JOURNAL OF THE AMERICAN WATER RESOURCES ASSOCIATION

Vol. 43, No. 6 AMERICAN WATER RESOURCES ASSOCIATION December 2007

APPROACHES TO EVALUATE WATER QUALITY MODEL PARAMETER UNCERTAINTY FOR ADAPTIVE TMDL IMPLEMENTATION[1]

Craig A. Stow, Kenneth H. Reckhow, Song S. Qian, Estel Conrad Lamon III, George B. Arhonditsis, Mark E. Borsuk, and Dongil Seo[2]

ABSTRACT: The National Research Council recommended Adaptive Total Maximum Daily Load implementation with the recognition that the predictive uncertainty of water quality models can be high. Quantifying predictive uncertainty provides important information for model selection and decision-making. We review five methods that have been used with water quality models to evaluate model parameter and predictive uncertainty. These methods (1) Regionalized Sensitivity Analysis, (2) Generalized Likelihood Uncertainty Estimation, (3) Bayesian Monte Carlo, (4) Importance Sampling, and (5) Markov Chain Monte Carlo (MCMC) are based on similar concepts; their development over time was facilitated by the increasing availability of fast, cheap computers. Using a Streeter-Phelps model as an example we show that, applied consistently, these methods give compatible results. Thus, all of these methods can, in principle, provide useful sets of parameter values that can be used to evaluate model predictive uncertainty, though, in practice, some are quickly limited by the "curse of dimensionality" or may have difficulty evaluating irregularly shaped parameter spaces. Adaptive implementation invites model updating, as new data become available reflecting water-body responses to pollutant load reductions, and a Bayesian approach using MCMC is particularly handy for that task.

(KEY TERMS: total maximum daily load; water quality model; ecological forecasting; uncertainty analysis; parameter estimation; adaptive management; Bayesian; Streeter-Phelps; equifinality; computational methods; optimization.)

Stow, Craig A., Kenneth H. Reckhow, Song S. Qian, Estel Conrad Lamon III, George B. Arhonditsis, Mark E. Borsuk, and Dongil Seo, 2007. Approaches to Evaluate Water Quality Model Parameter Uncertainty for Adaptive TMDL Implementation. *Journal of the American Water Resources Association* (JAWRA) 43(6):1499-1507. DOI: 10.1111/j.1752-1688.2007.00123.x

INTRODUCTION

Water quality models provide an essential framework for scientific assessment in support of water quality management and decisions such as total maximum daily load (TMDL) determinations (NRC 2001). Models allow decision makers to evaluate the logical outcomes of alternative management actions based on informed speculation about system behavior captured in a set of equations.

[1] Paper No. J06104 of the *Journal of the American Water Resources Association* (JAWRA). Received August 1, 2006; accepted March 30, 2007. © 2007 American Water Resources Association. **Discussions are open until June 1, 2008**.

[2] Respectively (Stow) Senior Scientist, NOAA Great Lakes Environmental Research Laboratory, 2205 Commonwealth Blvd, Ann Arbor Michigan 48105; (Reckhow, Qian, Lamon) Professor, Associate Research Professor, and Research Scientist, Nicholas School of the Environment and Earth Sciences, Duke University, Durham, North Carolina; (Arhonditsis) Assistant Professor, Department of Physical & Environmental Sciences, University of Toronto, Ontario, Canada; (Borsuk) Assistant Professor, Thayer School of Engineering, Dartmouth College, Hanover, New Hampshire; (Seo) Professor, Department of Environmental Engineering, Chungnam National University, 220 Gung-dong, Yuseong-gu, Daejeon, Korea 305-764 (E-Mail/Stow: craig.stow@noaa.gov).

Given a choice of models, a decision maker is likely to choose the model that predicts most accurately. If a model were available that was 100% accurate (i.e., the model predicts correctly 100% of the time), this model would be a clear choice over one that was, say, 80% accurate. With 100% accuracy, management actions could be chosen based only on the societal value of the consequences of those actions. Even models of relatively low predictive accuracy can be useful, if the predictive accuracy is appropriately quantified. A model with only 80% accuracy is still informative, but applying such a model requires hedging decisions by the relative probabilities of a range of possible outcomes and the societal value of those outcomes. Thus, model uncertainty quantification provides information useful in both model selection and application.

However, decision makers are often provided with models, or model results, and given no information regarding forecast uncertainty. How then, can these models be appropriately used for decision purposes?

Model uncertainty is typically quantified by inclusion of an error-term on the model, and estimating the model's structural and error-term parameter values. Often, however, modelers have little data to support rigorous parameter estimation or assess parameter uncertainty; thus, modelers employ "judicious diddling" (Hornberger and Spear, 1981) to select values of key model parameters, aided by the user's manual or other established precedent. Among experienced water quality modelers, it is well-recognized that many "sets" of parameter values will fit the model about equally well; similar predictions can be obtained by simultaneously manipulating several parameter values in concert. This is plausible in part because all models are approximations of actual ecosystem processes, and because all parameters represent aggregate processes (spatially and temporally averaged at some implicit scale) and are unlikely to be represented by a fixed constant across scales. Additionally, many mathematical structures impart extreme correlation among model parameters, even when the model is overdetermined. This condition, called "equifinality," is well-documented in the hydrologic sciences (Franks *et al.*, 1997), but the concept has rarely been discussed in the water quality modeling research literature. We believe that the recognition of equifinality should change the perspective of water quality modelers from seeking a single "optimal" value for each model parameter, to seeking a *distribution of parameter sets* that all meet a pre-defined fitting criterion (Spear, 1997). These acceptable parameter sets may then provide the basis for estimating model prediction error associated with the model parameters.

