102319

Stochastic Environmental Research and Risk Assessment 17 (2003) 291–305 © Springer-Verlag 2003
DOI 10.1007/s00477-003-0151-7

Maximum likelihood Bayesian averaging of uncertain model predictions

S. P. Neuman

Abstract. 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. 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 developed by Neuman and Wierenga (2003). This paper describes a key formal element of this much broader and less formal strategy that concerns rendering optimum hydrologic predictions by means of several competing deterministic or stochastic models and assessing their joint predictive uncertainty. The paper proposes a Maximum Likelihood Bayesian Model Averaging (MLBMA) method to accomplish this goal. MLBMA incorporates both site characterization and site monitoring data so as to base the outcome on an optimum combination of prior information (scientific knowledge plus data) and model predictions. A preliminary example based on real data is included in the paper.

Keywords: Uncertainty, Conceptual modeling, Model structure, Bayesian averaging, Maximum likelihood

1 Introduction

Hydrologic analyses typically rely on a single conceptual model of site hydrogeology. 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 Type I model errors, which

S. P. Neuman
Department of Hydrology and Water Resources,
University of Arizona, Tucson, Arizona 85721, USA
e-mail: neuman@hwr.arizona.edu

This work was supported by the U.S. Nuclear Regulatory Commission under contract NRC-04-97-056. Special credit is due to Thomas J. Nicholson and Ralph Cady of the USNRC for having conceived and championed this research. I am most grateful to Weihsueh Chiu of the U.S. Environmental Protection Agency for having brought to my attention the extensive work on BMA conducted by the "Seattle school" of statistics at the University of Washington, and for his helpful comments on this paper.

31114

arise 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.

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). The bias and uncertainty that result from reliance on an inadequate conceptual-mathematical model are typically much larger than those introduced through an inadequate choice of model parameter values. Yet most hydrologic uncertainty analyses ignore the former and focus exclusively on the latter. This often leads to overconfidence in the predictive capabilities of the model, which the available hydrologic data seldom justify.

It is argued by Beven and Freer (2001) "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 behavioural or acceptable in reproducing the observed behaviour 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 unfeasibly 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.

There is no established literature on ways to construct alternative conceptual–mathematical models of site or regional hydrology, select the best among them, use them jointly to render optimum hydrologic predictions, and assess the uncertainty of such predictions. Useful philosophies of model building under uncertainty may be found in Gauch (1993), Burnham and Anderson (1998) and Christakos (2000, 2002a,b, 2003a,b). 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 developed 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

(a) the conceptual framework that determines model structure, (b) 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 (c) the scaling behavior of hydrogeologic variables.

The strategy of Neuman and Wierenga (2003) encourages exploration of varied conceptual frameworks and assumptions at all stages of hydrogeologic model development through a comprehensive evaluation of a broad range of regional and site data, their translation into coherent and internally consistent conceptual–mathematical models, and computation and visualization based on these data and models. Included among these frameworks and assumptions are various model simplification and abstraction schemes. The strategy recognizes that site characterization and monitoring data are expensive and difficult to collect, leading to a ubiquitous scarcity of hard site information. It is therefore critically important to assess the role that such data play in rendering the hydrogeologic performance analysis credible. The strategy stresses the role of characterization and monitoring data in helping one identify and test alternative conceptual models, make rational choices among them, gauge and reduce model bias and uncertainty through proper model selection and calibration, assess the reliability of model predictions, and confirm the assessment through independent peer review as well as at least some degree of direct verification.

The report of Neuman and Wierenga (2003) includes 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 describes a key formal element of their much broader and less formal strategy, which concerns rendering optimum predictions by means of several competing deterministic or stochastic models and assessing their joint predictive uncertainty. The paper proposes a Maximum Likelihood Bayesian Model Averaging (MLBMA) method to accomplish this goal and illustrates the idea with a preliminary example based on real data. Like the underlying strategy, MLBMA incorporates both site characterization and site monitoring data so that the outcome is based not only on model predictions but also on prior information (scientific knowledge and data) entering into the model.

2
Prediction and uncertainty analysis based on single model

2.1
Deterministic model

A traditional approach to hydrologic prediction and uncertainty analysis has been to postulate a deterministic model structure and treat its parameters as being imperfectly known. To quantify this imperfect knowledge, one must postulate a prior parameter uncertainty model. When model parameters are measurable and available in sufficient quantity, one could postulate a Type A probabilistic model of prior parameter uncertainty based on statistics derived from these data. When no such data are available in statistically significant quantities, one has the option of postulating a Type B model of prior parameter uncertainty on the basis of subjective probabilities. Such a model should always be suspected of suffering from an unknown amount of statistical and personal bias. Intermediate between Type A and Type B parameter uncertainty models is the case where indirect information about the parameters is available, from which relevant prior statistics can be derived formally. Such information may include parameter measurements

in other similar hydrologic environments and/or surrogate data (for example, pedologic or geophysical data that are indicative of hydrologic parameters). Statistics derived from data collected elsewhere are potentially biased due to a lack of site- or region-specific information about mean parameter values and incompatibility of scale. The associated variance may be too small or too large, depending on the quantity and quality of such data. Statistics derived from surrogate data may suffer from poorly defined correlations and incompatibility of scale.

