UNCERTAINTY AND SENSITIVITY ANALYSIS FOR WATERSHED MODELS WITH CALIBRATED PARAMETERS

A Thesis

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Master of Science

by

Seunguk Lee

February 2010

ABSTRACT

This thesis provides a critique and evaluation of the Generalized Likelihood Uncertainty Estimation (GLUE) methodology, and provides an appraisal of sensitivity analysis methods for watershed models with calibrated parameters.

The first part of this thesis explores the strengths and weaknesses of the GLUE methodology with commonly adopted subjective likelihood measures using a simple linear watershed model. Recent research documents that the widely accepted GLUE procedure for describing forecasting precision and the impact of parameter uncertainty in rainfall-runoff watershed models fails to achieve the intended purpose when used with an informal likelihood measure (Christensen, 2004; Montanari, 2005; Mantovan and Todini, 2006; Stedinger et al., 2008). In particular, GLUE generally fails to produce intervals that capture the precision of estimated parameters, and the distribution of differences between predictions and future observations. This thesis illustrates these problems with GLUE using a simple linear rainfall-runoff model so that model calibration is a linear regression problem for which exact expressions for prediction precision and parameter uncertainty are well known and understood. The results show that the choice of a likelihood function is critical. A likelihood function needs to provide a reasonable distribution for the model errors for the statistical inference and resulting uncertainty and prediction intervals to be valid.

The second part of this thesis discusses simple uncertainty and sensitivity analysis for watershed models when parameter estimates result form a joint calibration to observed data. Traditional measures of sensitivity in watershed modeling are based upon a framework wherein parameters are specified externally to a model, so one can independently investigate the impact of uncertainty in each parameter on model output. However, when parameter estimates result from a joint calibration to observed data,

the resulting parameter estimators are interdependent and different sensitivity analysis procedures should be employed. For example, over some range, evaporation rates may be adjusted to correct for changes in a runoff coefficient, and vice versa. As a result, descriptions of the precision of such parameters may be very large individually, even though their joint response is well defined by the calibration data. These issues are illustrated with the simple abc watershed model. When fitting the abc watershed model to data, in some cases our analysis explicitly accounts for rainfall measurement errors so as to adequately represent the likelihood function for the data given the major source of errors causing lack of fit. The calibration results show that the daily precipitation from one gauge employed provides an imperfect description of basin precipitation, and precipitation errors results in correlation among flow errors and degraded the goodness of fit.

BIOGRAPHICAL SKETCH

Seung Uk Lee was born on January 2, 1977, in Busan, Korea. In February 2002, he received his Bachelor's degree in Civil Engineering, and completed his Master's degree in Civil Engineering from the Seoul National University in August 2004. In January 2007, he enrolled in Cornell University in the graduate program of the School of Civil and Environmental Engineering to pursue a Ph.D in water resource systems.

This thesis is dedicated to all my loved ones

- Particularly my parents and Na Young -

ACKNOWLEDGMENTS

First of all, I would like to thank Professor Jery R. Stedinger, my research advisor and special committee chair, for his expert guidance, enlightening discussion, and encouragement. I also wish to thank Professor Daniel Stephen Wilks, my minor advisor and special committee member, for his helpful advice and comments.

I would like to offer my appreciation and gratitude to Professor Young Oh Kim without his support and encouragement I would never have pursued a graduate degree. I also want to thank Alexander Juschitz for providing optimization results for the abc watershed model fit.

I need to thank my coworkers, Christopher Daniel Andrew Boone, Tom Wild, and Andrea Gruber Veilleux, and my Korean friends Chang Hoon Lee and Yong Sung Park for their friendship and support.

My warmest thanks to my family for all of their support, not only through the completion of this thesis, but also through the years leading up to this day.

I would like to thank my parents who instilled in me the confidence and inspiration to undertake this work. Also thanks to my sister and brother who provided needed distraction. Finally, this thesis is dedicated to my beautiful wife, Na Young, without her support and patience this would not have been possible.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

In recent years there has been a renewal of interest among hydrologic and environmental modelers in descriptions of the uncertainty in models of environmental processes. Currently, uncertainty has been addressed by several probabilistic approaches. Bayesian statistical inference has emerged as a popular tool for specifying hydrological models to characterize uncertainty. Many hydrological studies have included classical Bayesian schemes (Kuczera and Parent, 1998; Bates and Campbell, 2001; Thiemann et al., 2001; Thyer and Kuczera, 2003; Marshall et al., 2005a; Kavetski et al., 2006a, 2006b; Stedinger et al., 2008). Others have used the Generalized Likelihood Uncertainty Estimation (GLUE) which is a methodology based on Monte Carlo simulation for estimating the predictive uncertainty associated environmental models (Beven and Binley, 1992; Zak et al., 1997; Page et al., 2008; Viola et al., 2009). Wanger et al. (2003) proposed a DYNamic Identifiability Analysis (DYNIA) methodology which draws from elements of the Regional Sensitivity Analysis (RSA), and includes aspects of the GLUE approach and the use of Kalman filtering for hypothesis testing. Vrugt et al., 2005 introduced a Simultaneous Optimization and Data Assimilation (SODA) methodology which consists of an inner Kalman filter loop for recursive state estimation conditioned on an assumed parameter set, and an outer global optimization loop for batch estimation of the high probability region of the posterior density of the parameters.

Generalized Likelihood Uncertainty Estimation (GLUE) methodology was proposed by Beven and Binley (1992). It has severe limitations that are explored here. GLUE has been applied to a variety of hydrologic and environmental problems (Zak et al., 1997; Hellweger and Lall, 2004; Smith et al, 2005; Ruessink, 2005; Bianchini et al.

2006; Hassan et al., 2008; Page et al., 2008; Viola et al., 2009). Its popularity is due to the apparent success it has enjoyed in real-world applications, and that it appears to provide the needed characterization of uncertainty (Montanari, 2005). Blasone et al. (2008b) point to GLUE's conceptual simplicity, ease of implementation, and its flexibility with different sources of information that can be combined with different criteria to define a likelihood measure.

Recent studies show that the widely accepted GLUE procedure for describing forecasting precision and the impact of parameter uncertainty in rainfall-runoff watershed models fails to achieve the intended purpose when used with an informal likelihood measure (Christensen, 2004; Montanari, 2005; Mantovan and Todini, 2006; Stedinger et al., 2008). In particular, GLUE generally fails to produce intervals that capture the precision of estimated parameters, and the distribution of differences between predictions and future observations. This thesis illustrates these problems with GLUE using a simple linear rainfall-runoff model so that model calibration is a linear regression problem for which exact expressions for prediction precision and parameter uncertainty are well known and understood. The simple regression example enables us to clearly and simply illustrate GLUE deficiencies.

Another aim of this thesis is to review uncertainty and sensitivity analysis criteria for watershed models when parameters are calibrated. Sensitivity analysis for models characterizes the impact that changes in model parameters have on the model output. The traditional measures of sensitivity in watershed modeling are based upon a framework wherein parameters are specified externally to a model, so one can independently investigate the impact of uncertainty in those parameters on the output of the model. However, when parameter estimates results from a join calibration to observed data, errors may be significantly cross-correlated and different sensitivity analysis procedures should be employed.

These issues are explored in the second part of this thesis. We discuss the appropriated definitions of parameter uncertainty when data is used to calibrate several parameters jointly. Three sets of sensitivity indices are evaluated to illustrate the important of reflecting interdependences in calibrated parameters in a sensitivity analysis. The case study analyzes the importance of different parameters in a watershed model using a one-at-a-time sensitivity index and a sensitivity index considering parameter interactions with the simple acb rainfall-runoff model.

Rainfall input errors can be very important, resulting in misrepresentation of errors in the calibration process. In general, the precipitation input is based on point observations which are sometimes combined with indirect measurements such as radar or satellite information. With just a single point-rainfall gage or even the average of several gages, there is a potentially large error due to the fact that the exact rainfall variables are not known at every point of the catchment (Kavetski et al., 2006a, 2006b; Vrugt et al., 2008). In fitting the abc watershed model to data, our analysis explicitly accounts for rainfall measurement errors so as to adequately represent the likelihood function for the data given the major source of errors causing lack of fit.

CHAPTER 2

STATISTICAL ASSESSMENT FOR PARAMETER UNCERTAINTY ESTIMATION: BAYESIAN INFERENCE AND GLUE

2.1 Introduction

As watershed and other environmental simulation models become more widely used, there is greater need for procedures that generate realistic prediction intervals and other representations of uncertainty that describe the likely difference between actual flows and their forecasts, and between estimated parameters and their true values (if such true values exist). Uncertainty analysis has now become common practice in the application of environmental simulation models. This is also a primary goal of the Predictions in Ungauged Basins (PUB) initiative promoted by the International Association of Hydrological Sciences (IAHS, 2003) and a fundamental need of most end users (Montanari, 2007).

The Generalized Likelihood Uncertainty Estimation (GLUE) technique introduced by Beven and Binley (1992) is an innovative uncertainty method that is often employed with environmental simulation models. There are now over 600 citations to their original paper which illustrates its tremendous impact. GLUE's popularity can be attributed to its simplicity and its applicability to nonlinear systems, including those for which a unique calibration is not apparent. Montanari (2005) suggests that GLUE's popularity is due to the apparent success it has enjoyed in real-world applications, and that it appears to provide the needed characterization of uncertainty. Blasone et al. (2008b, pp. 20-21) point to GLUE's conceptual simplicity, ease of implementation, and its flexibility with different sources of information that can be combined with different criteria to define a likelihood measure.

Recent evaluations of GLUE by Christensen (2004), Montanari (2005), Mantovan and Todini (2006), Stedinger et al. (2008), and this study clearly demonstrate that prediction limits derived from GLUE can be significantly different from prediction limits derived from correct classical and widely accepted statistical methods. Beven (2006b) discussed these concerns, and called for additional studies. Mantovan and Todini (2006) and Mantovan et al. (2007), show that, with the 'less formal' likelihood functions generally adopted in most previous GLUE applications, estimates of prediction uncertainty will be what they call "incoherent and inconsistent," compromising valid statistical inference. In response, Beven et al. (2007, 2008) point out problems that result when a likelihood function overestimates the information content of data. The example in Beven et al. (2007, 2008) again reinforces a major point made by Mantovan and Todini (2006), by Stedinger et al. (2008), and by this study: if one wants to correctly understand the information content of the data, one needs to use a likelihood function that correctly represents the statistical sampling distribution of the data.

In GLUE's defense, Beven and Freer (2001) argue that GLUE prediction limits should not and will not coincide with limits based on classical statistics. More recently Beven (2006a) states that

> "These prediction limits will be conditional on the choice of limits of acceptability; the choice of weighting function; the range of models considered; any prior weights used in sampling parameter sets; the treatment of input data error, etc. …However, given the potential for input and model structural errors, they [the choices] will not guarantee that a specified proportion of observations, either in calibration or future predictions, will lie within the tolerance or prediction limits (the aim, at least, of a statistical approach to uncertainty). Nor is this necessarily an aim in the proposed framework."

If the aim of the GLUE framework is not to generate prediction and uncertainty intervals that contain the specified quantities with the prescribed frequency or probability, then we do not know what the purpose of the analysis is, or what GLUE advocates intend for their uncertainty intervals to represent. If GLUE provides a valid statistical analysis of environmental models when employed as recommended, then we contend that when applied to a very simple model with a classic model error structure, GLUE should reproduce the widely accepted uncertainty intervals generated using both classical and Bayesian statistical methods that provide the correct descriptions of uncertainty in that case. If as we show, GLUE does not generally reproduce the correct uncertainty intervals when applied to a wide range of simple problems, then there is little reason to believe it will provide reasonable results for difficult problems for which the correct solution is not known.

The statistical and probabilistic interpretation of GLUE analyses and the choice of a likelihood measure is the focus of this chapter. This chapter also shows how to correctly employ GLUE with simulation models to assure that uncertainty analyses produce reasonable prediction limits consistent with traditional statistical methods. In a broader perspective, this chapter reflects on the difference between reality and the claims made for GLUE with subjective likelihood measures as a model calibration and sensitivity analysis framework, and the validity of Beven's Equifinality Manifesto (Beven 2006a).

2.1.1 Previous Applications of GLUE

Beven and Binley's (1992) paper introducing GLUE for use in uncertainty analysis of watershed models has now been extended well beyond rainfall-runoff watershed models to flood inundation estimation (Romanowicz et al., 1996; Cameron,

2006; Viola et al., 2009), atmospheric transport models (Page et al., 2008), ecological models (Pinol et al, 2005), schistosomiasis transmission models (Liang et al, 2005), algal dynamics models (Hellweger and Lall, 2004), crop models (Tremblay and Wallach, 2004), integrated urban drainage models (Thorndahl, 2008; Freni et al., 2009), water quality models (Smith et al, 2005), acid deposition models (Page et al, 2004), geochemical models (Zak et al., 1997), offshore marine sediment models (Ruessink, 2005), groundwater modeling (Christensen, 2004; Hassan et al., 2008), wildfire prediction (Bianchini et al. 2006) and others. Given the widespread adoption of GLUE analyses for a broad range or problems, it is appropriate that the validity of the approach be examined with care. Christensen (2004), Montanari (2005), Mantovan and Todini (2006), Stedinger et al. (2008), and this study provide such reviews.

2.1.2 Organization of the Chapter

The companion paper for this chapter, Stedinger et al. (2008), showed how the GLUE methodology when properly implemented with a statistically valid likelihood function can provide prediction intervals for future observations which will agree with widely accepted and statistically valid analyses. Much of this chapter is published as part of Stedinger et al. (2008). However, materials not included in that article can be found in Figures 1, 2, 6, 9, 10, 11, 12, and 15. Figures 11 and 12 are particularly interesting because they illustrate the difference between the parameter distributions obtained with different likelihood measures.

This study begins by developing the GLUE methodology and comparing it with a Bayesian analysis using a formally correct likelihood function. Section 2.2 discusses 3 methods for parameter estimation for a linear regression model. The aim of section 2.3 is to evaluate GLUE using a linear rainfall/runoff model so that model

calibration is a linear regression problem for which exact expressions for uncertainty are well known and understood. It is common practice to test new methods and theories on old well-understood problems and special cases to see if the new proposals provide valid solutions and thus are credible. Simple cases are, after all, special cases of complicated situations: so one cannot logically claim a method works for complicated situations if it does not work for the simple situations that are special cases.

2.2 Three Descriptions of Parameter Uncertainty

This section presents 3 methods for parameter estimation for a linear regression model describing a watershed: classical least squares regression, Bayesian, and GLUE.

2.2.1 Classical Least Squares Regression

The method of least squares is the classical way of estimating the parameters in a linear regression equation. Suppose annual streamflow observations Q_t are related to annual rainfall observations P_t by the simple linear model

$$Q_t = \alpha + \beta \cdot P_t + \varepsilon_t \tag{1}$$

wherein α and β are model parameters, and ε_t are normal and independently distributed (NID) model errors. The vertical distance of the *t* th point from any straight line is labeled ε_t. The straight line for which $\sum \varepsilon_t^2$ is minimal is called the least squares line. The least squares line can be written as

$$\hat{Q} = a + b \cdot P \tag{2}$$

where

$$b = \frac{\sum_{t=1}^{n}(P_t - \overline{P})(Q_t - \overline{Q})}{\sum_{t=1}^{n}(P_t - \overline{P})^2} \tag{3}$$

and

$$a = \overline{Q} - b \cdot \overline{P}\,. \tag{4}$$

Here a and b are the least squares estimates of the corresponding parameters α and β (Devore, 2008).

Many classical statistics textbooks introduce the derivation of confidence intervals for parameters and prediction intervals describing uncertainty associated with (1). The variance of prediction describing the difference between that prediction $\hat{Q}$ and a possible future flow Q_f *for* a particular precipitation value P_0; it is given by

$$\sigma_P^2 = E\left[\left(Q_f - \hat{Q}\right)^2\right] = \sigma_\varepsilon^2\left(1 + \frac{1}{n} + \frac{(P_o - \overline{P})^2}{\sum_{t=1}^{n}\left(P_t - \overline{P}\right)^2}\right) \tag{5}$$

where σ_ε^2 is generally estimated by s_ε^2 the classic unbiased estimator of the model residual variance (Devore, 2008). The formula for the s_ε^2 is

$$s_\varepsilon^2 = \frac{\sum_{t=1}^{n}(Q_t - \hat{Q}_t)^2}{n-2}\,. \tag{6}$$

For large sample sizes, the uncertainty in the two model parameter estimates a and b slowly vanishes like 1/n in eqn (5). Thus for large n, σ_P^2 in eqn (5) reduces to just σ_ε^2 which is independent of n, because the dominant error is the inability of even the best model to perfectly forecast individual values of Q.

Similarly if one's interest is in a mean prediction based on the regression $\hat{Q} = a + b \cdot P_o$ so that one is interested in the precision of the forecasted value given the likely sampling error in both parameters, then the variance of concern is given by

$$E\left\{\left[(a+bP)-\hat{Q}\right]^2\right\}=\sigma_\varepsilon^2\left(\frac{1}{n}+\frac{(P_o-\overline{P})^2}{\sum_{t=1}^{n}\left(P_t-\overline{P}\right)^2}\right) \tag{7}$$

2.2.2 Bayesian Inference

A classical treatment of our regression problem seeks a point estimate of the unknown parameters α and β. Confidence intervals can describe uncertainty. In contrast, with the Bayesian approach we characterize the uncertainty in the parameter through a probability distribution over their possible values. Observations modify this distribution by virtue of Bayes' theorem, with the effect of the data being mediated through the likelihood function. When estimating parameters, Bayesian inference can be depicted as

prior distribution + observed data → posterior distribution

Our knowledge about the parameter is encapsulated in a probability distribution describing first our prior knowledge and then a posterior probability distribution reflecting the additional information provided by the data. The posterior distribution is affected not only by the specific conditions of the experiment, but also by all of the specifics of model the system.

Denote the set of streamflow observations Q_t for t=1,…,n to be used for model calibration by the vector Q. Using Bayesian theorem the posterior probability distribution function (pdf) of the parameter vector θ based on these observations is

$$f_{\theta|Q}[\theta \mid Q]=c\, f_{\varepsilon|\theta}[\varepsilon \mid \theta] f_\theta[\theta] \tag{8}$$

Here c is a normalization constant, θ is the parameter vector, and $f_\theta[\theta]$ denotes the prior pdf for the parameter vector θ, and ε is the vector of model errors computed as $\varepsilon_t = Q_t - \hat{Q}_t$ where $\hat{Q}_t = M(\theta)$ represents the stream flow model prediction vector

which depends on the parameter vector θ. In equation (8), the subscript $\theta|Q$ has been added to the posterior distribution to clearly emphasize that this is the posterior distribution of the model parameters θ given a particular calibration dataset Q. The pdf $f_\theta[\theta]$ is called the "prior" distribution of the model parameters. Because it is based on information one had pertaining to the parameters prior to model calibration. The function $f_{\varepsilon|\theta}[\varepsilon|\theta]$ is called the likelihood function and must represent the probability that the watershed model would produce the observed flow vector. The likelihood function is the probability of the actual data being obtained for any specific values of the parameters.

In our initial study, we consider the linear watershed model in equation (1) with errors ε_t that are normal and independently distributed (NID) with zero mean and unknown variance σ_ε^2. Then the likelihood function is the probability that the observed value Q_t equals the predicted value $\hat{Q}_t$:

$$f_{\varepsilon|\theta}[\varepsilon_t|\theta] = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left[-\frac{(Q_t - \hat{Q}_t)^2}{2\sigma_\varepsilon^2}\right] \tag{9}$$

wherein Q_t is the observed flow and $\hat{Q}_t$ is the predicted flow so that $\varepsilon_t = Q_t - \hat{Q}_t$. Substitution of (9) into (8) for each observation yields the correct posterior pdf

$$f_{\theta|Q}[\theta|Q] = c\, f_\theta[\theta] \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left[-\frac{(Q_t - \hat{Q}_t)^2}{2\sigma_\varepsilon^2}\right] \tag{10}$$

where *n* is the length of the streamflow record used to calibrate the parameters, and c is simply a constant term whose value may be determined by the requirement that the integral over all parameters of the posterior density function must equal one. Equation (10) describes the posterior distribution of the model parameters for the case for NID

model errors, and needs an estimate of σ_ε^2. In a full Bayesian analysis, σ_ε^2 and θ have a joint posterior distribution. However, if one has sufficient data, the value of σ_ε^2 will be very close to the maximum likelihood estimate. Now consider the "best" estimator to be the maximum likelihood estimator (MLE) $\hat{\sigma}_\varepsilon^2$ equal to

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n}\sum_{t=1}^{n}(Q_t - \hat{Q}_t^{MLE})^2 \tag{11}$$

where $\hat{Q}_t^{MLE}$ is the model predictions obtained using the MLE of the model parameters. The posterior distribution for the model parameters θ given NID errors can be obtained by substitution of (11) into (10) to obtain

$$f_{\theta|Q}[\theta \mid Q] = c(\hat{\sigma}_\varepsilon)^{-n}(2\pi)^{-n/2}\exp\left[-\frac{n}{2}\frac{\sum_{t=1}^{n}\left(Q_t - \hat{Q}_t\right)^2}{\sum_{t=1}^{n}\left(Q_t - \hat{Q}_t^{MLE}\right)^2}\right] f_\theta[\theta] \tag{12}$$

where $c(\hat{\sigma}_\varepsilon)^{-n}(2\pi)^{-n/2}$ is constant in a particular application. Using independent uniform priors on each parameter set, the terms $c(\hat{\sigma}_\varepsilon)^{-n}(2\pi)^{-n/2}$ and $f_\theta[\theta]$ may be combined to yield

$$\begin{aligned} f_{\theta|Q}[\theta \mid Q] &= \kappa\exp\left[-\frac{n}{2}\frac{\sum_{t=1}^{n}(Q_t - \hat{Q}_t)^2}{\sum_{t=1}^{n}(Q_t - \hat{Q}_t^{MLE})^2}\right] \\ &= \kappa\exp\left[-\frac{n}{2}\frac{ns_\varepsilon^2}{\hat{\sigma}_\varepsilon^2}\right] \\ &= \kappa\exp\left\{-\frac{n}{2}\frac{\left[1 - R^2(\theta)\right]}{\left[1 - R^2\left(\hat{\theta}_{MLE}\right)\right]}\right\} \end{aligned} \tag{13}$$

where

$$s_\varepsilon^2 = \sum_{t=1}^{n} (Q_t - \hat{Q}_t)^2 / n,$$

$$R^2(\theta) = \left[1 - \frac{s_\varepsilon^2}{s_Q^2} \right], \text{ and}$$

$$R^2(\hat{\theta}_{MLE}) = \left[1 - \frac{\hat{\sigma}_\varepsilon^2}{s_Q^2} \right].$$

where κ is a constant term whose value may be determined by the requirement that the integral over all the parameters of the density function be equal to one. Here s_Q^2 is the variance of the observations, which is unaffected by the model parameters.

Equations (13) exhibit several key and important features: (1) the best fitting model determines the standard against which other solutions are compared, (2) the probability assigned to each parameter set depends upon how well the model with those parameters matches the calibration data, and (3) the length *n* of the calibration data set has a very large impact on the importance assigned to a parameter set not providing the best possible fit. If *n* is small, then there is insufficient data to resolve the values of the best parameters. However if *n* is large, then the calibration data should be much more informative and our ability to discriminate between different sets of parameters should increase. In the other words, for large sample size, the posterior distribution would be influenced more by the data than by the assumed prior distribution.

To generate an uncertainty distribution for what might be a future observation, one needs to consider the uncertainty in the parameters described by $f_{\theta|Q}\left[\hat{\theta}_i \middle| Q\right]$ as well as the likely difference between the model prediction and an observed value. The latter difference is due to a range of errors including the simplicity of the model compared to

reality, and limitations in the input data reflecting possible measurement errors and their misrepresentation of the needed inputs (data are often point values when areal averages are needed). Beven (2006a) provides a very complete description of the many sources of error and the challenges they pose for parameter estimation. There are promising approaches which represent input data error explicitly, in addition to model errors and response-variable measurement errors. See Kavetski et al. (2002, 2006ab), Vrugt et al. (2005), Moradkhani et al. (2005ab), Kuzcera et al. (2006), Clark and Vrugt (2006), and Huard and Mailhot (2006).

If one wishes to generate an uncertainty interval for a future observation, the predictive distribution $f_{Q_f|Q}[Q_f \mid Q]$ for a future observation Q_f given the data vector Q should be employed, which is given by (Zellner, 1971) as

$$f_{Q_f|Q}[Q_f|Q] = \int f_{Q_f|\theta}[Q_f|\theta] f_{\theta|Q}[\theta|Q] d\theta \,. \tag{14}$$

In developing a predictive distribution for a future observation, one needs to consider the uncertainty in the parameters, and also the deviations of the observed flows Q_f from the best prediction which is generally more important (Mantovan and Todini, 2006, p. 373).

Here we use the term “uncertainty interval” to describe an interval intended to contain an uncertain parameter with a specified probability or frequency [often called credible regions in the Bayesian literature, (Zellner, 1971), and confidence intervals in classical statistics]. “Prediction interval” will describe an interval for a future observation which depends both on parameter uncertainty, and upon future data, model and output measurement errors.

2.2.3 Generalized Likelihood Uncertainty Estimation (GLUE)

The idea of the GLUE is to combine a priori knowledge of the model parameters captured by the prior pdf, with new information reflected in the observed data as represented by the likelihood measure to obtain a posterior pdf of the model parameters. Rejecting a traditional statistical basis for the likelihood function, Beven and Binley (1992, p. 281) introduced their own requirements on their likelihood measures arguing that "the choice of a likelihood measure will be inherently subjective." The calculation of the likelihood of a given set of models and parameters is the key feature of GLUE, and differs from the classical method of calibration and uncertainty estimation. Parameter sets that result in their goodness-of-fit/likelihood values below a certain threshold are termed 'non-behavioral' and are discarded. The remaining 'behavioral' parameter sets are assigned rescaled likelihood weights that sum to 1, and thus look like probabilities.