Herein, we discuss several techniques that might be used for evaluating plausible parameter sets, and compare their utility. We then illustrate the approaches using a simple Streeter-Phelps dissolved oxygen model. Though the rationale for uncertainty analysis in water quality modeling has been recognized for many years (Reckhow and Chapra, 1983; Beck, 1987), in practice rigorous uncertainty analysis is rare. Pappenberger and Beven (2006) suggested that one of the reasons modelers often fail to do uncertainty analysis is that there are many "competing methods" making it difficult to choose a method and interpret the results. A primary goal of this paper is to show that, though these techniques have origins in distinct disciplines, they will provide similar inference if they are consistently applied. Accordingly, we encourage water quality modelers to consider a refocus from single optimal parameter selection to estimation of complete parameter sets, leading to the multi-parameter distribution. Using the multi-parameter distribution to make predictions then provides a quantified estimate of predictive uncertainty.

Regionalized (Generalized) Sensitivity Analysis

The development of methods for identifying plausible parameter sets for large multi-parameter environmental models with limited observational data began with the work of Hornberger and Spear (1981). Their method, called regionalized (or generalized) sensitivity analysis (RSA), is a Monte Carlo sampling approach to assess model parameter sensitivity. Hornberger and Spear advocated the application of this method as a means to prioritize future sampling and experimentation for model and parameter improvements.

Regionalized sensitivity analysis is simple in concept, and is a useful way to use limited information to bound model parameter distributions. Given a particular model and a system (e.g., water body) being modeled, the modeler first defines the plausible range of certain key model response variables (e.g., chlorophyll *a*, total nitrogen) as the "behavior." Outside the range is "not the behavior." The modeler then samples from (often uniform) distributions of each of the model parameters and computes the values for the key response variables. Each complete sampling of all model parameters, leading to prediction, results in a "parameter set." All parameter sets that result in predictions of the key model response variables in the "behavior" range are termed "behavior generating" and thus become part of the model parameter distribution. The parameter sets that do not meet this behavior criterion are termed "nonbehavior generating."

Hornberger and Spear (1981) proposed that the cumulative distribution function (cdf) of each parameter distribution from these two classes of parameter sets (behavior generating and nonbehavior generating) be compared with evaluate model parameter sensitivity. For a particular parameter, if the behavior generating and nonbehavior generating distributions are substantially different, then prediction of the key response variables is sensitive to that parameter. Hence, resources devoted toward model improvement might be preferentially allocated toward improved estimation of that parameter.

In addition, we can consider the distribution of the behavior generating parameter sets as reflecting equifinality. Thus, the empirical distribution characterizes the error (variance and covariance) structure in the model parameters, conditional on the model and on the fitting criterion (the defined plausible range of key response variables).

Generalized Likelihood Uncertainty Estimation

The Generalized Likelihood Uncertainty Estimation (GLUE) approach is an extension of the original RSA; the binary system of acceptance/rejection of behavioral/nonbehavioral simulations is replaced by a "likelihood" measure that assigns different levels of confidence (weighting) to different parameters sets (Beven and Binley, 1992; Zak and Beven, 1999; Page *et al.*, 2004). Unlike Bayesian Monte Carlo (BMC), Importance Sampling (IS), and Markov Chain Monte Carlo (MCMC), the term *likelihood* has a very broad meaning in the GLUE methodology and it is specified as any measure of goodness-of-fit that can be used to compare observed responses and model predictions (Zak *et al.*, 1997). Herein, we will use "likelihood measure" to distinguish this concept from "likelihood function", a term that is well-defined and universally applied in the statistical literature. A wide variety of likelihood measures can be found in the GLUE literature [e.g., likelihood measures based on the sum of squared errors (Beven and Binley, 1992; Sorooshian and Gupta, 1995; Freer *et al.*, 1997), fuzzy measures (Franks *et al.*, 1998; Page *et al.*, 2004) or even qualitative measures for model evaluation (Beven, 2001)].

The GLUE procedure requires a large number of Monte Carlo model runs sampled from (usually) uniform distributions across plausible parameter ranges. Prior knowledge regarding the expected joint parameter distributions can be incorporated by assigning appropriate prior likelihood weights to each of the parameter sets (Schulz *et al.*, 1999). The behavioral runs are selected on the basis of a subjectively chosen threshold of the likelihood measure and are rescaled so that their cumulative total is 1.0. The weighting assigned to the retained behavioral runs is propagated to the model output and forms a likelihood-weighted cumulative distribution of the predicted variable(s), which are then used for estimating the prediction uncertainty ranges (Beven and Binley, 1992).