The traditional approach to reduce parameter bias and uncertainty has been to calibrate the model against observed system behavior by means of a suitable inverse method. The last thirty years have seen major advances in the development of theories and algorithms for the estimation of deterministic model parameters. Many (though not all) of these theories and algorithms are "statistical" in that they include analyses of parameter estimation uncertainty. Such analyses typically accept, but do not necessarily require, information about prior parameter statistics as input. The output includes *posterior* statistics of parameter estimation errors, which are generally less biased and smaller than the prior estimation errors. A recent summary and comparison of various statistical inverse methods for groundwater flow models has been published by Zimmerman et al. (1998). A detailed set of guidelines for the effective calibration of deterministic groundwater flow models has been prepared by Hill (1998).

The most common way to propagate input errors through an otherwise deterministic model is by means of Monte Carlo simulations. This is done by generating multiple, equally likely sets of randomized inputs; computing deterministically a set of corresponding model outputs for each; and analyzing the resultant multiple, equally likely random output sets statistically. Another approach is to associate the predictions with approximate error bounds, or confidence limits, computed on the basis of linear regression theory applied to the (typically nonlinear) inverse model (Hill, 1998).

2.2 Stochastic model

Hydrologic parameters exhibit both systematic and random spatial (and possibly temporal) variations on a multiplicity of scales. Traditional deterministic models capture at best the larger-scale, systematic components of these variations. They however fail to resolve smaller scale variations or account for their uncertain nature. The emphasis is therefore shifting from deterministic to probabilistic methods that are better suited for these needs. The trend has become to describe the spatial variability and scaling of hydrologic parameters geostatistically, and to analyze hydrologic systems stochastically. This trend has been documented in a number of recent books including those by Dagan and Neuman (1997) and Zhang (2001).

The most common method of stochastic analysis is high-resolution computational Monte Carlo simulation that produces a large number of equally likely results. These nonunique results are summarized in terms of statistically averaged quantities, their variance–covariance, and perhaps higher moments of the corresponding sample probability distributions. Results that honor measured values of medium properties are said to be conditioned on these data. Upon conditioning the simulations on measured values of parameters in space, one obtains (among others) conditional mean flow and transport variables that constitute optimum unbiased predictors of these unknown random quantities. One also

obtains conditional second moments (variance–covariance) that provide a measure of the associated prediction errors. To condition the predictions on observed system behavior, one must either discard random simulations that do not reproduce the observations, or employ an inverse procedure of the kind developed for this purpose by Gómez-Hernández et al. (1997).

Monte Carlo analysis requires knowing the multivariate probability distribution of relevant hydrologic parameters, which is difficult to infer from commonly available data. To achieve a high space–time resolution of relevant stochastic phenomena, it requires the use of large space–time grids with very small discretization intervals. To yield sample statistics that converge to their theoretical (ensemble) counterparts requires numerous repetitions (realizations). The net result is a large amount of computational time and storage, which are considered uneconomical for many practical applications.

This has given impetus to the development of alternative stochastic methods that allow one to compute the conditional mean, variance and covariance of hydrologic system states directly, without Monte Carlo simulation. This is done on the basis of moment equations conditional on parameter measurements as illustrated for example by Guadagnini and Neuman (1999) and Ye et al. (2002). Conditioning additionally on observed state variables requires an inverse procedure (Hernandez et al., 2003).

3
Prediction and uncertainty analysis based on multiple models

3.1
Previous hydrologic approaches

Carrera and Neuman (1986b) have noted that an inadequate model structure (conceptualization and mathematical description) is far more detrimental to its predictive ability than is a suboptimal set of model parameters. This helps explain why the National Research Council (1999) has listed as second among four recommended research directions in subsurface science the development of tools and methodologies for conceptual modeling with emphasis on heterogeneity, scale and uncertainty bounds on the basis of field experimental data.

Recently, a panel was convened by the National Research Council (2001) to describe the process through which conceptual models of flow and transport in the fractured vadose zone are developed, tested, refined and reviewed. The panel concluded that development of the conceptual model is the most important part of the modeling process. The conceptual model is the foundation of the quantitative mathematical representation of the field site (i.e., the mathematical model), which in turn is the basis for the computer code used for simulation. Reasonable alternative conceptualizations and hypotheses should be developed and evaluated. In some cases, the early part of a study might involve multiple conceptual models until alternatives are eliminated by field results.

