Ecological Modelling, 62 (1992) 149–162
Elsevier Science Publishers B.V., Amsterdam

Development of Bayesian Monte Carlo techniques for water quality model uncertainty

David W. Dilks [a], Raymond P. Canale [b] and Peter G. Meier [c]

[a] *Limno-Tech, Inc., 2395 Huron Parkway, Ann Arbor, MI 48104, USA*
[b] *Civil Engineering Department, University of Michigan, Ann Arbor, MI 48109, USA*
[c] *School of Public Health, University of Michigan, Ann Arbor, MI 48109, USA*

(Accepted 11 December 1990)

ABSTRACT

Dilks, D.W., Canale, R.P. and Meier, P.G., 1992. Development of Bayesian Monte Carlo techniques for water quality model uncertainty. *Ecol Modelling*, 62: 149–162.

A new technique, Bayesian Monte Carlo (BMC), is used to quantify errors in water quality models caused by uncertain parameters. BMC also provides estimates of parameter uncertainty as a function of observed data on model state variables. The use of Bayesian inference generates uncertainty estimates that combine prior information on parameter uncertainty with observed variation in water quality data to provide an improved estimate of model parameter and output uncertainty. It also combines Monte Carlo analysis with Bayesian inference to determine the ability of random selected parameter sets to simulate observed data. BMC expands upon previous studies by providing a quantitative estimate of parameter acceptabilty using the statistical likelihood function. The likelihood of each parameter set is employed to generate an n-dimensional hypercube describing a probability distribution of each parameter and the covariance among parameters. These distributions are utilized to estimate uncertainty in model predictions. Application of BMC to a dissolved oxygen model reduced the estimated uncertainty in model output by 72% compared with standard Monte Carlo techniques. Sixty percent of this reduction was directly attributed to consideration of covariance between model parameters. A significant benefit of the technique is the ability to compare the reduction in total model output uncertainty corresponding to: (1) collection of more data on model state variables, and (2) laboratory or field studies to better define model processes. Limitations of the technique include computational requirements and accurate estimation of the joint probability distribution of model errors. This analysis was conducted assuming that model error is normally and independently distributed.

Correspondence to: D.W. Dilks, Limno-Tech, Inc., 2395 Huron Parkway, Ann Arbor, MI 48104, USA.

INTRODUCTION

Mathematical models of aquatic systems are being used with increasing frequency to predict the effect of alternative pollution control strategies. It is now recognized that such models can have a large degree of uncertainty associated with their projections, and that this uncertainty can significantly impact the utility of the model predictions.

Model uncertainty is caused by a combination of two factors (Burges and Lettenmaier, 1975). First, water quality models are simplistic representations of the real world. Error can be introduced by using a model framework that incompletely describes the true system. The second type of error is caused by uncertain or incorrect model parameters. This error arises because it is not possible to determine model parameters exactly.

Considerable research has focused upon quantifying errors caused by uncertain model parameters. One of the more popular techniques for quantifying this error has been Monte Carlo analysis. With Monte Carlo analysis, the uncertainty in model parameters is represented by statistical frequency distributions. Models are run for several iterations, with the uncertain parameter values for each iteration being randomly selected from their pre-specified distributions. Tabulation of model output for each iteration allows construction of a frequency distribution for any model output variable. The primary limitation of Monte Carlo analysis involves the lack of information available for specifying frequency distributions for uncertain model parameters. Often, little or no site-specific data are available that describe model parameters. The only information typically available for many parameters is a range of values obtained from published studies. To compound this difficulty, Monte Carlo analysis also requires specification of the covariance structure among uncertain model parameters.