Bayesian Approaches – General

Bayesian approaches begin with the realization that model predictions will contain error; thus, a term representing this error is explicitly incorporated in the model. This prediction error is often written as an additive term (though other error structures are possible),

$$Y = g[x, \theta] + \varepsilon, \tag{1}$$

where Y is the response variable (such as dissolved oxygen or chlorophyll a), g is a general model form (such as a Streeter-Phelps dissolved oxygen model), x represents one or more state variables (such as temperature or nutrient concentration), θ represents one or more model parameters (such as rate coefficients), and ε is the model error. Often ε is assumed to be normally distributed with mean (denoted μ) = 0, and variance = σ^2. Under the assumption that ε is distributed normally, the likelihood function for this model is

$$f(y|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{(Y - g[x,\theta])^2}{-2\sigma^2}\right], \tag{2}$$

where n is the number of observations. In this function, θ is regarded as an unknown quantity that can be predicted from the observed data.

Bayes theorem combines Equation (2) with any prior information a modeler has about the value of θ resulting in

$$\pi(\theta|y) = \frac{\pi(\theta)f(y|\theta)}{\int_\theta \pi(\theta)f(y|\theta)\mathrm{d}\theta}, \tag{3}$$

where $\pi(\theta\,|\,y)$ is the posterior probability of θ (the probability of the parameter vector, θ, after observing the data, y), $\pi(\theta)$ is the prior probability of θ, (the probability of θ before observing y), and $f\,(y\,|\,\theta)$ is the likelihood function. In water quality modeling $\pi(\theta)$ is often represented by a single fixed value based on the prior knowledge of the modeler, or chosen from the literature or a compendium of such values (Bowie *et al.*, 1985). In contrast, noninformative priors are typically used if little prior knowledge about the parameter values is available, or if the modeler prefers that the parameter values be estimated using only information conveyed by the data. When noninformative priors are used Bayesian approaches provide results consistent with maximum likelihood results or, if the model error term is additive and normally distributed, least-squares estimation. However, Bayesian approaches emphasize inference using the entire posterior parameter distribution, whereas maximum likelihood and least-squares methods emphasize the choice of a single optimal value for each parameter.

Bayesian Monte Carlo

The BMC approach (Dilks *et al.*, 1992) is similar to the Hornberger-Spear algorithm, but carries the additional assumptions of an additive, normally distributed error term, with mean = 0 and variance = σ^2 (Equation 1). Acceptable model behavior can be implicitly constrained *a priori* by setting the value of σ^2. Then, the modeler samples from uniform distributions were chosen to represent plausible ranges of values for each parameter. However, rather than grouping parameter sets into two categories, "behavior generating" and "nonbehavior generating", parameter sets are weighted using the likelihood function. Parameter sets that result in more likely model predictions (closer to the maximum of the likelihood function) are weighted more heavily than those resulting in unlikely predictions. The result is analogous to a multivariate probability density function for the model parameters.

Importance Sampling

The Hornberger-Spear algorithm, GLUE, and the BMC all run the risk of becoming limited by the "curse of dimensionality"; in high-dimensional models (models with many unknown parameters) the plausible parameter space can become an extremely small proportion of the space defined *a priori* by a set of independent uniform distributions. When this occurs sampling may be, at best, inefficient and, at worst, ineffective. Additionally, some combinations of parameter values may provide plausible model results, though these combinations may include values for the individual parameters that would not be deemed plausible when the parameters are considered one at a time. This latter situation is particularly problematic when the parameters are highly correlated (Figure 1). In this case, the joint parameter space defined *a priori* by uniform distributions (solid box) for each individual parameter may exclude important regions in the tails of the parameter space (ellipse). Enlarging the space by increasing the width of each of the uniform distributions may incorporate these regions (dashed box), but this approach exacerbates the curse of dimensionality. Using this tactic, the volume of the parameter space to be sampled is likely to increase more rapidly than the important parameter space, making it even less likely that the plausible region will be sampled effectively.

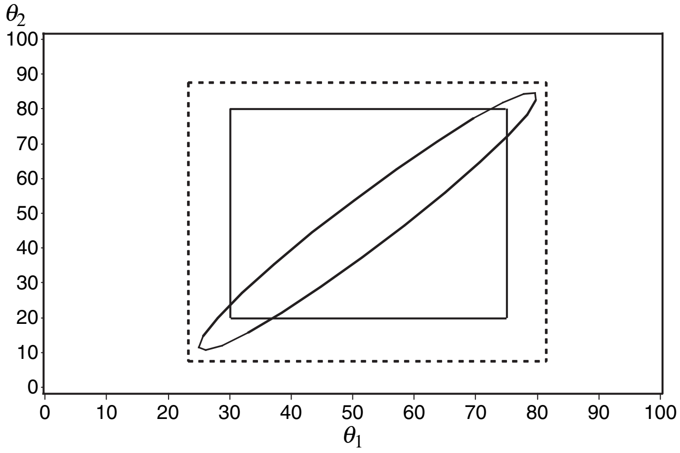

FIGURE 1. Illustration of How *a Priori* Independent Uniform Distributions Can Miss Important Regions of the Parameter Space. Ellipse depicts important parameter space of two positively correlated parameters, θ_1 and θ_2 while the solid box shows the area encompassed by *a priori* plausible ranges of 30-75 for θ_1 and 20-80 for θ_2. The ellipse encompasses only a small proportion of the box indicating that random sampling within the box will be inefficient. Concurrently, random sampling within the box will miss the upper and lower tails of the ellipse. If the box is enlarged (dashed box) to capture the tails, then the efficiency of random sampling will be further reduced because the area of the box increases more rapidly than the additional area included in the tails.