To obtain uncertainty intervals for parameters and model predictions using these rescaled likelihood weights, the parameter values on the model outputs are ranked so that the rescaled likelihood weights can be used to form a cumulative distribution. From that distribution, quantiles are selected to provide uncertainty intervals for the variable of concern.

According to Beven and Binley (1992) only two conditions must be satisfied by the 'likelihood measure':

1. "It should be zero for all simulations that are considered to exhibit behavior dissimilar to the system under study."
2. "It should increase monotonically as the similarity in behavior increases."

Furthermore, Beven and Binley (1992) argue that the likelihood function can be chosen from "many of the goodness-of-fit indices used in the past." Beven and Binley

(1992) acknowledge that the choice of likelihood function will greatly influence the resulting uncertainty intervals and so argue that this choice must be made explicit so they can be the "subject of discussion and justification" (Beven and Freer, 2001, p. 18).

Several likelihood measures have been proposed and used in previous applications of the GLUE methodology. Table 1 in Beven et al. (2000) provides a summary. A popular likelihood measure, the inverse error variance, was introduced by both Beven (1989) and Beven and Binley (1992), where

$$L_{IV} = \left[s_\varepsilon^2\right]^{-N} \tag{15}$$

Here s_ε is the standard deviation of the model errors, and N is called the 'shaping factor' by Beven and Freer (2001). Beven and Binley (1992) used N=1 but suggested that the shaping factor can be chosen by the user. As expected, different values of $N>0$ lead to different descriptions of uncertainty (Ratto et al., 2001). Increasing N gives greater weight to model parameters which yield a better 'goodness of fit'. As N approaches infinity the best parameter set that is generated will be given a weight of 1, while all other parameter sets will be discarded. As N approaches zero, all parameter sets receive equal weight.

The likelihood measure adopted most frequently employs the efficiency index introduced by Nash and Sutcliffe (1970), so that

$$L_{NS} = \left[1 - \frac{s_\varepsilon^2}{s_Q^2}\right]^N \tag{16}$$

where s_ε is the standard deviation of the errors, s_Q is the standard deviation of the observations and again, N is a "shaping parameter". Examples of the application of GLUE with this efficiency index can be found in Kinner and Stallard (2004), Hossain and Anagnostou (2005), Uhlenbrook and Sieber (2005) and a variety of other studies summarized in Table 1 of Beven et al. (2000). This index only makes sense if in the

computation of s_ε^2 one assumes that the errors have zero mean. Otherwise systematic bias would be ignored. Freer et al. (1996) used (15) with N = 1 and 30.

The third likelihood measure in Table 1 of Beven and Freer (2001) is

$$L_{EXP} = \exp\left[- N \frac{s_\varepsilon^2}{s_Q^2} \right]. \tag{17}$$

We have not used L_{EXP} because it is very similar to L_{NS} for s_ε^2 small relative to s_Q^2, and fails to go to zero as is desirable when $s_\varepsilon^2 \to s_Q^2$.

Others introduced their own likelihood measures based on the requirements suggested by Beven and Binley (1992). For example, the inverse of the mean square error model was adopted by Mertens et al. (2004)

$$L_{1/MSE} = \frac{1}{MSE} = \left[\frac{1}{n} \sum_{t=1}^{n} (\varepsilon_t^2) \right]^{-1} \tag{18}$$

which is close to (15) with N=1; the difference is that (18) clearly includes a penalty for bias $\{E[\varepsilon] \neq 0\}$. Eqn. (15) has no penalty for bias unless s_ε is computed assuming the mean value of ε_i is zero, which we suspect has generally been the case. Other likelihood measures have been developed for particular modeling applications (Page et al., 2004; Mertens et al., 2004), and several methods have been proposed for combining likelihood measures (Engeland and Gottschalk, 2002; Uhlenbrook and Sieber, 2005; Page et al., 2004; Mo and Beven, 2004; Beven et al., 2008).

2.2.4 Comparison of Likelihood Functions

In our experiments, we examine how closely GLUE likelihood measures resemble a statistically correct likelihood function for the NID model. The likelihood measures in (15) and (16) are employed in our experiments. Here the sample variance

of the errors ε_i is computed assuming the true mean is zero, which provides a penalty for bias. Using standard notation, the GLUE likelihood measures can be written to resembles (13) as follows

$$f_{IV}[\theta \mid Q] = \kappa_1 \left[s_\varepsilon^2 \right]^{-N} = \kappa_1 \left[\left(1 - R^2(\theta_i)\right) s_Q^2 \right]^{-N} \tag{19}$$

$$f_{NS}[\theta \mid Q] = \kappa_2 \left[1 - \frac{s_\varepsilon^2}{s_Q^2} \right]^N = \kappa_2 \left[R^2(\theta_i) \right]^N \tag{20}$$

where again the subscript i denotes the ith parameter set, κ_1 and κ_2 are constant terms chosen to make $\sum_{i=1}^{T} f(\theta_i \mid Q) = 1$ across all T parameter sets, s_Q^2 describes the variance of the observed streamflows and $R^2(\theta_i)$ is defined in the pervious section.

The GLUE likelihood measures in (19)-(20) are different from the form of the Bayesian likelihood functions in (8)-(13). The correct likelihood function depends critically upon $R^2(\hat{\theta}_{MLE})$ which reflects how well the model really can fit the data, as well as upon the length *n* of the calibration sample upon which the analysis is based; more data should provide more information. Neither of these factors appears in (19) and (20). The GLUE likelihood measure most commonly used in the past, and given in (19) and (20), depends on s_Q^2 which describes how much variation there is in the data, not how well the best models can reproduce the data, or how long a sample one has to estimate the model parameters, as does the correct likelihood function in (13). Suppose one had a long calibration data set and the best models were almost perfect ($\sigma_\varepsilon^2 = 0, R^2(\hat{\theta}_{MLE}) \approx 1$), one would then find as entirely unreasonable a parameter set with $R^2(\theta_i) = 0.80$. On the other hand, if *n* is small and the best model only achieves $R^2(\hat{\theta}_{MLE}) = 0.81$, then a parameter set with $R^2(\theta_i) = 0.80$ is probably just as credible as the optimal parameter set. The informal likelihood functions that are so

often used with GLUE fail to recognize this critical message.

2.3 Case Study of Linear Watershed Model

In this section, we have chosen a simple linear rainfall/runoff model to compare the three estimator procedures. It enables us to compare the uncertainty intervals based on a commonly adopted GLUE likelihood measures with Bayesian inference and with exact confidence intervals based on classical statistical theory.

2.3.1 Linear Watershed Model

Consider the simple univariate normal linear model in equation (1). Use of this model enables us to compare the uncertainty intervals generated using commonly adopted GLUE likelihood measures, with exact confidence intervals based on available classical statistical theory. In equation (1), the annual streamflow Q_t is related to annual rainfall observations P_t by a simple linear equation where α and β are model parameters, and ε_t are NID model errors with zero mean and constant σ_ε^2.

For simplicity, we can imagine that the marginal distributions of Q and P are fixed. Suppose that mean values of P and Q are 100 cm and 60 cm respectively, and their respective standard deviations are 20 and 15 cm. Thus coefficients of variation of P and Q are 0.20 and 0.25, respectively. We will consider a range of cases distinguished by the sample size n, the place at which we predict Q, and the model precision determined by the model error variance σ_ε^2. If we describe model precision by R_{Model}^2, then we have

$$\sigma_\varepsilon^2 = (1 - R_{Model}^2)\sigma_Q^2 \tag{21}$$

where the corresponding values of the parameters are

$$\beta = R \cdot (\frac{\sigma_Q}{\sigma_P}) = 0.75 \cdot R^2_{Model} \tag{22}$$

$$\alpha = \mu_Q - \beta \cdot \mu_P = 60 - 0.75 \cdot R^2_{Model} \tag{23}$$

To define a prior distribution of the parameters we must consider the possible values of α and β. If all the rain runs off, then β=1. If it all evaporates, β would be zero. However, our rain gauge may not be accurate in that it might not yield the average rainfall for the basin. We can consider as our prior on β the interval (0, 2).

For the prior on α, if the rain gauge is uncorrelated (β=0) with the runoff for the basin, then α would be the mean runoff. If we assume the mean rainfall is 60 cm, then α could be as large as 60cm; however, if our rainfall gauge is inaccurate, then α could be larger; perhaps 200 cm. To be safe and to reflect possible losses in the basin and possible values of β, we might assume a lower bound for α to be -100cm. Thus an uncertainty range could be (-100, 200). We do not really believe independent and uniform distributions over these ranges are reasonable. In fact our discussion reveals the linkage between the two parameters. Still, we follow the logic GLUE applications adopt to provide a reasonable implementation of that method.

For the purposes of streamflow simulations with both good and poor parameter estimates, the estimate of the flow with the linear model is taken to be:

$$\hat{Q} = \max[0, a + b \cdot P] \tag{24}$$

where *a* and *b* are ordinary least squares (OLS) estimates of the parameters α and β, respectively.

2.3.2 Monte Carlo analysis

A Monte Carlo analysis was conducted to compare GLUE results with the

exact confidence and prediction intervals for this situation obtained with (5) and (7). Our experimental procedure is as follows:

1. For the linear model $Q_t = \alpha + \beta \cdot P_t + \varepsilon_t$ with R^2_{Model} = 0.90, compute the model error variance, and the values of α and β using (21), (22), and (23).
2. Generate a single sample of precipitation P_t, model errors ε_t , and corresponding streamflow Q_t observations t = 1,…n (=40). The idea here is to mimic the problem faced by a hydrologist who typically has only a single record of n observations.
3. Compute the ordinary least squares estimates a and b for model parameters α and β. In this case, the OLS estimators are the MLE's.
4. Compute the exact 95% confidence intervals for the mean flow associated with precipitation P_o (90^{th} percentile of P) using (7) and the 95% prediction intervals for an observed flow given each precipitation value for a simple linear regression based on n=40 observation; see Figure 5.
5. Following the standard GLUE procedures, generate m = 10,000 parameter sets (a_i, b_i) drawing from uniform distributions over the intervals (-100, 200) and (0, 2).
6. For each set of parameters (a_i, b_i) compute model predictions $\hat{Q}$ for n= 1, …40 using the n precipitation P_t observations.
7. Compare each of the m sets of $\hat{Q}$ to the observations Q_t, to compute the goodness-of-fit statistic, $R^2(\theta_i)$ for i=1,…,m.
8. Using the m values of $R^2(\theta_i)$ for i=1,…m along with n, $R(\hat{\theta}_{MLE})$ and s_Q, compute the Bayesian likelihood function (13), and the GLUE likelihood measures L_{NS} with shaping factor N=1 and 30, and L_{IV} with N=1.
9. If a behavioral threshold is adopted, reject non-behavioral parameter sets.
10. For each set of parameters compute probabilities $p(a_i, b_i)$, i=1,…,m, using the

different likelihood functions and behavioral thresholds.

11. Each set of parameters is used to generate one estimate of the streamflow Q associated with precipitation P_o. The probability associated with each parameter set is assigned to the flow estimate it produces.

12. Sort the flows with their corresponding probabilities to create the pdf for forecast uncertainty, and use these to generate uncertainty intervals.

2.3.3 Impact of GLUE Likelihood Measures on the Parameter Sets

Clearly Beven, Binley, Freer and others who have advanced this scheme do not trust their likelihood measure to be able to distinguish between realistic (behavioral) and unrealistic (nonbehavioral) data sets, and thus impose an independent 'behavioral' threshold criterion. If the statistical analyses were correct, it should be able to distinguish between behavioral and non-behavioral solutions without the imposition of an arbitrary and rigid cutoff. As we will show in this section, a correct statistical analysis does just that in our example.

Figure 1 illustrates the comparison of the probabilities assigned to different parameter sets employing four GLUE likelihood measures as a function of model goodness of fit $R^2(\theta)$: The likelihood function for normal and independent distributed (NID) model errors in (9), the Inverse Variance (IV) likelihood measure in (15) with N=1, and the Nash-Sutcliffe (NS) efficiency index in (16) with N=1 and 30. There is no behavioral threshold. Each parameter set can be described by the goodness-of-fit value $R^2(\theta)$ computed using equation (13). Equations (13) for NID, equation (19) for IV, and equation (20) for NS express the three likelihood functions in terms of $R^2(\theta)$. With 10,000 randomly generated parameter sets, the best sets come very close to the best possible $R^2(\theta)$ value for this data set of almost 0.90, which corresponds to the

vertical line in the Figure 1. Some 90% of the generated parameter sets had negative values of $R^2(\theta)$. A correct statistical analysis using the NID likelihood function indicates that parameter sets with $R^2(\theta)$ values less than 0.85 have miniscule probabilities: that is with n = 40 observations, we can be enormously certain that such parameter sets do not represent the true values of the parameters.

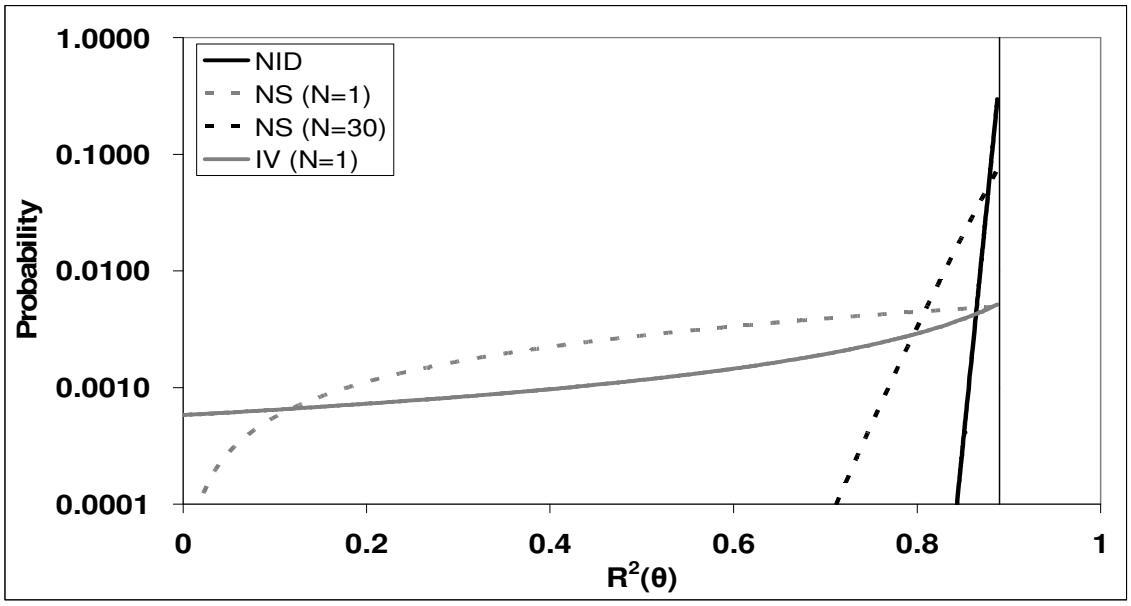


Figure 1: Comparison of the probabilities assigned to different parameter sets employing GLUE likelihood measures as a function of model goodness of fit $R^2(\theta)$, for the proposed parameter vector $\boldsymbol{\theta}$ for n=40

Figure 2 describes the cumulative distribution functions assigned to the parameter sets. These represent the uncertainty assigned to the parameter sets by the four likelihood functions. The cumulative distribution function obtained using the NID likelihood function has the smallest variance and 95% probability intervals: the realistic range for parameter sets are those with $R^2(\theta)$ values greater than 0.85.

One sees from the cumulative distribution function obtained with IV where

N=1 that the probability assigned to parameter sets with $R^2(\theta) \leq 0$ is 0.43. It seems unreasonable to an assign probability of over 40 % to parameter sets with $R^2(\theta)$ values less than 0. The problem is that the analysis does not employ a valid likelihood function, and as a result the probabilities assigned to the different parameters sets are insensitive to how well the model fits the data.

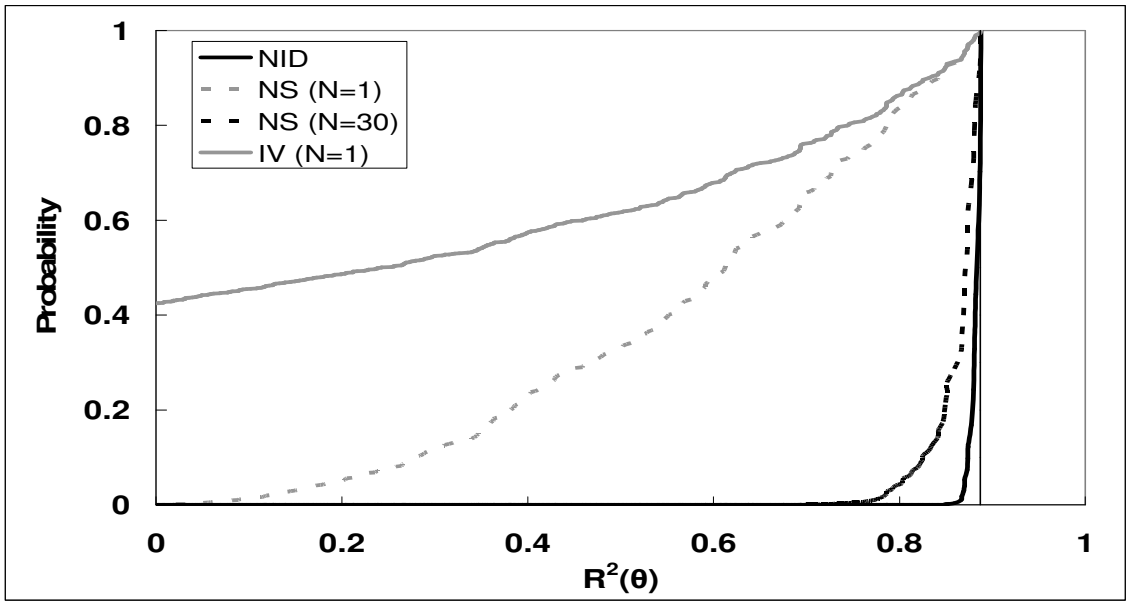


Figure 2: Cumulative distribution functions of the probabilities assigned to different parameter sets employing GLUE likelihood measures as a function of model goodness of fit $R^2(\theta)$, for the proposed parameter vector $\boldsymbol{\theta}$ for *n*=40

To obtain reasonable results with the IV likelihood measures, behavioral thresholds of $R^2(\theta) = 0$ and $R^2(\theta) = 50\%$ were adopted resulting in cases IV00 (*N*=1) and IV50 (*N*=1). Figure 3 displays the comparison of the probabilities assigned to different parameter sets employing GLUE likelihood measures as a function of model goodness of fit $R^2(\theta)$. This Figure is different from Figure 1 in that it includes IV00 (*N*=1) and IV50 (*N*=1), instead of IV (*N*=1) which had no behavioral threshold. IV50

is only employed for models with a true $R^2(\theta)>0.50$.

One can see that probabilities obtained with NS (N=1) and IV00 (N=1) are similar, though IV00 probabilities are more peaky while larger probabilities also are assigned to very poor models until its behavioral threshold of $R^2(\theta) = 0$ is reached. GLUE with the NS and IV objectives finds many parameter sets to be plausible when they are actually beyond the realm of credibility using the true likelihood function NID. This explains the need in many GLUE analyses to impose a behavioral constraint on parameters sets. No such constraint is needed with the correct likelihood function for NID data.

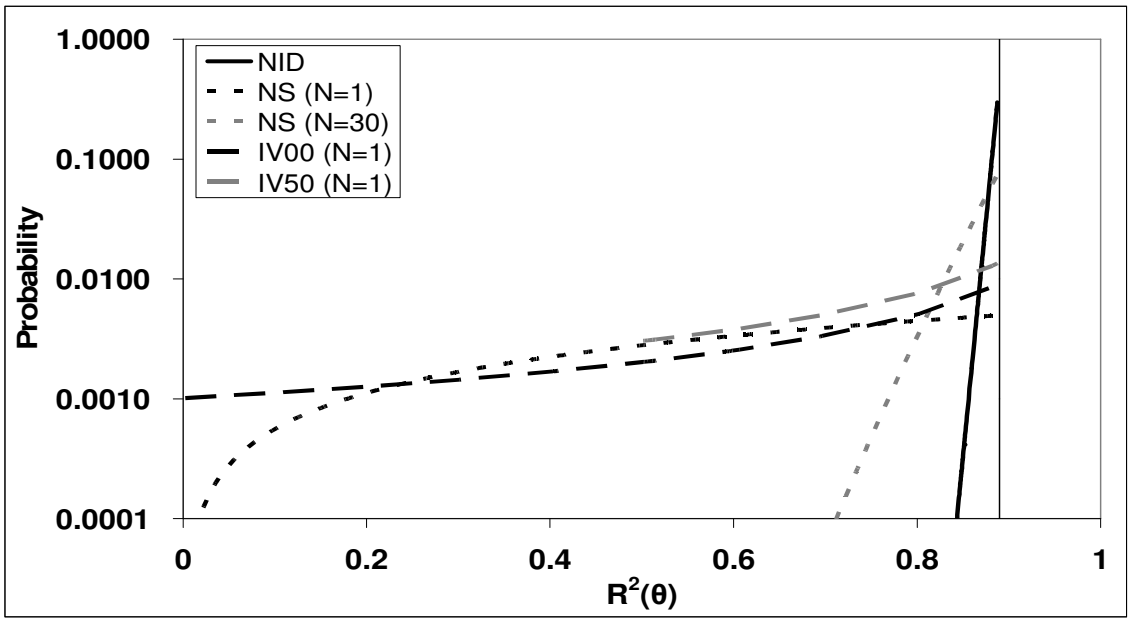


Figure 3: Comparison of the probabilities assigned to different parameter sets employing GLUE likelihood measures as a function of model goodness of fit $R^2(\boldsymbol{\theta})$, for the proposed parameter vector $\boldsymbol{\theta}$ with n=40 when R^2_{Model}=0.90

2.3.4 Impact of Likelihood Functions on Uncertainty Intervals

This section compares the uncertainty intervals generated by GLUE assuming different likelihood measures, with confidence intervals based on classical regression theory. First 95% confidence intervals for the model mean and 95% prediction intervals for a future observation for different P are displayed in the Figure 4. Figure 4 is based on a single sample of length n=40 generated from (1). Clearly the GLUE methodology should yield similar intervals when applied to this sample with the linear model because equations (5) and (7) are the correct answer (Devore, 2008). Using a correct Bayesian analysis with a noninformative prior yields essentially the same result, though the Bayesian interpretation would be different (Zellner, 1971).

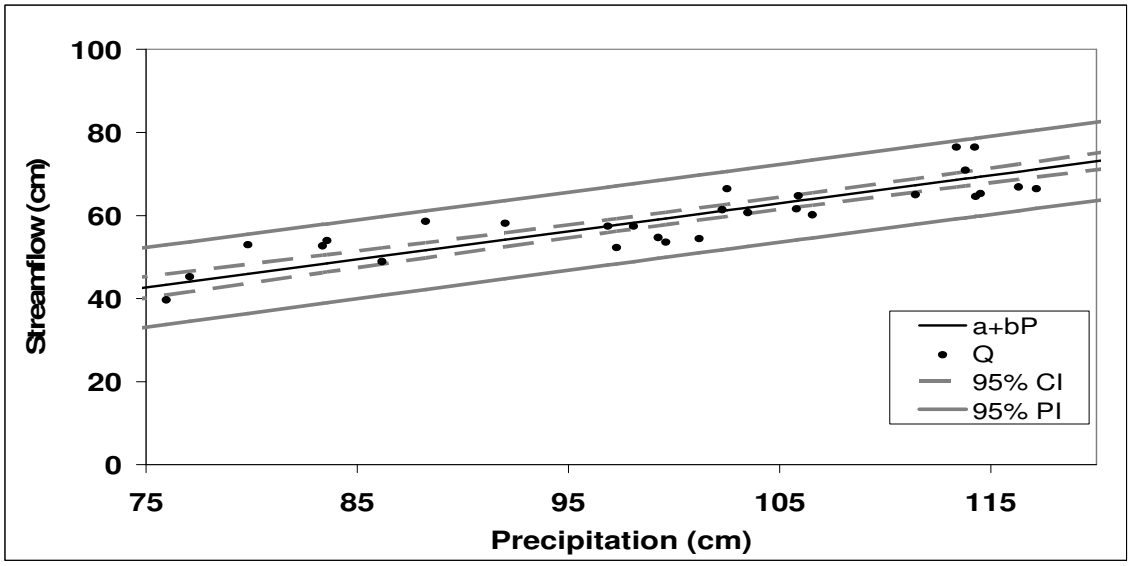


Figure 4: Example of confidence (uncertainty) intervals (CI) for the mean flow associated with each precipitation value and prediction intervals (PI) for an observed flow given each precipitation value for a simple linear regression based on n=40 observations computed using equations (5) and (7) with the Students t distribution with n-2 degrees of freedom to reflect uncertainty in the sample variance

Figure 5 compares the posterior probability distribution for the mean flow associated with precipitation P_o = 125.6 cm generated by 10,000 GLUE repetitions using four likelihood measures. In order of the maximum value of the probability density function, one obtains NID, NS (N=30), IV50, IV00 and finally NS (N=1) which is very similar to IV00. For the uncertainty distribution of the mean flow associated with P_o = 125.6 cm, illustrated in Figure 5, the NID likelihood is the correct description of the error distribution, so those results provide the correct posterior distribution for the mean flow conditional on the precipitation value with the adopted prior. One observes that the posterior distributions generated by the inverse-variance (IV) likelihood with both thresholds are different, and both are much wider than the correct posterior distribution obtained using the true likelihood function for NID model errors. The Nash-Sutcliffe (NS) results with the commonly used shaping factor N=1 are very similar to those for IV00, and both grossly overestimates the uncertainty in the mean flow. With N = 30, the NS likelihood generates a posterior distribution that better resembles the correct distribution. Clearly the choice of likelihood measure and the shaping factor N do matter.

