Accounting for conceptual model uncertainty via maximum likelihood Bayesian model averaging

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

Abstract Analyses of groundwater flow and transport typically rely on a single conceptual model of site hydrogeology. Yet hydrogeological environments are open and complex, rendering them prone to multiple interpretations. Adopting only one of these may lead to statistical bias and underestimation of uncertainty. A comprehensive strategy for constructing alternative conceptual-mathematical models, selecting the best among them, and using them jointly to render optimum predictions under uncertainty is being developed by the author. This paper proposes a maximum likelihood Bayesian model averaging approach, MLBMA, to rendering optimum predictions by means of several competing models and assessing their joint predictive uncertainty.

Key words Bayesian; conceptual models; maximum likelihood; model uncertainty; predictive uncertainty

INTRODUCTION

Analyses of groundwater flow and transport typically rely on a single conceptual model of site hydrogeology. Yet hydrogeological environments are open and complex, rendering them prone to multiple interpretations. This is true regardless of the quantity and quality of available site data. Focusing on interpretation of only one site may lead to Type I model errors, which 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 framework. Indeed, critiques of hydrogeological analyses, and legal challenges to such analyses, typically focus on the validity of the underlying conceptual 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 hydrogeological 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 undersampling of the relevant model space). The bias and uncertainty that result from reliance on an inadequate conceptual model are typically much larger than those introduced through an inadequate choice of model parameter values. Yet most uncertainty analyses of flow and transport ignore the former and focus exclusively on the latter. This often leads to overconfidence in the predictive capabilities of the model, which the available site data do not justify.

It is argued by Beven & Freer (2001) that, “given current levels of understanding and measurement technologies, it may be endemic to mechanistic modelling 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 & 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 (Beven, 2000) to 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 models of site hydrogeology, select the best among them, use them jointly to render optimum predictions of groundwater flow and transport, and assess the uncertainty of such predictions. A comprehensive strategy for doing so is being developed by the author. This paper focuses on one aspect of the strategy, which addresses the question how to render optimum predictions by means of several competing deterministic or stochastic models and how to assess their joint predictive uncertainty.

**PREDICTION AND UNCERTAINTY ANALYSIS BASED ON SINGLE MODEL**

**Deterministic model**

The traditional approach to hydrological 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 sufficient site parameter values are available, 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 off-site parameter measurements and/or surrogate data. Statistics derived from off-site data are potentially biased due to a lack of site-specific information about mean parameter values and incompatibility of geology and 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 behaviour by means of a suitable inverse method. The last 30 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 analysing 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) groundwater inverse model (Hill, 1998).

Stochastic model

Hydrogeological medium properties exhibit both systematic and random spatial variations on a multiplicity of scales. Traditional deterministic models capture at best the larger-scale, systematic components of these variations. However, they 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 hydrogeological medium properties geostatistically, and to analyse subsurface fluid flow and solute transport stochastically. This trend has been documented in a number of recent books including those by Dagan & 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 non-unique results are summarized in terms of statistically averaged quantities, their variance-covariance, and perhaps higher moments of the corresponding sample probability distributions. Results that honour 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 system monitoring data, 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 hydrogeological properties, 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 groundwater flow and transport variables directly, without Monte Carlo simulation. This is done on the basis of moment equations conditional on hydraulic parameter measurements as illustrated for example by Guadagnini & Neuman (1999) and Ye et al. (2002). Conditioning additionally on observed hydraulic head values requires an inverse procedure (Hernandez et al., 2002).

PREDICTION AND UNCERTAINTY ANALYSIS BASED ON MULTIPLE MODELS

Previous hydrological approaches

Carrera & Neuman (1986b) have noted that an inadequate model structure (conceptualization) 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 modelling 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 modelling 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 modelling endeavour, 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.

We have 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 hydrological 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 modelling bias and underestimation of model uncertainty.
An example of how one could account quantitatively for structural model uncertainties was given by James & Oldenburg (1997). They investigated the uncertainty of simulated tetrachloroethane (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) non-isothermal 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 & 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 & 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 evaluation of 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 & Binley (1992) proposed a strategy which they refer to 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 hydrological behaviours and to compare the results with monitored system behaviour during the same period. Likelihood measures are defined to gauge the degree of correspondence between each simulated and observed record of system behaviour. 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 behaviour 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 & Freer (2001).
A Bayesian approach to the quantification of errors in a single groundwater model was recently proposed by Gaganis & 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 hydrological predictions in an absolute sense, only in a conditional or relative sense.

Proposed approach

Bayesian model averaging At the heart of the approach we 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 in a quantity one wants to predict, then its posterior distribution given a discrete set of data is:

$$p(A|D) = \sum_{k=1}^{K} p(M_k|D) p(A|M_k)$$

where is the set of all models (or hypotheses) considered, at least one of which must be valid. In other words, is the average of the posterior distributions under each model, weighted by their posterior model probabilities . The posterior probability for model is given by Bayes’ rule:

$$p(M_k|D) = \frac{p(D|M_k) p(M_k)}{\sum_{i=1}^{K} p(D|M_i) p(M_i)}$$

where:

$$p(D|M_k) = \int p(D|\theta_k, M_k) p(\theta_k|M_k) d\theta_k$$

is the integrated likelihood of model , is the vector of parameters associated with model , is the prior density of under model , is the joint likelihood of model and its parameters , and is the prior probability that is the correct model. All probabilities are implicitly conditional on \( \mathcal{M} \).
The posterior mean and variance of $\Delta$ are (Draper, 1995):

$$E[\Delta|D] = \sum_{i=1}^{k} E[\Delta|D, M_i] p(M_i|D)$$

(4)

$$\text{var}[\Delta|D] = \sum_{i=1}^{k} \left[ \text{var}[\Delta|D, M_i] + E[\Delta|D, M_i]^2 \right] p(M_i|D) - E[\Delta|D]^2$$