According to the panel, it is important to recognize that model predictions require assumptions about future events or scenarios, and are subject to uncertainty. Meaningful quantification of uncertainty should be considered an integral part of any modeling endeavor, as it establishes confidence bands on predictions given the current state of knowledge about the system. A suite of predictions for a range of different assumptions and future scenarios is more useful than a single prediction.

It was noted earlier that there is uncertainty not only about the parameter values that should enter into a given model (as characterized by its structure), but also about the very structure (conceptual and mathematical) of the model that should represent the hydrologic system of interest. The traditional approach to model uncertainty analysis, which considers only a single deterministic model structure, fails to adequately sample the complete space of plausible hydrologic models. As such, it is prone to modeling bias and underestimation of model uncertainty.

An example of how one could account quantitatively for structural model uncertainties was given by James and Oldenburg (1997). They investigated the uncertainty of simulated TCE concentrations, at the point of potential human exposure, due to uncertainty in the parameters (permeability, porosity, diffusivity, solubility, adsorption) and variations in the conceptual-mathematical model (injection rate of TCE source; initial TCE source saturation; regional groundwater flow; heterogeneity of permeability). The authors used the three-dimensional code T2VOC to simulate three-phase (gas, aqueous, NAPL), three-component (air, water, VOC) nonisothermal flow based on an actual site with a 25 m thick vadose zone and a saturated zone. To assess parameter uncertainty associated with a given model, they used the inverse code ITOUGH2. Their final step was to assess the range of outcomes that one obtains with the entire set of alternative conceptual–mathematical models. James and Oldenburg found that uncertainties in their model outcomes span orders of magnitude, and that both parameter and model uncertainty contribute significantly to this wide range of outcomes. They concluded that "risk assessment and remediation selection ... is meaningful only if analysis includes quantitative estimates of ... uncertainty" in both the parameters and the conceptual–mathematical models.

A similar approach has been advocated more recently by Samper and Molinero (2000). The authors consider the main uncertainties in predicting groundwater flow and transport to be those associated with the selection of future scenarios, choice of model structure and assignment of model parameters. The authors consider parameter uncertainty to be minor in comparison to structural (i.e. conceptual) model errors. They suggest to evaluate model predictive uncertainty by calibrating a number of conceptual–mathematical models against available monitoring data, to retain those calibrated models that can adequately reproduce past observations, to assess the predictive uncertainty of each model due to the uncertainty of its parameters, to treat the predictive uncertainty of each model as being equally likely, and to produce a single combined range of predictive uncertainties.

Rather than relying on model calibration and treating the outcomes of different structural models as being equally likely, 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).

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."

It must be understood that the set of predictions one produces with any given choice of alternative structural models and parameter sets, by whatever method, is conditional on the choice of models and the data used to support them. As such, these predictions do not represent all possibilities but only a limited range of such possibilities, associated with these models and data. Any change in the latter would generally lead to a different assessment of predictive model uncertainty. There thus appears to be no way to assess the uncertainty of hydrologic predictions in an absolute sense, only in a conditional or relative sense.

4
Proposed approach

Bayesian model averaging

At the heart of the approach I propose is the concept of Bayesian Model Averaging (BMA), described with clarity in a recent tutorial by Hoeting et al. (1999). According to these authors, "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, at least one of which must be valid. 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

$$p(\mathbf{D}|M_k) = \int p(\mathbf{D}|\theta_k, M_k) p(\theta_k|M_k)\, \mathrm{d}\theta_k \tag{3}$$

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

The posterior mean and variance of Δ are (Draper, 1995)

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

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

According to Hoeting et al. (1999), there is considerable empirical evidence that averaging over *all* models in this fashion provides better average predictive ability than relying on a single model, M_k, conditional on **M**. However, they list a number of factors that render the application of BMA to complex systems (such as those encountered in hydrology) difficult: (a) The number of potentially feasible models may be exceedingly large, rendering their exhaustive inclusion in (1) infeasible; (b) integrals of form (3) may be hard to compute; and (c) the specification of prior model probabilities $p(M_k)$ is challenging, having received little attention.

A practical way to eliminate the first difficulty is to average over a manageable subset of models that are supported by the data. The strategy being developed by the author promotes the idea of Occam's window (Madigan and Raftery, 1994) according to which averaging is limited to a relatively small set of the most parsimonious models that are most strongly supported by the data while remaining hydrologically plausible.