The result in the Figure 6 shows cumulative distribution functions of each case in Figure 5. A cumulative distribution function better shows the uncertainty range. Certainly, NS and IV (IV00 and IV50) results with N=1 yield 95 % uncertainty intervals that are much wider than NID's. It is clear that the GLUE likelihood measures result in different uncertainty intervals and a gross misrepresentation of the model and parameter uncertainty.

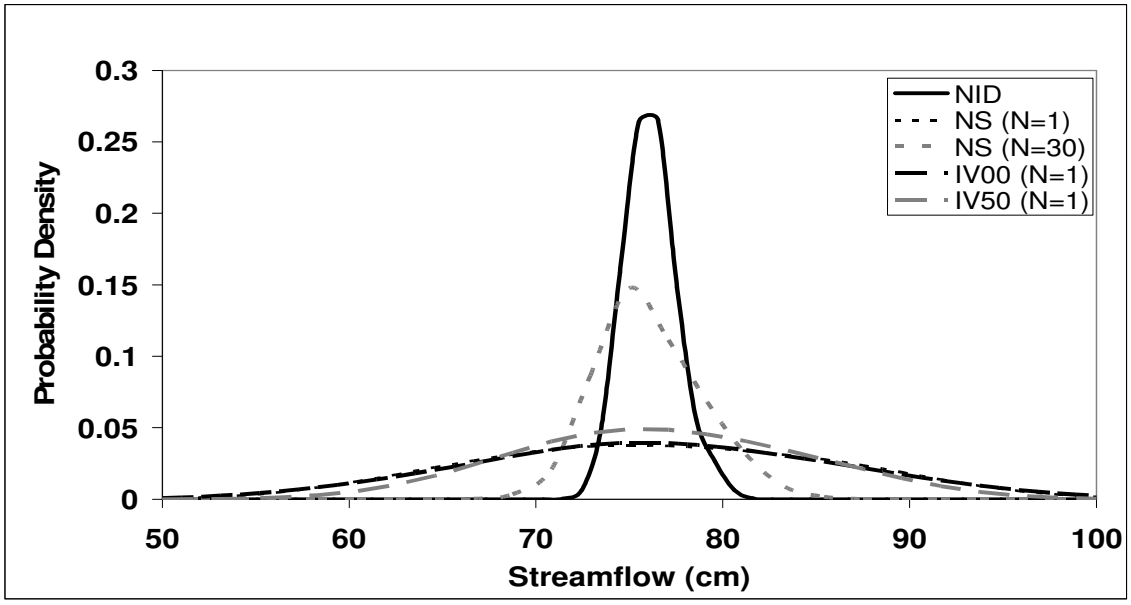


Figure 5: Posterior probability density functions for the mean flow associated with precipitation value P_o = 125.6 cm generated by GLUE with several likelihood measures (n=40)

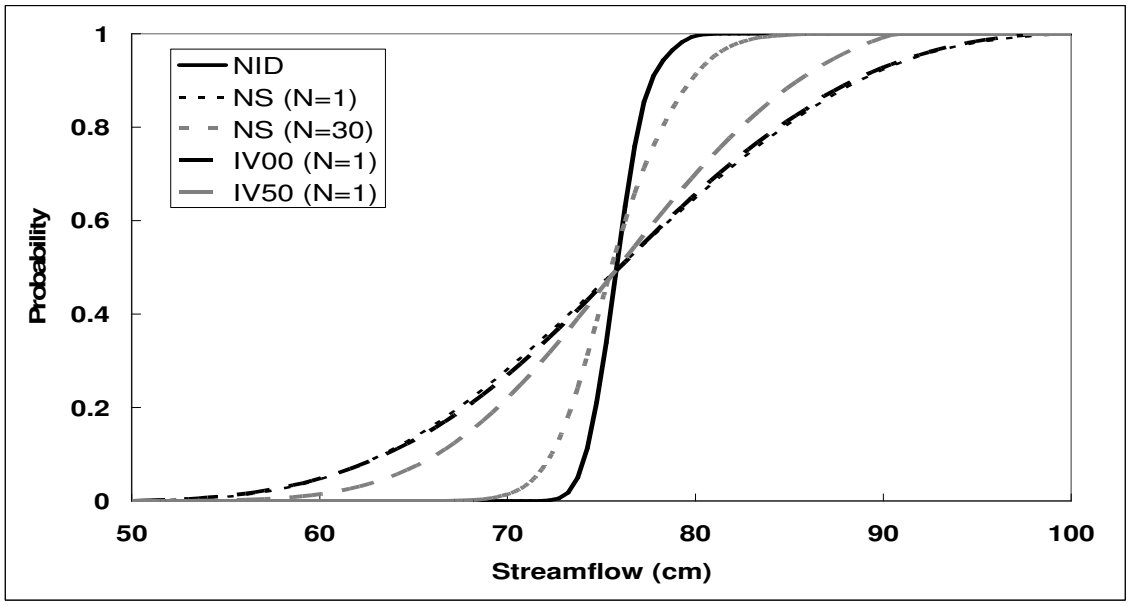


Figure 6: Posterior cumulative distribution functions for the mean flow associated with precipitation value P_o = 125.6 cm generated by GLUE with several likelihood measures (n=40)

Figure 7 illustrates the effect of the shaping factor *N*. The figure shows the end points of a 95% uncertainty interval for the mean flow associated with a precipitation value of P_o = 125.6 cm obtained using equation (7) and classical regression theory (REGR), GLUE with the NID likelihood which matches those results, and GLUE with the IV (IV00 and IV50) and NS likelihoods with *N* values from 1 to 100. For *N*=1, REGR and NID have the smallest intervals and are almost indistinguishable. IV50 is next followed by NS and then IV00. For *N*>2, NS yields the widest intervals. With IV, the GLUE values match the correct result for *N* values in the 20-30 range; with NS, the GLUE values would only match the correct result for *N* greater than 100. For *N* < 10, NS, IV00 and IV50 generated intervals are much too large. Figure 7 illustrates how GLUE yields arbitrary results using arbitrary likelihood measures.

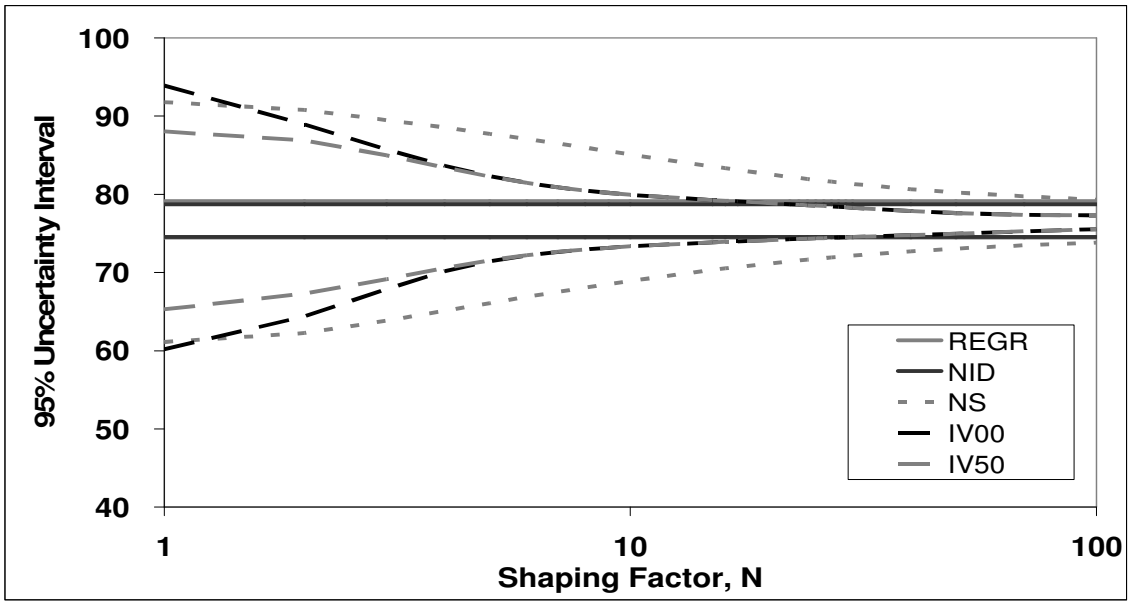


Figure 7: Effect of shaping factor *N* with NS and IV likelihoods on generated uncertainty intervals for the mean flow associated with precipitation value P_o = 125.6 cm, relative to the confidence intervals computed for the simple linear regression using equation (7) with the Students *t* distribution (*n*=40)

As noted above, sample size is very important in determining the precision of parameter estimators. Figure 8 is the same as Figure 7, except the sample size changes from 10 to 100 while R^2_{Model} has a fixed value of 0.90. (Here we consider samples corresponding to the first 10 observations through all 40 observations plus 60 additional observations. Thus different n values correspond to different samples.) While NS with N =30 comes close to the correct result with $n < 20$, it is far away for $n > 60$. NS and IV results with $N = 1$ are just absurd.

A serious concern with subjective GLUE likelihood measures is their lack of dependence on sample size n. One expects the confidence intervals for a regression model to become narrower as more observations are available for calibration (Mantovan and Todini, 2006). As Figure 8 demonstrates, the uncertainty intervals for the mean flow associated with P_o = 125.6 cm generated using GLUE likelihood measures remain constant regardless of sample size. This makes absolutely no sense: more data increases the precision of parameter estimators – the problem here is that arbitrary likelihood measures NS and IV fail to reflect the value of sample information.

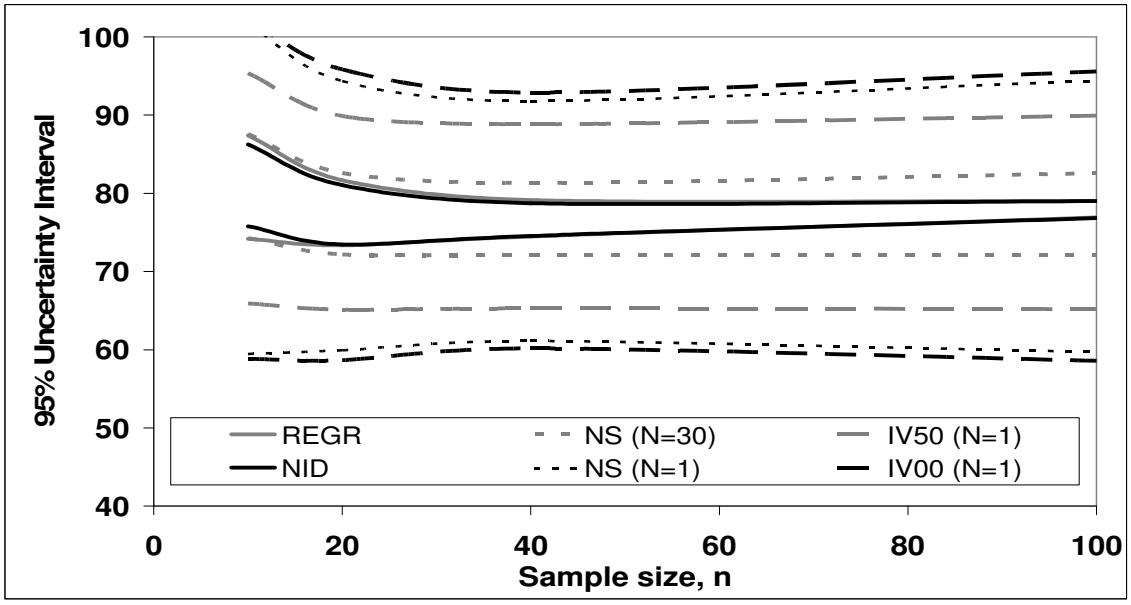


Figure 8: Effect of sample size *n* on uncertainty intervals for the mean flow associated with precipitation value P_o = 125.6 cm for simple regression using equation (7) with the Students *t* distribution compared with the corresponding uncertainty intervals obtained using various likelihood measures. R^2_{Model}=0.90

The impact of the threshold for the generated parameter sets on the uncertainty intervals is an extremely important phenomenon and it only occurs with the GLUE procedure when the informal likelihood measures are used. Figure 9 illustrates how the behavioral threshold for NS and IV with *N*=1 impact the 95% uncertainty intervals on the GLUE methodology, with threshold values from 0 to 0.8. As anticipated, 95% uncertainty intervals become narrower as higher threshold is available; however, NS and IV with *N*=1 results never match the correct 95% uncertainty intervals within threshold range from 0 to 0.8. There would be no way of determining a proper threshold which would lead to correct uncertainty intervals, if the modeler adopted IV or NS with *N*=1.

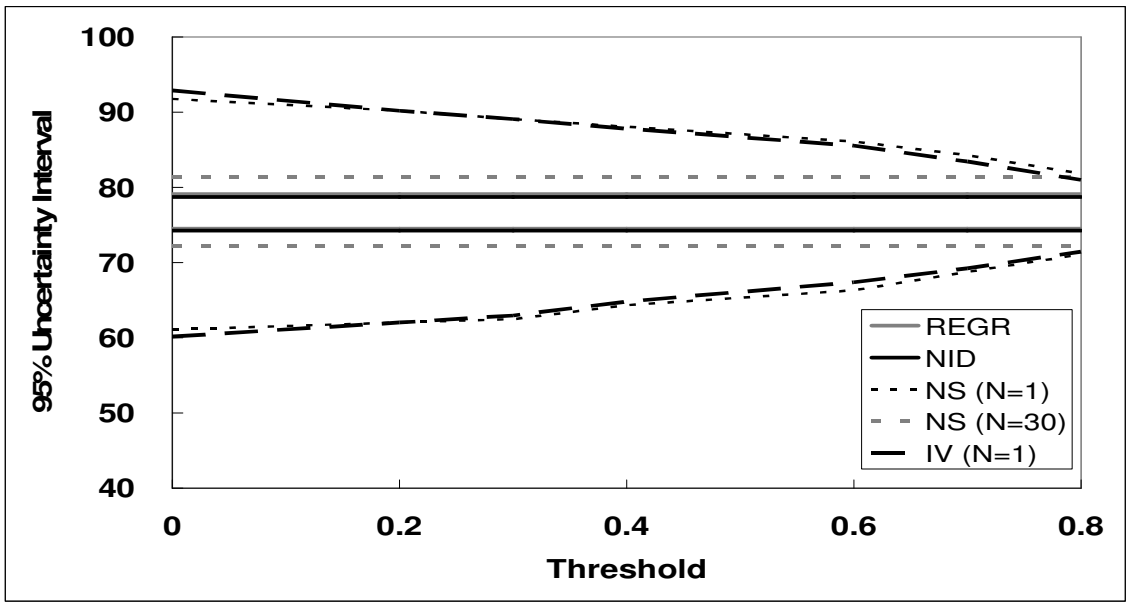


Figure 9: Effect of threshold on uncertainty intervals for the mean flow associated with precipitation value P_o = 125.6 cm for simple regression using equation (7) with the Students *t* distribution compared with the corresponding uncertainty intervals obtained using various likelihood measures. n=40, R^2_{Model}=0.90

2.3.5 Impact of Likelihood Functions on Posterior Distributions

This section compares the uncertainty distribution of parameters generated by GLUE with different likelihood measures, with uncertainty distribution of parameters based on Bayesian inference.

The posterior distributions for parameter *a* generated by GLUE with 5 likelihood functions are plotted in Figure 10. The shapes of all posterior distributions for parameter *a* are quite similar to Figure 5. The NID likelihood yields a much smaller variance. The posterior distributions resulting from IV with both thresholds are wider than with the correct posterior distribution obtained with the NID likelihood function. IV00 results with N=1 resembles NS (N=1) results. Compared to IV00

(N=1) results, the posterior distribution obtained using the NS (N=1) likelihood has lower peaks and heavier tails.

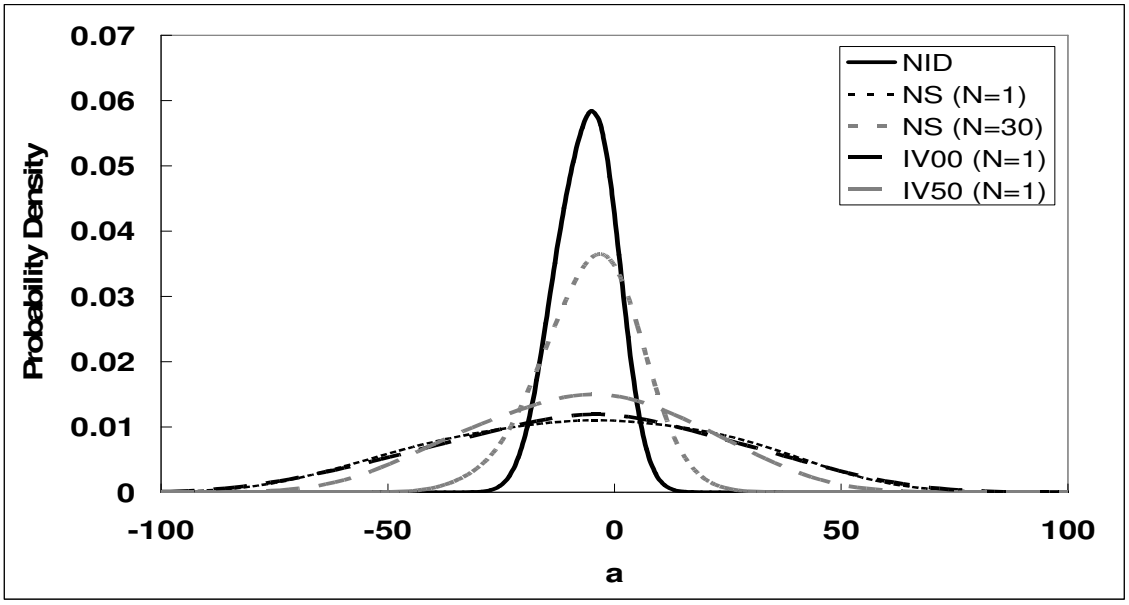


Figure 10: Posterior probability density functions for parameter *a* generated by GLUE with several likelihood measures

Figure 11 illustrates the resulting posterior distributions for parameter *b*. Again, one observes that the posterior distributions generated by NS and IV00 with N=1 likelihood measures are much wider than NID likelihood results. With the shaping factor N=30, the NS likelihood yields a distribution that better resembles the correct posterior distribution obtained using the true likelihood function for NID model error. Clearly, use of the GLUE likelihood measures overestimates the uncertainty of parameter *b* with a larger variance than the NID results in Figure 11. The posterior interval of IV00, IV50 and NS with N=1 results even include zero in their posterior intervals.

Figure 12 shows the relationship between parameters *a* and *b* in a form of 90% contour plots based on a multivariate normal approximation to their joint distribution. Each ellipse encloses the smallest area in a Bayesian analysis that has a 90 % probability of containing the unknown parameters. A negative correlation between the parameters is obvious.

As expected, the 90% contour generated by NID likelihood results has the smallest and a very different region. The 90% probability region of IV00 (N=1) results is approximately equal to the NS (N=1) contour and both still overestimate the parameter uncertainty. Figure 12 indicates that GLUE likelihood measures fail to represent what is necessary and what is not necessary about the likely value of model parameters.

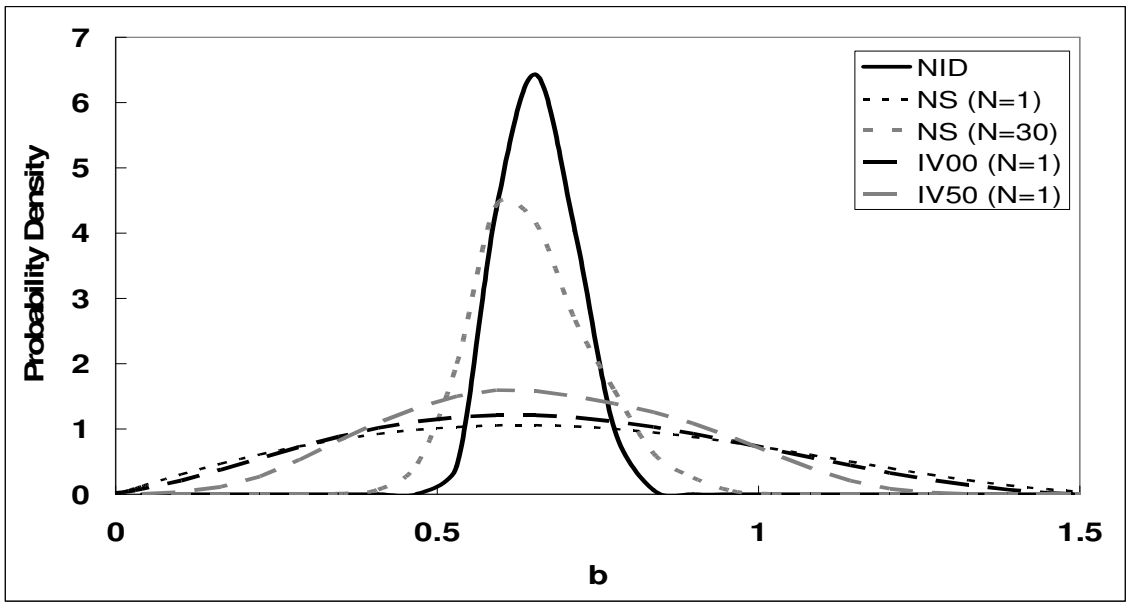


Figure 11: Posterior probability density functions for parameter *b* generated by GLUE with several likelihood measures

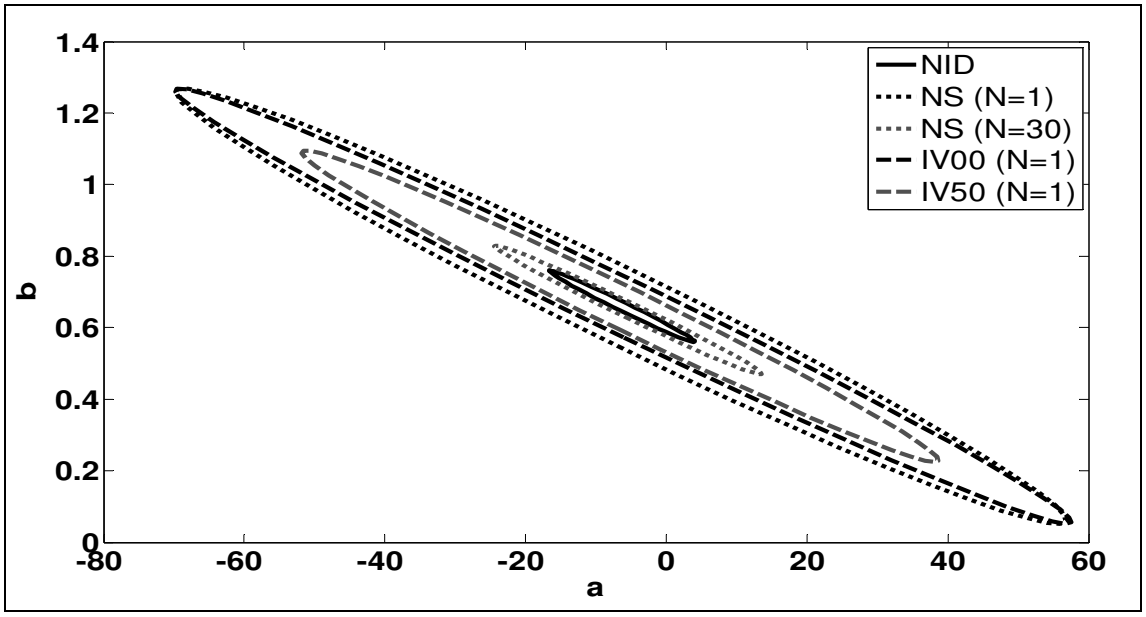


Figure 12: 90% probability ellipses for posterior density distributions of parameters *a* and *b* generated by GLUE with several likelihood functions. The extent of 90% ellipses is determined by the 90 % quantile of the χ^2 distribution

2.3.6 Effect of R^2_{Model}

Figures 5-9 consider the case where the true model had an R^2_{Model} of 0.90. Figure 13 is like Figures 7-8 except different values of R^2_{Model} are considered. Thus we explore the use of GLUE to solve a range of possible problems wherein the precision of the model changes. Figure 13 describes the effect of the value of R^2_{Model} on 95% uncertainty intervals computed with different likelihood measures. In Figure 13 one sees that if the correct NID likelihood function is adopted, the 95% uncertainty intervals for the mean flow associated with interested precipitation collapse to a point as R^2_{Model} goes to 1, as is expected from equation (7). If the model is perfect, it only takes a few points to determine the values of the coefficients exactly. However, that truth is only honored with REGR and NID.

Use of a subjective likelihood measures NS and IV results in intervals that do not collapse to a point as R^2_{Model} goes to 1. The width of the uncertainty intervals obtained with NS and IV are insensitive to the true value of R^2_{Model}. The quality of the fitted model does not seem to matter in a GLUE analysis with NS or IV, unless a solution falls below a non-behavioural threshold which is imposed on the analysis. But how can it be true that the precision of the estimated parameters does not depend on the precision of the model?