Thus, IS and its variations (Sampling/Importance Resampling – SIR) is premised on the idea that sampling effectiveness can be increased by choosing a sampling distribution (for which pseudorandom number generators exist) that more closely approximates the important region of the parameter space. In a Bayesian context, this means choosing a surrogate, such as a multivariate normal or *t*-density that closely approximates the posterior parameter distribution. Often this can be done by first finding the maximum of the posterior distribution, and then using Fisher information (the negative expectation of the Hessian of the log of the posterior) to estimate the parameter covariance structure (Geweke, 1989).

Like BMC, the SIR algorithm often includes a normally distributed additive error term, but the parameters of this error term are included in the set of model parameters to be estimated. SIR is most useful when a good surrogate exists to the posterior distribution, when this surrogate is easy to sample, and when a limited number of samples is desired (Rubin, 1988). The SIR algorithm takes more samples than needed (say M) from the surrogate distribution, then resamples from this finite sample of size M, based on the ratio of the true posterior to the surrogate, to obtain m final draws (where $m \ll M$).

Markov Chain Monte Carlo

An historical limitation in the application of Bayesian approaches was that, for many model forms, using the posterior parameter distribution required solving analytically intractable integrals. Importance sampling addresses this limitation by using a surrogate to provide a sample from the posterior distribution; MCMC estimation (prediction) uses cleverly written algorithms to draw samples directly from the posterior distribution (more accurately – these samples will converge, in distribution to the posterior) allowing precise numerical approximation of any function of the posterior distribution (Gelfand and Smith, 1990; Smith and Roberts, 1993). There are several algorithms available; the Metropolis-Hastings algorithm (Chib and Greenberg, 1995) is general but less numerically efficient, while the Gibbs Sampler (Casella and George, 1992), a special case of Metropolis-Hastings, can take advantage of structural regularities present in some models to converge more efficiently. Selecting the most appropriate algorithm is dependent on the model form and the distributional structure chosen to represent the stochastic terms. Fortunately, there is freely available software for this task; WinBUGS incorporates MCMC algorithms into a straightforward programming environment (Gilks *et al.*, 1994).

Summary

These five approaches can be thought as approximately evolutionary, facilitated by the availability of fast, inexpensive computers (Figure 2). The RSA approach is completely general, assumes no structure associated with model error and serves as a screening approach to identify plausible regions of model parameter values. RSA requires *a priori* determination of the behavior-generating region for the response variables. This determination is very important and can be based on either expert-judgment, or more empirically derived like it is using the other four procedures. The BMC builds on the RSA approach by adding assumptions regarding model error structure and uses that added structure to delimit plausible parameter regions. GLUE is similar to BMC (and in some cases can be the same) but permits a broader range of functions that define the model error structure. GLUE can also be "updated", much like a Bayesian procedure (Beven and Binley, 1992). IS recognizes the problems associated with the "curse of dimensionality" that can limit the effectiveness of sampling using RSA, BMC, and GLUE and employs a well-chosen surrogate distribution instead of sampling parameter values from independent uniform distributions. MCMC employs a full Bayesian framework and uses clever algorithms to choose a sample that approaches the posterior density function in distribution.

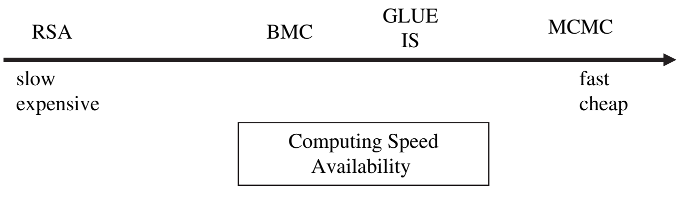

FIGURE 2. Conceptual Timeline Depicting the Availability of Fast, Cheap Computing and Parameter Evaluation Methods.

EXAMPLE USING THE STREETER-PHELPS DISSOLVED OXYGEN MODEL

To illustrate and compare these five approaches, we simulated a dataset using the following form of the Streeter-Phelps stream dissolved oxygen model (Streeter and Phelps, 1925)

$$\mathrm{DO} = \mathrm{DO_s} - \frac{k_1 \mathrm{BOD_u}}{k_2 - k_1}\left(\mathrm{e}^{-k_1 \frac{x}{v}} - \mathrm{e}^{-k_2 \frac{x}{v}}\right) - D_i \mathrm{e}^{-k_2 \frac{x}{v}}, \tag{4}$$

where DO is the dissolved oxygen concentration (mg/l), $\mathrm{DO_s}$ is the saturation oxygen concentration, k_1 is the BOD decay coefficient (1/day), k_2 is the reaeration coefficient (1/day), $\mathrm{BOD_u}$ is the ultimate BOD (mg/l), x is the downstream distance (km), v is stream velocity (km/day), and D_i is the initial DO deficit (mg/l). Using simulated data allows comparison of estimated (predicted) parameters with the true parameter values that generated the data. For this example we set $\mathrm{DO_s} = 8.0$, $k_1 = 0.25$, $k_2 = 0.8$, $\mathrm{BOD_u} = 35$, $v = 10$, and $D_i = 1.0$. A random normal error with $\mu = 0$ and $\sigma^2 = 0.6$ was added to each observation. Thirteen x values between 0 and 100 km were randomly generated from a uniform distribution resulting in a set (Figure 3) with observed DO ranging from 1.9 to 7.8 mg/l and a minimum at ~20 km. For straightforward depiction on bivariate plots, we