Maximum likelihood Bayesian model averaging

To render BMA computationally feasible, I adopt a suggestion by Taplin (1993) that $p(\Delta|M_k, \mathbf{D})$ in (1) be approximated by $p\left(\Delta \middle| M_k, \hat{\boldsymbol{\theta}}_k, \mathbf{D}\right)$ where $\hat{\boldsymbol{\theta}}_k$ is the maximum likelihood estimate of $\hat{\boldsymbol{\theta}}_k$ based on the likelihood $p\left(\mathbf{D} \middle| \hat{\boldsymbol{\theta}}_k, M_k\right)$. Hoeting et al. (1999) note that Draper (1995), Raftery et al. (1996) and Volinsky et al. (1997) have shown this to be useful in the BMA context.

Methods to evaluate $\hat{\boldsymbol{\theta}}_k$ by calibrating a deterministic model M_k against hydrogeologic data **D**, which may include prior information about the parameters, are described by Carrera and Neuman (1986a,b) 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. (2003). The approach yields a negative log likelihood (NLL) criterion that includes two weighted square residual terms: a generalized sum of squared differences between simulated and observed state variables, and a generalized sum of squared differences between posterior and prior parameter estimates. 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. The corresponding constants of proportionality are estimated jointly with other model parameters by maximum likelihood. Including prior information in the calibration criterion allows conditioning 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).

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 $p\left(\Delta\middle|M_k, \hat{\boldsymbol{\theta}}_k, \mathbf{D}\right)$ by Monte Carlo simulation of Δ through random perturbation of the parameters. The simulation also yields corresponding approximations $E\left[\Delta\middle|M_k, \hat{\boldsymbol{\theta}}_k, \mathbf{D}\right]$ of $E[\Delta|M_k, \mathbf{D}]$, and $Var\left[\Delta\middle|M_k, \hat{\boldsymbol{\theta}}_k, \mathbf{D}\right]$ of $Var[\Delta|M_k, \mathbf{D}]$, in (4) and (5). If M_k is a stochastic model based on moment equations, it can yield $E\left[\Delta\middle|M_k, \hat{\boldsymbol{\theta}}_k, \mathbf{D}\right]$ and $Var\left[\Delta\middle|M_k, \hat{\boldsymbol{\theta}}_k, \mathbf{D}\right]$ directly without Monte Carlo simulation (Hernandez et al., 2002).

To eliminate the need for computer intensive integration according to (3), I propose to evaluate the weights $p(M_l|\mathbf{D})$ in (1) and (4)–(5) based on a result due to Kashyap (1982). The author considers a set $M_1, \ldots, M_K$ of mutually exclusive models so that any set of observational data could have originated from only one of them. I interpret this to mean that only one of the models is correct even in the event that some yield similar predictions for a given set of data (in which event the degeneracy could be resolved by prior information about which model "makes most sense," parsimony, and/or additional data for which the predictions would differ). The models may be linear or nonlinear, Gaussian or non-Gaussian. Kashyap proves that, under some fairly standard conditions,

$$\ln\ p(M_k|\mathbf{D}) = \ln\ C_k + \ln\ p\left(\mathbf{D}\middle|\hat{\boldsymbol{\theta}}_k, M_k\right) + \ln\ p\left(\hat{\boldsymbol{\theta}}_k|M_k\right) + \frac{N_k}{2}\ln\left(\frac{2\pi}{N}\right) - \frac{1}{2}\ln\left|\mathbf{F}_k\left(\mathbf{D}\middle|\hat{\boldsymbol{\theta}}_k, M_k\right)\right| + R(N) \tag{6}$$

where $C_k = cp(M_k)$, $p(M_k)$ being the prior probability of model M_k and c a constant that can be evaluated from

$$\sum_{l=1}^{K} p(M_l|\mathbf{D}) = 1\ , \tag{7}$$

N_k is the dimension of $\hat{\boldsymbol{\theta}}_k$, N is the dimension of $\mathbf{D}$, $\mathbf{F}_k$ is the normalized (by N) observed (as opposed to ensemble mean) Fisher information matrix having components

$$F_{k,ij} = -\frac{1}{N}\frac{\partial^2 \ln\ p\left(\mathbf{D}\middle|\hat{\boldsymbol{\theta}}_k, M_k\right)}{\partial\theta_i\partial\theta_j} \tag{8}$$

and $NR(N)$ tends to a constant almost surely as $N \to \infty$.

Kashyap (1982) suggests that, in the absence of any contrary information, the models be assigned equal prior probabilities, yielding $C_k = C = constant$ for all k. The assumption that all models are a priori equally likely is considered by Hoeting et al. (1999) to be a "reasonable 'neutral' choice" when there is insufficient prior reason to prefer one model over another. Draper (1999) and George (1999) express concern that if two models are near equivalent as concerns 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 non-collinear. Otherwise, one should consider reducing (diluting) the prior probabilities assigned to models that are deemed closely related.