This paper describes a technique which combines Monte Carlo analysis with Bayesian inference to overcome the problems associated with specifying model input parameter distributions. The technique, termed Bayesian Monte Carlo (BMC), uses Monte Carlo analysis to sample from preliminary estimates of parameter distributions. The statistical likelihood function is employed to evaluate the ability of any given set of model parameters to describe observed data on model state variables. Preliminary (prior) information on parameter distributions is combined with measurements of state variables to provide improved estimates of parameter distributions. This technique directly accounts for covariance among uncertain parameters by storing the parameter distributions in an n-dimensional hypercube. For the purpose of this paper, the technique is applied to a dissolved oxygen model of the Grand River near Grand Rapids, MI.

METHODOLOGY DEVELOPMENT

The original concept for the Bayesian Monte Carlo was first described by Hornberger and Spear (1980) and Spear and Hornberger (1980), and was developed as a technique to identify research needs in the absence of extensive data. Spear and Hornberger's approach to the problem defined acceptable ranges for each parameter based upon literature values, and then characterized each parameter as a uniform statistical distribution over the specified range. Next, they conducted a Monte Carlo-type analysis by randomly selecting parameter values from these distributions and performing model simulations. The authors qualitatively compared the output of each model calculation with the expected response of the system. The parameter values for each simulation were then stored in one of two matrices, depending upon whether or not the particular model simulation was able to qualitatively describe the behavior of the data.

Fedra (1980) stated that the matrix of parameter values that satisfactorily approximates the system behavior serves to define the best understanding of each parameter. If a wide range of parameter values all led to satisfactory simulation of system behavior, this indicated that little information was gained on the true parameter value. Fedra also expanded this technique from a method of determining the uncertainty in model inputs to a method of predicting the uncertainty in model results. He accomplished this by performing a second set of simulations using as inputs only those parameter sets that satisfactorily simulated the system behavior. Tabulation of these results provided a description of the uncertainty in model projections based upon the uncertainty in model inputs. A major advantage of this technique is that it automatically accounts for covariance between model input parameters by selecting whole parameter sets that satisfactorily simulated the behavior.

Statistical basis

Although the above technique is intuitively pleasing, its underlying statistical basis had not been addressed. The method follows a Bayesian statistical approach for improving preliminary estimates of parameter distributions. Bayes Theorem can be used to obtain an improved estimate of parameter distributions by mathematically combining previously known general information about those parameters with site-specific measured field data that describe system behavior. Bayes Theorem can be interpreted for a model with a single uncertain parameter as (Benjamin and Cornell, 1970):

$$P(\theta \mid x) = \frac{P(x \mid \theta) P(\theta)}{P(x)} \tag{1}$$

where θ is the uncertain model parameter, x the observed data on model state variables, $P(\theta|x)$ the probability of the parameter value being accurate given the observed data (aka "posterior" distribution), $P(\theta)$ the preliminary estimate for the probability of each parameter value being accurate, $P(x|\theta)$ the probability of data occurring, given that the value of i is accurate ("likelihood"), and $P(x)$ the probability of the data occurring.

Equation (1) provides the probability that any given parameter value, θ, is accurate based upon: (i) the likelihood of the data occurring, given that parameter value (i.e. how well the parameter allows the model to describe the data); and (ii) a prior assessment of the probability of a particular parameter value occurring.

The term $P(x)$ in equation (1) represents the joint likelihood of the data and parameter occurring over all values of θ and will be constant for all values of θ. This term is dropped from consideration, because equation (1) is being used only to calculate the relative likelihoods for each value of θ. This leads to:

$$\underset{\substack{\text{Posterior (improved)}\\ \text{estimate of parameter}\\ \text{probability}}}{P(\theta|x)} = c * \underset{\substack{\text{Likelihood of}\\ \text{the parameter value}\\ \text{given the observed data}}}{P(x|\theta)} * \underset{\substack{\text{Prior (preliminary)}\\ \text{estimate of parameter}\\ \text{probability}}}{P(\theta)} \qquad (2)$$

where c is the normalizing constant.

Equation (2) provides an improved probability estimate for each parameter value as a function of the observed data and the prior assessment of the parameter's probability.