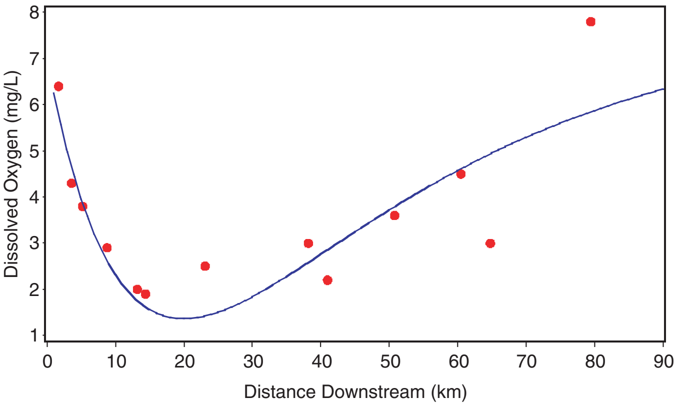

FIGURE 3. Depiction of the Example Streeter-Phelps Model (blue line) and Simulated Observations (red dots).

treated k_1 and k_2 as the unknown model parameters, though various combinations of the other model inputs could also be predicted (estimated) from the data.

To illustrate the application of RSA, we defined the plausible DO range as ≥0 mg/l. No upper bound on DO was necessary because this form of the Streeter-Phelps model has no oxygen source term that would push DO above saturation, thus no combination of values for k_1 and k_2 will cause model predictions to exceed the 8 mg/l saturation value.

Choosing candidate ranges for k_1 and k_2 was somewhat trickier; Bowie *et al.* (1985) listed k_1 values ranging from 0.004 to ~5 and k_2 values from ~0.01 to ~100, while in our experience, values between 0 and 1.0 are most common for each. Selecting different parameter spaces can strongly affect the inference made; the parameter ranges suggested by Bowie *et al.* (1985) result in an acceptable parameter region ≃95% of the total parameter space (Figure 4a), whereas ranges from 0 to 1 for k_1 and k_2 result in an acceptable region ≃22% of the total space (Figure 4b). Considering the larger parameter space, we would conclude that the model was more sensitive to k_2, as indicated by large difference (relative to k_1) between the cdfs for the behavior and nonbehavior generating sets (Figure 5, panels a and b). Conversely, when the parameter space for both parameters is constrained to range from 0 to 1, the behavior and nonbehavior generating cdfs are more similar to each other for k_2 than for k_1 indicating that the model will be sensitive to choices of k_1 (Figure 5, panels c and d).

We chose to illustrate the GLUE procedure using an error sum of squares likelihood measure defined as

$$\mathrm{ess} = \sum_{i=1}^{n} (P_i - O_i)^2, \tag{5}$$

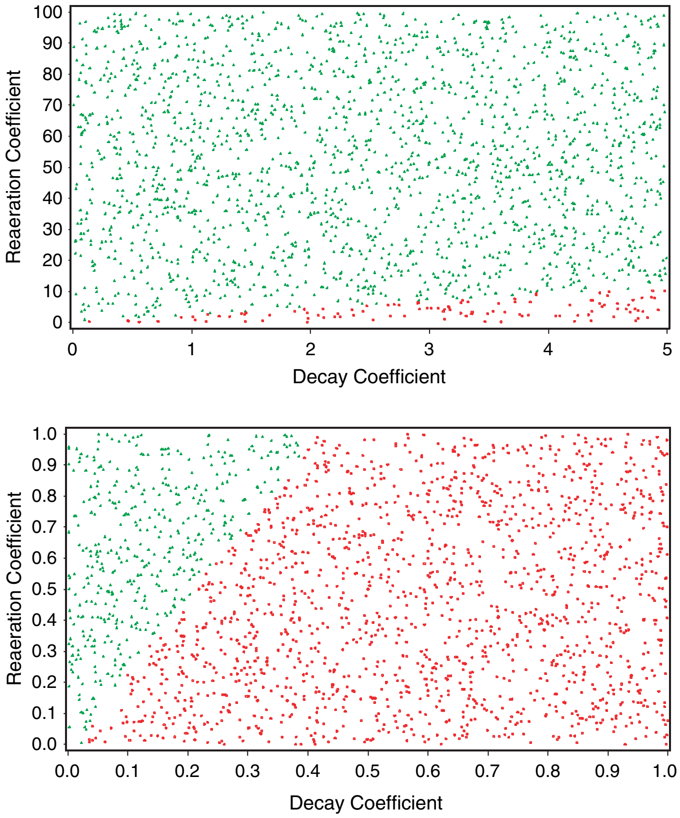

FIGURE 4. Behavior-Generating (green) and Nonbehavior-Generating Regions Using RSA Using the Streeter-Phelps Example. Top panel depicts *a priori* ranges for reaeration and decay coefficients that are very wide, and bottom panel depicts narrower ranges.

where ess is error sum of squares, n is the number of observations, P_i is the ith of n predicted values, and O_i is the ith of n observed values. Using the error sum of squares provides a result that is closely analogous to Bayesian estimation with a normal, additive model error, and a noninformative prior distribution. The result (Figure 6) is consistent with the RSA result (Figure 5), but provides more information about the location of the most likely parameter values. The plot contours depict parameter sets that are equally likely, given the chosen likelihood measure. While the most likely values are near the center of the contour ellipse, many other sets that are almost as likely are also identified.