Kashyap's (1982) purpose in developing (6) was to derive an optimum decision rule for selecting one among several competing models, unrelated to BMA. Since the first term on the right hand side is constant and the last is asymptotically zero, Kashyap proposed to select that model which minimizes the KIC criterion

$$\hat{d}_k = -\ln\ p\left(\mathbf{D}\middle|\hat{\boldsymbol{\theta}}_k, M_k\right) - \ln\ p\left(\hat{\boldsymbol{\theta}}_k|M_k\right) - \frac{N_k}{2}\ln\left(\frac{2\pi}{N}\right) + \frac{1}{2}\ln\left|\mathbf{F}_k\left(\mathbf{D}\middle|\hat{\boldsymbol{\theta}}_k, M_k\right)\right| \tag{9}$$

Increasing the number of parameters N_k allows $-\ln\ p\left(\mathbf{D}\middle|\hat{\boldsymbol{\theta}}_k, M_k\right)$ 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 $\hat{d}_k$ grows while $p(M_k|\mathbf{D})$ in (6) diminishes. This means that a more parsimonious model with fewer parameters is preferred by (9) and assigned a higher probability by (6). On the other hand $-\ln\ p\left(\mathbf{D}\middle|\hat{\boldsymbol{\theta}}_k, M_k\right)$ 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.

The last term in (9) reduces the relative emphasis on model fit as the information content of the data diminishes. As illustrated by Carrera and Neuman (1986b), it 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. The term tends to a constant as N becomes large, so that $\hat{d}_k$ becomes asymptotically equivalent to the Bayes information criterion BIC derived by Akaike (1977), Rissanen (1978) and Schwarz (1978) on the basis of different considerations. Raftery (1993) proposed adopting the asymptotic BIC approximation

$$\ln\ p(\mathbf{D}|M_k) \approx \ln\ p\left(\mathbf{D}\middle|\hat{\boldsymbol{\theta}}_k, M_k\right) - \frac{N_k}{2}\ln\ N \tag{10}$$

for BMA (see also Raftery et al. 1996; Volinsky et al. 1997; Hoeting et al. 1999).

To my knowledge, the nonasymptotic expression (6) has not been previously incorporated into BMA. I propose to do so because environmental models seldom satisfy the assumption that N is large. To render the use of (6) in BMA computationally feasible, I propose to follow the approach of Carrera and Neuman (1986a) who incorporate in $\mathbf{D}$ both observational data such as head and prior estimates of the parameters, and treat the two sets as being mutually uncorrelated while allowing internal correlations between members of each set. This allows them to incorporate $\ln\ p\left(\hat{\boldsymbol{\theta}}_k|M_k\right)$ into the log likelihood function $\ln\ p\left(\mathbf{D}\middle|\hat{\boldsymbol{\theta}}_k, M_k\right)$ and to compute $\mathbf{F}_k\left(\mathbf{D}\middle|\hat{\boldsymbol{\theta}}_k, M_k\right)$ in a straightforward manner (also Carrera et al. 1997).

5
Preliminary example: alternative geostatistical models of spatially varying data

The following example illustrates preliminary steps toward a potential application of MLBMA. The example concerns single-hole pneumatic injection tests conducted by Guzman et al. (1994, 1996) in unsaturated fractured tuff at the Apache Leap Research Site (ALRS) in central Arizona. Spatially distributed log air permeability data obtained from a steady state interpretation of 184 such tests, in 1-m borehole intervals (Fig. 1) along 6 vertical and inclined (at 45°) boreholes, were subjected to geostatistical analysis by Chen et al. (2000). The authors compared three different conceptual–geostatistical models of spatial variability

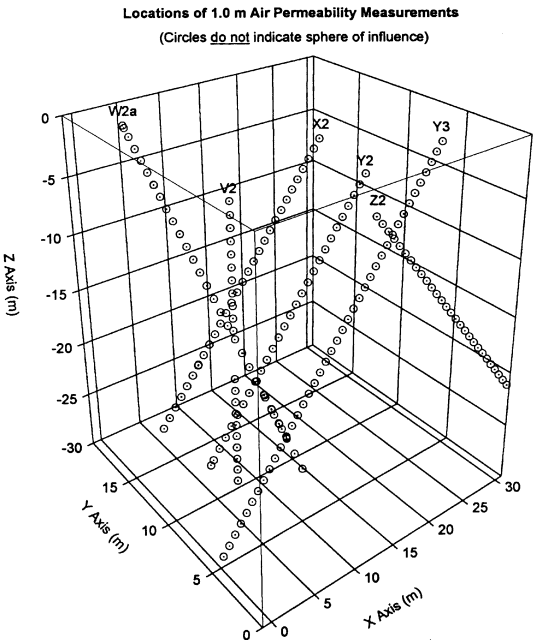

Fig. 1. Center locations of 1-m single hole pneumatic test intervals; overlapping circles indicate retested locations (after Chen et al., 2000)