(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 equation (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 & 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, we adopt a suggestion by Taplin (1993) that $p(\Delta|M_k, D)$ in equation (1) be approximated by $p(\Delta|M_k, \hat{\theta}_k, D)$ where $\hat{\theta}_k$ is the maximum likelihood estimate of $\theta_k$ based on the likelihood $p(D|\theta_k, M_k)$. 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 $\hat{\theta}_k$ by calibrating a deterministic model $M_k$ against hydrogeological data $D$, which may include prior information about the parameters, are described by Carrera & 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. (2002). These methods also yield an approximate covariance matrix for the estimation errors of $\hat{\theta}_k$. Upon considering the parameter estimation errors of a calibrated deterministic model $M_k$ to be Gaussian or log Gaussian, one easily determines $p(\Delta|M_k, \hat{\theta}_k, D)$ by Monte Carlo simulation of $\Delta$ through random perturbation of the parameters. The simulation also yields corresponding approximations $E(\Delta|M_k, \hat{\theta}_k, D)$ of $E[\Delta|M_k, D]$, and $\text{var}(\Delta|M_k, \hat{\theta}_k, D)$ of $\text{var}[\Delta|M_k, D]$, in equations (4) and (5). If $M_k$ is a stochastic model based on moment equations, it can yield $E(\Delta|M_k, \hat{\theta}_k, D)$ and $\text{var}(\Delta|M_k, \hat{\theta}_k, D)$ directly without Monte Carlo simulation (Hernandez et al., 2002).
To eliminate the need for computer intensive integration according to equation (3), I propose to evaluate the weights $p (M_i | D)$ in equations (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:

$$\ln p (M_i | D) = \ln C_i + \ln p (D | \bar{\theta}_i, M_i) + \ln p (\bar{\theta}_i | M_i) + \frac{N_i}{2} \ln \left( \frac{2\pi}{N} \right)$$

$$- \frac{1}{2} \ln | F_{i,ij} (D | \bar{\theta}_i, M_i) | + R(N)$$

where $C_i = c p (M_i)$, $p (M_i)$ being the prior probability of model $M_i$ and $c$ a constant that can be evaluated from:

$$\sum_{i=1}^{K} p (M_i | D) = 1$$

$N_i$ is the dimension of $\Theta_i$, $N$ is the dimension of $D$, $F_{i,ij}$ is the normalized (by $N$) observed (as opposed to ensemble mean) Fisher information matrix having components:

$$F_{i,ij} = - \frac{1}{N} \frac{\partial^2 \ln p (D | \theta_i, M_i)}{\partial \theta_i \partial \theta_j}$$

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

Kashyap (1982) suggests that, in the absence of any contrary information, the models be assigned equal prior probabilities, yielding $C_i = C = \text{constant}$ 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 equation (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 criterion:

$$\hat{d}_k = - \ln p (D | \bar{\theta}_k, M_k) - \ln p (\bar{\theta}_k | M_k) - \frac{N_k}{2} \ln \left( \frac{2\pi}{N} \right) + \frac{1}{2} \ln | F_{i,ij} (D | \bar{\theta}_i, M_i) |$$
Increasing the number of parameters $N_k$ allows $-\ln p\left(D|\hat{\theta}_k, M_k\right)$ 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 $p\left(M_k|D\right)$ in equation (6) diminishes. This means that a more parsimonious model with fewer parameters is preferred by equation (9) and assigned a higher probability by equation (6). On the other hand, $-\ln p\left(D|\hat{\theta}_k, M_k\right)$ 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 equation (9) reduces the relative emphasis on model fit as the information content of the data diminishes. As illustrated by Carrera & 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:

$$\ln p\left(D|M_k\right) = \ln p\left(D|\hat{\theta}_k, M_k\right) - \frac{N_k}{2} \ln N$$

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

To my knowledge, the non-asymptotic 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 equation (6) in BMA computationally feasible, I propose to follow the approach of Carrera & Neuman (1986a) who incorporate in $D$ 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 $\ln p\left(\hat{\theta}_k|M_k\right)$ into the loglikelihood function $-\ln p\left(D|\hat{\theta}_k, M_k\right)$ and to compute $F_k\left(D|\hat{\theta}_k, M_k\right)$ in a straightforward manner (also Carrera et al., 1997).

CONCLUSION

Bayesian model averaging 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 & Neuman (1986a).

Acknowledgement This work was supported by the US Nuclear Regulatory Commission under contract NRC-04-97-056. Special credit is due to Thomas J. Nicholson and Ralph Cady of the US NRC for having conceived and championed this research. I am most grateful to Weihsueh Chiu of the US 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.

REFERENCES


Accounting for conceptual model uncertainty via maximum likelihood Bayesian model averaging 313