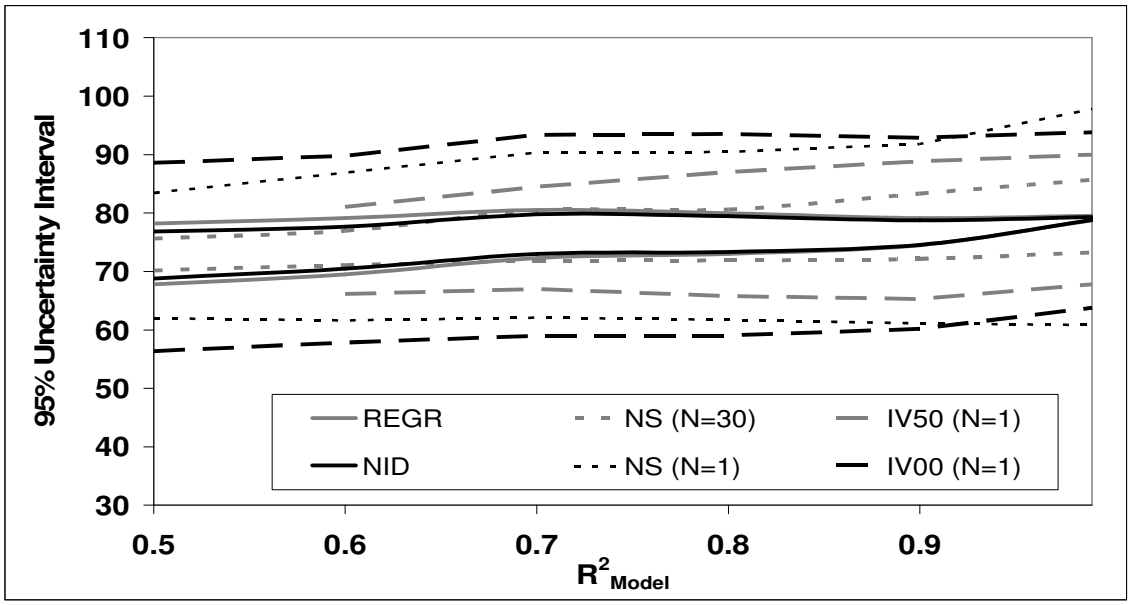


Figure 13: Effect of R^2_{Model} on 95% uncertainty intervals for the mean flow when P_o = 125.6 cm obtained with different likelihood functions

2.3.7 Prediction Intervals for Future Observations

Most applications of GLUE have used the generated parameters with their assigned probabilities to construct intervals which the investigators have asserted will contain future observations with the specified frequency. For example, Beven and

Freer, (2001 p. 24) observe that the GLUE prediction limits generally bracket the observations, suggesting that GLUE output provides an appropriate description of the range within which individual observations may occur. They state in the abstract that, "Any effects of model nonlinearity, covariation of parameter values and errors in model structure, input data or observed variable, which the simulations are compared, are handled implicitly within this procedure." How is it possible that such simple subjective likelihood measures can understand and represent all these issues? As has been observed above, this is clearly an inappropriate expectation because previous GLUE analyses have generally ignored the model error (ε_t in eqn 1) that describes the likely diffference between the observations Q_t and their mean values $\alpha + \beta P_t$. Of course, Bayes Theorem can be used to compute the predictive distribution of an observation, which was illustrated in (14). Here we illustrate how this can be done in a Monte Carlo simulation.

The most rigorous approach would be to compute the convolution of the uncertainty distribution for the parameters with the error distribution describing the probability of different errors ε_t in (1). For example, if one performed a GLUE Monte Carlo analysis with a statistically valid likelihood function, one would generally find the distribution of the predictions $a + b P_t$ due to parameter error was approximately normally distributed. Combining this with a normal distribution for the errors ε_t in our example, would if done correctly result in the distribution in (5), so we include those prediction intervals as our standard.

To show that an essentially equivalent result can be obtained by a more general and less restrictive procedure, we also provide the results for a GLUE Monte Carlo procedure. In this case, using uniform priors on the parameters and the correct likelihood function for our regression model, sets of parameters and their associated probabilities were generated. Those parameters were then used to develop estimates of

the streamflow Q_t for each of the observed rainfalls P_t (t = 1,…n, with n = 40). Then to include the distribution of possible errors ε_t, 50 different zero-mean random normal variates with variance $\hat{\sigma}_\varepsilon^2$ were added to capture the possible model errors that could be associated with the prediction of the mean obtained with each observed P. There is no magic to the choice of 50, but in practice having gone to the work of generating model parameters and simulating a system to determine the likelihood function value for those parameters, it would be computationally efficient to generate a number of possible errors.

Combining the 10,000 different possible sets of parameters that provide predictions of Q for each observed P, with 50 normal replicates, yielded 500,000 possible values for each Q. These were used to construct the 95 % prediction intervals in Figure 15. Figure 15 describes that the effect of R^2_{Model} on 95% prediction intervals for a future flow Q with P_o = 125.6 cm obtained with different likelihood functions. One can see that they agree very well with the correct statistical result obtained with equation (5). Also shown are the GLUE intervals with NS and IV, which in some instances are wider than they should be and in other instances too narrow. The result clearly shows that the correct effect of R^2_{Model} is not reproduced at by GLUE with the NS and IV likelihood measures. Again, when R^2_{Model} approaches unity, the predictions intervals should collapse to a point, which again only occurs with REGR and NID. Christensen (2004) and Montanari (2005) illustrated the same problem with GLUE intervals that ignore model errors.

Table 1 provides another way to represent the results in Figure 14. It reports the number of observations K in the three n=40 period records (for R^2_{Model} =0.80, 0.90, and 0.95 respectively) that fall outline of the 90% prediction intervals for each record. Normally for 90 % prediction intervals with n=40, one would expect E[K]=4, neglecting the fact that the prediction intervals are based on an analysis of each record

with which they are compared. Clearly REGR and NID provide very reasonable results as expected with $3 \leq K \leq 5$. NS (N=1) and IV 00 (N=1), intervals are absurdly wide with no observations falling outside the prediction interval. The NS (N=30) intervals are much too narrow for R^2_{Model}=0.80-0.90 resulting in K=18-19 observations outside the prediction intervals. IV50 (N=1) intervals go from being too narrow for $R^2_{Model} < 0.80$ to too wide for $R^2_{Model} > 0.90$.

Figure 14 and Table 1 illustrate once again that use of the correct likelihood function with the GLUE methodology will lead to prediction intervals which perform as expected, whereas use of an arbitrary likelihood measure will lead to arbitrary intervals without statistical validity, and which fail to include future observations with the target frequency.

Table 1. Numbers of observations K in the three n=40 period records that fall outside the 90% prediction intervals for each record [anticipate $K \approx 4$]

Model	R^2_{Model}		
	0.80	0.90	0.95
REGR	3	4	5
NID	3	4	4
NS (N=1)	0	0	0
NS (N=30)	19	18	2
IV00 (N=1)	0	0	0
IV50 (N=1)	9	3	1

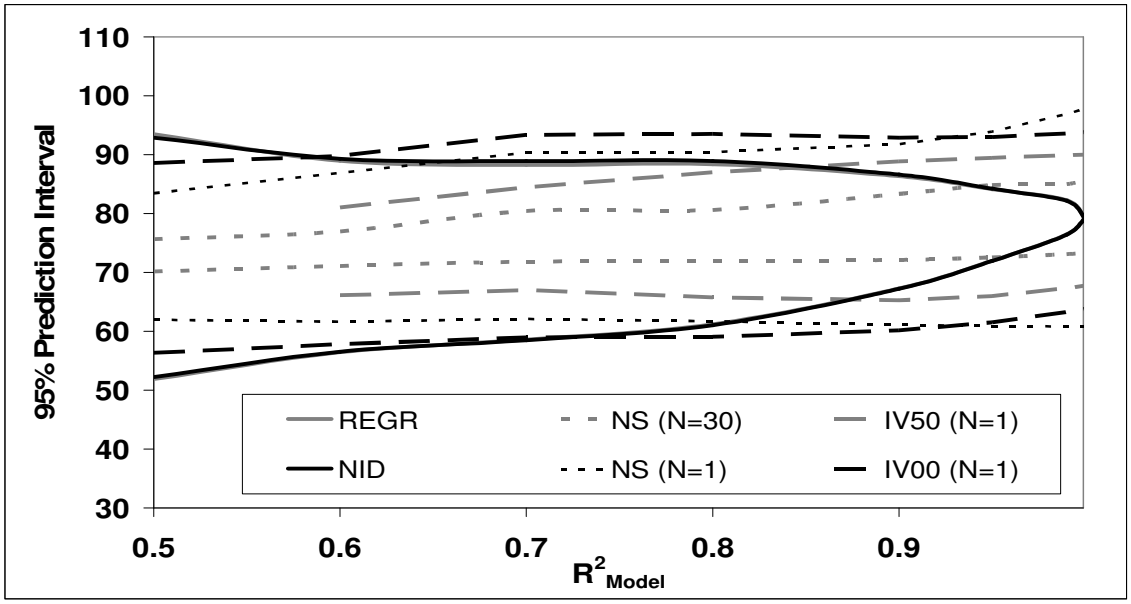


Figure 14: Effect of R^2_{Model} on 95% prediction intervals for the future flow when P_o = 125.6 cm obtained with different likelihood functions for n=40

2.3.8 Effect of Autocorrelation in a Watershed Model

One of common concerns in hydrologic modeling is that residuals exhibit temporal persistence. A variety of methods are available including (1) use of seasonal autoregressive moving average (ARMA) models to describe the serial dependence structure of residuals (Salas et al. 2006), (2) thinning the hydrologic record by only considering say, every fifth day, (3) through the use of moving-blocks (weekly/monthly volumes), because weekly or monthly flows exhibit far less serial correlation than daily flows, and (4) state-space error updating to address correlation introduced by input and some model errors (Vrugt et al., 2005; Moradkhani et al., 2005b; Bulygina and Gupta 2009). For example, a simple autoregressive model of residuals was employed by Sorooshian and Dracup (1980), Duan et al. (1988),

Romanowicz et al. (1994, 1996), and Beven and Freer (2001). Duan et al. (1988) derive an MLE approach for watershed model parameter estimation when model residuals follow a lag-one autoregressive process which can be employed with data collected at unequally spaced time intervals.

In practice, environmental simulation models are far more complex than the simple linear model with normal and independently distributed (NID) errors used in our analysis. Thus error models should be adopted that address non-normal, heteroscedastic and serially correlated residuals as well as other complexities. The previous sections of this chapter have been devoted to NID model errors in the watershed model. The model with autocorrelated errors is considered in this section.

Assuming the errors of watershed model in (1) follow a first-order autoregressive process, a Bayesian estimate of autocorrelation is derived. Zellner (1971) analyzed this problem with a disturbance term generated by a first-order autoregressive process, which is given by

$$Q_t = \alpha + \beta P_t + \varepsilon_t \tag{25a}$$

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t \qquad \text{for t=1,2,…,40} \tag{25b}$$

where ρ is the autocorrelation parameter, and u_t is an independent and identically distributed error term with zero mean and unknown variance σ_u^2.

From (25a) and (25b), the model may be rewritten as

$$Q_t = \rho Q_{t-1} + \beta(P_t - \rho P_{t-1}) + \alpha(1-\rho) + u_t \tag{26}$$

Now the new model (26) has normally distributed (NID) model errors u_t with zero mean and constant variance σ_u^2. With this NID error, the likelihood function becomes

$$f_{u|\theta}\left[u_t|\theta\right] = \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\left[-\frac{u_t^2}{2\sigma_u^2}\right] \tag{27}$$

where $u_t = Q_t - \hat{Q}_t$, and $\hat{Q}_t = \rho Q_{t-1} + b(P_t - \rho P_{t-1}) + a(1-\rho)$.

With sufficient data, the best estimator of σ_u^2 is the maximum likelihood

estimate (MLE) for σ_u^2 equal to

$$\hat{\sigma}_u^2 = \frac{1}{n}\sum_{t=1}^{n}(Q_t - \hat{Q}_t^{MLE})^2 \tag{28}$$

where $\hat{Q}_t^{MLE}$ denotes the model predictions obtained using the MLE of the model parameters, which are those model parameters which led to the minimum value of σ_u^2 for the likelihood function in (27). By combining (27) and a prior distribution, we obtain the posterior distribution for parameter θ given NID error u_t

$$f_{\theta|Q}[\theta \mid Q] = c\, f_\theta[\theta] \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\left[-\frac{(Q_t - \hat{Q}_t)^2}{2\sigma_u^2}\right]$$

$$= \kappa \exp\left[-\frac{n}{2}\frac{\sum_{t=1}^{n}(Q_t - \hat{Q}_t)^2}{\sum_{t=1}^{n}(Q_t - \hat{Q}_t^{MLE})^2}\right] = \kappa \exp\left[-\frac{n}{2}\frac{s_u^2}{\hat{\sigma}_u^2}\right] \tag{29}$$

where $s_u^2 = \sum_{t=1}^{n}(Q_t - \hat{Q}_t)^2 / n$. κ is a constant term whose value may be determined by the requirement that the integral over all the parameters of the density function be equal to one.

To illustrate the results of these computations, we generated a new sample Q_t from (25). First, we inferred n values of u_t from the original observation P_t and Q_t, t=1,2,…, n (n=40) which is generated in the previous section when ρ = 0. Then n values of Q_t with fixed P_t and u_t were generated from (25) by varying ρ from 0 to 0.9. Finally, using the new samples of precipitation P_t and streamflow Q_t, 95% uncertainty intervals for the mean flow associated with precipitation value P_0 = 125.6 cm were generated by GLUE with several likelihood measures (n=40).

Figure 15 illustrates the effect of autocorrelation parameter ρ on the 95% uncertainty intervals for the mean flow associated with precipitation value P_0 = 125.6 cm. The results indicated that the likelihood function results with first-order autocorrelated (N-AR(1)) model errors are quite sensitive to changes in ρ for large ρ.

The subjective GLUE likelihood measure results are completely insensitive to such changes. The problem with GLUE is that the likelihood measures only depend on the residual mean square errors s_ε^2 and ignore correlation among the residual errors. Thus, while each function, depending upon the shaping factor *N*, assigns different likelihoods to different parameter sets, the rank ordering of models corresponding to those sets is exactly the same with all three likelihood measures and all positive values of the shaping factor *N*. Thus, use of the resulting subjective likelihood measures cannot do anything to fundamentally reflect the high correlations in errors from period-to-period in low flow periods, and less correlation in high-flow periods, but greater dependence on input errors and less accurate measurement of flows.

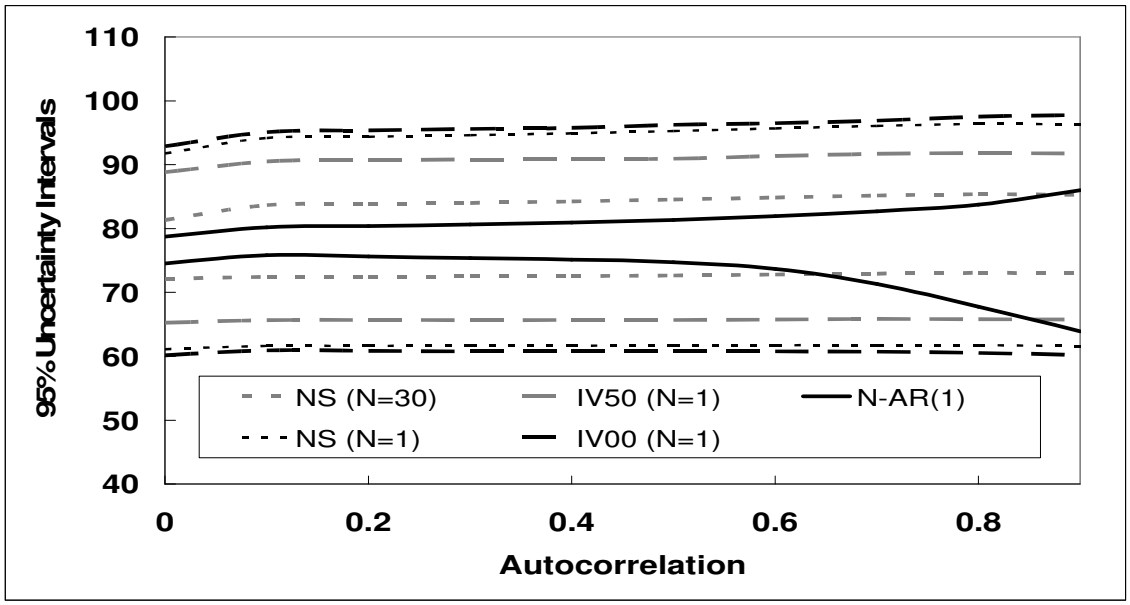


Figure 15: Effect of autocorrelation factor ρ on 95% uncertainty intervals for the mean flow associated with precipitation value P_0= 125.6 cm generated by GLUE with several likelihood measures. n=40, R^2_{Model}=0.90

2.4 Conclusion

This chapter explores the strengths and weaknesses of the GLUE methodology with commonly adopted subjective likelihood measures using a simple linear watershed model. The results of our experiments showed why the choice of a likelihood function is critical and needs to depend upon a reasonable distribution for the model errors for the statistical inference and resulting uncertainty and prediction intervals to be valid. Our findings document that in order to generate uncertainty intervals using the GLUE methodology which agree with classical and Bayesian statistical theory, the assumed likelihood function must be based on the actual statistical distribution of the errors, or at least a good approximation. When one employs the likelihood function for NID errors, the resulting uncertainty intervals become narrower as the sample size increases or the precision of the model is increased. We also showed that these relationships are not honored when traditional goodness-of-fit likelihood measures recommended by Beven and Binley (1992) (and many others) are employed with GLUE.

Beven and Binley (1992), Beven and Freer (2001), and others have suggested that the choice of the likelihood measure used in GLUE should be subjective as is the method for combining likelihood measures. These recommendations are made because to do otherwise, would require specification of a particular error model structure which Beven et al. (2000) and others are unable to justify. Their argument goes as follows: "There would appear to be no reason why subjective likelihood measures should not be precluded from use in the conditioning process in cases where the theoretical rigor of a truly objective likelihood function may be difficult to achieve for all behavioral models" (Beven et al. 2000). Unfortunately, despite this claim, because the absolutely correct likelihood function may be difficult to construct, it is not the

case that any function one subjectively selects and calls a likelihood measure will yield probabilities with any statistical validity. What is hopefully made clear in this study is that many recommended choices for a likelihood measure for use with GLUE lead to prediction intervals for model predictions entirely inconsistent with classical or Bayesian statistics which are known to correctly represent the model uncertainties, nor do the generated intervals reflect common sense or the actual uncertainty in estimated parameters or in model predictions. Blasone et al. (2008a, pp. 632) observe that "…the GLUE derived parameter distributions and uncertainty bounds are entirely subjective and have no clear statistical meaning." Montanari (2007) suggests that "GLUE should not be considered a probabilistic method, but instead should be considered a weighted sensitivity analysis. Therefore the confidence limits provided by GLUE could be better named sensitivity envelopes." Imposition of an arbitrary and sharp behavioral threshold does not solve the problem, even if the threshold is calibrated (Stedinger et al., 2008, page 43).

The use of likelihood functions to evaluate model fit is the key feature of the GLUE. However, if an arbitrary likelihood measure is adopted, then GLUE generates arbitrary results without statistical validity. The conclusion that should be drawn from this work is that if the correct likelihood function is employed to properly account for parameter uncertainty and additional extensions described here are made to account for prediction uncertainty, then the GLUE methodology should be a valuable tool for estimating model uncertainty with all the advantages that have made it so popular for generating uncertainty intervals for model simulations. If an arbitrary likelihood function is adopted that does not reasonably reflect the sampling distribution of the model errors, then GLUE generates arbitrary results without statistical validity that should not be used in scientific work.

CHAPTER 3

UNCERTAINTY AND SENSITIVITY ANALYSIS FOR MODELS WITH CALIBRATED PARAMETERS

3.1 Introduction

Traditionally the parameters of most watershed and ecological models were estimated as part of an extensive calibration process conducted by experienced models. Currently, the trend in the profession is to make much greater use of automated calibration procedures (Yapo et al., 1996; Boyle et al., 2000; Duan et al., 2003; Bekele and Nicklow, 2007). This eliminates the tedious trial-and-error exercises that modelers conducted in earlier years. However, without the need to conduct such trial-and-error analyses, a model user can lose insight into how well different parameters can be resolved, and how they interact. Uncertainty and sensitivity analyses can provide such insight for model development and calibration, and well as the appropriate confidence to attach to model predictions.

A key input of both sensitivity and uncertainty analyses is a description of the precision with which each parameter has been resolved. When parameters are specified with external data, then measures of precision should also come with those parameters. But when parameters are the result of a calibration exercise, then that calibration exercise should also provide measures of parameter precision. Typical measures of precision are the standard error or standard deviation of a parameter, or an interval thought to contain the parameter with some prescribed probability. Both of these measures of precision represent the sampling error in parameter estimates that results from randomness in the data.

Many hydrological studies have embraced the Bayesian credo and have

derived different model specification schemes that attempt to characterize existing model uncertainties and to reflect this in confidence limits on the simulated model outputs. These approaches include classical Bayesian schemes (Kuczera and Parent, 1998; Bates and Campbell, 2001; Thiemann et al., 2001; Thyer and Kuczera, 2003; Marshall et al., 2005a; Yang et al., 2007).

In this chapter, we consider the uncertainty and sensitivity analysis for watershed models when parameters are calibrated. In section 3.2, we discuss traditional sensitivity and uncertainty analyses measures. Section 3.3, 3.4, 3.5 and 3.6 discuss the appropriated definitions of parameter uncertainty when data is used to calibrate several parameters jointly. Finally, in section 3.7 and 3.8, we illustrate several approaches with the simple abc watershed model to illustrate different sensitivity analysis approaches and a posterior Bayesian evaluation of parameter uncertainty.

In fitting the abc watershed model to data, our analysis explicitly accounts for rainfall measurement errors so as to adequately represent the likelihood function for the data given the major source of errors causing lack of fit. Rainfall input errors can be very important, resulting in misrepresentation of errors in the calibration process. In general, the precipitation input is based on point observations which are sometimes combined with indirect measurements such as radar or satellite information. With just a single point-rainfall gage or even the average of several gages, there is a potentially large error due to the fact that the exact rainfall variables are not known at every point of the catchment (Kavetski et al., 2006a, 2006b; Vrugt et al., 2008).

3.2 Traditional Measures of Sensitivity

Morgan and Henrion (1990) provide a definition of traditional sensitivity analysis and uncertainty analysis:

"It is usually not immediately obvious which assumptions and uncertainties may significantly affect the conclusions. The purpose of sensitivity and uncertainty analysis is to find out. *Sensitivity analysis* is the computation of the effect of changes in input values or assumptions (including boundaries and model functional form) on the outputs. *Uncertainty analysis* is the computation of the total uncertainty induced in the output by quantified uncertainty in the inputs and models, and the attributes of the relative importance of the input uncertainty in terms of their contributions. Failure to engage in systematic sensitivity and uncertainty analysis leaves both analysts and users unable to judge the adequacy of the analysis, and the conclusions reached."

Several metrics have been employed by different investigators for sensitivity and uncertainty analysis; they include:

- Deterministic, one-at-a-time analysis of each factor holding all others constant at nominal values.
- Deterministic join analysis, changing the value of more than one factor at a time.
- Parametric analysis, moving one or a few inputs across reasonably selected ranges such as from low to high values in order to examine the shape of the response.
- Probabilistic analysis, using correlation, rank correlation, regression, or other means to examine how much of the uncertainty in conclusion is attributable to which input.

Morgan and Henrion (1990, pp. 174-176) provide several metrics that may be used to quantitative describe sensitivity. Other authors provide similar definitions (for example; Saltelli et al, 2000; Frey and Patil, 2002;' Lall et al., 2002; Loucks et al. 2005; Benaman et al., 2005).

If Y is the output or model performance index of interest, and β_i are parameters, then the near universal definition of sensitivity is

$$S_i = \frac{\partial Y}{\partial \beta_i} \tag{28}$$

which needs to be computed for each variable β_i. These derivatives are normally evaluated as a nominal or best estimate of the parameters, and can be computed with analytical equations, or with numerical finite-difference formulas. These sensitivities S_i have units, and thus can be difficult to compare directly. Thus one can employ normalized or dimensionless sensitivities, which are called the elasticity

$$e_i = \left(\frac{\beta_i}{Y} \right) \frac{\partial Y}{\partial \beta_i} \tag{29}$$

While the elasticity is dimensionless and thus compares across variables, they still fail to take an important dimension of the problem into consideration: the relative uncertainty in each parameter. A variable with a small elasticity, but a large uncertainty, may be more important than a variable with a large elasticity that is known relative precisely. Thus the first-order or Gaussian approximation of the uncertainty, also called the importance uncertainty, is defined to be

$$u_i = \sigma_i \frac{\partial Y}{\partial \beta_i} \tag{30}$$

where σ_i is the standard deviation describing the uncertainty in each parameter β_i (Morgan and Henrion, 1990). A natural definition of importance uncertainty follows from the first order approximation of the total uncertainty in Y that results from uncertainty in each β_i; to first order and assuming the errors are uncorrelated:

$$\mathrm{Var}(y) = \sum_{i=1}^{n} \left(\frac{\partial Y}{\partial \beta_i} \right)^2 \sigma_i^2 = \sum_{i=1}^{n} u_i^2 \tag{31}$$

Thus, using this first order approximation, the relative importance of variable i is

$$r_i = \frac{u_i^2}{\sum_{i=1}^{n} u_i^2} \tag{32}$$

However, when parameters are estimated by calibration of the model to data, the resulting parameter estimates are often correlated. As a result, to first order the variance of Y is given by

$$\mathrm{Var}(Y) = \sum_{i=1}^{n}\left(\frac{\partial Y}{\partial \beta_i}\right)^2 \sigma_i^2 + 2\sum_{i=1}^{\mathrm{n\text{-}1}}\sum_{j=j+1}^{\mathrm{n}}\left(\frac{\partial Y}{\partial \beta_i}\right)\left(\frac{\partial Y}{\partial \beta_j}\right)\mathrm{Cov}(\beta_i, \beta_j) \tag{33}$$

where $\mathrm{Cov}(\beta_i, \beta_j)$ is the covariance of β_i and β_j (Morgan and Henrion, 1990). These are basic definitions of sensitivity metrics and the commonly given motivation for each.