for these data: (1) a fractal power-law variogram model, (2) an exponential variogram model with linear drift, and (3) an exponential variogram model with quadratic drift. Variogram parameters were estimated using a maximum likelihood cross-validation (MLCV) method developed for this purpose by Samper and Neuman (1989a, b), coupled with a generalized least squares drift removal approach due to Neuman and Jacobson [1984]. MLCV estimates variogram parameters by maximizing the likelihood of kriging (geostatistical interpolation) cross validation errors. As MLCV assumes that the variogram model is known, it leads to optimum parameters for a given model structure without regard to the question how well this model represents the real system. All three models yielded comparable kriged estimates of log permeability but very different estimation variances in three dimensions. This is illustrated along a vertical cross-section for models 1 and 3 in Fig. 2. The negative log likelihood model fit criterion, NLL, in Table 1 is seen to be very similar for all three variogram models, implying that they fit the data almost equally well.

A number of model discrimination criteria have been developed in the context of maximum likelihood estimation theory, which allow one to rank the three variogram models. Among these criteria are AIC (Akaike, 1974), BIC (Akaike, 1977), HIC (Hannan, 1980) and KIC (Kashyap, 1982). BIC and KIC are defined, respectively, in (10) and (9). All four criteria support the principle of parsimony in that, everything else being equal, they favor the model with the smallest number of parameters. Kashyap's criterion favors the model that is least probable (in an average sense) of being incorrect. Stated otherwise, the criterion minimizes the average probability of selecting the wrong model among a set of alternatives. While it favors the simpler model, KIC nevertheless allows considering models of growing complexity as the database improves in quantity and quality. In other words, KIC 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 then prefer a simpler model with fewer parameters,

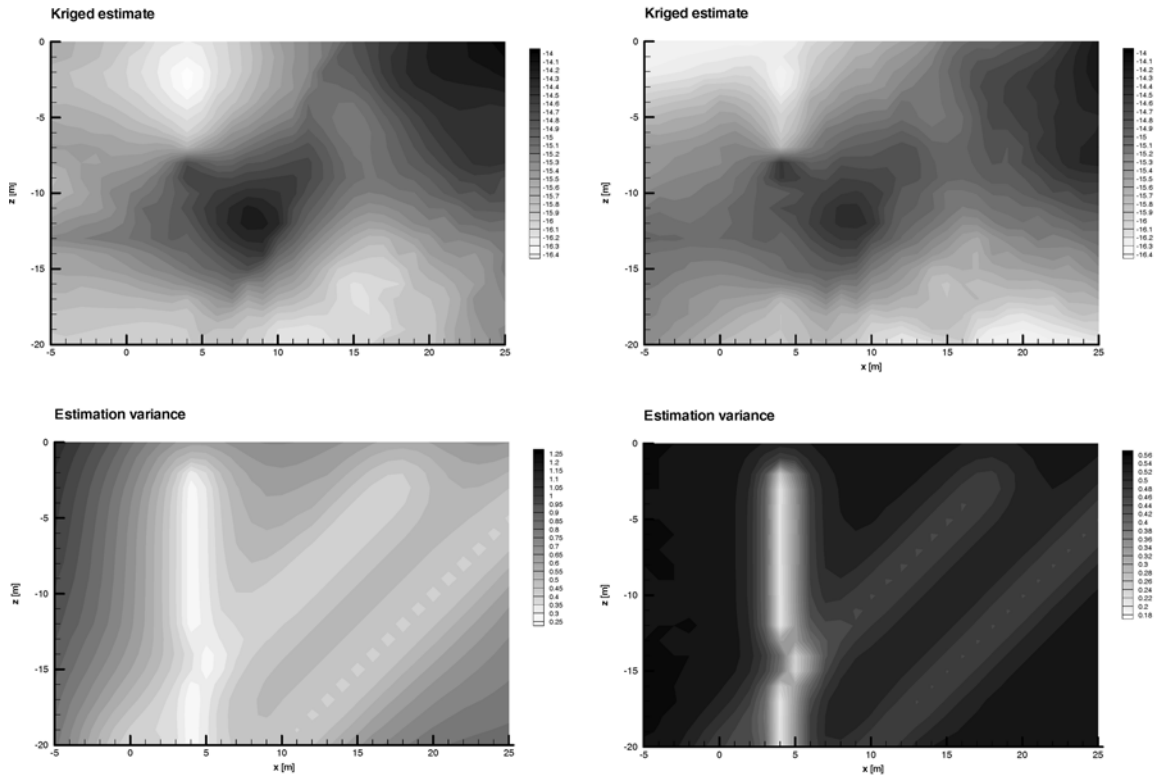

Fig. 2. Kriged estimates and variance of base-10 log permeability at $y = 7$ m using power model (left) and exponential model with second order drift (right). After Chen et al. (2000)