The technique provides the benefit of Bayesian inference in that information from two separate sources is combined to provide an improved estimate of the true parameter value. Information on site-specific observed data is combined with prior information on parameter distributions. Bayesian theory states that the posterior distribution will contain less uncertainty than either of the two sources used in its determination. Therefore, high quality field data combined with little prior knowledge of parameter distributions will result in posterior distributions based primarily on the field data, but improved by whatever prior information was available. If strong prior information is combined with poorer field data, the posterior distribution will primarily reflect the prior distribution.

The analysis described in equation (2) is generalized in BMC to simultaneously consider the uncertainty in any number (n) of parameters. Monte Carlo analysis is used to repeat this calculation over the entire n-dimensional range of parameters. Prior probabilities for each joint set of parameters are not explicitly calculated, but are incorporated using Monte Carlo analysis by sampling for all values in direct proportion to their (prior)

assumed frequency of occurrence. This allows improved probability distributions (and covariance structures) to be generated for all uncertain parameters. The resulting matrix can be analyzed to generate improved (posterior) marginal distributions for each uncertain parameter, and can also be used as input to a second analysis to determine the uncertainty in model predictions.

Calculation of the likelihood

Monte Carlo application of equation (2) provides an improved estimate of parameter probability distributions to be obtained from: (1) prior estimates $[P(\theta)]$; and (ii) the likelihood of each parameter value, given the observed data $[P(x \mid \theta)]$. The work of Spear and Hornberger (1980) and Fedra (1980) both qualitatively estimated the likelihood through criteria which determined whether or not a given parameter set was capable of describing the observed data; for their application, $P(x \mid \theta) = 1$ or $P(x \mid \theta) = 0$. This qualitative estimate is a potential limitation of their technique for more rigorous application, due to the subjective and arbitrary nature in which the criteria are determined.

The term $P(x \mid \theta)$ in equation (2) is called the sample likelihood function, $L(\theta \mid x)$. It provides the likelihood of having observed the data given that the parameter values in θ are correct. The likelihood function can be calculated for independent variables as (Benjamin and Cornell, 1970):

$$L(\theta \mid x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} f_x(x_i \mid \theta) \tag{3}$$

where $f_x(x_i \mid \theta)$ is the probability density function of x, given θ.

The likelihood function can be directly calculated for water quality model application, as long as the probability density function of the observed data can be defined. To allow application of the technique, the assumption is made that the errors in the data are normally and independently distributed with a mean of zero. The error term is defined as;

$$e_i = x_i - u_i \tag{4}$$

where e_i is the model error for data point i, x_i the value of data point i, and u_i the model prediction at data point i.

The probability density function at each individual data point x_i (error value e_i) is easily defined, and results in a calculation of the likelihood function of:

$$L(\theta \mid x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{e_i}{\sigma}\right)^2\right] \tag{5}$$

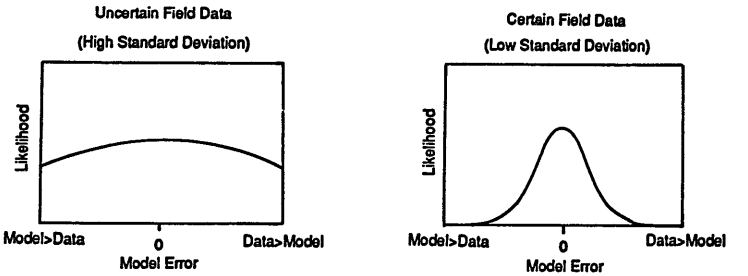

Fig. 1. Likelihood distribution at different levels of uncertainty in observed data.

where n is the number of observed data point, and σ the standard deviation for the data error.

Equation (5) allows the likelihood function to be calculated directly from the model output and the observed data. The likelihood is seen to vary not only as a function of the data error, but also as a function of the number of data values and the standard deviation of the data error. Figure 1 illustates the dependence of the likelihood function on the standard deviation of the data error. With highly uncertain field data, likelihood values remain relatively constant over a wide range of data error. As the standard deviation decreases, the likelihood value decreases sharply as the relative error becomes non-zero. This is consistent with the benefit of Bayesian inference, in that data with greater certainty have a stronger impact on the posterior distribution than highly uncertain data.