3.3 Sensitivity Analysis with Calibrated Parameters

The traditional measures of sensitivity defined in section 2 are based upon the concept that parameters are determined externally to a model, so one can independently investigate the impact of uncertainty in those parameters on the output of the model. Figure 16 represents a conceptualization of a modeling process when this is appropriate.

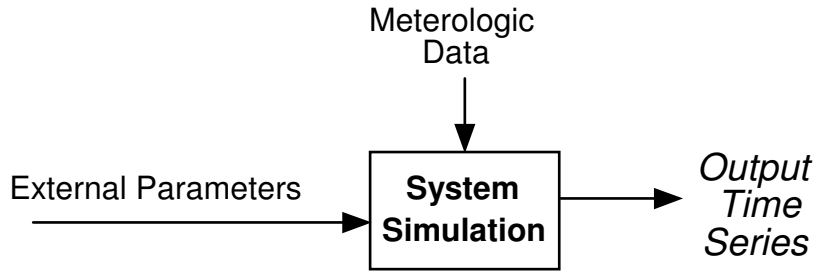


Figure 16: Simple conceptualization of modeling process

However, it is important to recognize that hydrologic models have many key parameters that are not determined by external data. Instead they are estimated in an involved calibration process. That means that these "calibrated parameters" are estimated as best one can to make the model output time series match observed flow and stage series when the model is run with the corresponding historical meteorological data. This is important for at least three reasons (Madsen, 2000; Kavetski et al. 2003):

1. Any errors in external parameters may have been compensated for by adjustment of calibrated parameters so that the model output matches observed values. For example, if depression storage is over-estimated, the runoff coefficient may be underestimated in the calibration process so that the resulting model better estimates observed runoff. Thus calibrated parameters may have errors that are correlated with any errors in external parameters.
2. Because calibrated parameters are estimated jointly, they may be significantly cross-correlated. For example, it may be difficult for the model to exactly estimate a detention storage parameter and a runoff coefficient: over some range almost the same model output can be obtained with a higher detention storage parameter and a lower runoff coefficient, and vice versa. Thus when the calibration procedure attaches to detentions storage a standard deviation σ_{DS}, (as described by the procedures in Appendix A), that value implicitly assumes that such variations in detentions storage goes with corresponding and compensating variations in the runoff coefficient σ_{RC}. Thus it is no longer appropriate in a sensitivity analysis to independently vary detention storage by $\pm 2\sigma_{DS}$ without making the corresponding adjustments in other parameters.

3. The mathematical model is not a perfect representation of reality. And the calibration data providing flow and stage data, and some groundwater elevations, is not without measurement error and does not exactly match the spatial and time averages with which it is compared. As a result, the match between the model predictions and the calibration data is not perfect making it impossible to precisely resolve the values of the calibrated parameters.

At this point before serious calibration of the model, traditional sensitivity analysis that varies one parameter at a time can have an important role revealing or documenting when, where, and which parameter have a large impact on model predictions. Francos et al. (2001), van Griensven et al. (2002), and Shoemaker (2004) discuss procedures for identifying in this framework which parameters are relatively important to determining how a model will perform. However, after a model is calibrated a different concept of sensitivity analysis is appropriate reflecting the critical information provided by the calibration data set.

3.4 Calibration and Parameter Uncertainty

When many of the parameters are calibrated with available data describing watershed behavior, an appropriate concept of the modeling process is shown in Figure 17.

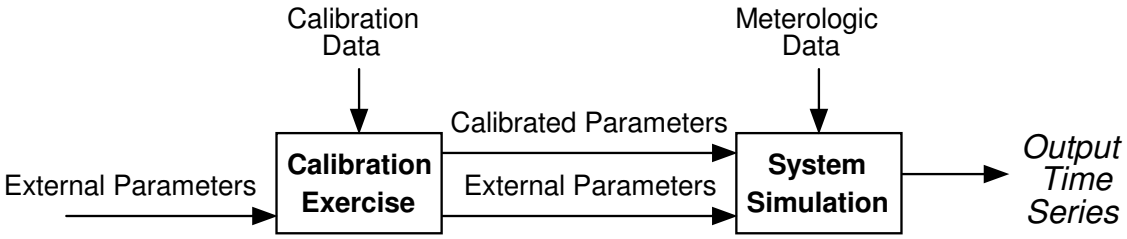


Figure 17: Appropriate conceptualization of modeling process when calibration data is employed to estimate some parameters

Given the values of external parameters, calibration data is never sufficient to accurately resolve all of the parameters of complex watershed and environmental models. Thus in considering the conceptualization of the modeling process above, one recognizes that conditional on the assumed model structure and the external parameter values, there is significant uncertainty in the calibrated parameters. And as mentioned above, the errors in different parameters of complex models are often interrelated. Thus we desire to consider the uncertainty in the calibrated parameters, and the resulting uncertainty in the simulation model output time series and statistics based upon that time series.

Several methods are available to extend classical sensitivity analysis to address the uncertainty in calibrated parameters when there is significant cross-correlation.

1) Conceptually, the simplest is to use the profile likelihood function (Coles, 2003). This corresponds to adjusting all of the parameters except one so as to maximize the goodness-of-fit objective, and then see how the goodness-of-fit objective or a forecast of system performance changes as the value of the one key parameter is varied. Thus we see if uncertainty in that key parameter (given that the other

parameters are adjusted so the model still matches the calibration data as best it can) has an impact on the performance statistic of interest.

2) Another simple approach is the first-order uncertainty analysis methods described in section 2 wherein the cross-correlation among the estimates is included. Thus given the estimated variances and covariances among the parameters, we have for the total prediction error in equation (33) reflecting the uncertainty in the parameters. The importance of the uncertainty in any parameter β_i can be described by the decrease in Var(Y) that would occur were the value of β_i specified, and the model recalibrated.

3) A classical solution to their problem is an uncertainty analysis that would yield an approximation of the sampling covariance of the estimated parameters given the calibration data sets, the functional form of the model, and the residual error variance.

4) The most complete approach to sensitivity analysis would be a full Bayesian analysis that would yield the complete posterior distributions of all the parameters. Section 3.5 below describes the more complete Bayesian approach in more detail.

3.5 Bayesian Approach

The Bayesian conceptualization of the modeling process is described in Figure 18. Here no distinction need be made between external and calibrated parameters. What were before external parameters are now parameters for which external data is thought to provide very precise information; as a result their values will not change during the calibration process: the size of canals, the dimension of cells, and the area of land in different crops would be examples. Still the values of these parameters could be adjusted during the calibration process if a much better fit could be obtained,

though it is also possible to prohibit such adjustments.

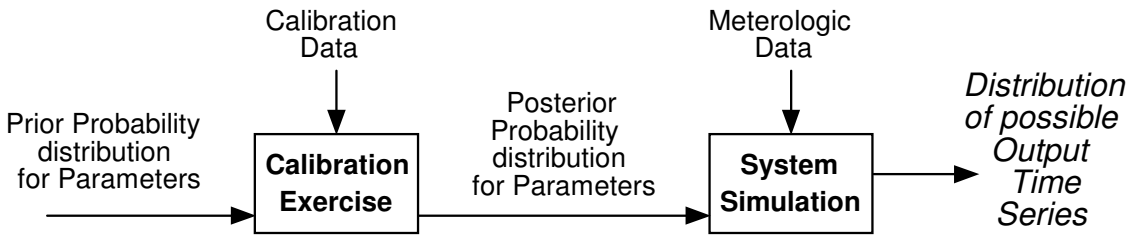


Figure 18: Bayesian conceptualization of modeling process when calibration data is employed to improve the ones understanding of the likely values of the parameters

Moreover, there are now no internal parameters whose value is only determined by calibration. Such calibrated parameters initially have some prior distribution representing conceptual reasonableness, physical appropriateness, and experience elsewhere with similar natural systems. However, if such prior information does not result in a relatively precise resolution of the values of such parameters, then the information about system performance introduced in the calibration process is likely to be the major determinant of the value of such parameters. As described in Figure 18, having derived the posterior distribution for the model parameters, the results of system simulation is a distribution of possible system performance time series. This distribution represents how the system might operate with the ranges of parameters that are reasonable given both prior information and the information added in the calibration process. In practice the distribution of parameters is often described by a large number of generated values, perhaps with different weights (probabilities), which yield an equal number of simulation time series, with the same weights (probabilities).

The complex, non-linear nature of many hydrological models means that the posterior distribution usually cannot be derived analytically, making Bayesian inference difficult to apply in most modeling situations. Markov Chain Monte Carlo (MCMC) technique has become increasingly popular as a general method that provides a solution to these difficult problems of sampling from a high dimensional posterior distribution. Standard Monte Carlo methods produce a set of independent simulated values according to some desired probability distribution. MCMC methods produce chains in which each of the simulated values is mildly dependent on the preceding values. The basic principle is that once this chain has run sufficiently long enough it will yield the desired posterior distribution of interest.

Calibration schemes have since evolved to better characterize the uncertainties in the modeling process. In recent years, many studies have focused on developing automatic calibration techniques that need not be used for the entire optimization process, or the entire model considered at the same time. Trial-and-error is still appropriate to determine appropriate ranges for parameters, perhaps considering areas within the modeled region where model performance is sensitive to specific parameters. However, if the model has multiple local optima, then this approach may not be sufficient and Monte Carlo procedures discussed in the this section may be the best course of action.

Kuczera (1988) and Kuczera and Parent (1998) discuss concerns with the accuracy of such first-order methods. Such schemes depend upon the calibration objective having a unique and well defined minimum, which may not be the case, as illustrated by Duan et al. (1992). The analysis also assumed that the parameter errors are multivariate normal whose covariance matrix can be computed from the second derivative of the likelihood function. However, Engeland et al. (2004) found that descriptions of parameter uncertainty and uncertainty in system performance indices

obtain with first-order methods based upon maximum likelihood analysis were almost identical to those obtain by an MCMC analysis on 25 basins in Sweden.

For very complicated watershed models with multiple sub-surface layers or storage volumes, there can be many alternative parameter sets that yield almost equal performance. This has long been a concern of those who have attempted to automate the calibration of complex multi-layer conceptual watershed models (Gupta et al., 1998, 2003). Beven and his co-authors have articulated these concerns very well (Beven and Binley, 1992; Beven 1993, 1996, 2002ab). With a given calibration data set exhibiting a limited range of hydrologic conditions, it may not be possible to resolve the parameters used to represent several soil layers, or different surface runoff mechanisms. This issue is illustrated by Gan and Burges (1991).

Sophisticated Bayesian methods have been developed and can address model predictions when parameters cannot be clearly resolved. The basic idea is to generate multiple sets of parameters that are all physically plausible, and then to assign to them weights or probabilities reflecting their relative likelihood. This can be seen as abandoning the vision that there is one best parameter set, which we can identify with reasonable precision. In an over-parameterized model (given a particular calibration data set), there will be alternative sets of parameters that for all intents and purposes appear equally valid (see for example, Beven and Binley, 1992).

The GLUE method (Generalized uncertainty estimate; Beven and Binley, 1992; Freer, Beven, and Peters, 2003) has seen many uses for this purpose. Given a prior probability for all of the uncertain parameters (generally uniform over specified ranges), possible sets of parameters are generated randomly. Sets whose performance fails to meet some minimum standard are viewed as not being behavioral and are discarded. The remaining sets are assigned probabilities using a goodness-of-fit objective. Unfortunately, the goodness-of-fit objectives that have been used are not

always valid likelihood functions, which cause serious problems (Batchelder, 2005; Batchelder et al., 2005; Stedinger et al., 2008). And as Kuczera and Parent (1998) explain, use of a simple and non-informative uniform prior can result in an algorithm that after billions of model evaluations may not have generated even one good solution.

More efficient and consistent procedures are based upon use of a statistical valid likelihood functions and Bayesian inference procedures. The use of efficient search procedures is illustrated by examples in Duan et al. (2003), Vrugt et al. (2003), and Tolson and Shoemaker (2005). These procedures sample around parameter sets which have previously been determined to be relatively good, thereby increasing the chances that even better solutions are found.

Good examples of sound statistical attacks on this problem using efficient procedures are provided by both Kuczera and Parent (1998), Bates and Campbell (2001), and Marshall et al. (2004). All three papers employ Markov Chain Monte Carlo (MCMC) procedures which incorporates a balanced probabilistic sampling methodology what moves from one parameter set to another based upon how well both parameter set performs (Carlin and Louis, 2000, and Gelman et al., 1995): as a result regions of the parameter space that result in good performance are explored extensively, whereas regions with poorly performing parameters are sampled less frequently correctly reflecting their implausibility.

Kuczera and Parent (1998) also discuss solving the problem of inefficient sampling by using intelligent importance sampling procedures: the likelihood function is used to identify approximately where good solutions will be found. One can then sample from that probability distribution and weight the generated parameter sets to correct for the error in their approximation. This is called importance sampling in the statistical literature (Morgan and Henrion, 1990), and is a well known method that is effective for many problems.

3.6 The Sensitivity Index

Three sets of sensitivity indices were evaluated in the examples here to illustrate the important of reflecting interdependences in calibrated parameters in a sensitivity analysis.

1) *Simple sensitivity index*: This represents a simple conceptualization of sensitivity analysis as described in Figure 16. A dimensionless one-at-a-time sensitivity index can be defined as

$$e(i) = \frac{\beta_i}{Q} \cdot \frac{\partial Q}{\partial \beta_i}\bigg|_{\beta_j \ fixed} \tag{34}$$

where $\partial\beta_i$ is relative change in parameter *i* from the nominal value and ∂Q is the corresponding relative flow change in the output of interest. In this study, the outputs of interest were considered to be the lowest streamflow, the mean streamflow, the peak flow, and the sum of square error. In the case of the sum of square error, sensitivity index (34) is generally not of interest, because at the optimized parameter equation (34) is zero. Thus the critical statistic that would be of interest is the second derivative

$$d^2_{SS}(i) = \frac{\beta_i^2}{SS} \cdot \frac{\partial^2 SS}{\partial \beta_i^2} \tag{35}$$

where ∂SS is relative change of sum of square error which describes how well a model fits a calibration data set.

2) *Sensitivity index with calibrated parameters*: This is an appropriate sensitivity index when parameters are calibrated as described in Figure 17. External and calibrated parameters are linked by how a model fits the calibration data, so one is not free to change those values independently. Thus, the sensitivity index with calibrated

parameters can be expressed as:

$$e_C(i) = \frac{\beta_i}{Q} \cdot \left. \frac{\partial Q}{\partial \beta_i} \right|_{\beta_j \ optimized} \tag{36}$$

This index describes the sensitivity of parameter β_i when other parameters β_j ($j \neq i$) are adjusted to retain the best fit possible; thus other parameters β_j will be re-optimized to consider the interrelation of different parameters of the model. One can also compute the value of d_{SS}^2 when the other parameters are simultaneously re-optimized. The sensitivity index of $d_{SS}^2(i)$, $d_{SS_c}^2(i)$, is expressed as:

$$d_{SS_c}^2(i) = \frac{\beta_i^2}{SS} \cdot \left. \frac{\partial^2 SS}{\partial \beta_i^2} \right|_{\beta_j \ optimized} \tag{37}$$

3) *Sensitivity index using a multivariate normal approximation*: If all parameters of abc watershed model follow normal distributions with each mean and variance from the posterior distribution of MCMC, the multivariate normal distribution approximation can be applied. Let β_i represent an interested or critical parameter, and β_j the other calibrated parameters. Furthermore, let μ_{β_i} and μ_{β_j} represent the mean of the two quantities, and $\Sigma_{\beta_i\beta_j}$ and $\Sigma_{\beta_j\beta_j}$ represent the covariance matrix of the indicated quantities. Then, using the properties of multivariate normal distributions, the conditional distribution of β_i given β_j is normal and has mean

$$E[\beta_i \mid \beta_j] = \mu_{\beta_i} + \Sigma_{\beta_i\beta_j}\Sigma_{\beta_j\beta_j}^{-1}(\beta_j - \mu_{\beta_j}) \tag{38}$$

Thus, the sensitivity index using a multivariate normal approximation can be expressed as:

$$e_{Bayes}(i) = \frac{\beta_i}{Q} \cdot \left. \frac{\partial Q}{\partial \beta_i} \right|_{\beta_j = E[\beta_j \mid \beta_i]} \tag{39}$$

which reflects the impact of variations in the other parameters β_j when an interested parameter β_i is simultaneously changed.

3.7 Case Study: abc Watershed Model

Two case studies involving the abc watershed model are presented. The first illustrates different sensitivity analysis approaches and a posterior Bayesian evaluation of parameter uncertainty with the abc watershed model using software and data provided by G. Kuczera (Kuczera and Parent, 1998). The second involves use of the abc watershed model with rainfall errors: in fitting the abc watershed model to data from the Pigeon River Watershed (North Carolina), our analysis explicitly accounts for rainfall measurement errors so as to adequately represent the likelihood function for the data given the major source of errors causing lack of fit.

3.7.1 The abc Watershed Model

The abc watershed model was presented by Fiering (1967) primarily for simulation of monthly streamflow records. The schematic in Figure 19 describes the structure of the model. There are three model parameters: a, b, and c. In addition, S_0 represents the initial groundwater stroage; it must be treated as a quantity internal to the model. The model was used in a previous study (Kuczera and Parent, 1998) to assess parameter uncertainty using the Metropolis algorithm.

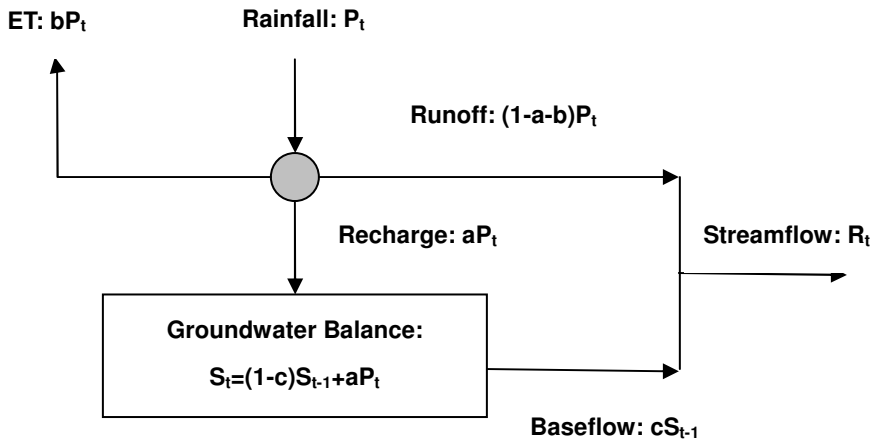


Figure 19: Schematic of the abc watershed model

The abc watershed model uses two linear equations. The model is driven by the rainfall P_t:

$$Q_t = (1-a-b)P_t + cS_{t-1} \tag{40}$$

$$S_t = (1-c)S_{t-1} + aP_t \tag{41}$$

Here Q_t is the streamflow output and S_t is the groundwater storage at given time t. The abc model partitions rainfall P_t into evapotranspiration bP_t, recharge to the groundwater store aP_t, and quickflow to the stream $(1\text{-}a\text{-}b)P_t$. The parameter c is the fraction of groundwater storage released as baseflow in each period.

3.7.2 Data

In this case study, the streamflow and rainfall data which were used in the previous work of Kuczera and Parent (1998) are employed. They generated sreamflow data using the "true" parameter values (a=0.10, b=0.75, c=0.05, S_0=500). The abc

watershed model is calibrated to the streamflow data using a software program called NLFIT provided by G. Kuczera (Kuczera and Parent, 1998).

3.7.3 Parameter Uncertainty

In this study 12,000 iterations were performed with the Metropolis algorithm. The first 2,000 samples generated were ignored to forget the initial parameter set using the NLFIT. As the models were applied using a Bayesian framework, an estimate of the parameter uncertainty can be made by examining the MCMC chains for each parameter in the model.

Figure 20 presents the posterior probability distribution obtained for the parameters a (Recharge), b (Evaportranspiration, ET), c (Baseflow), and S_0 (Initial groundwater storage, Init_GW) with the Metropolis algorithm. The posterior distribution represents the uncertainty in the model's parameters and can be propagated through a Monte Carlo simulation to assess the uncertainty in the model's output attributable to the parameters' uncertainties. From the histogram of S_0, one sees that the initial groundwater storage parameter, S_0, has a wide range.

Figure 21 describes the relationships between each pair of parameters. From the scatter plots of baseflow parameter, both recharge and ET parameters have a wider range for small baseflow parameter. We also see that there is a strongly linear relationship between the recharge and ET parameters, and a very strong non-linear relationship between baseflow and initial groundwater storage parameters.

Figure 22 and Figure 23 display posterior distributions and scatter plots using log transformation for c and S_0. The relationship between baseflow and initial aquifer storage parameters in Figure 21 is nonlinear. A linear relationship can be seen in Figure 23. In Figure 22, the posterior distributions of log c and log S_0 are more normal

distribution than those for c and S_0 in Figure 20. Next section 3.7.4 applies a sensitivity index using a multivariate normal approximation using the results in Figure 22 and Figure 23. It would enable us to provide reasonable sensitivity describing the relationships among parameters when calibrated parameters are estimated jointly.

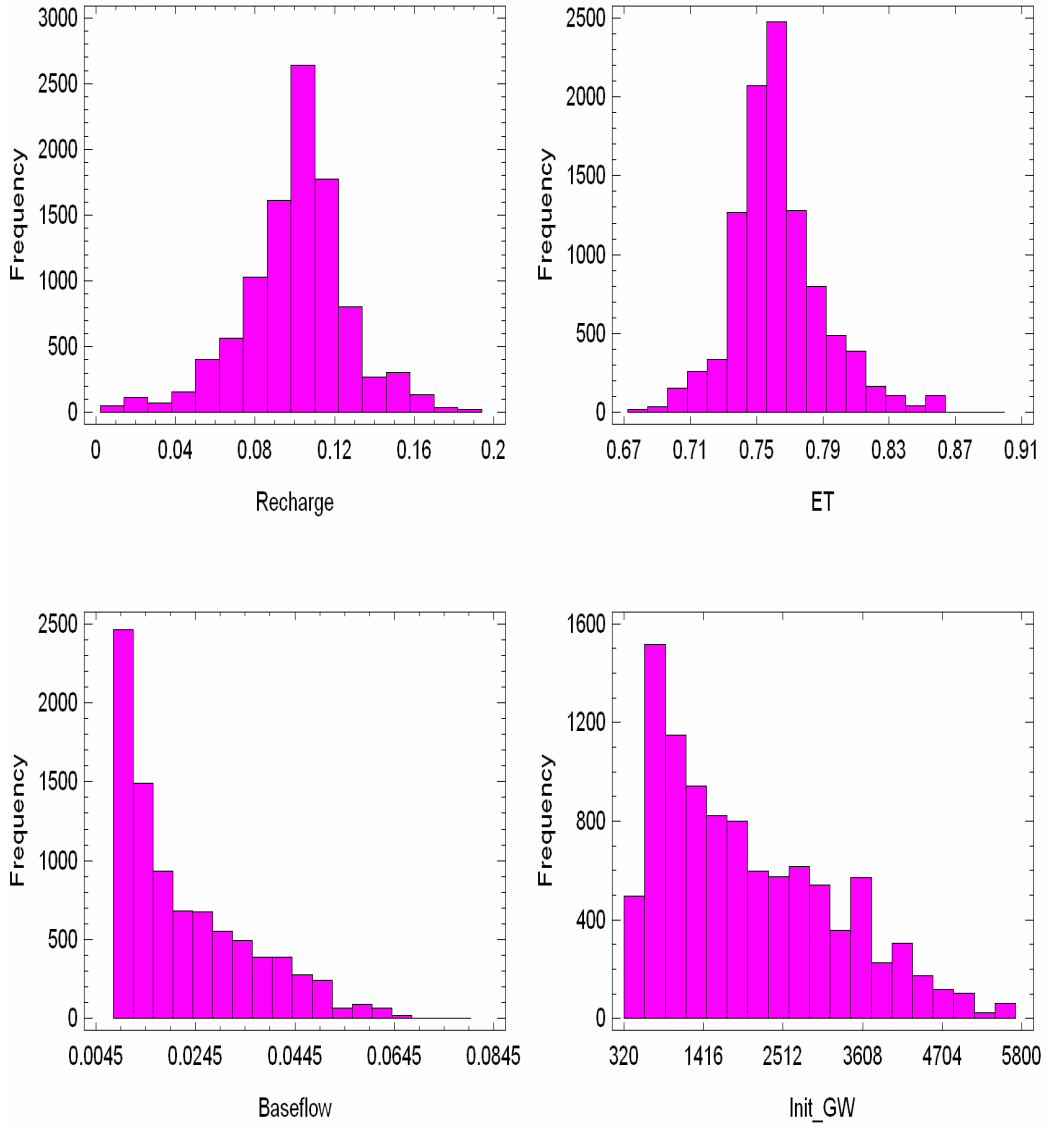


Figure 20: Posterior distribution of 4 parameters generated by the Metropolis algorithm

