negligibly small posterior probabilities and used the remaining models to project the measured log permeabilities by kriging onto a rock volume containing the six boreholes. *Ye et al.* [2004] then averaged these four projections, and associated kriging variances, using the posterior probability of each model as weight. Finally, they cross validated the results by ignoring 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. The authors found MLBMA to contain more information (have a smaller log score) and exhibit better predictive performance (show wider predictive coverage) than any individual model among those considered.

[4] In both BMA and MLBMA one postulates a set **M** of K alternative models, M_k, the prior probabilities of which sum up to one, $\sum_{k=1}^{K} p(M_k) = 1$. Theoretically, this implies that all possible models of relevance are included in **M** and that all models in **M** differ from each other sufficiently to be considered mutually exclusive (the joint probability of two or more models being zero). *Ye et al.* [2004] interpreted the prior model probabilities, $p(M_k)$, to be subjective values reflecting the analyst's (or a group of analysts') perception about how plausible each alternative model (or a group of models) is relative to other models based on their apparent (qualitative, a priori) consistency with available knowledge and data. The analyst's perception, degree of reasonable belief [*Jeffreys*, 1957], or confidence [*Zio and Apostolakis*, 1996] in a model are ideally based on expert judgment, which *Bredehoeft* [2005] considers to be the basis of conceptual model development. Hence we view integrating expert judgment in BMA and MLBMA (by specifying subjective prior probabilities) to be strength rather than a weakness. According to this view, the models included in **M** must be those (and only those) that experts consider to be of potential relevance to the problem at hand.

[5] There is considerable ambiguity about what renders a set of models mutually exclusive (for example, it is not entirely clear to us that an exponential and a spherical variogram model, both without a drift, necessarily have less in common than do two exponential models, one with and one without a drift, as some might intuit). We take the attitude that in standard hydrologic practice, one typically selects a single model at the exclusion of all other models. Accordingly, we consider each model in the set **M** to be a potential candidate for exclusive selection at the expense of all other candidates, in the sense that setting its prior probability equal to 1 would require setting the prior probabilities of all other candidates equal to zero (in order to insure that the probabilities sum up to 1). This means that the joint probability of any two models is zero, i.e., they are mutually exclusive in the above sense.

[6] Statisticians have been concerned with the fact that the number of potentially feasible models may be exceedingly large, rendering their exhaustive inclusion in **M** infeasible [*Hoeting et al.*, 1999]. Some have suggested that a practical way to eliminate the difficulty is to adopt the idea of Ockham's window [*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]. For example, *Poeter and Anderson* [2005] eliminated from consideration a priori models deemed by them to have unreasonable zonation patterns of hydraulic conductivity and models that had failed to converge; *Ye et al.* [2004] discarded a posteriori models having very low posterior probabilities even though they had been assigned relatively large prior probabilities. 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.

[7] The question of how to assign prior probabilities $p(M_k)$ to models M_k in a given set **M** has received little attention in the statistical literature and remains largely open [*Kass and Wasserman*, 1996; *Clyde*, 1999; *Hoeting et al.*, 1999]. A common practice is to adopt a "reasonable 'neutral' choice" [*Hoeting et al.*, 1999] according to which all models are initially equally likely, there being insufficient prior reason to prefer one over another. *Draper* [1999] and *George* [1999] have expressed concern that if two models are near equivalent as regards predictions, treating them as separate, equally likely models amounts to assigning double weight to a single model of which there are two slightly different versions, thereby "diluting" the predictive power of BMA (and, we add, MLBMA). One way to minimize this effect is to eliminate at the outset models that are deemed potentially inferior; *Ye et al.* [2004] did so post facto by deleting models whose posterior probability turned out to be negligibly small in comparison to that of other models. Another way 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, an idea that *Ye et al.* [2004] explored through an example. The neutral choice (with or without dilution) ignores prior knowledge of the system to be modeled and thus reflects maximum ignorance [*Jaynes*, 2003].

[8] 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. Just like their prior counterparts, posterior probabilities are valid only in a comparative, not in an absolute, sense. They are conditional on the choice of models (as well as the available data) and may be sensitive to the choice of prior model probabilities as demonstrated by *Ye et al.* [2004]. This sensitivity is expected to diminish with increased level of conditioning on data. Whereas there have been numerous sensitivity analyses of Bayesian parameter estimators (see *Insua et al.* [2000] for a summary), we are not aware of such studies in the context of prior model probabilities. As sensitivity analyses must be site specific, one purpose of our paper is to address the question, How sensitive are lead

statistics of posterior log permeability predictions to prior probabilities associated with alternative variogram models at the Apache Leap Research Site?

[9] We noted earlier that the neutral choice of prior model probabilities ignores expert knowledge of the system to be modeled, thereby reflecting maximum ignorance on the part of the analyst. It has been demonstrated by *Madigan et al.* [1995] and *Zio and Apostolakis* [1996] that assigning variable prior probabilities to alternative models on the basis of expert judgment may improve the correspondence between model predictions and measurements. This raises the question of whether it might be possible, and feasible, to embed such expert knowledge formally in MLBMA. We address this question in the present paper by exploring the feasibility of quantifying prior model probabilities through maximization of Shannon's entropy [*Shannon*, 1948] subject to constraints reflecting a single analyst's (or group of analysts') prior perception about how plausible each alternative model (or a group of models) is relative to others, and selection of the most likely (and possibly also the smallest) among such maxima corresponding to alternative perceptions of various analysts (or groups of analysts). We also discuss another way of selecting among alternative prior model probability sets, which, however, is not guaranteed to yield optimum predictive performance (though it did so in our example) and would therefore not be our preferred option: a minimum-maximum (min-max) approach according to which one selects a priori the set corresponding to the smallest value of maximum entropy. Whereas maximizing H subject to the prior perception of a single analyst maximizes the potential for further information gain through conditioning (as explained later), selecting the smallest among such maxima gives preference to the most informed prior perception among those of several analysts (or groups).

[10] Following a brief introduction of MLBMA in section 2, we examine in section 3 the sensitivity of MLBMA posterior predictions (model probability, mean, variance, and risk) to prior model probabilities through a case example concerning alternative variogram models of log air permeability in unsaturated fractured tuff. We then propose in section 4 a constrained maximum entropy approach to the assessment of prior model probabilities, which differs in both purpose and detail from the way this concept has been applied to the estimation of parameter probabilities [*Woodbury and Ulrych*, 1998; *Jaynes*, 2003]. We close by applying our proposed approach to the above case example in section 5, demonstrating that it yields better predictive performance than (1) the noninformative neutral case in which all prior model probabilities are set equal to each other or (2) an informed case which however violates Ockham's razor (the principle of parsimony). Our conclusions are summarized in section 6.

2. Maximum Likelihood Bayesian Model Averaging (MLBMA)

[11] To render our paper complete and self-contained, we start with a brief description of MLBMA; for additional details the reader is referred to *Neuman* [2003] and *Ye et al.* [2004]. If Δ is a quantity one wants to predict given a set $\mathbf{M}$ of K alternative models, then its posterior distribution, given a discrete set $\mathbf{D}$ of site data, is provided according to BMA by

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

where $p(\Delta|M_k, \mathbf{D})$ is the posterior distribution of Δ under model M_k and $p(M_k|\mathbf{D})$ is the posterior probability of M_k. The corresponding posterior mean and variance are

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

$$Var[\Delta|\mathbf{D}] = \sum_{k=1}^{K} Var[\Delta|\mathbf{D}, M_k] p(M_k|\mathbf{D}) + \sum_{k=1}^{K} (E[\Delta|\mathbf{D}, M_k] - E[\Delta|\mathbf{D}])^2 p(M_k|\mathbf{D}). \tag{3}$$

The weights $p(M_k|\mathbf{D})$ are given by Bayes' rule, which, in MLBMA, is approximated as [*Ye et al.*, 2004]

$$p(M_k|\mathbf{D}) = \frac{p(\mathbf{D}|M_k)p(M_k)}{\sum_{l=1}^{K} p(\mathbf{D}|M_l)p(M_l)} \approx \frac{\exp\left(-\frac{1}{2}KIC_k\right)p(M_k)}{\sum_{l=1}^{K} \exp\left(-\frac{1}{2}KIC_l\right)p(M_l)}, \tag{4}$$

where $p(\mathbf{D}|M_k) \approx \exp(-KIC_k/2)$ is the likelihood of model M_k (a measure of its consistency with data) and KIC_k is the corresponding Kashyap information criterion [*Kashyap*, 1982]. In the absence of suitable conditioning data $\mathbf{D}$, dropping the latter from all terms in (1)–(3) results in an unconditional model averaging process [*Apostolakis*, 1990].