Determining error variance

The term σ in equation (5) represents the expected deviation of the data caused by imperfect measurement techniques. This number can be determined in one of two fashions. First, the standard deviation may be known a priori based upon previous statistical analysis of field sampling and/or laboratory measurement error. The standard deviation of the data error can also be estimated during the Monte Carlo analysis using the maximum likelihood theory. Using this method, the standard deviation of the data error is estimated as the minimum value of the standard deviation across all iterations.

Multiple state variables

The proposed technique is relatively straightforward for cases where only a single state variable is being simulated. This is rarely the case in

water quality models, which typically simulate several variables simultaneously. For a model with m state variables, each variable will have its own error variance term σ_j. The likelihood function thus becomes:

$$L(\theta \mid e_1, e_2, \ldots e_n) = \prod_{i=1}^{n} \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi\sigma_j}} \exp\left[-\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\frac{e_{ij}}{\sigma_j} \right)^2 \right] \quad (6)$$

where e_{ij} is the error term for the ith prediction of variable j.

Again, the σ_j can be specified or can be estimated as the minimum standard deviation for each parameter across all simulations. This requires an additional assumption that errors not be correlated across state variables. Correlated errors will require consideration of the covariance between variables. This will cause the complexity of the mathematics to increase, but the technique will still remain applicable.

APPLICATION OF BMC

The Bayesian Monte Carlo technique was applied to a model developed to evaluate the dissolved oxygen impacts of combined sewer overflows on the Grand River near Grand Rapids, Michigan (Limno-Tech, Inc., 1982). The goal was to determine if combined sewer overflows would lead to violation of the local dissolved oxygen standard. The model used is one-dimensional and purely advective and has four state variables: dissolved oxygen, two forms of carbonaceous biochemical oxygen demand (CBOD), and nitrogenous biochemical oxygen demand (NBOD). Water quality data are available that describe the concentration of these variables at five locations in the river. The model begins at the point where the combined sewer overflows enter the river. The available data indicate that combined sewer overflows have a significant impact on CBOD and NBOD concentrations, but are less clear as regarding dissolved oxygen impacts.

Dissolved oxygen concentrations are modeled as a function of reaeration, deoxygenation, nitrification, and net community productivity (algae and sediments). The differential equation for dissolved oxygen over time (or distance) is:

$$\mathrm{d[DO]}/\mathrm{d}t = k_a \times (\mathrm{CS} - \mathrm{[DO]}) - k_d \times \mathrm{[CBOD]} - k_n \times \mathrm{[NBOD]} + P \quad (7)$$

where [DO] is the dissolved oxygen concentration (mg/l), t the time of travel, x/u (days), x the distance downstream (miles), u the average river velocity (miles/day), k_a the reaeration rate (1/day), CS the saturation concentration of dissolved oxygen (mg/l), k_d the CBOD deoxygenation rate (1/day), [CBOD] the total CBOD concentration (mg/l), k_n the

nitrification rate (1/day), [NBOD] the NBOD concentration (mg/l), and P the net community productivity (mg/l/day).

CBOD was divided into two components for modeling purposes: a particulate portion ($CBOD_p$) to represent settleable sources of CBOD associated with the combined sewer overflows, and a dissolved portion ($CBOD_d$) incapable of settling. The differential equation for each type of CBOD is:

$$d[CBOD_p]/dt = (k_s + k_d) \times [CBOD_p] \quad (8)$$

$$d[CBOD_d]/dt = k_d \times [CBOD_d] \quad (9)$$

where k_s is the particulate settling rate (1/day).

Field data are available on total CBOD only, so $CBOD_d$ and $CBOD_p$ are combined for comparison with data.