Table 1. Discrimination among log permeability variogram models. After Chen et al. (2000)

Drift Model	No Drift	1st order polynomial	2nd order polynomial
NLL	665.8	665.1	655.8
Variogram model of residuals	Power	Exponential	Exponential
Number of parameters	2	6	12
Variance (scaling coefficient for power model)	0.2715	0.5807	0.495
Integral scale (power for power model)	0.4475	1.665	1.2602
AIC (Rank)	669.8 (1)	677.1 (2)	679.8 (3)
BIC (Rank)	677.2 (1)	696.4 (2)	718.4 (3)
HIC (Rank)	672.4 (1)	684.9 (2)	695.5 (3)
KIC (Rank)	680.0 (1)	690.1 (2)	700.9 (3)

which nevertheless reflects adequately the underlying structure of the rock and the corresponding flow and transport regime.

Table 1 shows that, whereas the exponential variogram model with a quadratic drift fits the data best (as measured and implied by the smallest negative log likelihood model fit criterion, NLL), all four model discrimination criteria (AIC, BIC, HIC, KIC) consistently rank the power model as best, and the former model as least acceptable. The reason is that whereas all three models fit the data almost equally well, the power model is most parsimonious with only two parameters, and the exponential variogram model with second-order drift is least parsimonious with twelve parameters.

This consistent ranking notwithstanding, the model discrimination criteria in Table 1 are very close to each other and it is therefore not clear that one variogram model must be preferred over the others. An alternative approach would be to consider all three, and perhaps additional, variogram models jointly in the context of MLBMA. Work along these lines is underway.

6 Conclusion

Bayesian model averaging (BMA) provides an optimal way to combine the predictions of several competing conceptual-mathematical models and to assess their joint predictive uncertainty. It can be made computationally feasible by basing it on a maximum likelihood approximation due to Kashyap (1982) and the parameter estimation method of Carrera and Neuman (1986a). The resulting maximum-likelihood Bayesian model averaging approach (MLBMA) incorporates both site characterization and site monitoring data in a way that bases the outcome on an optimum combination of prior information (scientific knowledge plus data) and model predictions.

References

Akaike H (1977) On entropy maximization principle. In: Applications of Statistics, Krishnaiah PR (ed.), pp. 27–41. North Holland, Amsterdam, The Netherlands

Beven KJ (1993) Prophecy, reality and uncertainty in distributed hydrological modelling. Adv. Water Resour. 16: 41–51