Qian *et al.* (2003) indicated that *a priori* specification of a precise value for σ^2, using BMC, can strongly influence the variance of the posterior parameter distribution and thus prediction variance. However, if σ^2 is treated as a parameter to be estimated from the data, then the main difference between BMC and IS is that IS uses a well-chosen surrogate for the posterior distribution, to concentrate sampling effort near the most probable parameter values. To compare these two approaches, we

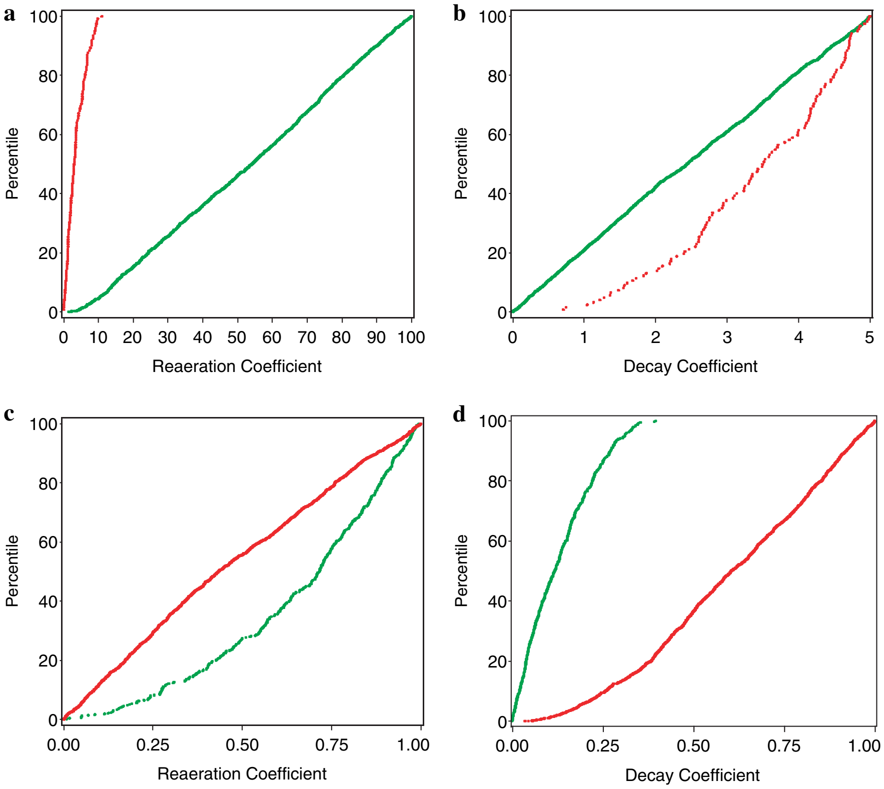

FIGURE 5. Cumulative Density Functions of the Behavior Generating (green) and Nonbehavior Generating (red) Parameter Values for the Two RSA Sets of Results. Panels a and b depict *a priori* ranges for reaeration and decay coefficients that are very wide, and bottom panel depicts narrower ranges.

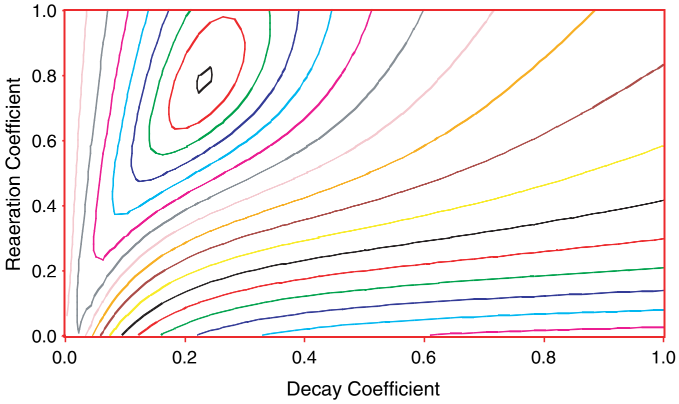

FIGURE 6. GLUE Results for the Streeter-Phelps Example Using Error Sum of Squares Likelihood Measure. Each successive contour from the inner circle represents and interval of 1.5 times the previous contour.