NBOD was defined as 4.57 times the total ammonia concentration. The only kinetic process affecting NBOD was first-order loss due to nitrification. The differential equation for NBOD is:

$$d[NBOD]/dt = -k_n \times [NBOD] \quad (10)$$

This model requires specification of a total of nine spatially constant uncertain parameters and forcing functions. The uncertain parameters and forcing functions are upstream concentration of dissolved oxygen, $CBOD_p$, $CBOD_d$ and NBOD; reaeration rate, CBOD deoxygenation rate, nitrification rate, net productivity, and particulate settling rate.

RESULTS

BMC was applied to the Grand River dissolved oxygen model to determine frequency distributions for the nine uncertain model parameters. These parameter distributions were subsequently used to predict the overall uncertainty in model projections. The model output variable used to measure this uncertainty was the predicted minimum dissolved oxygen concentration. Uniform distributions were employed to initially describe each parameter, because little prior information was available. The range for each parameter was selected to cover all believable values.

Insufficient observed data were available to estimate the error variance (standard deviation) for any of the model state variables. Values were selected based upon past experience. The standard deviations selected were 2 mg/l for dissolved oxygen, 1.0 mg/l for NBOD and 25 mg/l for CBOD.

Selected marginal posterior distributions calculated by the Bayesian Monte Carlo technique for 100 000 iterations are shown in Figs. 2 and 3

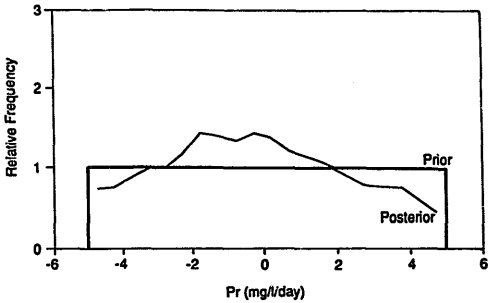

Fig. 2. Marginal distributions for net productivity.

and are compared with the initial (prior) uniform distributions. Figure 2 shows the posterior distribution for net productivity. It was found to have a slightly negative impact on dissolved oxygen concentrations on the average, with a mean posterior value of -1.0 mg/l/day. The posterior distribution for the nitrification rate (Fig. 3) was significantly different from the prior distribution. The nitrification rate remained within the originally specified range of 0–2 day, with a value of 0.9/day three times more probable than originally assumed.

The posterior distributions determined during the first step of the BMC procedure were then used to calculate the overall uncertainty in model predictions. The Monte Carlo method can provide a frequency distribution for every location of model output; however, results are more easily examined if a single measure of model output is chosen. Minimum dis-

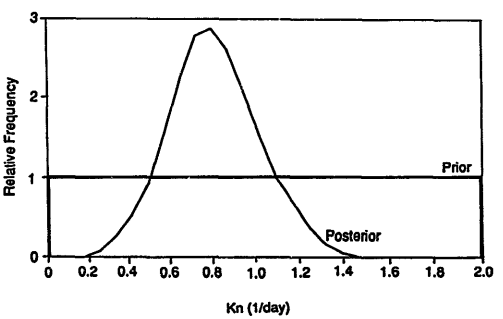

Fig. 3. Marginal distributions for nitrification rate.

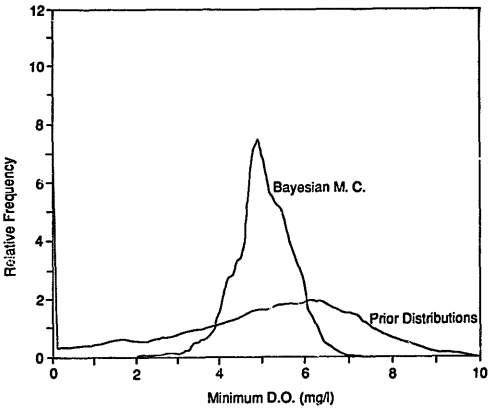

Fig. 4. Predicted frequency distribution for minimum dissolved oxygen: BMC vs. prior distributions.