[12] KIC_k is given by $KIC_k = NLL_k + N_k \ln(N) - N_k \ln(2\pi) + \ln|\boldsymbol{F}_k(\mathbf{D}|\hat{\boldsymbol{\theta}}_k, M_k)|$, where N_k is the number of parameters associated with model M_k (dimension of the parameter vector $\boldsymbol{\theta}_k$), N is the number of measurements used to compute the likelihood of M_k (dimension of the data vector $\mathbf{D}$), $\hat{\boldsymbol{\theta}}_k$ is the vector of maximum likelihood (ML) parameter estimates (obtained through ML calibration of the model against the data $\mathbf{D}$), NLL_k is the corresponding negative log likelihood of M_k (closely related to a generalized least squares model calibration criterion), and F_k is a Fisher information matrix (obtained as a by-product of ML calibration) [*Kashyap*, 1982]. Detailed mathematical expressions for NLL_k and F_k are given by *Carrera and Neuman* [1986a], *Neuman* [2003], and *Ye et al.* [2004]. The first two terms on the right-hand side of KIC_k constitute the Bayesian information criterion (BIC_k), which dominates the expression asymptotically as N becomes large. The last two terms render KIC_k applicable to the nonasymptotic case where N is finite (statements to the contrary by *Poeter and Anderson* [2005, pp. 601, 604] notwithstanding), which is critical for hydrologic (and in particular hydrogeologic) models that are often based on sparse data.

[13] In fact, 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

Table 1. Three Prior and Posterior Probability Sets Corresponding to Seven Variogram Models of Log Permeability at the Apache Leap Research Site

		*Pow*0	*Exp*0	*Exp*1	*Exp*2	*Sph*0	*Sph*1	*Sph*2
KIC_k		369.6	370.1	369.5	416.7	390.5	378.1	424.6
Prior	$p(M_k)$, %	14.29	14.29	14.29	14.29	14.29	14.29	14.29
set 1	$p(M_k\|\mathbf{D})$, %	35.30	26.58	37.61	0	0	0.51	0
Prior	$p(M_k)$, %	5.00	5.00	5.00	5.00	5.00	5.00	70.0
set 2	$p(M_k\|\mathbf{D})$, %	35.30	26.58	37.61	0	0	0.51	0
Prior	$p(M_k)$, %	14.75	5.00	5.00	14.75	5.00	8.25	47.25
set 3	$p(M_k\|\mathbf{D})$, %	61.56	15.72	22.23	0	0	0.49	0

with fewer parameters is assigned a higher posterior probability. As illustrated by *Carrera and Neuman* [1986b], KIC_k recognizes that when the database 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. KIC_k may thus 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. 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. Stated otherwise, the last term gauges the information content of the available data, associating higher and higher posterior probabilities with more and more complex models as the database improves in quantity (N) and quality.

[14] *Carrera and Neuman* [1986b], *Samper and Neuman* [1989b], and *Ye et al.* [2004] have shown through synthetic and real data that *KIC* is superior to *BIC* or *AIC*, the *Akaike* [1974] information criterion, in model selection and multimodel inference. *Burnham and Anderson* [2002, p. 295] make clear (statements to the contrary by *Poeter and Anderson* [2005, pp. 597, 601, 604] notwithstanding) that neither *BIC* nor *KIC* requires a "true" model (which seldom exists in hydrology) to be included in the set **M** of models being considered, a point we have already made earlier.

[15] *Ye et al.* [2004] applied MLBMA to seven alternative geostatistical models of log permeability variations in unsaturated fractured tuff at the Apache Leap Research Site in central Arizona. Their variogram models, listed in Table 1, included (1) power (*POW*0), (2) exponential without a drift (*EXP*0), (3) exponential with a linear drift (*EXP*1), (4) exponential with a quadratic drift (*EXP*2), (5) spherical without a drift (*SPH*0), (6) spherical with a linear drift (*SPH*1), and (7) spherical with a quadratic drift (*SPH*2). Table 1 also lists the KIC_k value obtained by *Ye et al.* [2004] for each model and, under prior set 1, the neutral prior model probabilities $p(M_k)$ considered by the authors. Our rational for including prior sets 2 and 3 in Table 1 will become clear below.

3. Sensitivity of MLBMA to Prior Probabilities of Log Permeability Variogram Models in Unsaturated Fractured Tuff

[16] For purposes of sensitivity analysis we adopt the variogram models as well as corresponding values of the information criterion KIC_k, likelihood $p(\mathbf{D}|M_k)$, mean $E[\Delta|\mathbf{D}, M_k]$, and variance $Var[\Delta|\mathbf{D}, M_k]$ from *Ye et al.* [2004]. We note that all of these quantities are independent of $p(M_k)$. As these variogram models are nonlinear, it would be difficult to predict or explain the sensitivities of associated statistics to prior model probabilities theoretically. Instead, we obtain discrete prior model probability sets that satisfy $\sum_{k=1}^{K} p(M_k) = 1$ by disregarding probabilities smaller than 5%, thereby limiting the maximum probability to $100\% - 6 \times 5\% = 70\%$. We then subdivide the probability space [5%, 70%] of each model into 20 equal intervals of 3.25% and associate with each a uniform probability equal to the lower limit of the interval, yielding a total of 168,322 nonzero prior probability sets. We compute the sensitivity of various model statistics to each of these discrete sets numerically.

3.1. Sensitivity of Posterior Model Probability

[17] Posterior probabilities $p(M_k|\mathbf{D})$ of models *EXP*2, *SPH*0, and *SPH*2 are zero for all the 168,322 prior probability combinations listed in Table 1, including the neutral combination (with equal prior probabilities) labeled Prior Set 1 and two combinations labeled Prior Sets 2 and 3. This lack of sensitivity to prior probabilities is due to the relatively small likelihoods of *EXP*2, *SPH*0, and *SPH*2 in light of the available data, which renders the corresponding KIC_k values relatively large regardless of the analyst's prior perceptions about the relative merits of the seven models. Because of their zero posterior probabilities, models *EXP*2, *SPH*0, and *SPH*2 contribute nothing to posterior mean and variance in (2)–(3). Table 1 suggests that posterior probabilities of the other four models are sensitive to the choice of prior model probabilities. For example, the posterior probability of model *POW*0 corresponding to Prior Set 3 (61.56%) is about twice as large as that corresponding to Prior Set 1 (35.30%) even though the prior probabilities of *POW*0 in these two prior sets are almost the same (14.75% and 14.29%, respectively). It is interesting to note that posterior probabilities associated with these four models are identical for Prior Sets 1 and 2. This is so because in both cases the four models have equal prior probabilities, which cancel out in (4) and thus yield $p(M_k|\mathbf{D}) \approx \exp(-KIC_k/2)/\sum_{l=1}^{K}\exp(-KIC_l/2)$. Consequently, the posterior probabilities depend solely on KIC_k, which are independent of priors.

[18] Figure 1 plots posterior versus prior probabilities for *POW*0 (Figure 1a), *EXP*0 (Figure 1b), *EXP*1 (Figure 1c), and *SPH*1 (Figure 1d). Each column in Figure 1 represents posterior probabilities corresponding to all combinations of discrete prior model probabilities one can generate for a given (on the horizontal axis) prior probability of a particular model. Posterior probabilities of *POW*0, *EXP*0, and *EXP*1, which are associated with very similar KIC_k values, are seen to be much more sensitive to prior probabilities than is *SPH*1 whose KIC_k is distinctly larger. Indeed (4) indicates that for a given set of prior probabilities, posterior probability tends to diminish as KIC_k increases. Diamonds representing the neutral choice lie at or near midcolumn close to the lower end of each probability spectrum.