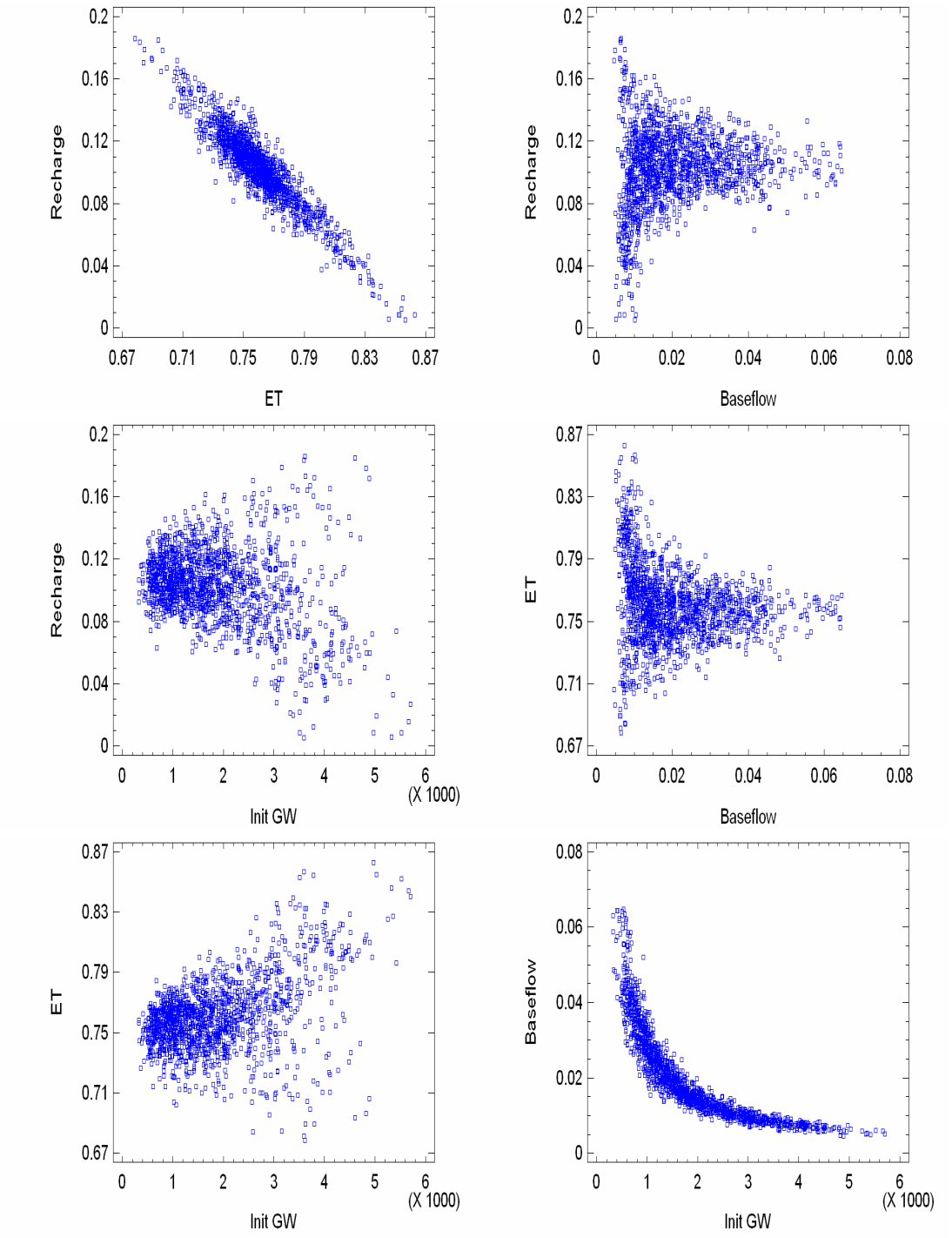


Figure 21: Scatter plots of each pair of parameter values generated by the Metropolis algorithm

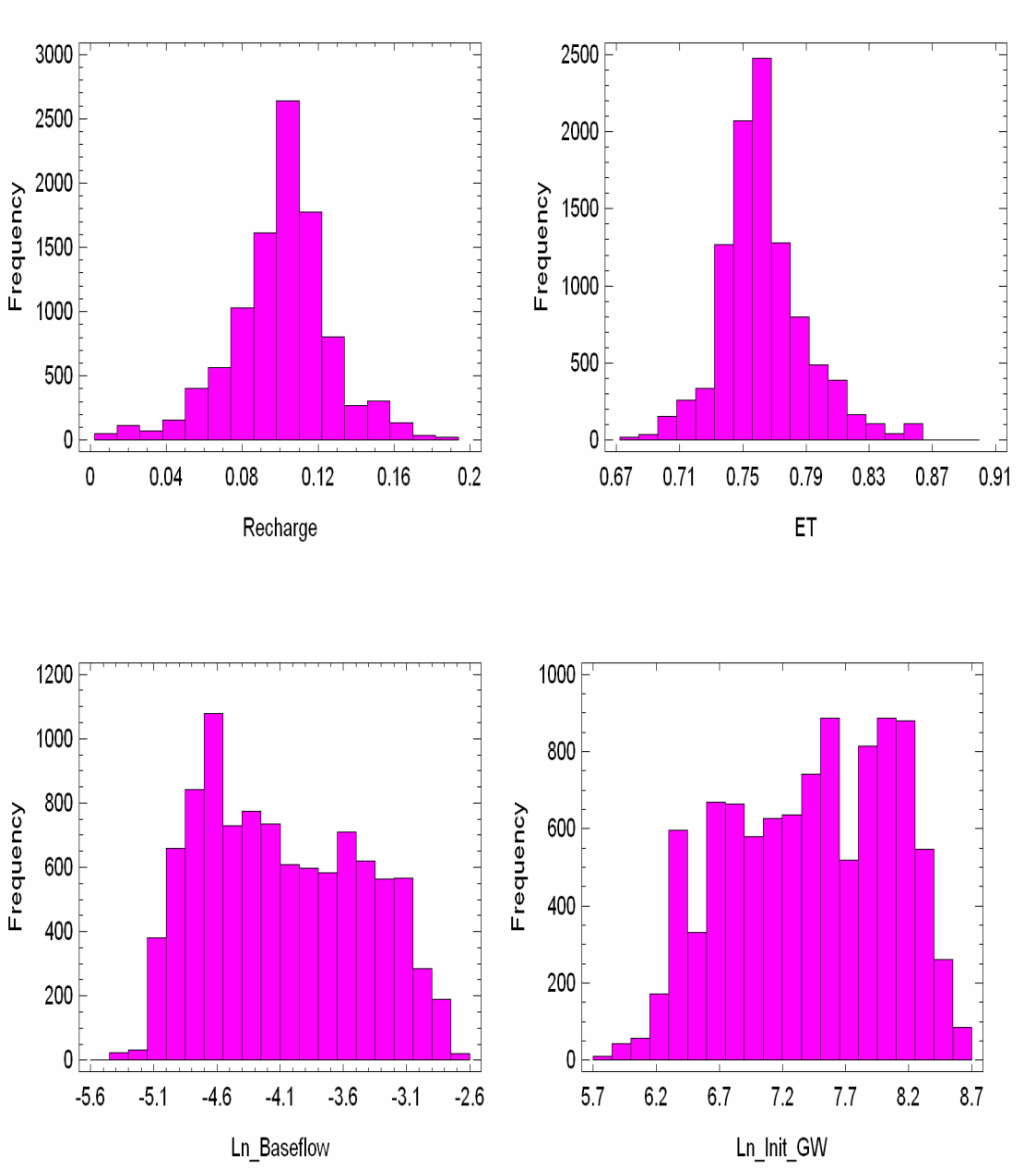


Figure 22: Posterior distribution of 4 parameters generated by the Metropolis algorithm using log transformation for c and S_0

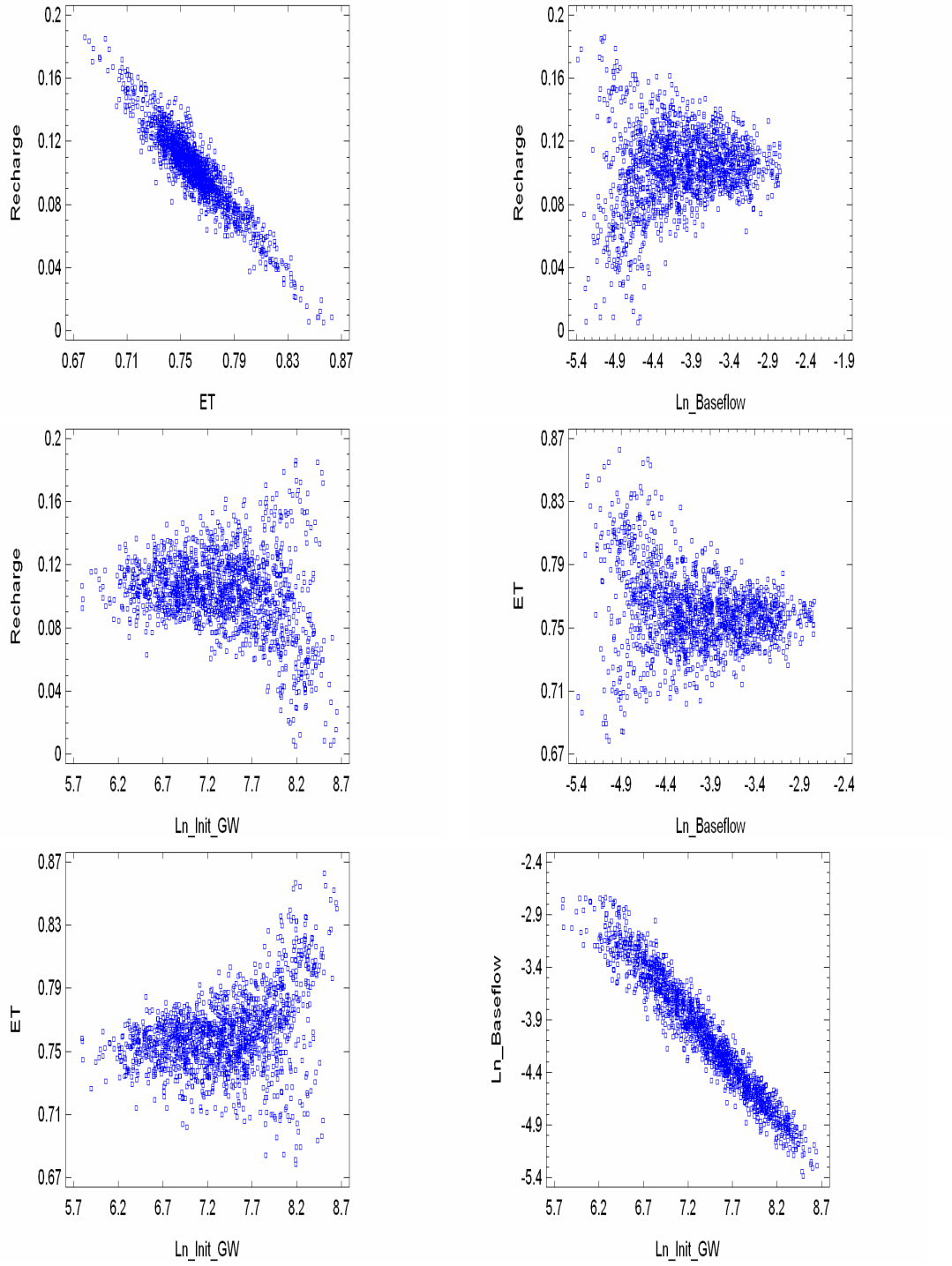


Figure 23: Scatter plots of each pair of parameter values generated by the Metropolis algorithm using log transformation for c and S_0

3.7.4 Sensitivity Analysis Results

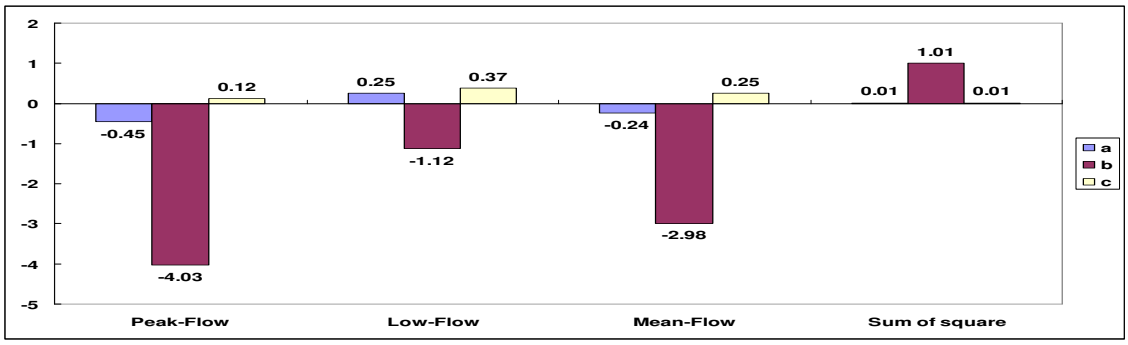
Sensitivity analysis can consider basic input data, model structure, initial conditions (S_0) or weather parameters. The sensitivity analyses reported here consider the lowest flow, mean flow, the peak flow, and the sum of squares. The indices of sensitivity were measured with +/-1 % changes in each parameter.

Figure 24 displays sensitivity indices defined in the section 3.6. The parameter *b* has the greatest sensitivity index values for peak flow, low flow, mean flow and the sum of square in Figure 24a. In Figure 24b, *b* is still the most important, but the sensitivity index values are much smaller. Figure 24a shows that *a*'s sensitivity index is negative for the peak flow and mean flow; it becomes positive due to adjustment of other calibrated parameters in Figure 24b. The traditional statistics indicate *c* is critical to peak flow, low flow, and also mean flow. However, recalibration of *a* and *b* compensates for variations in *c* so that model predictions are unaffected by errors in *c* in Figure 24. These results show that the simple sensitivity analysis does not reflect parameter interactions when parameters are calibrated.

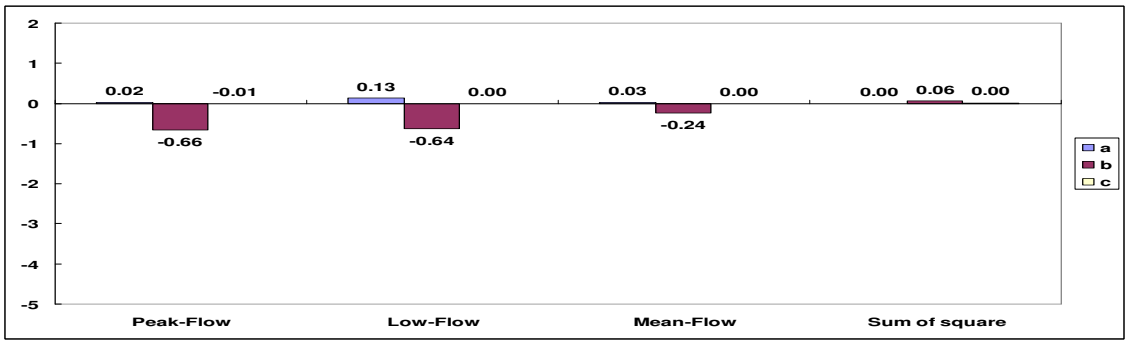
Sensitivity indices using a multivariate normal approximation is presented in Figure 24c. Figure 24c displays the sensitivity indices defined in equation (39). It was hoped that the results in Figure 24c based on a Bayesian description of uncertainty would closely resembled those in Figure 24b. In general that was not the ease.

This analysis is highly illustrative of the values of different sensitivity coefficients. However the next section uses real data for Pigeon creek in North Carolina. The Kuczera and Parent (1998) example lacked realism in that the rainfall was given without errors, and the use of the abc watershed model did not introduce any model error. Moreover, with the parameter values employed by Kuczera and Parent (1998) a hydrologically unrealistic flows sequence was obtained wherein the

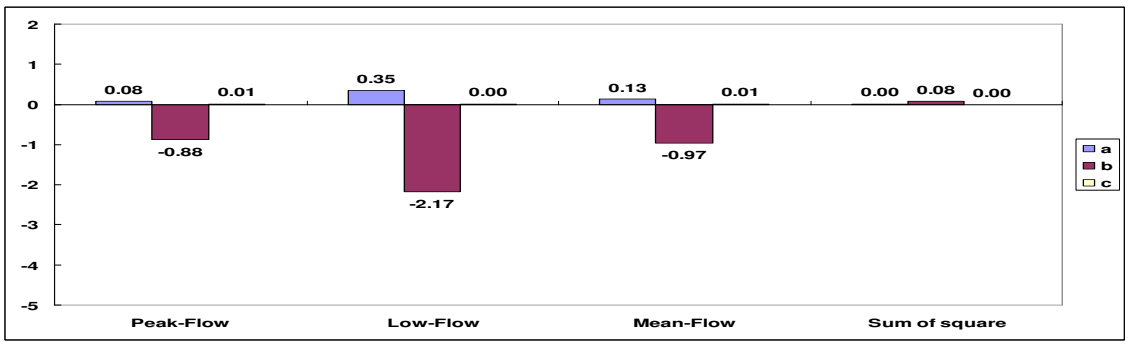
ratio of high to low flow months was at most 1:4, and generally less than 1:3 (See Figure 25). This is mostly due to the choice of parameters, but also due to the simple linear runoff model employed by the abc watershed model.



a) Simple sensitivity analysis with $e(i)$ and $d_{SS}^2(i)$



b) Sensitivity analysis with calibrated parameters with $e_C(i)$ and $d_{SS_C}^2(i)$



c) Sensitivity analysis using multivariate normal approximation with $e_{Bayes}(i)$

Figure 24: Sensitivity results for the abc watershed model parameters

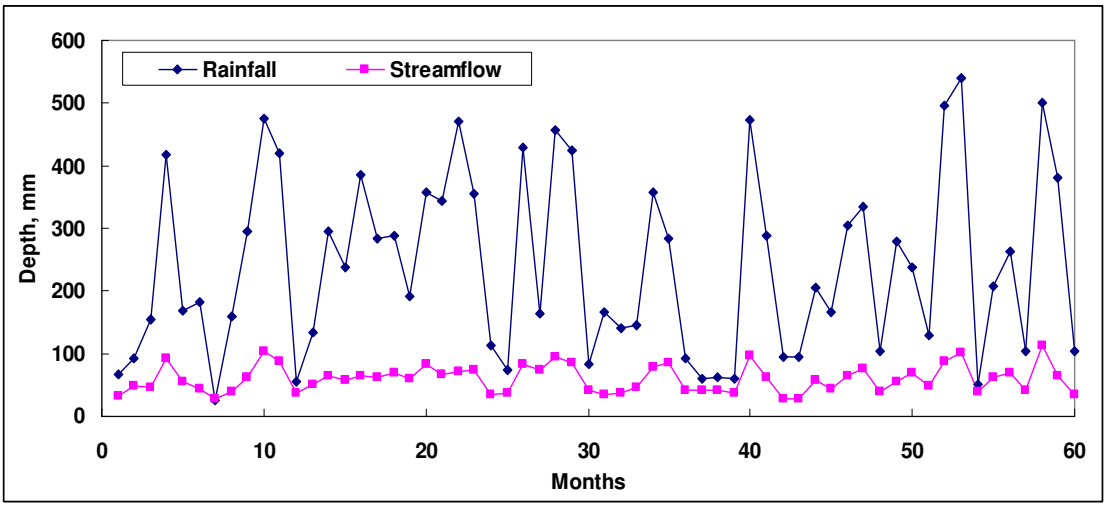


Figure 25: Time series plot of streamflow and rainfall data used in the previous work of Kuczera and Parent (1998)

3.8 The Rainfall Error and abc Watershed Model

Environmental models use relatively simple mathematical equations to conceptualize complex hydrological and environmental processes. For this reason, the model parameters often do not represent measurable entities and must therefore be estimated using measurements of the system inputs and outputs. Typically, hydrological models are calibrated to make the streamflow predicted by the model with observed precipitation match the recorded discharge. However, precipitation in particular is subject to considerable sampling and measurement uncertainty, because rainfall fields are highly variable in time and space (Chaubey et al., 1999; Bonta, 2004; Clark and Slater, 2006; Morin et al., 2006; Clark et al., 2008; Ekström and Jones, 2009). Moreover, precipitation measurement errors depend heavily on wind speed, type of precipitation (rain or snow), temperature, and gauge designs (Hanna, 1995; Chang and Harrison, 2005; Sieck et al., 2007). These spatial rainfall variability and measurement errors significantly can increase uncertainty about the estimation of

hydrological model parameters and consequently estimation (Kuczera and Williams, 1992; Arnaud et al., 2002; Clark and Vrugt, 2006). Thiemann et al. (2001), Kavetski et al. (2002, 2006ab), Kuzcera et al. (2006), Clark and Vrugt (2006), Huard and Mailhot (2006), Ajami et al. (2007), and Vrugt et al. (2008) discuss promising analyses which have the potential to capture better the combined impacts of parameter uncertainty and model error, as well as both input and output measurement errors on the overall uncertainty of environmental simulation model predictions.

Sequential data assimilation algorithms can be used to estimate model states and parameters simultaneously, due to the ability to explicitly treat the various sources of uncertainty (Reichle et al., 2002ab; Evensen, 2006; Moradkhani et al., 2005a; Vrugt et al., 2005). Moradkhani et al. (2005ab) showed the impact on parameter estimation and model predictions that result from use of sequential data assimilation techniques. Vrugt et al. (2005) presented a combined parameter and state estimation method, entitled SODA (Simultaneous Optimization and Data Assimilation), for improved treatment of input, output, parameter and model structural errors during model calibration. The basic idea of SODA is that model errors accumulate and persist in model state variables, and updating (or correcting) model state variables during the optimization process will improve the identifiability of model parameters.

Several studies have analyzed precipitation uncertainty using a precipitation multiplier model (Kavetski et al., 2002, 2006ab; Kuczera et al., 2006; Vrugt et al., 2008; Salamon and Feyen, 2009). Kavetski et al. (2002) introduced the BAyesian Total Error Analysis (BATEA) to address theses issues. This approach explicitly accounts for precipitation measurement errors by introducing a storm multiplier for each observed storm. By calibrating not only model parameters but also precipitation multipliers, which represent the systematic error in rainfall forcing data, Vrugt et al. (2008) demonstrated that for the two watersheds analyzed the predictive capabilities of

the hydrological model were significantly improved when accounting additionally for precipitation uncertainty. In this section, we adopt a similar approach to quantify precipitation uncertainty.

With just a single point-rainfall gage or even the average of several gages in or near a watershed, there will exist a potentially large error in basin-wide average precipitation. This is especially important for large storm events which provide the majority of the volume of the water entering this hydrologic system (Kavetski et al., 2006a; Vrugt et al., 2008). Errors in this storm rainfall input can be very important, resulting in misrepresentation of streamflows, and yield correlated residual errors as observed in the calibration process. In order to reflect likely rainfall errors in large storms, the abc watershed model was rewritten to include multiplicative rainfall errors as:

$$Q_t = (1-a-b)P_t \exp(\delta_i) + cS_{t-1} \tag{42}$$
$$S_t = (1-c)S_{t-1} + aP_t \exp(\delta_i)$$

where δ_i is a normally distributed error for the rainfall on day t, P_t ; δ_i is assumed to have mean 0 and standard deviation σ_δ. This expanded model was fit as a real data set which included real rainfall prediction errors, model errors, and streamflow measurement errors. This should be a better test of our model calibration and sensitivity analysis procedures than the idealized situation in section 3.7 that used a synthetic dataset created by Kuczera and Parent (1998) which did not include rainfall errors.

3.8.1 Data

A 2-year record (water year: 1 Jan. 2003-31 Dec. 2004) for the Pigeon watershed, North Carolina (USGS Cataloging Unit: 06010106), was employed consisting of daily precipitation from a rainfall USGS raingage station 03455773, and mean daily streamflow discharge rate from USGS streamflow station 03456500. Some precipitation was shifted a day to obtain a better fit to the daily flow data.

3.8.2 Calibration and Rainfall Errors

To proceed with the calibration process it was necessary to identify the days in which rainfall errors would be included in the model as described in equation (42). On days without rain (P_t=0), there is no need to include a rainfall error because $P_t\exp(\delta_i)$ will always be zero. Moreover, if P_t is very small, then $P_t\exp(\delta_i)$ is also likely to be small and it may not be worthwhile increasing the complexity of the model by adding a rainfall error for those days. Cleary, days with large recorded rainfall depths are likely to be important because those events provide the water that results in streamflow from the watershed. Thus rainfall errors are introduced into the model corresponding to critical days with observed rainfall depths exceeding a threshold of either 30 mm/day or a lower threshold of 22 mm/day. In the two-year daily record, 26 and 36 critical days are identified respectively for the two cases.

Estimates of a, b, c for the abc watershed model and the streamflow prediction errors, $\varepsilon_t = \ln Q_t - \ln \hat{Q}_t$, and also simultaneous estimates of the rainfall errors $\{\delta_i\}$ are developed. The objective function is to maximize the likelihood function; based on the assumption that the flow and rainfall measurement errors are lognormally distributed. The negative of the log-likelihood function is:

$$\mathrm{L} = \sum_{t=1}^{n} \frac{(\ln Q_t - \ln \hat{Q}_t)^2}{2\sigma_e^2} + \frac{n}{2} \ln 2\pi\sigma_e^2 + \sum_{i=1}^{I} \frac{\delta_i^2}{2\sigma_\delta^2} + \frac{I}{2} \ln 2\pi\sigma_\delta^2 \tag{43}$$

where n is the number of days of data available for calibration, I is the number of days modeled as having a rainfall error, σ_δ^2 is a rainfall error variance, and σ_e^2 is a model error variance.

Three calibration runs using different numbers of rainfall errors (0, 26, 36 rainfall errors) were undertaken for evaluating an effectiveness of accounting rainfall errors in the calibration process. The values of the rainfall error parameters $\{\delta_i\}$ were estimated simultaneously with the abc model parameters (a, b, c) by minimizing the mismatch between streamflow observations and simulated streamflows, and sum of squared δ_i rainfall errors. This minimum was obtained using a Levenberg-Marquardt algorithm with analytical derivation (Press et al., 1986, pp. 521-526).

Note that on the margin, days with positive rainfall errors are more likely to be modeled as having an error than days with negative errors δ_i. Whereas in a traditional analysis one would estimate model parameters and the associated streamflow prediction errors, our analysis also estimates likely rainfall measurement errors. Our hypothesis is that the large and important errors are in the measured rainfall as a representation of the true average precipitation over the watershed, rather than measurements of streamflow.

The Mean Square Error (MSE) statistic and the correlation of simulated and observed streamflows are used as a measure of model accuracy in the calibration results. MSE is estimated as:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (\ln Q_t - \ln \hat{Q}_t)^2 \tag{44}$$

where Q_t is the observed streamflows, $\hat{Q}_t$ is the simulated streamflows, and n is the number of days of data available for calibration.