solved oxygen is utilized to represent model output in the uncertainty calculations, because it is the single most important model result. The frequency distribution for minimum dissolved oxygen is shown in Fig. 4, having a mean value of 5.1 mg/l and a standard deviation of 0.69. Application of BMC to the observed data translated into a significant reduction in variability from the results predicted by the initial distributions alone. Using only the initial distributions for all input parameters resulted in the second distribution shown in Fig. 4. This distribution has a mean of 4.6 mg/l and a standard deviation of 2.5 mg/l.

Conclusions

Results of this analysis indicate that measured values of the state variables, in conjunction with the model, provided sufficient information to determine posterior distributions that were significantly different from the prior assumptions. This, in turn, indicates that each of the nine parameters was significantly correlated to the ability of the model to simulate the measured data.

BMC was able to reduce the uncertainty (as measured by the standard deviation) in the predicted minimum dissolved oxygen concentration by 72%, from 2.5 to 0.69 mg/l. This three-fold change in model output variance clearly demonstrates the advantages of explicitly considering the ability of each random Monte Carlo selection to simulate the measured data.

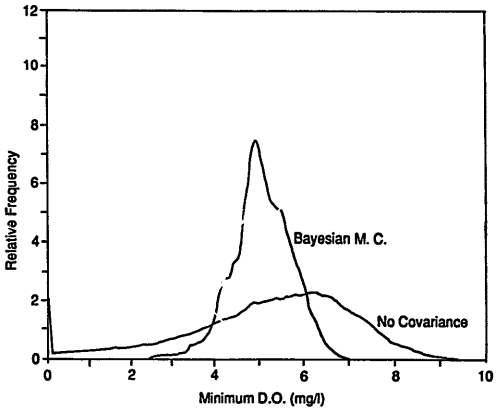

Fig. 5. Predicted frequency distribution for minimum dissolved oxygen, BMC vs. no covariance.

An interesting sidelight to these results was the large effect that covariance among input parameters had on model output uncertainty. A separate analysis of model output uncertainty was conducted using the marginal posterior distributions generated by BMC, but ignoring covariance among parameters. The results of this analysis (Fig. 5) indicate a standard deviation of 1.8 mg/l for the marginal distributions case. The majority (60%) of the reduction in model output uncertainty gained from BMC resulted from use of covariance between input parameters. Clearly, Monte Carlo analyses which ignore the covariance between model parameters may provide estimates of model uncertainty which are very misleading.

DISCUSSION

Application of BMC to the Grand River water quality model provided several insights into the practical application of the BMC method, and to its applicability for model error analysis.

Limitations

The Bayesian Monte Carlo technique, although highly promising, is not without limitations. Specifically, the technique can have potentially extensive computer requirements, and requires several assumptions on the behavior of the model and the data. A total of one hundred thousand

(100 000) iterations were used to generate the output presented herein, requiring 200 CPU seconds. A larger number of iterations would be required to generate more precise estimates of the frequency distributions. The model tested has relatively low computational requirements, and is therefore more amenable to Monte Carlo analysis than a more computationally intensive model. Clearly, the technique is limited to models for which running thousands of iterations per analysis are feasible.

The second limitation of BMC pertains to accurate estimation of the likelihood function. Halfon (1985) described the likelihood function as one potential measure of model goodness of fit, but pointed out that its use requires information on the covariance between model errors. This application of BMC assumes that all data error is normally and independently distributed, a reasonable first assumption which allows for ready application of the method. Insufficient data were available for the Grand River data set to allow for rigorous testing either for independence or normality. This assumption was found to be partially violated when BMC was applied to a phosphorus model of Green Bay, Lake Michigan (Dilks, 1987), but was resolved by stratifying the observed field data by sampling station to provide a more independent data set. While BMC can be readily adapted to use a more rigorous application of the likelihood function, lack of information on the joint probability density function of model errors will be a practical limitation in this regard.