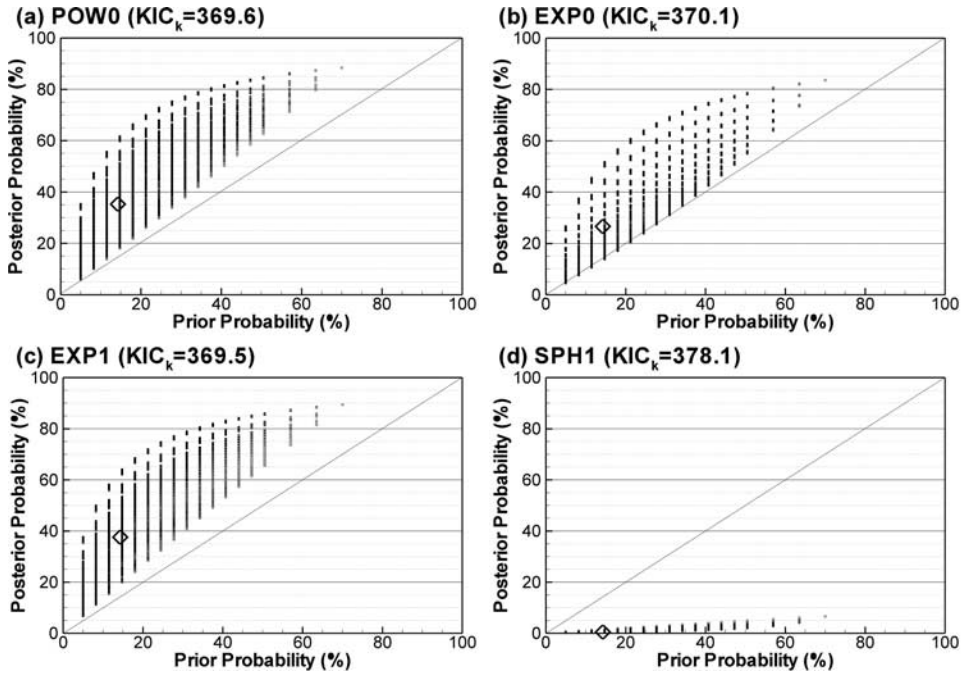

Figure 1. Posterior model probability of models (a) *POW*0, (b) *EXP*0, (c) *EXP*1, and (d) *SPH*1 for various discrete combinations of prior model probabilities. Diamonds represent neutral choice of equal prior probabilities.

[19] Figure 2 depicts contours of posterior probability *Post*(*POW*0) of model *POW*0 in the space of prior model probabilities *Pr*(*EXP*0) and *Pr*(*EXP*1) when *Pr*(*POW*0) = 14.75%, close to the neutral case in which all priors are equal to 14.29%. The contour corresponding to the column *Pr*(*POW*0) = 14.75% in Figure 1a shows how the posterior probability of model *POW*0 varies with the prior probabilities of models *EXP*0 and *EXP*1; the variation of *Pr*(*SPH*1) is not shown due to its small effect on *Post*(*POW*0). As expected, when *Pr*(*EXP*0) and *Pr*(*EXP*1) increase, *Post*(*POW*0) decreases at a rate (a measure of sensitivity) that diminishes with *Pr*(*EXP*0) and *Pr*(*EXP*1). More important, *Post*(*POW*0) is seen to be more sensitive to *Pr*(*EXP*1) than to *Pr*(*EXP*0) because the former has a smaller *KIC* value (as pointed out earlier). For example, a decrease in *Post*(*POW*0) from 0.46% to 0.42% corresponds to an increase in *Pr*(*EXP*0) of 0.04% but a smaller change in *Pr*(*EXP*1).

3.2. Sensitivity of Posterior Mean

[20] Like *Ye et al.* [2004], we use each of the four variogram models *POW*0, *EXP*0, *EXP*1, and *SPH*1 to 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 × 40 × 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 of their Figure 1. If one thinks of Δ as a random value of log permeability in a given grid block, then the kriging estimates represent ML approximations of the posterior mean $E[\Delta|M_k, \mathbf{D}]$, and the kriging variances stand for posterior variance $Var[\Delta|M_k, \mathbf{D}]$, associated with variogram model M_k.

[21] Figure 3 shows how the maximum, minimum, and grid-averaged MLBMA posterior mean of log air permeability, obtained from (2) using the posterior probabilities in Figure 1 as weights, depend on each of the 168,322 prior probability combinations associated with variogram models *POW*0 (Figure 3a), *EXP*0 (Figure 3b), *EXP*1 (Figure 3d), and *SPH*1 (Figure 3d); diamonds represent the neutral choice of equal prior probabilities. Each column in Figure 3 represents MLBMA posterior mean values corresponding to

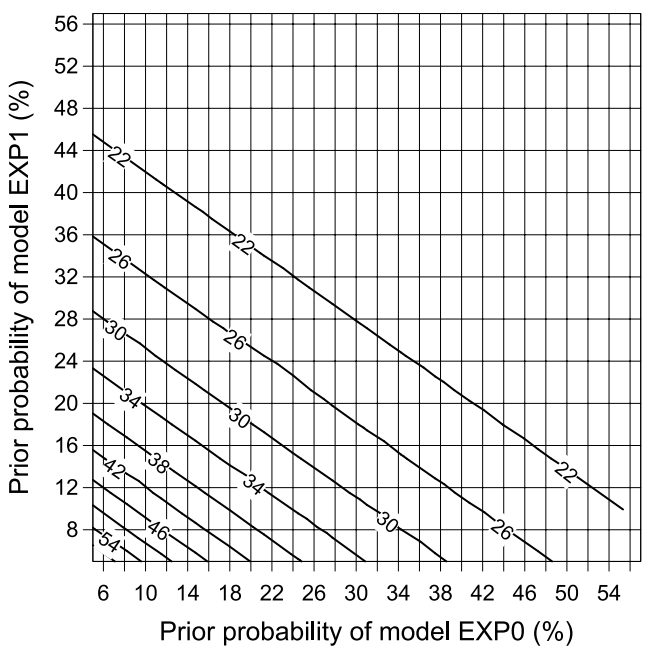

Figure 2. Contours of posterior probability *Post*(*POW*0) of model *POW*0 in the space of prior model probabilities *Pr*(*EXP*0) and *Pr*(*EXP*1) when *Pr*(*POW*0) = 14.75%, close to the neutral case in which all priors are equal to 14.29%.

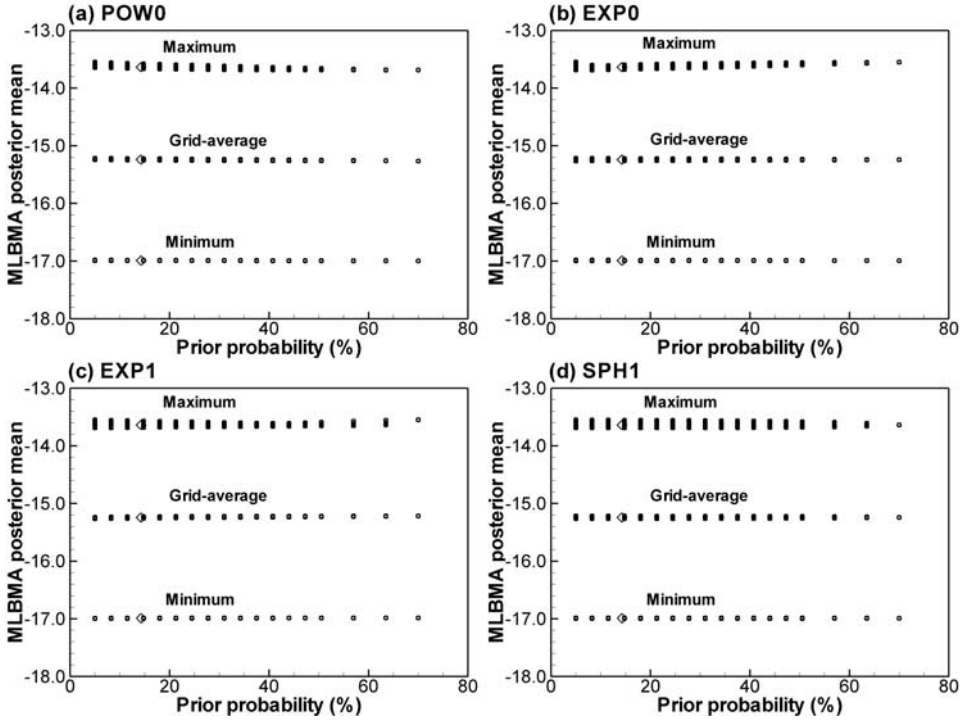

Figure 3. Maximum, minimum, and grid-averaged MLBMA posterior mean of log air permeability for various discrete combinations of prior probabilities corresponding to variogram models (a) *POW*0, (b) *EXP*0, (c) *EXP*1, and (d) *SPH*1. Diamonds represent neutral choice of equal prior probabilities.