Beven KJ (2000) Uniqueness of place and non-uniqueness of models in assessing predictive uncertainty. In: Computational Methods in Water Resources XIII, Bentley LR, Sykes JF, Gray WG, Brebbia CA, Pinder GF (eds.), pp. 1085–1091. Balkema, Rotterdam, The Netherlands
Beven KJ, Binley AM (1992) The future of distributed models: model calibration and uncertainty prediction. Hydrological Processes 6: 279–298
Beven KJ, Freer J (2001) Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. J. Hydrology 249: 11–29
Burnham KP, Anderson DR (1998) Model Selection and Inference, Springer, New York
Carrera J, Neuman SP (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, Neuman SP (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
Carrera J, Medina A, Axness C, Zimmerman T (1997) Formulations and computational issues of the inversion of random fields. In: Subsurface Flow and Transport: A Stochastic Approach, Dagan G, Neuman SP (eds.), pp. 62–79. Cambridge University Press, Cambridge, United Kingdom
Christakos G (2000) Modern Spatiotemporal Geostatistics, Oxford University Press, New York
Christakos G (2002a) On the assimilation of uncertain physical knowledge bases: Bayesain 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, Kovar K and Hrkal Z, IAHS Publ. No. 277–19
Christakos G (2003a) Critical conceptualism in environmental modeling and prediction. Envir. Sci. and Technol. (in press)
Christakos G (2003b) Multi-Disciplinary Systems in Uncertain Environments, Springer-Verlag, New York, NY (in press)
Dagan G, Neuman SP (1997) Subsurface Flow and Transport: A Stochastic Approach. Cambridge University Press, Cambridge, United Kingdom
Draper D (1995) Assessment and propagation of model uncertainty. J. Roy. Statist. Soc. Ser. B 57: 45–97
Draper D (1999) Comment. Statist. Sci. 14(4): 405–409
Gaganis P, Smith L (2001) A Bayesian approach to the quantification of the effect of model error on the predictions of groundwater models. Water Resour. Res. 37(9): 2309–2322
Gauch HG Jr. (1993) Prediction, parsimony and noise, American Scientist 81: 468–478
George EI (1999) Comment. Statist. Sci. 14(4): 409–412
Gómez-Hernández JJ, Sahuquillo A, Capilla JE (1997) Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data, 1. Theory. J. Hydrol. 203: 162–174
Guadagnini A, Neuman SP (1999) Nonlocal and localised analyses of conditional mean steady state flow in bounded, randomly nonuniform domains, 2, Computational examples. Water. Resour. Res. 35(10): 3019–3039
Hernandez AF, Neuman SP, Guadagnini A, Carrera J (2003) Conditioning mean steady state flow on hydraulic head and conductivity through geostatistical inversion. Stochastic Environ. Res. Risk Assess., 17(5): 329–338
Hill CM (1998) Methods and Guidelines for Effective Model Calibration. U.S. Geological Survey Water-Resources Investigations Report 98-4005, Denver, Colorado
Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: A tutorial. Statist. Sci. 14(4): 382–417
Hornberger GM, Speer RC (1981) An approach to the preliminary analysis of environmental systems. J. Environ. Mgmt. 12: 7–18
James AL, Oldenburg CM (1997) Linear and Monte Carlo uncertainty analysis for subsurface contaminant transport simulation. Water Resour. Res. 33(11): 2495–2508
Kashyap RL (1982) Optimal choice of AR and MA parts in autoregressive moving average models. IEEE Trans. Pattern Anal. Mach. Intel. PAMI 4(2): 99–104

Madigan D, Raftery AE (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. J. Amer. Statist. Assoc. 89: 1535–1546
National Research Council (1999) Groundwater & Soil Cleanup. Nat. Acad. Press, Washington, DC
National Research Council (2001) Conceptual Models of Flow and Transport in the Vadose Zone. Nat. Acad. Press, Washington, DC
Neuman SP, Wierenga PJ (2003) A Comprehensive Strategy of Hydrogeologic Modeling and Uncertainty Analysis for Nuclear Facilities and Sites, NUREG/CR-6805, prepared for U.S. Nuclear Regulatory Commission, Washington, DC
Raftery AE (1993) Bayesian model selection in structural equation models. In: Testing Structural Equation Models Bollen K, Long J (eds.), pp. 163–180. Sage, Newbury Park, California
Raftery AE, Madigan D, Volinsky CT (1996) Accounting for model uncertainty in survival analysis improves predictive performance. In: Bayesian Statistics, Bernardo J, Beerger J, Dawid A, Smith A (eds.), pp. 323–349. Oxford University Press, United Kingdom
Rissanen J (1978) Modeling by shortest data description. Automatica 14: 465–471
Samper J, Molinero J (2000) Predictive uncertainty of numerical models of groundwater flow and solute transport. In: AGU 2000 Fall Meeting Program, Washington, DC, F500
Schwarz G (1978) Estimating the dimension of a model. Ann. Statist. 6: 461–464
Taplin RH (1993) Robust likelihood calculation for time series. J. Roy. Statist. Soc. Ser. B 55: 829–836
Volinsky CT, Madigan D, Raftery AE, Kronmal RA (1997) Bayesian model averaging in proportional hazard models: assessing the risk of a stroke. J. Roy. Statist. Soc. Ser. C 46: 433–448
Ye M, Neuman SP, Guadagnini A, Tartakovsky DM (2002) Nonlocal and localized analyses of conditional mean transient flow in bounded, randomly nonuniform domains. In: Computational Methods in Water Resources XIV, Hassanizadeh M, Schotting RJ, Gray WG, Pinder GF, Brebbia CA (eds.), pp. 1155–1162. Elsevier, Amsterdam, The Netherlands
Zhang D (2001) Stochastic Methods for Flow in Porous Media. Academic Press
Zimmerman DA, De Marsily G, Gotway CA, Marietta MG, Axness CL, Beauheim RL, Bras R, Carrera J, Dagan G, Davies PB, Gallegos DP, Galli A, Gómez-Hernández J, Gorelick SM, Grindrod P, Gutjahr AL, Kitanidis PK, Lavenue AM, McLaughlin D, Neuman SP, Rama Rao BS, Ravenne C, Rubin Y (1998) A comparison of seven geostatistically-based inverse approaches to estimate transmissivities for modeling advective transport by groundwater flow. Water Resour. Res. 34(6): 1373–1413