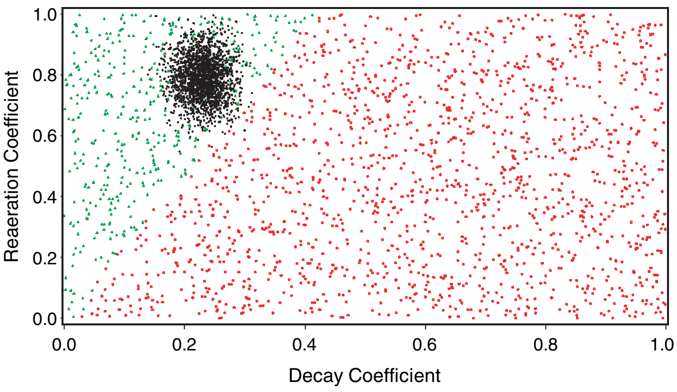

FIGURE 7. IS Sample (black dots) and BMC Sample (green and red dots). BMC sample is depicted in two colors to illustrate correspondence with the RSA behavior-generating (green) and nonbehavior-generating (red) regions.

chose 2000 samples from two uniform [0,1] distributions for the BMC, and two normal distributions, estimated from the example data using nonlinear least squares, for the IS distribution. The results (Figure 7) indicate the relative inefficiency of the BMC, with only about 4% of the BMC samples falling within the area IS sampled. This inefficiency is exacerbated when the parameters are highly correlated, particularly in higher dimensional models (Qian *et al.*, 2003). Similarly, a poor choice for the IS surrogate can cause inefficient or nonrepresentative sampling. In this example, we used independent, normal distributions; though incorporating parameter correlation into the IS sampling distribution can increase

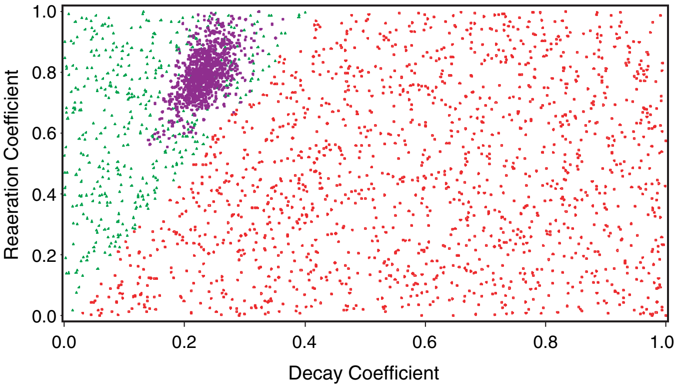

FIGURE 8. MCMC Sample (purple dots) and BMC Sample (green and red dots). BMC sample is depicted in two colors to illustrate correspondence with the RSA behavior-generating (green) and nonbehavior-generating (red) regions.

efficiency and accuracy. However, choosing a good surrogate can be difficult for high-dimensional models or highly nonlinear models, where the tails of the posterior distribution are often irregularly shaped.

Comparison of Figures 6 and 7, however, reveals that these methods provide consistent results, with the most likely values for k_1 and k_2 near the true values that were used to generate the dataset. An MCMC sample (Figure 8) using a noninformative prior distribution, generated using WinBUGS, is also similar to the IS sample (Figure 7) and the most likely region of the GLUE (Figure 6). The advantage of using MCMC is that a well-written algorithm quickly converges to provide a sample from the posterior parameter distribution and does not require independent information regarding a surrogate distribution to sample. This is particularly advantageous when extreme parameter correlation and nonlinear model structure make choosing a good surrogate distribution difficult.

CONCLUSIONS

Our simple two-dimensional Streeter-Phelps example illustrates the capabilities and limitations of these methodologies. RSA is completely general, but only separates parameter sets into two groups: in or out. Adding structural assumptions about the model error term, either implicitly, applying GLUE, or explicitly, using Bayesian approaches, yields considerably more information; the resultant parameter sets are expressed probabilistically. MCMC methods make it feasible to generate large samples from these probabilistic parameter sets, which can be used in model predictions, thus resulting in a straightforward calculation of model prediction uncertainty. We deliberately chose an example using a simple low dimensional model, where all properties of the model are known, for easy depiction. In real applications and higher dimensional models, the concepts are analogous but problems resulting from the "curse of dimensionality" become more difficult. Thus, using an approach capable of effectively and efficiently sampling the appropriate parameter space becomes increasingly important.

The National Research Council (NRC 2001) TMDL report recommended "Adaptive Implementation" of TMDLs, an approach based on the "Adaptive Management" concept (Holling, 1978). Using adaptive implementation water quality models are an integral component of the TMDL assessment phase in which alternative management actions are evaluated based on the probability of attaining water quality standards. To fully implement this NRC recommendation, it will be imperative to routinely incorporate uncertainty analysis approaches, such as those we have reviewed, into model development. Within the Adaptive Management framework, TMDL implementation is regarded as a "learning by doing" opportunity – an ecosystem-scale experiment (Carpenter *et al.*, 1995), that can provide data and information about system behavior not available by other means. Bayesian methods are particularly useful for model development under adaptive management because they provide a straightforward, rigorous basis for data assimilation and model updating using Bayes theorem.

ACKNOWLEDGMENTS

This work was partially supported by EPA STAR Grant No. R830883. This is GLERL contribution number 1448.

LITERATURE CITED

Beck, M.B., 1987. Water Quality Modeling: A Review of the Analysis of Uncertainty. Water Resources Research 23:1393-1442.

Beven, K.J., 2001. Rainfall-Runoff Modeling: The Primer. John Wiley & Sons Ltd, West Sussex, England, pp. 360.

Beven, K. and A. Binley, 1992. The Future of Distributed Models-Model Calibration and Uncertainty Prediction. Hydrological Processes 6:279-298.