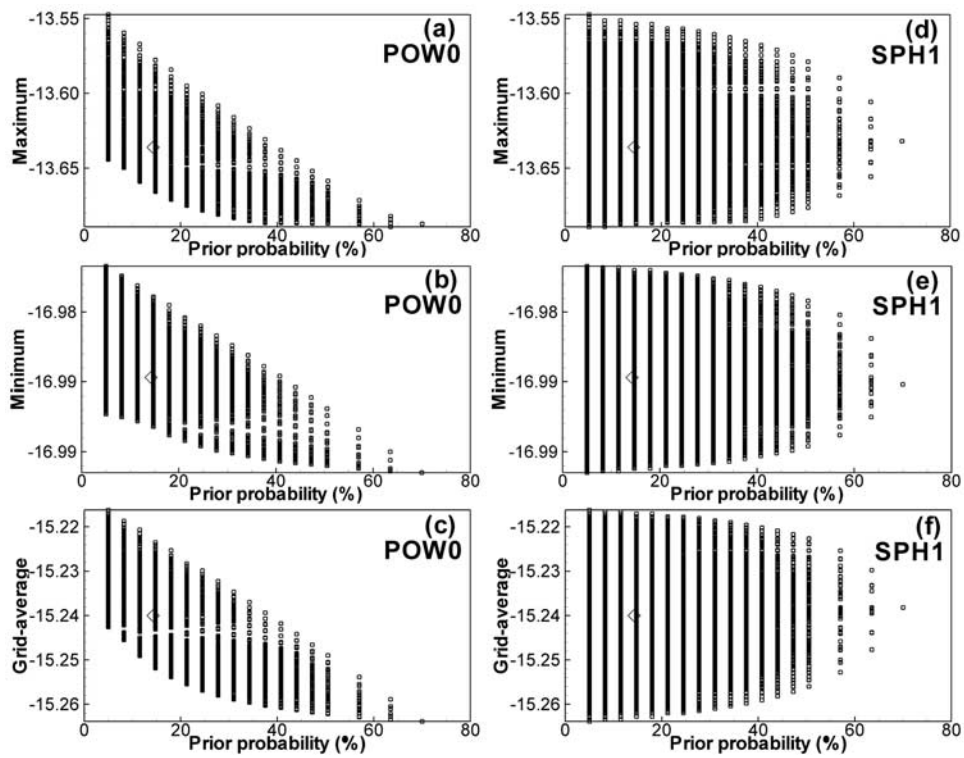

Figure 4. Expanded view of Figures 3a and 3d: (a and d) maximum, (b and e) minimum, and (c and f) grid-averaged MLBMA posterior mean of log air permeability for various discrete combinations of prior probabilities corresponding to variogram model *POW*0 and *SPH*1. Diamonds represent neutral choice of equal prior probabilities.

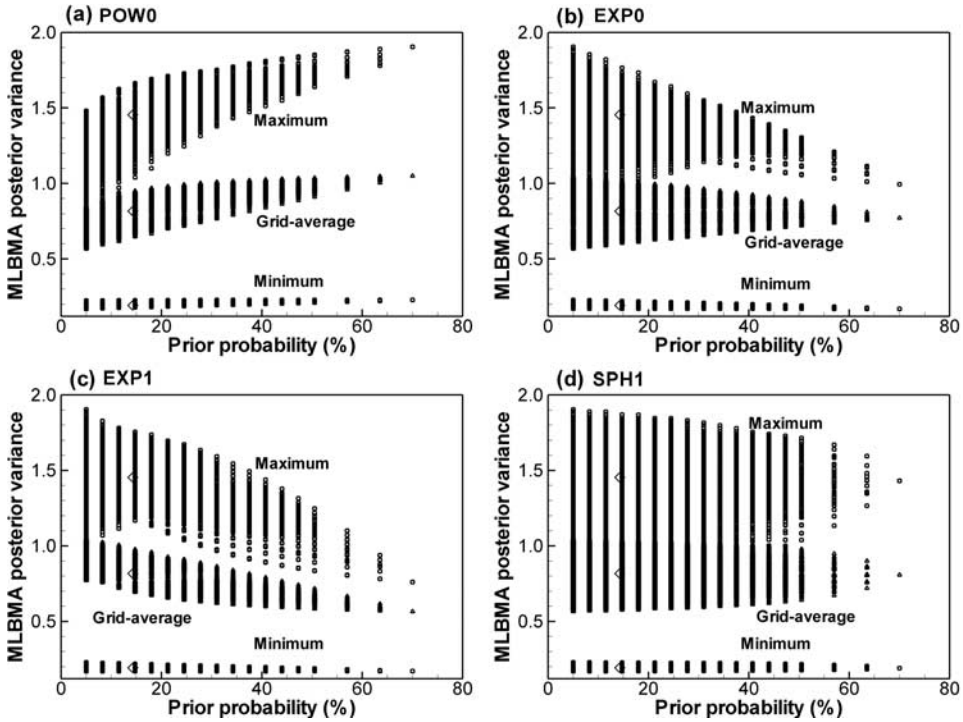

Figure 5. Maximum, minimum, and grid-averaged MLBMA posterior variance of log air permeability for various discrete combinations of prior probabilities corresponding to variogram models (a) *POW*0, (b) *EXP*0, (c) *EXP*1, and (d) *SPH*1. Diamonds represent neutral choice of equal prior probabilities.

all combinations of discrete prior model probabilities one can generate for a given (on the horizontal axis) prior probability of a particular model. Figures 4a–4c depict these columns on an expanded scale for discrete combinations of prior probabilities corresponding to variogram model *POW*0, and Figures 4d–4f do the same for model *SPH*1. On the scale of Figure 3 the statistics of MLBMA posterior mean exhibit very little sensitivity to prior model probabilities, indicating robustness of the MLBMA predictor with respect to these priors. On the expanded scales of Figure 4 one detects different patterns of variability for the two models. Whereas the statistics of *POW*0 diminish monotonically with increasing prior probability of this model down to lower values than those obtained with the neutral choice, the statistics of *SPH*1 converge toward central values that are very close to those obtained with the neutral choice.

3.3. Sensitivity of Posterior Variance

[22] Figure 5 shows how the maximum, minimum, and grid-averaged MLBMA posterior variance of log air permeability, obtained from (3) using the posterior probabilities in Figure 1 as weights, depend on each of the prior probability combinations associated with variogram models *POW*0 (Figure 5a), *EXP*0 (Figure 5b), *EXP*1 (Figure 5c), and *SPH*1 (Figure 5d), diamonds representing the neutral choice. Each column in Figure 5 represents MLBMA posterior variance values corresponding to all combinations of discrete prior model probabilities one can generate for a given (on the horizontal axis) prior probability of a particular model. The maximum MLBMA posterior variance is seen to be most sensitive to prior model probabilities and the minimum is least sensitive. The patterns of variability are seen to be different for different models. For example, the maximum and mean MLBMA posterior variances increase with *POW*0 prior probability but decrease with *EXP*1 prior probability. Overlaps between maximum and grid-averaged MLBMA posterior variances imply that the maximum corresponding to some prior sets may be smaller than the grid-average corresponding to other prior sets. MLBMA posterior variances are seen to be much more sensitive to prior model probabilities than are MLBMA posterior means, due to the quadratic term $(E[\Delta|\mathbf{D}, M_k] - E[\Delta|\mathbf{D}])^2$ in (3). The effect of this term is depicted in Figure 6, which shows contours of maximum MLBMA posterior variance in the space of prior model probabilities $Pr(EXP0)$ and $Pr(EXP1)$ when $Pr(POW0) = 14.75\%$, close to the neutral case in which all priors are equal to 14.29%. The contour corresponding to the column $Pr(POW0) = 14.75\%$ in Figure 5a shows how the maximum posterior variance varies with the prior probabilities of models *EXP*0 and *EXP*1; the variation of $Pr(SPH1)$ is not shown due to its small effect on posterior variance. As expected, when $Pr(POW0)$ is fixed, increasing the prior probabilities of model *EXP*0 or *EXP*1 brings about a decrease in the corresponding posterior variance, i.e., an increase in model uncertainty. The rate at which the maximum variance diminishes with $Pr(EXP0)$ and $Pr(EXP1)$, a measure of sensitivity to these priors, diminishes as they grow. Posterior variance is more sensitive to $Pr(EXP1)$ than to $Pr(EXP1)$ because the latter has a smaller *KIC* value.

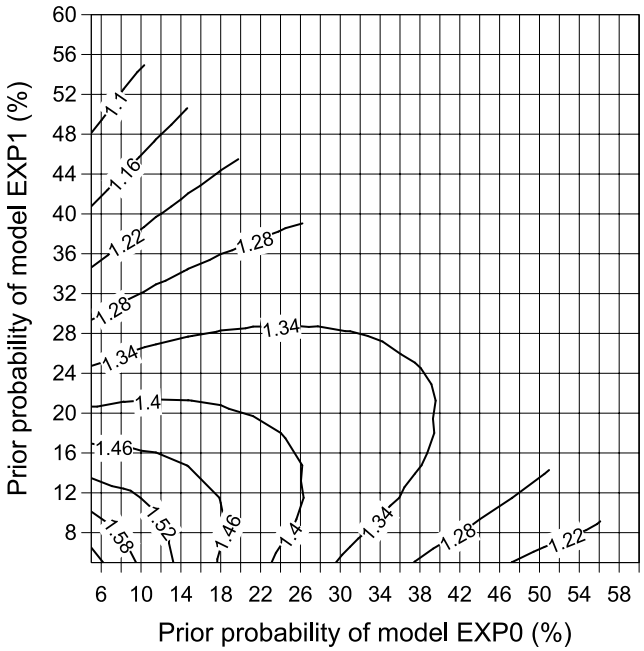

Figure 6. Contours of maximum MLBMA posterior variance in the space of prior model probabilities $Pr(EXP0)$ and $Pr(EXP1)$ when $Pr(POW0) = 14.75\%$, close to the neutral case in which all priors are equal to 14.29%.