Table 2 presents summary statistics of three rainfall error model calibration

studies. The results illustrate the best calibration performance (MSE) is obtained in the third calibration results, when 36 rainfall errors are employed. With 36 precipitation errors, the correlation between the observed and predicted streamflows is 0.777. Without the precipitation errors, the best calibration yields a correlation between observed and predicted flows of only 0.712.

With 36 precipitation errors, the autocorrelation of the prediction errors was 0.548, whereas without precipitation errors, the autocorrelation of the prediction errors as 0.783 which is considerately large. Thus introduction of rainfalls did decrease the serial correlation of the streamflow errors, as was hoped. The calibrated value of parameter a is 0.602 in the calibration run without rainfall errors, whereas in calibration run with 36 rainfall errors it is just 0.360. These results show that rainfall uncertainty affects the estimated values of the calibrated parameters and yields correlated residual errors as observed in the calibration process. There is still correlation in the model residuals, most likely due to structural model errors that would result in correlation in output errors even if rainfall measurement errors are eliminated.

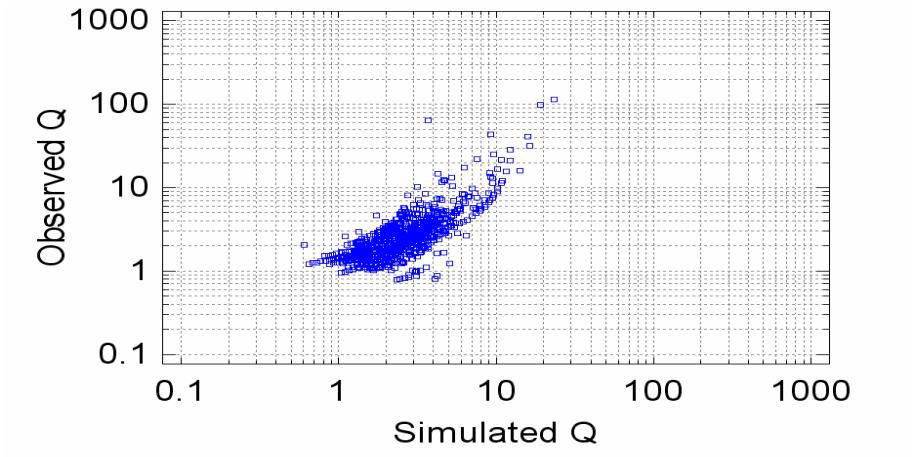
Figure 26 shows comparisons of streamflow observations and simulated streamflows. Here one observes that the rainfall error calibration approach produces flow predicts that accurately match the observed streamflow data using 4 estimated parameters, and 36 storm errors for the days with the largest precipitation totals. The correlation between the observed and predicted streamflows is 0.777. Table 2 and Figure 26 illustrate that to correctly understand the data and to reliable estimate model parameters fitting method needs to address rainfall errors, and accounting for rainfall errors estimation gives better calibration results.

Figure 27 shows the scatter plots of the rainfall error parameters, δ_i versus associated observed precipitations. The estimated parameter δ_i exhibits no significant

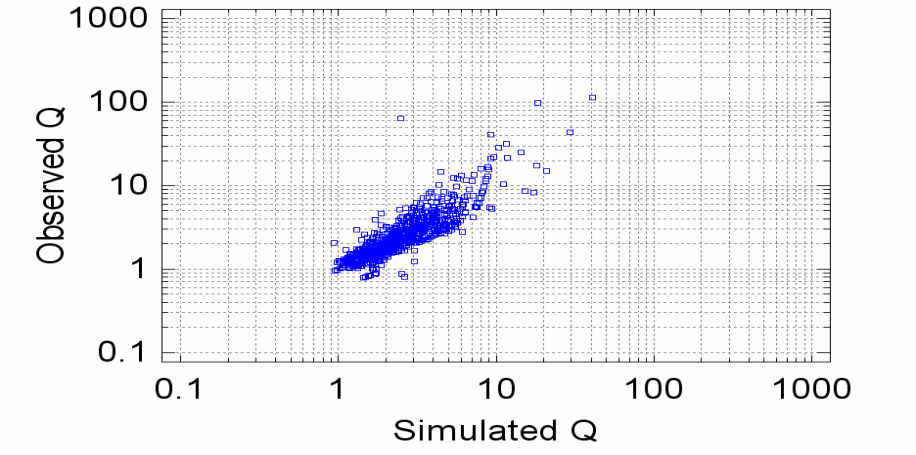
relationship with observed precipitations; however, they appear to be biased upward in both cases. In practice, the magnitude of errors depends heavily on wind speed, sitting characteristics, type of precipitation (rain or snow), and temperature. Rain gage measurement is difficult in a variety of settings, including mountain ridges, forecasts, and water bodies. The rainfall measurements can be biased by theses sources of errors. Even if our rainfall error analysis provides better calibration results, our assumption that the parameter δ_i has mean 0 and standard deviation σ_δ can be relaxed to allow a non-zero mean for δ_i which would result in a correction for bias in observed rainfall rates.

Table 2. Summary of calibration statistics

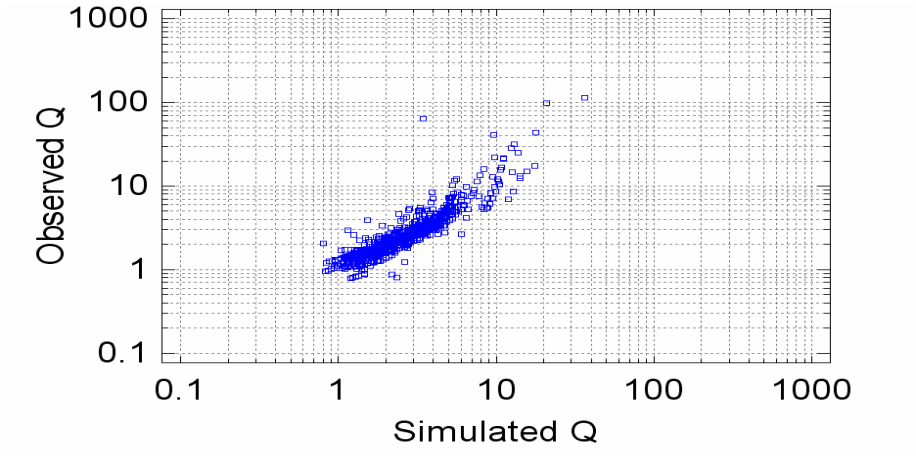
Model	MSE	Correlation (Q_t, $\hat{Q}_t$)	Auto-correlation of prediction errors	Variance of rainfall errors	Calibrated parameter values		
					a	*b*	*c*
abc model (0 error)	0.182	0.712	0.783		0.602	0.298	0.099
abc model (26 errors)	0.109	0.758	0.656	0.730	0.394	0.349	0.095
abc model (36 errors)	0.077	0.777	0.548	1.001	0.360	0.578	0.063



a) abc watershed model with 0 rainfall error

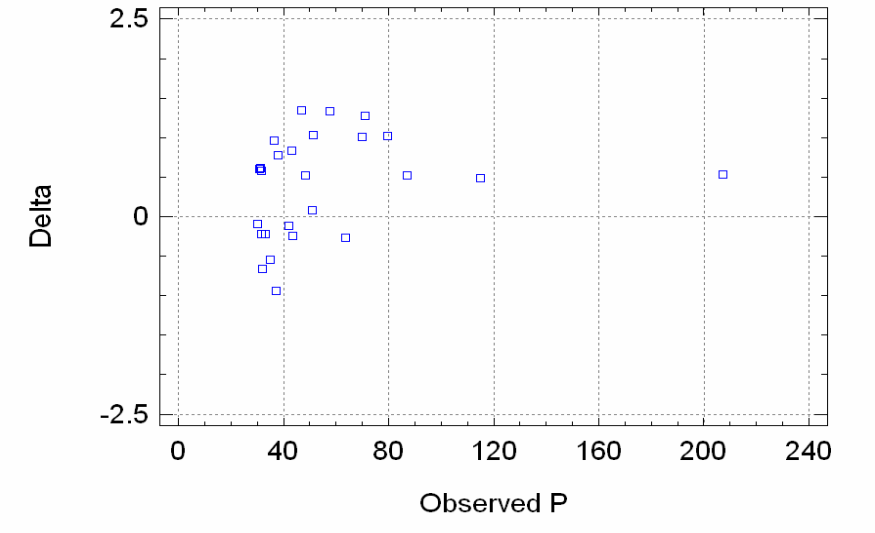


b) abc watershed model with 26 rainfall errors

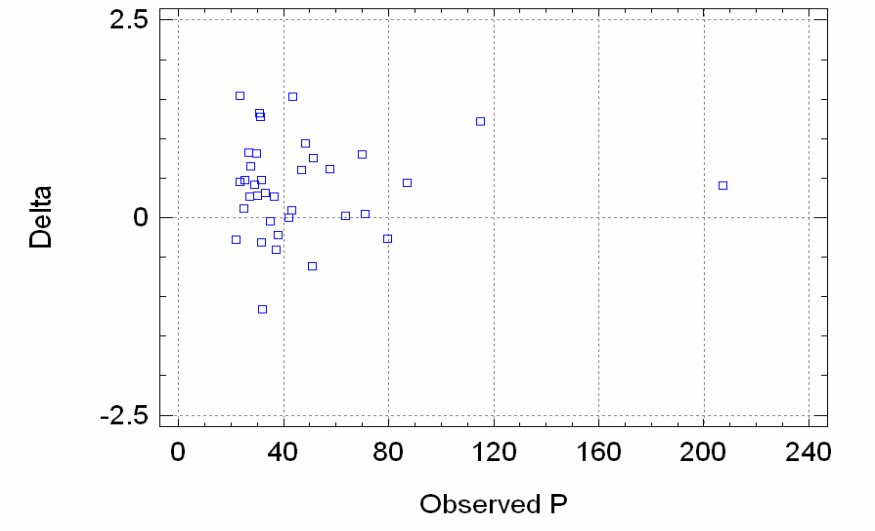


c) abc watershed model with 36 rainfall errors

Figure 26: Comparisons between observed and simulated streamflows



a) abc watershed model with 26 rainfall errors



b) abc watershed model with 36 rainfall errors

Figure 27: Scatter plots of observe precipitations versus the rainfall error parameters, δ_i

3.8.3 Parameter Uncertainty through MCMC

In this section the Bayesian posterior distribution of the parameters of abc watershed model with 26 rainfall errors was represented using 10,000 iterations of a Monte Carlo Markov Chain Gibbs sampler algorithm (Geman and Geman, 1984). The first 1,000 samples generated were removed to forget the initial parameter set. Figure 28 describes the relationships between two pairs of parameters, *a* and *b*, and *b* and *c*. From the scatter plot of *b* and *c*, one sees that the evapotranspiration parameter, *b*, has a wider range for small baseflow parameter *c*. We also see that there is a strong linear relationship between the recharge and evapotranspiration parameters, *a* and *b*. These scatter plots show the relationships among parameters when calibrated parameters are estimated jointly. Table 3 summarizes posterior correlations among the three estimated parameters.

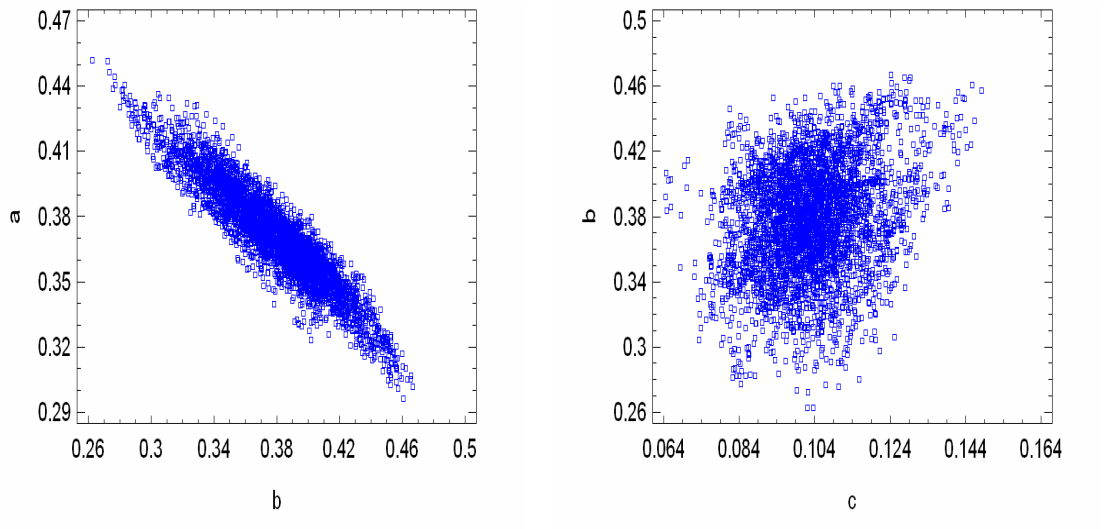


Figure 28: Scatter plots of parameter pairs generated by the Bayesian MCMC analysis for the abc watershed model with 26 rainfall errors

Table 3. Correlations between the estimated parameters, MCMC posterior moments, MLE parameter estimates (the abc watershed model with 26 rainfall errors)

Parameter	Correlation			Mean	St. Dev.	MLE
	a	*b*	*c*			
a	1			0.371	0.023	0.394
b	-0.930	1		0.379	0.034	0.349
c	-0.435	0.232	1	0.103	0.012	0.095

3.8.4 Sensitivity Analysis Results

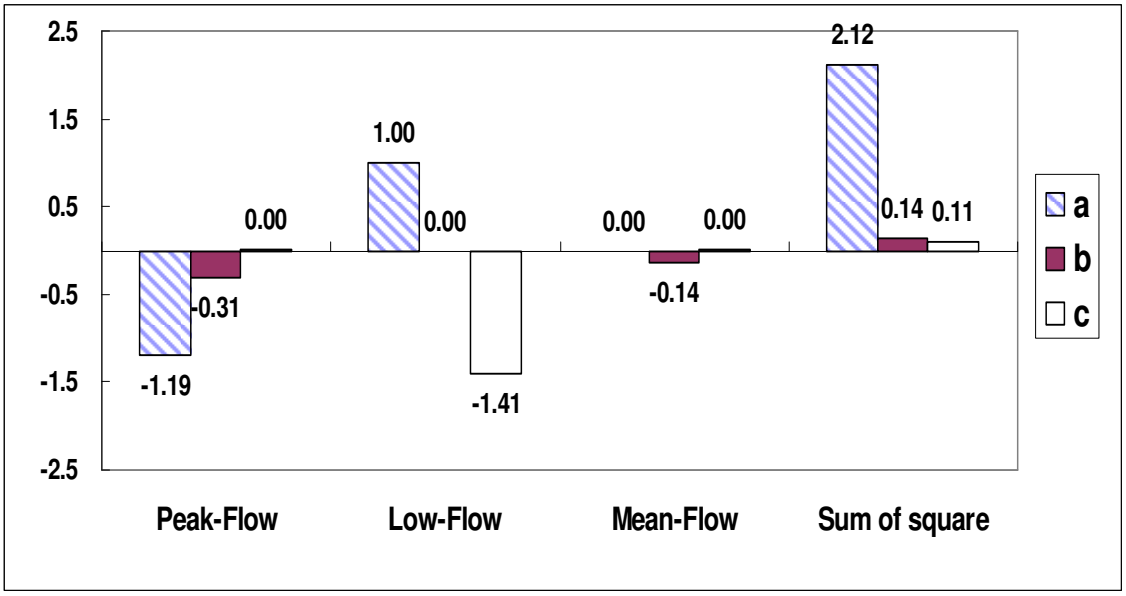
Sensitivity analysis can consider basic input data, model structure, initial conditions (S_0) or weather parameters. Here our focus is on calibrated model parameters, specifically *a*, *b* and *c*.

Figure 29 provides the results of the two sensitivity analyses for the 3 parameters of the abc watershed model with 26 rainfall errors. The sensitivity analyses consider the lowest flow, mean flow, the peak flow, and the sum of squares. The lowest and peak flows were considered as the averages of the lowest and highest in each of the two years respectively. The indices of sensitivity were measured with +/- 1 % changes in each parameter (*a*, *b*, *c*) where the rainfall error parameters δ_i were fixed at the estimated values. Thus the analysis can be viewed as employing the corrected or adjusted rainfall depths, which were fixed for the sensitivity analsyis.

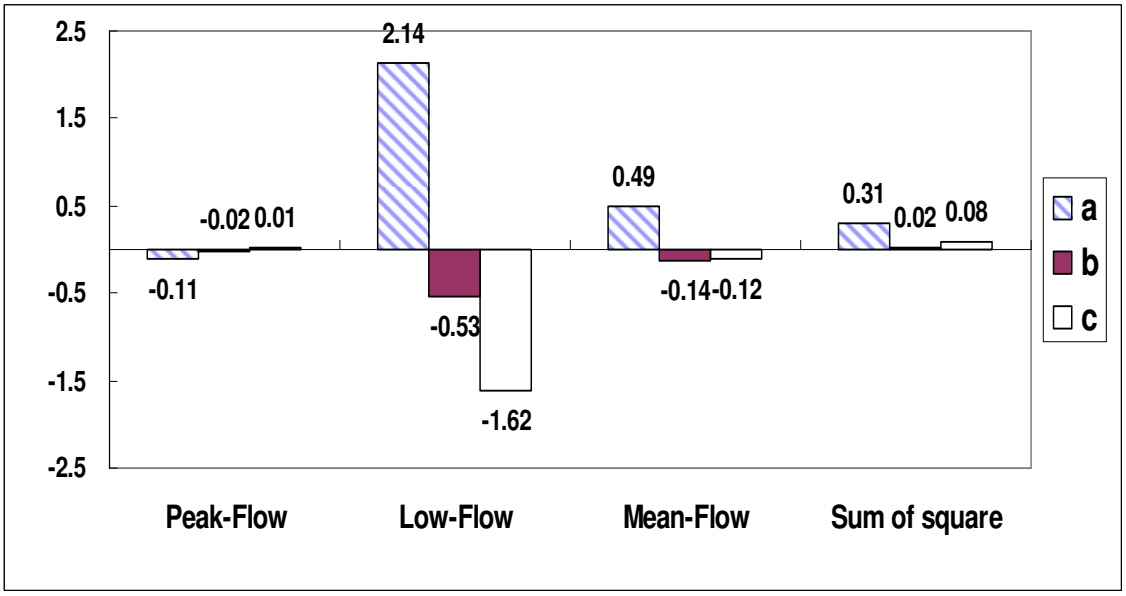
Both sensitivity analysis statistics indicate that *a* (recharge from rainfall) is

potentially the most important parameter. Both statistics indicate *a* is critical to low flows, though the new statistic attributes to *a* twice the impact; however, they disagree about *a*'s impact on high flow. The traditional statistics indicate *a* is important for floods, as a simplistic analysis would indicate. But recalibration of *b* compensates for variations in *a* so that model predictions of floods are relatively unaffected by errors in *a*. However, such changes in *a* have a critical impact on low flows and also the mean flow. The traditional sensitivity analysis indices miss entirely the link between errors in *a* and the mean flow that result from the calibration process what the impact of recalibration is included, instead of mean flow being unaffected by *a*. *a* becomes clearly the most important parameter.

Parameter *c* (base flow from storage) is the next most important, and both statistics indicate that it has a critical impact on low flows. However, in Figure 28a parameter *c* is more important than parameter *a* for low flows whereas in Figure 28b parameter *c* is less important parameter *a*. Overall, as shown by the sensitivity analysis indices for the sum of squares, *a* is the most important parameter, *c* is next most critical, and *b* (evapotranspiration) has the smallest impact on model performance. The sensitivity analysis index for the sum of squares without recalibration ranks *b* before *c*, but this makes little sense. These results show that the simple sensitivity analysis that does not reflect parameter interactions when parameters are calibrated fails to capture several critical relationships and the relative importance of different parameters.



a) Simple sensitivity analysis with $e(i)$



b) Sensitivity analysis with calibrated parameters with $e_C(i)$

Figure 29: Sensitivity results with two different sensitivity indices for the abc watershed model with 26 rainfall errors

3.9 Conclusion

This study analyzed the importance of different parameters in a watershed model using a one-at-a-time sensitivity index and a sensitivity index considering parameter interactions. The results of the sensitivity analysis showed that one-at-a-time sensitivity analysis can misrepresent critical issues when parameters are jointly calibrated. One-at-a-time sensitivity analysis does not reflect the importance of the calibration process. Our new sensitivity analysis indices capture the critical issues reflecting the interaction among the parameters in the calibration process. The second case studies also considered the impact of rainfall measurement errors.

If the least-squares analysis procedure is unable to provide a good description of the uncertainty in calibrated parameters due to multiple optima, or the failure of the likelihood function to be well approximated by a quadratic function, then the task will require more care and effort. To improve the reliability of the model and its ability to be understood, it is important to be more explicit about the precision and uncertainty throughout the model calibration and verification process. This includes recognizing the precision of the data, which are generally point measurements used to represent spatial and time averages. Clearly the daily precipitation from one gauge employed in the second case study provided an imperfect description of basin precipitation, and precipitation errors resulted in correlation among flow errors and degraded the goodness of fit. As part of the calibration process, quantitative measures of the precision of estimated parameters should be developed. Similarly, sensitivity and uncertainty analyses can be used to quantify the impact of parameter uncertainty on the precision of simulated measures of system performance.

CHAPTER 4

CONCLUSION

This thesis has two major chapters which addressed different issues. The first part of the thesis, Chapter 2, provides an appraisal of the Generalized Likelihood Uncertainty Estimation (GLUE) methodology using a simple linear watershed model. The second part of the thesis, Chapter 3, reviews uncertainty and sensitivity analysis for watershed models with calibrated parameters and rainfall errors.

Chapter 2 explores the strengths and weaknesses of the GLUE methodology with commonly adopted subjective likelihood measures using a simple linear watershed model. The results show that the choice of a likelihood function is critical. A likelihood function needs to provide a reasonable distribution for the model errors for the statistical inference and resulting uncertainty and prediction intervals to be valid.

The use of likelihood functions to evaluate model fit is the key feature of the GLUE. However, if an arbitrary likelihood measure is adopted, then GLUE generates arbitrary results without statistical validity. If the correct likelihood function is employed to properly account for parameter uncertainty, and additional extensions described there are made to account for prediction uncertainty, then the GLUE methodology should be a valuable tool for estimating model uncertainty with all the advantages that have made it so popular for generating uncertainty intervals for model simulations. If an arbitrary likelihood function is adopted that does not reasonably reflect the sampling distribution of the model errors, then GLUE generates arbitrary results without statistical validity that should not be used in scientific work.

Chapter 3 focuses on sensitivity analysis for watershed models when model parameters are jointly estimated. When calibrated parameters are estimated jointly,

they may be significantly cross-correlated. After a model is calibrated a different concept of sensitivity analysis is appropriate reflecting the critical information provided by the calibration data set. Our case study analyzes the importance of different parameters in a watershed model using a one-at-a-time sensitivity index and a sensitivity index considering parameter interactions. The results of the sensitivity analysis show that one-at-a-time sensitivity analysis does not reflect parameter interaction and as a result, fails to represent sensitivity. Our new sensitivity analysis indices capture these critical issues reflecting the interaction among the parameters in the calibration process.

As part of the calibration process, the precision of estimated parameters and data should be developed. Most hydrological simulation models suffer from the problem of having input measurements (i.e. rainfall and potential evapotranspiration) whose accuracy is significantly lower than the output (streamflow) measurements used for model calibration. In fitting the abc watershed model to data from the Pigeon River Watershed, North Carolina (USA), our analysis explicitly accounts for rainfall measurement errors so as to adequately represent the likelihood function for the data given the major source of errors causing lack of fit. Clearly, the daily precipitation from one gauge employed provides an imperfect description of basin precipitation, and precipitation errors results in correlation among flow errors and degraded the goodness of fit. The calibration process will be more efficient and effective if we correctly recognize the errors in input measurements.

APPENDIX A. PARAMETER CALIBRATION AND NONLINEAR LEAST SQUARES

If a calibration objective is selected for which the errors in different periods and at different sites are essentially independent over time, than the nonlinear least squares methods are very attractive for automatic calibration and specification of the precision of the parameters. While daily estimates of stage and flow are likely to have errors that are clearly correlated, the errors for weekly or monthly averages are likely to be fairly independent.

Similarly, if an urban site clearly has stage values that deviate from the average from a cell, then a cell bias parameter should be added to the model and estimated as part of the calibration procedure. Addition of such parameters where physically appropriate will appropriately reduce the residual mean square error of the model, thus yielding a more realistic description of the precision of estimated parameters.

Recall in ordinary least squares regression, one considers the model

$$Y = X\beta + E \tag{45}$$

where Y is a vector of values (monthly stages or flows at different sites and different times) that are to be predicted, β is the vector of parameters to be estimated, X is called the design matrix and contains as columns the values of covariates used to explain Y, and E is a vector of residual errors. The optimal estimator of the parameter vector β is

$$b = \left(X^T X\right)^{-1} X^T Y \tag{46}$$

The covariance matrix of the sample estimator b describing the precision with which it can be estimated with the available data (Y, X) is

$$Var[b] = \Sigma[b] = \sigma^2 (X^T X)^{-1} \tag{47}$$

where σ^2 is the variance of the components of the errors in *E*. The correlation between two elements *i* and *j* of *b* is

$$Corr[b_i, b_j] = \frac{\Sigma[b]_{ij}}{\sqrt{\Sigma[b]_{ii}\Sigma[b]_{jj}}} \tag{48}$$

In nonlinear regression problem where a function $f_i(b)$ provides the model estimates of values y_i, one wishes to estimate *b* by minimizing

$$J = \sum_i [y_i - f_i(b)]^2 \tag{49}$$

Here the $\{y_i\}$ would be groundwater levels, stage and/or flow values at different locations for different months, and b represents the estimate of all of the parameters which are to be estimated by calibration.