Bowie, G.L., W.B. Mills, D.B. Porcella, C.L. Campbell, J.R. Pagenkopf, G.L. Rupp, K.M. Johnson, P.W.H. Chan, S.A. Gherini, and C.E. Chamberlin, 1985. Rates, Constants, and Kinetic Formulations in Surface Water Quality Modeling. EPA/600/3-85/040, U.S. Environmental Protection Agency, Washington, D.C.

Carpenter, S.R., S.W. Chisolm, C.J. Krebs, D.W. Schindler, and R.F. Wright, 1995. Ecosystem Experiments. Science 269:324-327.

Casella, G. and E.I. George, 1992. Explaining the Gibbs Sampler. American Statistician 46:167-174.

Chib, S. and E. Greenberg, 1995. Understanding the Metropolis-Hastings Algorithm. American Statistician 49:327-335.

Dilks, D.W., R.P. Canale, and P.G. Meijer, 1992. Development of Bayesian Monte Carlo Techniques for Water Quality Model Uncertainty. Ecological Modelling 62:149-162.

Franks, S.W., K.J. Beven, P.F. Quinn, and I.R. Wright, 1997. On the Sensitivity of Soil-Vegetation-Atmosphere Transfer (SVAT) Schemes: Equifinality and the Problem of Robust Calibration. Agricultural and Forest Meteorology 86:63-75.

Franks, S.W., P. Gineste, K.J. Beven, and P. Merot, 1998. On Constraining the Predictions of a Distributed Moder: The Incorporation of Fuzzy Estimates of Saturated Areas Into the Calibration Process. Water Resources Research 34:787-797.

Freer, J., J. McDonnell, K.J. Beven, D. Brammer, D. Burns, R.P. Hooper, and C. Kendal, 1997. Topographic Controls on Subsurface Storm Flow at the Hillslope Scale for Two Hydrologically Distinct Small Catchments. Hydrological Processes 11:1347-1352.

Gelfand, A.E. and A.F.M. Smith, 1990. Sampling Based Approaches to Calculating Marginal Densities. Journal of the American Statistical Association 85:398-409.

Geweke, J., 1989. Bayesian Inference in Econometric Models Using Monte Carlo Integration. Economtrica 57:1317-1339.

Gilks, W.R., A. Thomas, and D.J. Spiegelhalter, 1994. A Language and Program for Complex Bayesian Modelling. The Statistician 43:169-177.

Holling, C.S., 1978. Adaptive Environmental Assessment and Management. International Institute for Applied Systems Analysis. Blackburn Press, Caldwell, New Jersey.

Hornberger, G.M. and R.C. Spear, 1981. An Approach to the Preliminary Analysis of Environmental Systems. Journal of Environmental Management 12:7-18.

NRC (National Research Council), 2001. Assessing the TMDL Approach to Water Quality Management. National Research Council, National Academy Press, Washington, DC.

Page, T., K.J. Beven, and J.D. Whyatt, 2004. Predictive Capability in Estimating Changes in Water Quality: Long-Term Responses to Atmospheric Deposition. Water Air and Soil Pollution 151:215-244.

Pappenberger, F. and K.J. Beven, 2006. Ignorance is Bliss: Or Seven Reasons Not to Use Uncertainty Analysis. Water Resources Research 42: WO5302, doi: 10.1029/2005WR004820.

Qian, S.S., C.A. Stow, and M.E. Borsuk, 2003. On Monte Carlo Methods for Bayesian Inference. Ecological Modelling 159:269-277.

Reckhow, K.H. and S.C. Chapra, 1983. Engineering Approaches for Lake Management, Vol. 1 and 2. Butterworth Publishers. Boston, Vols. 1 and 2.

Rubin, D.B., 1988. Using the SIR Algorithm to Simulate Posterior Distributions. *In:* Bayesian Statistics 3, J.M. Bernardo, M.H. DeGroot, D.V. Lindley, and A.F.M. Smith (Editors). Oxford University Press, Oxford, pp. 395-402.

Schulz, K., K. Beven, and B. Huwe, 1999. Equifinality and the Problem of Robust Calibration in Nitrogen Budget Simulations. Soil Science Society of America Journal 63:1934-1941.

Smith, A.F.M. and G.O. Roberts, 1993. Bayesian Computation Via the Gibbs Sampler and Related Markov-Chain Monte-Carlo Methods. Journal of the Royal Statistical Society Series B-Methodological 55:3-23.

Sorooshian, S. and V.K. Gupta, 1995. Model Calibration. *In:* Computer Models of Watershed Hydrology, V.P. Singh (Editor). Water Resources Publications, Highlands Ranch, Colorado, pp. 23-68.

Spear, R.C., 1997. Large Simulation Models: Calibration, Uniqueness and Goodness of fit. Environmental Modelling and Software 12:219-228.

Streeter, H.W. and E.B. Phelps, 1925. A Study in the Pollution and Natural Purification of the Ohio River. U.S. Public Health Service, Public Health Bulletin No. 146, Washington, DC.

Zak, S.K. and K.J. Beven, 1999. Equifinality, Sensitivity and Predictive Uncertainty in the Estimation of Critical Loads. Science of the Total Environment 236:191-214.

Zak, S.K., K. Beven, and B. Reynolds, 1997. Uncertainty in the Estimation of Critical Loads: A Practical Methodology. Water Air and Soil Pollution 98:297-316.