3.4. Sensitivity of Posterior Risk

[23] In a manner similar to that of *Ruggeri and Sivaganesan* [2000] we define the MLBMA posterior risk, $R(\Delta|\mathbf{D})$, as

$$R(\Delta|\mathbf{D}) = \left[E_p(\Delta|\mathbf{D}) - E_0(\Delta|\mathbf{D})\right]^2 + Var_p(\Delta|\mathbf{D}) = \left[E_p(\Delta|\mathbf{D}) - E_0(\Delta|\mathbf{D})\right]^2 + \left[Var_p(\Delta|\mathbf{D}) - Var_0(\Delta|\mathbf{D})\right] + Var_0(\Delta|\mathbf{D}). \quad (5)$$

This risk is a combined measure of bias and error variance introduced by adopting an MLBMA posterior mean corresponding to some prior model probability set P_0 when the optimal (say, in the sense of predictive performance as discussed below) prior model probability set is $P \neq P_0$. In (5), (E_0, Var_0) and (E_p, Var_p) are the MLBMA posterior mean and variance corresponding to prior model probability sets P_0 and P, respectively. In (5), $(E_p - E_0)^2$ is a quadratic loss of reliability due to adopting E_0 instead of E_p as posterior mean (a quadratic measure of bias), and ($Var_p - Var_0$) is loss due to increased predictive error variance resulting from the adoption of Var_0 instead of Var_p as posterior variance. Setting P_0 to the neutral choice of equal prior probabilities we plot in Figure 7 the maximum, minimum, and grid-averaged MLBMA posterior risk for each of the prior probability combinations associated with variogram models *POW*0 (Figure 7a), *EXP*0 (Figure 7b), *EXP*1 (Figure 7c), and *SPH*1 (Figure 7d), diamonds representing the neutral choice. Comparing Figure 7 with Figure 5 reveals that in our case, posterior risk is only slightly larger than posterior variance. This is so because posterior mean is relatively insensitive to the choice of prior

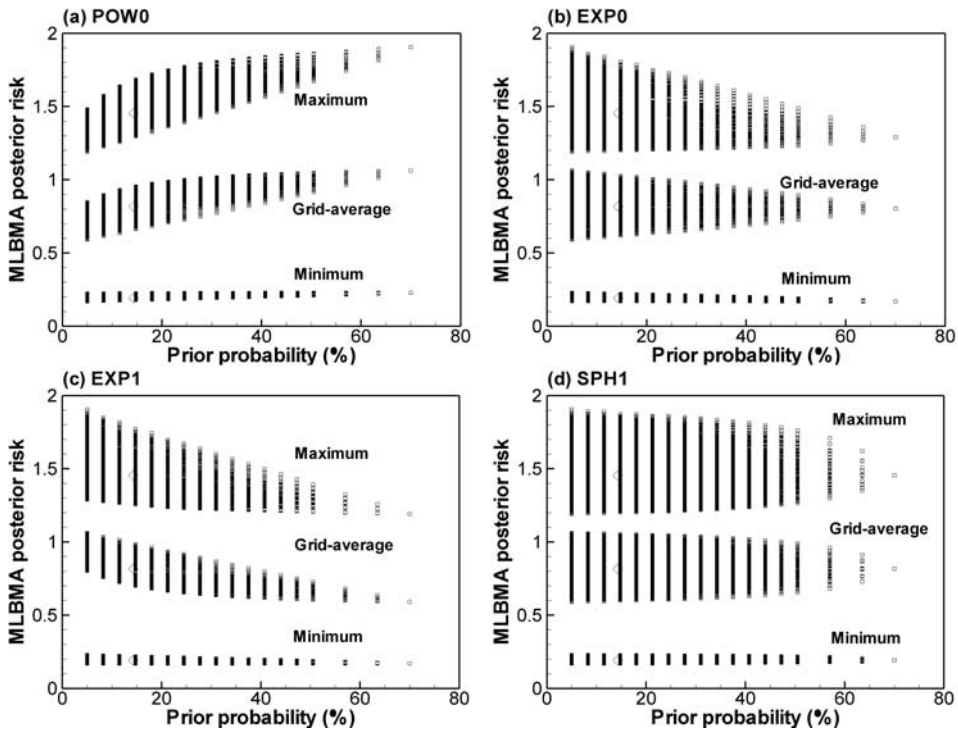

Figure 7. Maximum, minimum, and grid-averaged MLBMA posterior risk for various discrete combinations of prior probabilities corresponding to variogram models (a) *POW*0, (b) *EXP*0, (c) *EXP*1, and (d) *SPH*1. Diamonds represent neutral choice of equal prior probabilities.

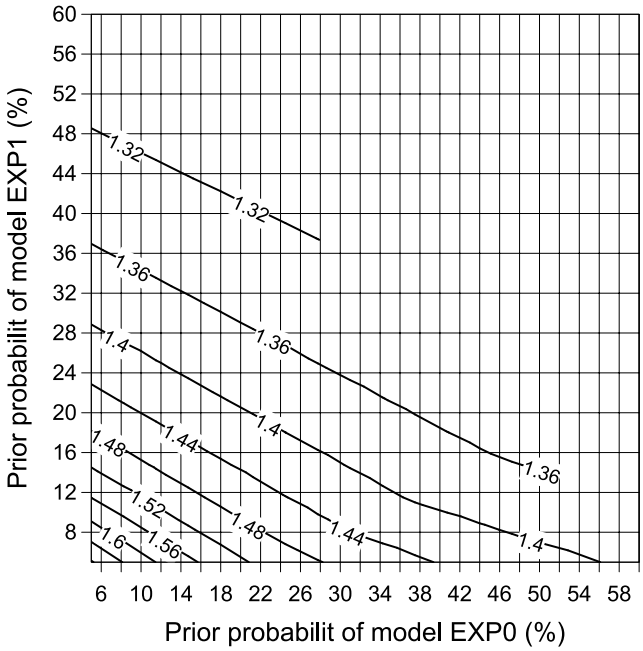

Figure 8. Contours of maximum MLBMA posterior risk in the space of prior model probabilities $Pr(EXP0)$ and $Pr(EXP1)$ when $Pr(POW0) = 14.75\%$, close to the neutral case in which all priors are equal to 14.29%.

model probabilities (Figure 3), rendering $(E_p - E_0)^2$ small in comparison to Var_p.

[24] Figure 8 depicts contours of maximum MLBMA posterior risk in the space of prior model probabilities $Pr(EXP0)$ and $Pr(EXP1)$ when $Pr(POW0) = 14.75\%$, close to the neutral case. The contour, corresponding to the maximum posterior risk column $Pr(POW0) = 14.75\%$ in Figure 7a, shows how the posterior risk of using equal prior model probabilities varies with $Pr(EXP0)$ and $Pr(EXP1)$; the variation of $Pr(SPH1)$ is not shown. As expected, when $Pr(EXP0)$ and $Pr(EXP1)$ increase, $R(\Delta|\mathbf{D})$ decreases at a rate (a measure of sensitivity) that diminishes with $Pr(EXP0)$ and $Pr(EXP1)$. Because of the quadratic term $(E_p - E_0)^2$, the linear pattern of maximum MLBMA posterior risk variation in Figure 8 resembles that of the posterior probability in Figure 2 more closely than that of the nonlinear maximum MLBMA posterior variance variation in Figure 6. The similarity, however, has no known significance as the two figures represent different model statistics.