A common method for searching for the optimal *b* is to linearize the function $f_i(b)$ at the current estimate of *b* to obtain a sensitivity matrix S_f that is substituted for the *X* matrix in the equation above. This yields an iterative solution procedure for the optimal estimator of *b* (Draper and Smith, 1998; Weisstein, 2005). The basic equation

$$y_i = f_i(\beta) + e_i \tag{50}$$

is replaced with

$$y_i \cong f_i(b_t) + S_f(b_{t+1} - b_t) + e_i \tag{51}$$

yielding the revised estimate of *b*:

$$b_{t+1} = b_t + (S_f^T S_f)^{-1} S_f^T (Y - f(b_t)). \tag{52}$$

wherein σ^2 is the estimated variance of the independent errors e_i. Dennis and Schabel (1983) discuss the relative merits of this approach to solving this nonlinear least squares problem (also called the Gauss-Newton method) versus more sophisticated approaches.

An estimate of the precision of the estimator *b* is then given by

$$Var[b] \sim \sigma^2 (S_f^T S_f)^{-1} \tag{53}$$

This is a reasonable measure of the precision that is required, wherein S_f describes the partial derivates of each the observed stage or flow at each site at each time point considered (likely a week or month), with respect to each parameter estimated in this step.

REFERENCE

Ajami, N., Q. Duan, and S. Sorooshian, An integrated multi-model ensemble prediction approach to account for total uncertainty, *Water Resour. Res.*, 43, W01403, doi:10.1029/2005WR004745, 2007.

Arnaud, P., C. Bouvier, L. Cisneros, and R., Influence of rainfall spatial variability on flood prediction, *Journal of Hydrology*, 260, 216-230, 2002.

Batchelder, R., R.M. Vogel and J. R. Stedinger, *A Correction to GLUE: Generalized Likelihood Uncertainty Estimation*, AGU Spring Meeting, New Orleans, 2005.

Batchelder, Rebecca, *A Case for the use of a Bayesian Likelihood Function in the Application of GLUE (Generalized Likelihood Uncertainty Estimation),* Master of Science Thesis, Tufts University, Medford, Mass, 2005.

Bates, B. C. and E. Campbell, A Markov chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall-runoff modeling, *Water Resources Research*, 37, 937-947, 2001.

Bekele E. G., and J. W. Nicklow, Multi-objective automatic calibration of SWAT using NSGA-II, *Journal of Hydrology*, 341, 165-176, 2007.

Benaman, J., C.A. Shoemaker, and D.A. Haith, Calibration and Validation of the Soil and Water Assessment Tool on the Cannonsville Reservoir Watershed, *J. of Hydrologic Engineering, 10*(5), pp. 363-374, 2005.

Beven, K J., Towards a coherent philosophy for environmental modelling, *Proc. Roy. Soc. Lond.,*A460, 458, 2465-2484, 2002a.

Beven, K. J. and Binley, A. M., The future of distributed models: model calibration and uncertainty prediction, *Hydrological Processes*, 6, 279-298, 1992.

Beven, K. J., Equifinality and Uncertainty in Geomorphological Modelling, in B L Rhoads and C E Thorn (Eds.), *The Scientific Nature of Geomorphology*, Wiley: Chichester, 289-313, 1996.

Beven, K. J., Prophecy, reality and uncertainty in distributed hydrological modelling, *Adv. Water Resourc.*, 16, 41-51, 1993.

Beven, K. J., Towards an alternative blueprint for a physically-based digitally simulated hydrologic response modelling system, *Hydrol. Process.*, 16, 2002b.

Beven, K., J. Freer, B. Hankin and K. Schulz, The use of generalized likelihood measures for uncertainty estimation in high-order models of environmental systems, in *Nonlinear and Nonstationary Signal Processing*, eds: W.J. Fitzgerald, R.L. Smith, A.T. Walden and P. Young, Cambridge University Press, 2000.

Beven, K., P. Smith, and J. Freer, So just why would a modeler choose to be incoherent?*, Journal of Hydrology*, doi:10.1016/j.hydrol.2008.02.007, 2008.

Beven, K.J., "Interflow', in Morel-Seytoux, H.J. (Ed.), *Unsaturated flow in hydrological modeling*, D. Reidel, Dordrecht, 1989.

Beven, K.J., A.M. Binley, The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes,* 6: 279-298, 1992.

Beven, K.J., and J. Freer, Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, *Jour. of Hydrol.,* 249: 11-29, 2001.

Beven, K.J., *Rainfall-Runoff Modelling – The Primer*, John Wiley and Sons, New York, 360p, 2001.

Bianchini, G., A. Cortes, T. Margalef and E. Luque, Improved prediction methods for wildfires using high performance computing: A comparison, *Lecture Notees in Computer Science*, 3991: 539-546, 2006.

Blasone, R-S., H. Madsen, and D. Rosbjerg, Uncertainty assessment of integrated distributed hydrological models using GLUE with Markov chain Monte Carlo sampling *J. of Hydrology*, 353, 18-32, 2008b.

Blasone, R-S., J.A. Vrugt, H. Madsen, D. Rosbjerg, B.A. Robinson, and G.A. Zyvoloski, Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov chain Monte Carlo sampling, *Advances in Water Resources* , 31, 630–648, 2008a.

Bonta J.V., PRECIPITATION, WATERSHED ANALYSIS, *Encyclopedia of Soils in the Environment*, pp 314-322, 2004.

Boyle, D. P., H. Gupta, V. Hoshin, and S. Sorooshian, Toward improved calibration of hydrologic models: combining the strengths of manual and automatic methods, *Water Resources Research*, 36, 3663-3674, 2000.

Bulygina, N. and H. Gupta, Estimating the uncertain mathematical structure of a water balance model via Bayesian data assimilation, *Water Resour. Res.,* VOL. 45, W00B13, doi:10.1029/2007WR006749, 2009.

Cameron, D., An application of the UKCIP02 climate change scenarios to flood estimation by continuous simulation for a gauged catchment in the northeast of Scotland, UK (with uncertainty), *Journal of Hydrology*, Volume 328, Issues 1-2, Pages 212-226, 2006.

Carlin, B. P. and T. A. Louis, *Bayes and Empirical Bayes Methods for Data Analysis*. 2nd Edition. Chapman and Hall, CRC, New York, NY, 2000.

Carrera J, A. Alcolea, J.Hidalgo, and L.J. Slooten, The inverse problem in hydrology, *Jour. of Hydrol.,* 13: 206-222, 2005.

Carrera J, A. Alcolea, J.Hidalgo, and L.J. Slooten, The inverse problem in hydrology, *Jour. of Hydrol.,* 13: 206-222, 2005.

Chang, M., and L. Harrison, Field assessments on the accuracy of spherical gauges in rainfall measurements, *Hydrol. Processes.*, 19, 403–412, 2005.

Chaubey, I., C. T. Haan, S. Grunwald, and J. M. Salisbury, Uncertainty in the model parameters due to spatial variability of rainfall, *Journal of Hydrology*, Volume 220, Issues 1-2, Pages 48-61, 1999.

Christensen, S., A synthetic groundwater modeling study of the accuracy of GLUE uncertainty intervals *Nordic Hydrology,* 35(1) 45-59, 2004.

Clark M. P. and J. A. Vrugt, Unraveling uncertainties in hydrologic model calibration: Addressing the problem of compensatory parameters, *Geophys. Res*., Lett., 33, L06406, doi:10.1029/2005GL025604, 2006.

Clark, M.P., and A.G. Slater, Probabilistic Quantitative Precipitation Estimation in Complex Terrain. *Journal of Hydrometeorology*, 7, 3–22, 2006.

Clark, M.P., A.G. Slater, D.E. Rupp, R.A. Woods, J.A. Vrugt, H.V. Gupta, T. Wagener, and L.E. Hay, Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models. *Water Resources Research*, 44, 2008.

Coles, S., *An Introduction to Statistical Modeling of Extremes,* Springer-Verlag London, 2001.

Dennis, J.E, and R.B Schnabel, *Numerical Methods for unconstrained optimization and nonlinear equations,* Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1983.

Devore, J. L., *Probability and statistics for engineering and the sciences, seventh edition*, Thomson Brooks/Cole, Belmont, CA, 2008.

Draper, N. R., and H. Smith**,** *Applied Regression Analysis,* 3rd Edition, ISBN: 0-471-17082-8, J. Wiley and Sons, New York, 1998.

Duan, Q., H. Gupta, S. Sorooshian, A. N. Rousseau, and R. Turcotte, *Calibration of Watershed Models*, AGU Books, Washington, 2003.

Duan, Q., S. Sorooshian, V. Gupta, Effective and efficient global optimization for conceptual rainfall-runoff models, *Water Resources Res., 28* (4), 1015 – 1031, 1992.

Ekström M., and P. D. Jones, Impact of rainfall estimation uncertainty on streamflow estimations for catchments Wye and Tyne in the United Kingdom, *International Journal of Climatology*, 29(1), 79, 2009.

Engeland K., and L. Gottschalk, Bayesian estimation of parameters in a regional hydrological model. *Hydrology and Earth System Sciences,* 6: 883-898, 2002.

Engeland, K., C-Y, Xu, and L. Gottschalk, Assessing uncertainties in a conceptual water balance model using Bayesian methodology, *Hydrological Sciences Journal 50*(1), 45-63, 2004.

Evensen, G., *Data assimilation: The Ensemble Kalman Filter*, Springer (2006) 280 pp, 2006.

Francos, A., F.J. Elorza, F. Bouraoui, L. Galbiati, and G.. Bidoglio, Sensitivity analysis of hydrological models at the catchment scale: a two-step procedure, *Reliability Engineering and System Safety*, 79, pp. 205-218, 2001.

Freer, J., Beven, K.J. and Ambroise, B., Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach. *Water Resour. Res.*, 32(7): 2161-2173, 1996.

Freer, J., K. Beven, and N. Peters, Multivariate seasonal period model rejection within the generalized likelihood uncertainty estimation procedure, in *Calibration of Watershed Models*, edited by Duan Q et al., pp. 69-87, AGU, 2003.

Freni, G., G. Mannina, and G. Viviani, Uncertainty assessment of an integrated urban drainage model, *Journal of Hydrology*, Volume 373, Issues 3-4, Pages 392-404, 2009.

Frey, H.C., and S.R. Patil, Identification and Review of Sensitivity Analysis Methods, *Risk Analysis 23*(3), 553-578, 2002.

Gan, T.Y. and S.J. Burges, An assessment of a conceptual rainfall-runoff model's ability to represent the dynamics of small hypothetical catchments: 2 Hydrologic responses for normal and extreme rainfall, *Water Resour. Res.*, 26(7), 1605-1619, 1990.

Gelman, A., J.B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, Chapman and Hall, CRC, Boca Raton, 1995.

Gupta, H.V., S. Soroooshian, T. S. Hogue, and D. P. Boyle, Advances in automatic calibration of watershed models, in *Calibration of Watershed Models*, edited by Duan Q et al., pp. 9-28, AGU, 2003.

Gupta, H.V., Sorooshian, S., and Yapo, P.O., Towards improved calibration of hydrologic models: multiple and incommensurable measures of information, *Water Resour. Res.*, 34, 751-763, 1998.

Hanna, E., How effective are tipping-bucket rain gauges?, A review, *Weather*, 50, 336–342, 1995.

Hassan, A. E., H. M. Bekhit, and J. B. Chapman, Uncertainty assessment of a stochastic groundwater flow model using GLUE analysis, *Journal of Hydrology*, Volume 362, Issues 1-2, Pages 89-109, 2008.

Huard, D., and A. Mailhot, A Bayesian perspective on input uncertainty in model calibration: Application to hydrological model ‘‘abc,’’ *Water Resour. Res.*, 42, W07416, doi:10.1029/2005WR004661, 2006.

Hellweger F.L. and U. Lall, Modeling the effect of algal dynamics on arsenic speciation in Lake Biwa *Environmental Science & Technology,* 38: 6716-6723, 2004.

Hogue, T. S., H. V. Gupta, S. Sorooshian, and C. D. Tomkins, A multi-step automatic calibration scheme for watershed models, in *Calibration of Watershed Models*, edited by Duan Q et al., pp. 165-174, AGU, 2003.

Hossain F. and E.N. Anagnostou, Assessment of current passive-microwave- and infrared-based satellite rainfall remote sensing for flood prediction, *Journal of Geophysical Research-Atmospheres,* **110:** Art. No. D06115, 2005.

Jia, Y.B., and T.B. Culver, Uncertainty analysis for watershed modeling using Generalized Likelihood Uncertainty Estimation with multiple calibration measures, *Journal of Water Resources Planning and Management*, Vol. 134, No. 2, doi: 10.1061/(ASCE)0733-9496(2008)134:2(97), 2008.

Kanso, A., G. Chebbo and B. Tassin, Application of MCMC-GSA model calibration method to urban runoff quality modeling, *Reliability Engineering and System Safety*, 91, pp. 1398–1405, 2006.

Kavetski, D.N., Franks, S.W. and G. Kuczera, Confronting input uncertainty in environmental modelling, in *Calibration of Watershed Models*, edited by Q. Duan, H. Gupta, S. Sorooshian, A. N. Rousseau, and R. Turcotte, pp. 49-68, AGU Books, 2002.

Kavetski, D.N., G. Kuczera and S.W. Franks, Bayesian analysis of input uncertainty in hydrological modelling: 1. Theory, *Water Resources Journal,* 42, W03407; 1–9, 2006a.

Kavetski, D.N., G. Kuczera and S.W. Franks, Bayesian analysis of input uncertainty in hydrological modeling: 2. Application, *Water Resources Research*, 42, W03408; 1–10, 2006b.

Kinner D.A. and R.F. Stallard, Identifying storm flow pathways in a rainforest catchment using hydrological and geochemical modelling, *Hydrological Processes,* **18**: 2851-2875, 2004.

Kuczera, G., and B.J. Williams, Effect of rainfall errors on accuracy of design flood estimates. *Water Resour. Res.*, 28, 4, pp. 1145–1153, 1992.

Kuczera, G. and E. Parent, Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm, *Jour. of Hydrol.,* 211: 69-85, 1998.

Kuczera, G., and E. Parent, Monte Carlo assessment of parameter uncertainty in conceptual catachment models: the Metropolis algorithm, *Journal of Hydrology*, 211, 69-85, 1998.

Kuczera, G., and M. Mroczkowski, Assessment of hydrologic parameter uncertainty and the worth of multiresponse data, *Water Resources Research*, 34, 1481-1489, 1998.

Kuczera, G., On the validity of first-order prediction limits for conceptual hydrologic models, *J. Hydrol., 103*, 229 – 247, 1998.

Kuczera, G., D. Kavetski, S. Franks, and M. Thyer, Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters, *Journal of Hydrology*, Volume 331, Issues 1-2, 161-177, 2006.

Lall, U., D.L. Phillips, K.H. Reckhow, and D.P. Loucks (chair), Quantifying and Communicating Model Uncertainty for Decision Making in the Everglades, Report of the Comprehensive Everglades Restoration Plan Model Uncertainty Workshop, May, 2002.

Liang S., R.C. Spear, E. Seto E, et al., A multi-group model of Schistosoma japonicum transmission dynamics and control: model calibration and control prediction *Tropical Medicine & International Health,* 10: 263-278, 2005.

Loucks, D. P., and E. van Beek, (with contributions by Jery R. Stedinger [chapters 7-8-9]), Jozef P.M. Dijkman, Monique T. Villars, *Water Resources Systems Planning and Management An Introduction to Methods, Models and Applications,* Studies and Reports in Hydrology, United Nations Educational, Scientific and Cultural Organization, UNESCO PUBLISHING, Paris, France, 2005.

Loucks, D.P., and J.R. Stedinger, *Sensitivity and Uncertainty Analysis in Hydrologic Simulation Modeling of the South Florida Water Management District*, Report of the Workshop on Reduction of Uncertainties in Regional Hydrologic Simulation Models, 10 June, 1994.

Mantovan, P. and E. Todini, Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology, *Journal of Hydrology*, 330, 368-381, doi:10.1016/j.hydrol.2006.04.046, 2006.

Marshall, L. D. Nott, and A. Sharma, Hydrological model selection: a Bayesian alternative, *Water Resources Research*, Vol. 41, 2005a.

Marshall, L., D. Nott, and A. Sharma, A comparative study of Markov chain Monte Carlo methods for conceptual rainfall-runoff modeling, *Water Resources Research*, 40, W02501, doi: 10.1029/2003WR002378, 2004.

Mertens J. et al., Including prior information in the estimation of effective soil parameters in unsaturated zone modeling *Journal of Hydrology,* **294**: 251-269, 2004.

Misirili, F., H. V, Gupta, S. Sorooshian, and M. Thiemann, Bayesian recursive estimation of parameter and output uncertainty for watershed models, in *Calibration of Watershed Models*, edited by Duan Q et al., pp. 113-124, AGU, 2003.

Mo X.G., and K.J. Beven, Multi-objective parameter conditioning of a three-source wheat canopy model, *Agriculture and Forest Meteorology,* **122**: 39-63, 2004.

Montanari, A., Large sample behaviours of the generalized likelihood uncertainty estimation (GLUE) in assessing the uncertainty of rainfall-runoff simulations, *Water Resour. Res.,* W08406, doi:10.1029/2004WR003826, 2005.

Morin, E., D. Goodrich, R.A. Maddox, X. Gao, H. Gupta, S. Sorooshian, Spatial patterns in thunderstorm rainfall events and their coupling with watershed hydrological response, *Advances in Water Resources*, 29(6): 843-860, 2006.

Moradkhani, H., K.-L. Hsu, H. Gupta, and S. Sorooshian, Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter, *Water Resour. Res.*, 41, W05012, doi:10.1029/2004WR003604, 2005a.

Moradkhani, H., Sorooshian S., Gupta, H.V., Houser, P, Dual State-Parameter Estimation of Hydrological Models using Ensemble Kalman Filter, *Advances in Water Resources*, 28, 2, 135-147, 2005b.

Morgan, M. G. and M. Henrion, *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*, Cambridge University Press, New York, 1990.

Mugunthan, P., and C.A. Shoemaker, Assessing the impacts of parameter uncertainty for computationally expensive groundwater models, *Water Resour. Res.*, 42, W10428, doi:10.1029/2005WR004640, 2006.

Nash J, and J. Sutcliffe, River flow forecasting through conceptual models, 1. A discussion of principles. *Journal of Hydrology,* **10**: 282–290, 1970.

National Research Council (NRC), *Risk Analysis and Uncertainty in Flood Damage Reduction Studies,* National Academy Press, Washington, D.C., 2000.

Page T., J.D. Whyatt, K.J. Beven, et al., Uncertainty in modelled estimates of acid deposition across Wales: a GLUE approach, *Atmospheric Environment,* 38: 2079-2090, 2004.

Page, T., J.D. Whyatt, S.E. Metcalfe, R.G. Derwent, and C. Curtis, Assessment of uncertainties in a long range atmospheric transport model: Methodology, application and implications in a UK context, *Environmental Pollution*, Volume 156, Issue 3, Pages 997-1006, 2008.

Pappenberger F, K.J. Beven, M. Horritt, and others, Uncertainty in the calibration of effective roughness parameters in HEC-RAS using inundation and downstream level observations. *Jour. of Hydrol.,* 302: 46-69, 2005.

Perrin, C. et al., Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments, *Journal of Hydrology*, 242, 275-301, 2001.

Pinol J., K.J. Beven, and D. Viegas, Modelling the effect of fire-exclusion and prescribed fire on wildfire size in Mediterranean ecosystems *Ecological Modelling,* 183: 397-409, 2005.

Press W.H., B. P. Flannery, S.A. Teukolsky, and W.T. Vetterling, *Numerical Recipes: The Art of Scientific Computing,* Cambridge University Press, 1986.

Reichle, R., D. McLaughlin and D. Entekhabi, Hydrologic data assimilation with the ensemble Kalman filter, *Mon. Weather Rev.*, 130 (1) (2002), pp. 103–114, 2002a.

Reichle, R., D. McLaughlin and D. Entekhabi, Extended versus ensemble Kalman filtering for land data assimilation, *J. Hydrometeorol*. 3 (6) (2002), pp. 728–740, 2002b.

Romanowicz, R., K.J. Beven, and J. Tawn, Bayesian calibration of flood inundation models, in M.G. Anderson and D.E. Walling and P.D. Bates (Eds.), *Floodplain Processes*, Wiley Chichester, 1996.

Ruessink, B.G., Predictive uncertainty of a nearshore bed evolution model, *Continental Shelf Research*, 25(9):1053-1069, 2005.

Salamon, P., and L. Feyen, Assessing parameter, precipitation, and predictive uncertainty in a distributed hydrological model using sequential data assimilation with the particle filter, *Journal of Hydrology*, 376, Issues 3-4, Pages 428-442, 2009.

Saltelli, A., K. Chan, and E.M. Scott, *Sensitivity Analysis,* J. Wiley and Sons, Ltd, Chichester, England, 2000.

Shoemaker, C.A., *Uncertainty Workshop Outline, Interagency Modeling Center, South Florida Water Management District*, Weest Palm Beach, FL, 24, 2004.

Sieck, L.C., S. J. Burges, and M. Steiner, Challenges in obtaining reliable measurements of point rainfall, *Water Resour. Res.*, 43, 2007.

Smith R.M.S., D.J. Evans and H.S. Wheater, Evaluation of two hybrid metric-conceptual models, for simulating phosphorus transfer from agricultural land in the river enborne, a lowland UK catchment *Journal of Hydrology,* 304: 366-380, 2005.

Stedinger, J.R., R.M. Vogel, S.U. Lee, and R. Batchelder. Appraisal of Generalized Likelihood Uncertainty Estimation (GLUE) Methodology, *Water Resour. Res.*, 44, WOOB06, doi:10.1029/2008WR006822, 2008.

Thiemann M., M. Trosset, H. Gupta, and S. Sorooshian, Bayesian recursive parameter estimation for hydrologic models, *Water Resources Research*, 37, 2521-2535, 2001.

Thorndahl, S., K.J. Beven, J.B. Jensen, and K. Schaarup-Jensen, Event based uncertainty assessment in urban drainage modelling, applying the GLUE methodology, *Journal of Hydrology*, Volume 357, Issues 3-4, Pages 421-437, 2008.

Thyer M., and G.. Kuczera, A hidden Markov model for modeling long-term persistence in multi-site rainfall time series 1. Model calibration using a Bayesian approach, *Journal of Hydrology*, Volume 275, Pages 12-26, 2003.

Tolson, B. A. and Shoemaker, C. A., Comparison of optimization algorithms for the automatic calibration of SWAT2000, SWAT 2005 3rd International Conference, Zurich, Switzerland, 2005.

Tolson, B.A. and C.A. Shoemaker, Improving the efficiency of practical uncertainty assessment in the calibration of environmental simulation models, *Water Resour. Res.*, under review, 2007.

Tremblay M. and D. Wallach, Comparison of parameter estimation methods for crop models *Agronomie,* 24: 351-365, 2004.

Turcotte, R., A. N. Rousseau, J. P. Fortin, and J. P. Villeneuve, A Process-oriented, multiple-objective calibration strategy accounting for model structure, in *Calibration of Watershed Models*, edited by Duan Q et al., pp. 153-164, AGU, 2003.

Uhlenbrook S., and A. Sieber, On the value of experimental data to reduce the prediction uncertainty of a process-oriented catchment model *Environmental Modelling & Software,* **20**: 19-32, 2005.

Van Griensven A., A. Francos and W. Bauwens, Sensitivity analysis and auto-calibration of an integral dynamic model for river water quality. *Water Sci. Technol*., 45(9), 325-332, 2002.

Viola, F., L.V. Noto, M. Cannarozzo, and G. La Loggia, Daily streamflow prediction with uncertainty in ephemeral catchments using the GLUE methodology, *Physics and Chemistry of the Earth*, Parts A/B/C, doi:10.1016/j.pce.2009.06.006, 2009.

Vrugt, J.A., H.V. Gupta, W. Bouten, and S. Sorooshian, A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrological model parameters, *Water Resour. Res.,* 39, 8, 1201, DOI: 10.1029/2002WR001642, 2003.

Vrugt J. A., C. G. H. Diks, H. V. Gupta, W. Bouten, and J. M. Verstraten, Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation, *Water Resour. Res.*, 41, W01017, doi:10.1029/2004WR003059, 2005.

Vrugt, J.A., J.F. Cajo, M.P. Clark, J.M. Hyman, and B.A. Robinson, Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, *Water Resour. Res.,* 44, DOI: 10.1029/2007WR006720, 2008.

Wagener, T., N. McIntyre, M. J. Lees, H. S. Wheater, and H.V. Gupta, Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiabilityanalysis. *Hydrological Processes*, 17(2), 455-476, 2003.

Weisstein, Eric W., "Nonlinear Least Squares Fitting." From *Mathworld* - A Wolfram Website Resource, http:// mathworld.wolfram.com/NonlinearLeastSquares Fitting.html, 2005.

Yang J., P. Reichert, K. C. Abbaspour, and H. Yang, Hydrological modelling of the Chaohe Basin in China: Statistical model formulation and Bayesian inference, *Journal of Hydrology*, Vol. 340, Pages 167-182, 2007.

Yapo, P. H. Gupta, and S. Sorooshian, Automatic calibration of conceptual rainfall-runoff modes: sensitivity to calibration data*, Journal of Hydrology*, 181, 23-48, 1996.

Zak, S., K.J. Beven and B. Reynolds, Uncertainty in the estimation of critical loads: A practical methodology, *Soil, Water and Air Pollution*, 98, 297-316, 1997.

Zellner, A., *An Introduction to Bayesian Inference in Econometrics*, John Wiley and Sons, New York, 1971.