Another assumption of the technique is that there be no model framework error. Framework error is difficult to measure, so this assumption cannot readily be tested. Future application of this technique must be limited to situations where parameter error can safely be assumed to be significantly greater than model framework error.

Use of the sample likelihood function as an estimate of model goodness-of-fit has removed much of the potential for subjective bias from the technique. There is still a subjective nature to the technique, in the specification of prior distributions. Different modelers may assume different prior distributions for the same data set, resulting in different output from the technique. The degree of subjectivity introduced will be greatest in cases where the observed data error is highly variable, as these are the cases where prior assumptions most affect the posterior distributions. The potential for subjective prior assumptions is inherent to many applications of Bayesian inference (Benjamin and Cornell, 1970), and is not limited to the technique developed here.

Benefits

BMC has the potential to be a valuable tool in the analysis of water quality model uncertainty for several reasons. The most important benefits

of the technique are those first described by Fedra (1980). It can provide improved estimates for model parameters not easily measured. More importantly, the technique also provides an estimate of the uncertainty in these parameters, based upon available data and prior assumptions (if any) on parameter distributions. These distributions are then directly incorporated into a prediction of model output uncertainty. This prediction of model uncertainty directly accounts for any covariance between model parameters, as the consideration given any projection is directly equivalent to the likelihood that the data were adequately described by the entire parameter set.

The technique provides the benefit of Bayesian inference in that information from two separate sources is combined to provide an improved estimate of the true system state. The technique will appropriately account for the degree of uncertainty in the observed field data. Field data with high confidence will lead to posterior distributions most representative of this data. Highly uncertain field data will lead to posterior distributions more representative of prior assumptions. Similarly, strong prior knowledge on parameter distributions will be directly reflected in posterior distributions; limited prior knowledge will have a lesser effect. The technique therefore provides the desirable ability to predict the amount of information that can be gained on parameter distributions based upon improved sampling of field data.

This research has provided two additional benefits with respect to future applicability of the technique. First, the statistical basis of the technique has been identified and verified as valid. This will allow any future application or modification of the technique to start with a firm theoretical basis, as opposed to the intuitive concepts previously provided in the literature. The second advantage of this research with respect to future use regards the straightforward method in which it can be applied. Previous descriptions of technique left specification of the calibration criteria open to the user. Results were highly susceptible to subjective interpretation of the data, because no rigorous methods were defined for specifying the criteria. The Bayesian Monte Carlo technique described herein provides an objective and statistically rigorous method of calculating the ability of a given set of parameters to simulate observed data, compared to additional field or laboratory studies on model parameters (better prior information).

REFERENCES

Benjamin, J.R. and Cornell, C.A., 1970. Probability, Statistics, and Decision for Civil Engineers. McGraw-Hill, New York.

Burges, S.J. and Lettenmaier, D.P., 1975. Probabilistic methods in stream quality management. Water Resour. Bull., 11: 115–130.

Dilks, D.W., 1987. Bayesian Monte Carlo, Ph.D. dissertation, University of Michigan, Ann Arbor, MI.

Fedra, K., 1980. Mathematical modeling – a management tool for aquatic ecosystems? Helgol. Meeresunters., 34: 221–235.

Halfon, E., 1985. Is there a best model structure? III. Testing the goodness of fit. Ecol. Modelling, 27: 15–23.

Hornberger, G.M. and Spear, R.C., 1980. Eutrophication in Peel Inlet – I. The problem-defining behavior and a mathematical model for the phosphorus scenario. Water Res., 14: 29–42.

Limno-Tech, Inc., 1982. Impact of Grand Rapids Combined Sewer Overflows on Grand River Water Quality. Ann Arbor, MI.

Spear, R.C. and Hornberger, G.M., 1980. Eutrophication in Peel Inlet – II. Identification of critical uncertainties via generalized sensitivity analysis. Water Res., 14: 43–49.