4. Constrained Maximum Entropy Approach to Assessing Prior Model Probabilities

4.1. Maximum Entropy of Model Probabilities

[25] We have seen that posterior model probabilities and MLBMA posterior variances as well as risks may be highly sensitive to prior model probabilities. To assess the latter quantitatively in an informative manner, we consider maximizing *Shannon*'s [1948] entropy subject to constraints representing the analyst's prior perception about how plausible each alternative model (or a group of models) is relative to others. We follow *Papoulis* [1991] by noting that the probability $p(\mathbf{B})$ of a single random event $\mathbf{B}$ can be interpreted as a measure of our uncertainty about the occurrence or nonoccurrence of $\mathbf{B}$ in a single trial: If $p(\mathbf{B}) \approx 0.999$, then we are almost certain that $\mathbf{B}$ will occur; if $p(\mathbf{B}) \approx 0.1$, then we are reasonably certain that $\mathbf{B}$ will not occur; our uncertainty is maximum if $p(\mathbf{B}) = 0.5$. In our case $\mathbf{B}$ is replaced by the set $\mathbf{M}$ of alternative models M_k ($k = 1, 2, \ldots K$), $p_k \equiv p(M_k)$ is a measure of our uncertainty about the plausibility or lack of plausibility of model M_k before it has been tested (conditioned on data) on experimental and/or observational data $\mathbf{D}$, and Shannon's entropy

$$H = -\sum_{k=1}^{K} p_k \log p_k \tag{6}$$

is the combined prior uncertainty measure of the set $\mathbf{M}$ (treating H as a comparative measure renders the base of the logarithm arbitrary; we use natural log in this paper). Since (6) is the expected value of $-\log p_k$, the latter can be viewed as a measure of prior uncertainty associated with model M_k. Once data are collected and $\mathbf{M}$ is conditioned on $\mathbf{D}$, our uncertainty about how plausible (or implausible) each model would turn out to be in light of $\mathbf{D}$ is generally reduced. This reduction in uncertainty is equivalent to a gain in information, suggesting that H can be viewed as a prior measure of information one may potentially gain (given sufficient conditioning data) about the set $\mathbf{M}$ upon conditioning (and $-\log p_k$ as a measure of information one may gain about model M_k). Absent any prior perception on the part of the analyst as to which model would be better than others, he or she may do well to maximize the potential gain of information by maximizing (6) subject to $\sum_{k=1}^{K} p(M_k) = 1$ or, equivalently,

$$L = -\sum_{k=1}^{K} p_k \log p_k + \lambda\left(\sum_{k=1}^{K} (p_k - 1)\right), \tag{7}$$

where λ is a Lagrange multiplier. Taking derivative with respect to p_k and setting it equal to zero yields

$$p_k = e^{\lambda - 1}, \tag{8}$$

which, upon substitution into $\sum_{k=1}^{K} p(M_k) = 1$, yields $p_k \equiv 1/K$ and $H = \log K$. This is the largest value that H can attain for a given K (the largest gain in information one can potentially accomplish through conditioning). The smallest value it can attain corresponds to perfect certainty on the part of the analyst, a priori, that model M_k associated with some k would prove to be correct so that $p_k = 1$ and, by virtue of $\sum_{k=1}^{K} p(M_k) = 1$, $H = 0$ (maximizing H subject to the constraints $p_k = 1$ and $\sum_{k=1}^{K} p(M_k) = 1$ yields zero potential gain in information). Whereas the first case represents the least amount of prior information about the relative merits of the models, the second case represent the greatest amount of such information; in this sense the neutral case is the least informative and the case of perfect certainty is the most informative; in the first case, conditioning has the potential of enhancing information

by an amount equal to log K (as measured by H), and in the second case it has zero potential for information gain (expressed in terms of posterior probabilities, H may be larger or smaller than its prior value given by (6), i.e., the actual gain in information achieved (objectively) by conditioning may possibly exceed the maximum gain one predicts (subjectively) a priori; this is most likely to happen when the prior probability assigned to one of the models is much larger than those assigned to all other models so that H, based on prior probabilities, is close to 0).

4.2. Constrained Maximum Entropy Method to Assess Model Probabilities

[26] Lying between the above two extremes are situations in which the analyst has neither complete lack nor perfect prior knowledge about the relative merits of the K models in **M**. If the analyst has a prior perception about how plausible each alternative model (or a group of models) is relative to others, he or she may be able to formulate this perception as a nonlinear constrained optimization problem [*Nelles*, 2001]:

$$\max_{p_k} H = -\sum_{k=1}^{K} p_k \log p_k \tag{9}$$

subject to

$$\begin{array}{ll} g_i \le 0 & i = 1, \ldots I \\ h_j = 0 & j = 1, \ldots J \end{array}, \tag{10}$$

where g_i and h_j are specified relationships between the various p_k values (in a manner related but not identical to the direct odds approach of *Bonano et al.* [1990]). If the constraints are logical in that they conform to Ockham's razor and to behavior expected on theoretical and/or empirical grounds (i.e., if one excludes arbitrary solutions that are not based on the principle of parsimony coupled with sound expert knowledge), then we prefer the corresponding informed solution over the noninformative neutral choice. As the former is associated with a smaller value of maximum entropy than the latter, our preference constitutes a min-max choice.

[27] If the analyst, or a group of analysts, is unable to select one set of constraints among several alternative sets all of which are based on the expert judgment of one or more analysts, then one possibility is to extend the min-max approach to all these alternatives by maximizing H subject to each set (maximizing the potential of each set for information gain through conditioning on data) and selecting that solution (those values of p_k) which yields the smallest value of maximum entropy (most informative among the prior sets). A potential difficulty with this min-max approach is the lack of a guarantee that it would lead to optimum predictive performance. We therefore prefer choosing among alternative expert opinions a posteriori, on the basis of posterior measures of model quality. If sufficient data are available to conduct a cross validation of the results in the manner of *Ye et al.* [2004] (as we do below), then we propose selecting the set of prior probabilities that yields optimum predictive performance as measured by criteria such as log score, predictive coverage, mean squared or mean absolute prediction error (defined below and in the cited reference). If there are not enough data to conduct a meaningful cross validation, then we suggest selecting the set of prior probabilities that maximizes the likelihood of **M** in light of the data, given by the normalizing term in (4) via

$$p(\mathbf{D}|\mathbf{M}) = \sum_{k=1}^{K} p(\mathbf{D}|\mathbf{M}_k) p(\mathbf{M}_k) \approx \sum_{k=1}^{K} \exp\left(-\frac{1}{2} KIC_k\right) p(\mathbf{M}_k). \tag{11}$$

Once KIC_k values, which we recall are independent of prior model probabilities $p(M_k)$, have been computed, then $p(\mathbf{D}|\mathbf{M})$ values corresponding to any admissible (i.e., arrived at via entropy maximization on the basis of prior expert judgments) prior probability set are easily calculated using (11).

[28] *Jefferys and Berger* [1992, p. 72] have proposed and demonstrated that absent other prior information about the relative merits of alternative models, it makes sense applying Ockham's razor by assigning higher probabilities to simpler models with fewer parameters. In the authors' words

> We have seen three different ways in which Ockham's razor can be interpreted in Bayesian terms: in the choice of the prior probabilities of hypotheses, using scientific experience to judge that simpler hypotheses are more likely to be correct; as a consequence of the fact that a hypothesis with fewer adjustable parameters will automatically have an enhanced posterior probability, due to the fact that the predictions it makes are sharp; and in the choice of parsimonious empirical models. All these are in agreement with our intuitive notion of what makes a theory powerful and believable. . . . This approach would lead us to try simpler laws first, only moving on to more complicated laws as we find that the simple ones are not adequate to represent the data.

[29] We take this to suggest that applying Ockham's razor to prior model probabilities is consistent with the principle of parsimony embodied in MLBMA, which computes posterior model probabilities on the basis of an information criterion (*KIC*) that (everything else being equal) favors parsimonious over complex models. Applying Ockham's razor to assess prior model probabilities and the principle of parsimony for posterior analysis is not redundant; if a relatively simple model is shown to render acceptable predictions, its posterior plausibility would be enhanced by the principle of parsimony; otherwise it would be falsified by the data.

[30] Our proposed constrained maximum entropy approach to the assessment of prior model probabilities differs in both purpose and detail from the way this concept has been applied to parameter probability estimation [e.g., *Woodbury and Ulrych*, 1998; *Jaynes*, 2003]. Whereas our purpose is to estimate prior probabilities of alternative conceptual-mathematical models, the purpose of the latter is to estimate prior parameter probabilities for a given model. Whereas we maximize Shannon's entropy subject to constraints representing (subjective) expert perceptions about the relative plausibility of alternative models without relying on any prior measurements, the latter maximizes Shannon's entropy expressed in terms of

Table 2. Statistics Corresponding to Three Prior and Posterior Probability Sets for Seven Variogram Models of Log Permeability at the Apache Leap Research Site, Obtained by Constrained Entropy Maximization[a]

		*Pow*0	*Exp*0	*Exp*1	*Exp*2	*Sph*0	*Sph*1	*Sph*2	Entropy	Likelihood Ratio
KIC		369.6	370.1	369.5	416.7	390.5	378.1	424.6		
Neutral case	$p(M_k)$, %	14.29	14.29	14.29	14.29	14.29	14.29	14.29	1.95	1.00
	$p(M_k\|\mathbf{D})$, %	35.29	26.58	37.61	0	0	0.51	0	1.11	
Case 1	$p(M_k)$, %	29.85	11.94	19.90	14.93	5.97	9.95	7.46	1.81	1.48
	$p(M_k\|\mathbf{D})$, %	49.59	14.94	35.23	0	0	0.24	0	1.01	
Case 2	$p(M_k)$, %	16.0	16.0	24.0	16.0	8.0	12.0	8.0	1.88	1.33
	$p(M_k\|\mathbf{D})$, %	29.74	22.40	47.54	0	0	0.32	0	1.07	

[a]Likelihood ratio is relative to neutral case.

moments (commonly up to second order) of (objective) prior parameter measurements.

5. Application of Constrained Maximum Entropy Approach to Log Permeability Variogram Models in Unsaturated Fractured Tuff

5.1. Case Examples

[31] Let $p_1, p_2, \ldots p_7$ be prior probabilities of the seven variogram models *POW*0, *EXP*0, *EXP*1, *EXP*2, *SPH*0, *SPH*1, and *SPH*2 we have previously discussed in connection with log air permeabilities at the Apache Leap Research Site. In Table 2 we list prior and posterior model probabilities corresponding to the neutral case in which one has no prior preference for any model, leading to equal prior probabilities of 1/7 (14.29%) and maximum entropy of $H = \log 7 = 1.95$.

[32] Next we use (1) Ockham's razor [*Jefferys and Berger*, 1992] to ascribe higher prior probabilities to models having fewer parameters, (2) prior generic knowledge [e.g., *Woodbury and Sudicky*, 1991] to ascribe higher prior probabilities to exponential than to spherical models having the same number of parameters, and (3) comparison of qualitative behavior with that of the data to prefer *POW*0 over *EXP*0, both of which have two parameters. We translate this prior knowledge subjectively into six inequality constraints embedded in the following nonlinear optimization problem:

$$\begin{aligned} &\max_{p_k} H = -\sum_{k=1}^{K} p_k \log p_k \\ &\sum_{k=1}^{K} p_k - 1 = 0 \\ &p_1 - 2.5p_2 \geq 0 \\ &p_1 - 1.5p_3 \geq 0 \\ &p_1 - 2.0p_4 \geq 0 \\ &p_2 - 2.0p_5 \geq 0 \\ &p_3 - 2.0p_6 \geq 0 \\ &p_4 - 2.0p_7 \geq 0 \end{aligned} \tag{12}$$

and solve the corresponding problem numerically using a sequential equality-constrained quadratic programming algorithm implemented in IMSL subroutine NNLPG (Visual Numerics, Inc.; http://www.vni.com/products/imsl/). The corresponding prior and posterior model probabilities are listed in Table 2 under case 1.

[33] The results listed in Table 2 under case 2 correspond to the constrained optimization problem

$$\begin{aligned} &\max_{p_k} H = -\sum_{k=1}^{K} p_k \log p_k \\ &\sum_{k=1}^{K} p_k - 1 = 0 \\ &p_1 - p_2 \geq 0 \\ &p_3 - 1.5p_1 \geq 0, \\ &p_1 - p_4 \geq 0 \\ &p_2 - 2.0p_5 \geq 0 \\ &p_3 - 2.0p_6 \geq 0 \\ &p_4 - 2.0p_7 \geq 0 \end{aligned} \tag{13}$$

which maintains the previous relationship between exponential and spherical models but reduces the plausibility of *POW*0 relative to all three exponential models, violating Ockham's razor by rendering *POW*0 less plausible than *EXP*1 and allowing *EXP*2 to be as plausible as *POW*0.

[34] Table 2 confirms that case 1, which we consider to be the most informative (by regarding Ockham's razor as prior knowledge), yields the largest likelihood (1.48 times that of the neutral case) and smallest maximum entropy (1.81) among the three cases, as we would anticipate. It also yields the least posterior entropy (1.01) or (equivalently) residual uncertainty after conditioning, thereby validating our choice of its priors as being the best among the three cases. In our case, there are enough data to provide additional validation of our approach to the selection of priors on the basis of predictive performance criteria, as discussed in the next section.

5.2. Predictive Performance

[35] To assess the predictive performance of MLBMA in each of the three cases listed in Table 2, we follow the cross-validation approach of *Ye et al.* [2004]. The approach consists of (1) splitting the data $\mathbf{D}$ into two parts, $\mathbf{D}^A$ and $\mathbf{D}^B$; (2) obtaining ML estimates of model parameters and posterior probabilities conditional on $\mathbf{D}^A$ for each case; (3) using these to render corresponding MLBMA predictions $\hat{\mathbf{D}}^B$ of $\mathbf{D}^B$; and (4) assessing and comparing the quality of the predictions. More specifically, we eliminate from consideration all log permeability data from one of the six boreholes at a time and predict 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 of *Ye et al.* [2004]. As

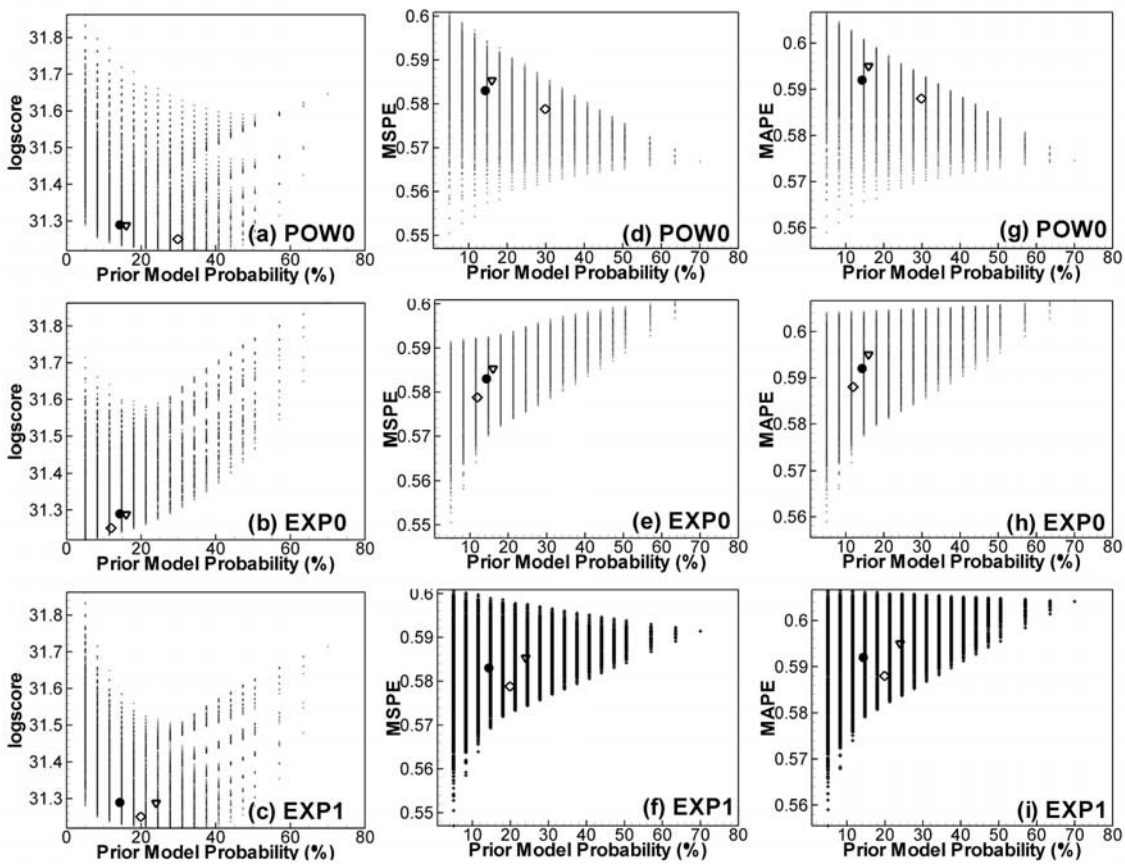

Figure 9. Values of (a–c) log scores, (d–f) MSPE, and (g–i) MAPE values of models *POW*0, *EXP*0, and *EXP*1 corresponding to various discrete combinations of prior model probabilities. Solid squares represent the neutral case, diamonds represent case 1, and inverted triangles represent case 2.

*SPH*1 has a very small posterior probability in comparison to *POW*0, *EXP*0 and *EXP*1 (Table 2), we limit the cross validation to the latter three models.

[36] To assess predictive performance, we compute the following three criteria [*Liang et al.*, 2001] for each borehole and average each of them over all six boreholes, the log score [*Ye et al.*, 2004]

$$-\ln p(\mathbf{D}^B|\mathbf{D}^A) = -\ln \sum_{k=1}^{K} p(\mathbf{D}^B|M_k, \mathbf{D}^A) p(M_k|\mathbf{D}^A), \qquad (14)$$

the mean squared prediction error

$$MSPE = \frac{1}{N^B} \sum_{d\in\mathbf{D}^B} \sum_{k=1}^{K} \left[\left(\hat{d}|M_k, \mathbf{D}^A\right) - d\right]^2 p(M_k|\mathbf{D}^A), \qquad (15)$$

and the mean absolute prediction error

$$MAPE = \frac{1}{N^B} \sum_{d\in\mathbf{D}^B} \sum_{k=1}^{K} \left|\left(\hat{d}|M_k, \mathbf{D}^A\right) - d\right| p(M_k|\mathbf{D}^A), \qquad (16)$$

where N^B is the number of cross-validation data and $\hat{d}$ is a prediction corresponding to cross-validation data d. The smaller are these criteria, the better is the predictive performance of MLBMA. In particular, the lower is the MLBMA predictive log score based on training data $\mathbf{D}^A$, the smaller is the amount of information lost upon eliminating $\mathbf{D}^B$ from the original data set $\mathbf{D}$ (i.e., the higher is the probability that MLBMA based on $\mathbf{D}^A$ would reproduce the lost data, $\mathbf{D}^B$).

[37] Figure 9 depicts MLBMA log scores (Figures 9a–9c), MSPE (Figures 9d–9f), and MAPE (Figures 9g–9i) values of models *POW*0 (Figures 9a, 9d, 9g), *EXP*0 (Figures 9b, 9e, 9h), and *EXP*1 (Figures 9c, 9f, 9i), respectively, for various discrete combinations of prior model probabilities. Solid squares represent values corresponding to the neutral case in Table 2, diamonds represent case 1, and inverted triangles represent case 2. In all three figures the performance criteria associated with case 1 are smaller than those associated with the other two cases, whereas those associated with case 2 are somewhat larger than those corresponding to the neutral case. In other words our selection of case 1 on the basis of both likelihood and min-max entropy has yielded the best predictive performance among all three cases we had considered. The three criteria vary over limited ranges due to their average nature and their being conditioned on the training data $\mathbf{D}^A$.

6. Conclusions

[38] 1. Previous application of maximum likelihood Bayesian model averaging (MLBMA, *Neuman* [2002, 2003]) to alternative variogram models of log air permeability data in fractured tuff has demonstrated its effective-

ness in quantifying conceptual model uncertainty and enhancing predictive capability [*Ye et al.*, 2004]. A question remained of how best to ascribe prior probabilities to competing models. We have shown in this paper that one answer is to (1) maximize *Shannon*'s [1948] entropy H subject to constraints reflecting a single analyst's prior perception about how plausible each model (or a group of models) is relative to others, and (2) select a posteriori the most likely among such maxima corresponding to alternative prior perceptions of various analysts or groups of analysts. Another way to select among alternative prior model probability sets, which, however, is not guaranteed to yield optimum predictive performance (though it did so in our example) and would therefore not be our preferred option, is a min-max approach according to which one selects a priori the set corresponding to the smallest value of maximum entropy. Whereas maximizing H subject to the prior perception of a single analyst maximizes the potential for further information gain through conditioning, selecting the smallest among such maxima gives preference to the most informed prior perception among those of several analysts.

[39] 2. We used the same variogram models and data as *Ye et al.* [2004] to demonstrate that our proposed approach leads to the least amount of posterior entropy (residual uncertainty after conditioning) and enhances predictive model performance as compared to (1) the noninformative neutral case in which all prior model probabilities are set equal to each other and (2) an informed case that nevertheless violates Ockham's razor (the principle of parsimony).

[40] 3. Upon considering the above models and data, we found that for a given set of prior probabilities, the posterior probability of a model tends to diminish as the corresponding value of *Kashyap*'s [1982] information criterion *KIC* increases; in fact, posterior probabilities of models associated with relatively large *KIC* values were zero regardless of priors. The sensitivity of MLBMA posterior mean predictions to prior model probabilities was much smaller than that of MLBMA predictive variance and risk (a combined measure of bias and error variance due to a nonoptimal choice of priors). The sensitivities of all three statistics tended to diminish with increasing prior probabilities.

[41] 4. Though we have not demonstrated it here, theory implies that the sensitivity of MLBMA posterior statistics to prior model probabilities would diminish with the degree of conditioning.

[42] **Acknowledgments.** This research was supported in part by the U.S. Department of Energy National Nuclear Security Administration Nevada Site Office under contract DE-AC52-00NV13609 with the Desert Research Institute; the National Science Foundation under grant EAR-0407123 to the University of Arizona; and the U.S. Nuclear Regulatory Commission Office of Nuclear Regulatory Research under contract JCN Y6465 with Pacific Northwest National Laboratory. The authors are thankful to Jenny Chapman and Greg Pohll of the Desert Research Institute for stimulating discussions and comments.

References

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

Apostolakis, G. (1990), The concept of probability in safety assessment of technological systems, *Science*, *250*, 1359–1364.

Bonano, E. J., S. C. Hora, R. L. Keeney, and D. von Winterfeldt (1990), Elicitation and use of expert judgment in performance assessment for high-level radioactive waste depositories, *NUREG/CR-5411*, U.S. Nucl. Regul. Comm., Washington, D. C.

Bredehoeft, J. (2005), The conceptualization model problem-surprise, *Hydrogeol. J.*, *13*, 37–46.

Burnham, K. P., and A. R. Anderson (2002), *Model Selection and Multiple Model Inference: A Practical Information-Theoretical Approach*, 2nd ed., Springer, 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*(2), 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*(2), 228–242.

Clyde, M. (1999), Comment, *Stat. Sci.*, *14*(4), 401–404.

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

Draper, D. (1999), Comment, *Stat. Sci.*, *14*(4), 405–409.

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

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

Insua, D. R., F. Ruggeri, and J. Martin (2000), Bayesian sensitivity analysis, in *Sensitivity Analysis*, edited by A. Saltelli et al., pp. 225–244, John Wiley, Hoboken, N. J.

Jaynes, E. T. (2003), *Probability Theory: The Logic of Science*, Cambridge Univ. Press, New York.

Jeffreys, H. (1957), *Scientific Inference*, 2nd ed., Cambridge Univ. Press, New York.

Jefferys, W. H., and J. O. Berger (1992), Ockham's razor and Bayesian analysis, *Am. Sci.*, *89*, 64–72.

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.

Kass, R. E., and L. Wasserman (1996), The selection of prior distributions by formal rules, *J. Am. Stat. Assoc.*, *91*(435), 1343–1370.

Liang, F., Y. K. Truong, and W. H. Wong (2001), Automatic Bayesian model averaging for linear regression and application in Bayesian curve fitting, *Stat. Sinica*, *11*, 1005–1029.

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

Madigan, D., J. Gavrin, and A. E. Raftery (1995), Eliciting prior information to enhance the predictive performance of Bayesian graphical models, *Comm. Stat. Theory Methods*, *24*, 2271–2292.

Nelles, O. (2001), *Nonlinear System Identification: From Classical Approaches to Neural Networks and Fuzzy Models*, Springer, New York.

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

Papoulis, A. (1991), *Probability, Random Variables, and Stochastic Processes*, 3rd ed., McGraw-Hill, New York.

Poeter, E., and D. Anderson (2005), Multimodel ranking and inference in groundwater modeling, *Ground Water*, *43*(4), 597–605.

Ruggeri, F., and S. Sivaganesan (2000), On a global sensitivity measure for Bayesian inference, *Indian J. Stat., Ser. A*, *62*, 110–127.

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*(3), 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*(3), 363–371.

Shannon, C. E. (1948), A mathematical theory of communication, *Bell Syst. Tech. J.*, *27*, 379–423, 623–656.

Woodbury, A. D., and E. A. Sudicky (1991), The geostatistical characteristics of the Borden aquifer, *Water Resour. Res.*, *27*(4), 533–546.

Woodbury, A. D., and T. J. Ulrych (1998), Minimum relative entropy and probabilistic inversion in groundwater hydrology, *Stochastic Hydrol. Hydraul.*, 12, 317–358.

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.

Zio, E., and G. E. Apostolakis (1996), Two methods for the structured assessment of model uncertainty by experts in performance assessments of radioactive waste repositories, *Rel. Eng. Syst. Safety*, *54*, 225–241.

P. D. Meyer, Pacific Northwest National Laboratory, Richland, WA 99352, USA.

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

K. Pohlmann and M. Ye, Desert Research Institute, 755 E. Flamingo Road, Las Vegas, NV 89119, USA.