Appraisal of the generalized likelihood uncertainty estimation (GLUE) method

Jery R. Stedinger,[1] Richard M. Vogel,[2] Seung Uk Lee,[1] and Rebecca Batchelder[2]

Received 9 January 2008; revised 18 June 2008; accepted 4 August 2008; published 1 November 2008.

[1] Recent research documents that the widely accepted generalized likelihood uncertainty estimation (GLUE) method 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. In particular, GLUE generally fails to produce intervals that capture the precision of estimated parameters, and the difference between predictions and future observations. This paper 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. Beven and others have suggested that the choice of the likelihood measure used in a GLUE computation is subjective and may be selected to reflect the goals of the modeler. If an arbitrary likelihood 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. The traditional subjective likelihood measures that have been used with GLUE also fail to reflect the nonnormality, heteroscedasticity, and serial correlation among the residual errors generally found in real problems, and hence are poor metrics for even simple sensitivity analyses and model calibration. Most previous applications of GLUE only produce uncertainty intervals for the average model prediction, which by construction should not be expected to include future observations with the prescribed probability. We show 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.

Citation: Stedinger, J. R., R. M. Vogel, S. U. Lee, and R. Batchelder (2008), Appraisal of the generalized likelihood uncertainty estimation (GLUE) method, *Water Resour. Res.*, *44*, W00B06, doi:10.1029/2008WR006822.

1. Introduction

[2] 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* [2003] and a fundamental need of most end users [*Montanari*, 2007].

[3] 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 500 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.

[4] Recent evaluations of GLUE by *Christensen* [2004], *Montanari* [2005], *Mantovan and Todini* [2006] 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*

[1]School of Civil and Environmental Engineering, Cornell University, Ithaca, New York, USA.
[2]Department of Civil and Environmental Engineering, Tufts University, Medford, Massachusetts, USA.

0043-1397/08/2008WR006822

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 from *Beven et al.* [2007, 2008] again reinforces a major point made by *Mantovan and Todini* [2006], 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.

[5] 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.''

[6] 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.

[7] The aim of this paper 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.

[8] The statistical and probabilistic interpretation of GLUE analyses and the choice of a likelihood function is the focus of this paper. This paper 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 paper 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].

1.1. Previous Applications of GLUE

[9] *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], 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], 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], 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] and this study provide such reviews.

1.2. GLUE Methodology

[10] The Beven-Binley GLUE method is a Monte Carlo approach which is an extension of Generalized Sensitivity Analysis (GSA) introduced by *Spear and Hornberger* [1980]. With GSA, ensembles of model parameters are sampled from distributions, typically with independent uniform or normal distributions for each parameter. The model is then run with many such parameter sets, producing multiple sets of model output. These are used together to generate uncertainty intervals for model predictions. *Spear and Hornberger* [1980] suggest a qualitative criterion to group the generated model parameters into two sets: (1) behavioral sets of model parameters that produce results consistent with the observations, and (2) nonbehavioral sets of model parameters that produce results that contradict the observations. Therefore, they implicitly weighted each model parameter set by giving nonbehavioral sets a probability of zero and all behavioral sets an equal nonzero probability.

[11] Like GSA, GLUE is based upon Monte Carlo simulation. Parameter sets may be sampled from any probability distribution, with most reported applications sampling from uniform distributions [*Beven*, 2001]. Each parameter set is used to produce model output; the acceptability of each model run is then assessed using a goodness-of-fit criterion which compares the predicted to observed values over some calibration period. The goodness-of-fit function is used to construct what *Beven and Binley* [1992, p. 283] call a likelihood measure. As with GSA, parameter sets that result in goodness-of-fit/likelihood values below a certain threshold are again termed ''nonbehavioral'' and are discarded. The remaining ''behavioral'' parameter sets are assigned rescaled likelihood weights that sum to 1, and thus look like probabilities. 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 nonbehavioral

solutions without the imposition of an arbitrary and rigid cutoff. As we will show, a correct statistical analysis does just that in our example.

[12] To obtain uncertainty intervals around model predictions using these rescaled likelihood weights, the model outputs are ranked so that the rescaled likelihood weights can be used to form a cumulative distribution for the output variable. From that distribution, quantiles are selected to provide uncertainty intervals for the variable of concern. Clearly this computation reflects only uncertainty arising from model parameter uncertainty. Nothing has been done in constructing the intervals to reflect the precision with which the model could reproduce observed values of the modeled variable over the calibration data set. In the previous quote, Beven referred to structural errors. Structural errors (or equivalently model errors) describe the inability, of even the best model with optimal parameters, to exactly reproduce the target output. *Kuczera et al.* [2006] provide a good example highlighting the fact that poorly determined parameters do not necessarily lead to high predictive uncertainty. Instead, they show that predictive uncertainty is often dominated by the model error component. We show that most previous GLUE applications have not handled this important model error component properly.

[13] Although GLUE is now very popular, it has frequently been criticized for its large computational demands. *Kuczera and Parent* [1998] note that GLUE "may require massive computing resources to characterize a highly dimensioned parameter space." *Jia and Culver* [2008] report generating 50,000 parameter sets to find 381 acceptable sets (just 0.8%) for their watershed study. As *Kuczera and Parent* [1998, p. 72] explain, use of a simple and uniform prior probability distribution of model parameters over a relatively large region, can result in an algorithm that, even after billions of model evaluations, may not have generated even one good solution. Others have noted that it is difficult to determine how great the computational demand will be, because there is no way of determining a priori how many parameters sets will be necessary to adequately characterize the model response surface [*Carrera et al.*, 2005; *Pappenberger et al.*, 2005). *Mugunthan and Shoemaker* [2006], *Tolson and Shoemaker* [2007], *Blasone et al.* [2008a, 2008b] and others, have developed computationally efficient approaches for performing calibration and uncertainty analysis of complex environmental simulation models. Our focus is on the validity of the GLUE statistical computation, and not its computational efficiency, though both are serious concerns.

[14] The next section develops a framework for describing model uncertainty so that the appropriate statistical computation for an uncertainty analysis using GLUE can be understood. Section 3 summarizes the various likelihood measures which have been employed in practice and their use of the residual mean square error. Section 4 describes an evaluation of GLUE performance using a simple linear regression model as an example for which exact and correct analytical uncertainty intervals are available. Section 5 summarizes results of our simulation experiments and shows how use of GLUE with a correct likelihood function can lead to meaningful uncertainty and prediction intervals. Section 6 raises questions regarding Beven's recent manifesto [*Beven*, 2006a] and finally, section 7 provides recommendations for future research for improving GLUE applications.

2. Bayes Theorem and the Likelihood Function

[15] In this section, we use Bayes Theorem to derive the posterior probability that should be assigned to different sets of parameters generated in a Monte Carlo simulation. *Romanowicz et al.* [1994, 1996], *Romanowicz and Beven* [1998], *Beven and Freer* [2001], *Beven et al.* [2008], and others have used GLUE as a correct Bayesian statistical procedure. Unfortunately, of the hundreds of previous GLUE applications, few have used a correct statistical error model instead of one of the informal likelihood functions traditionally adopted with GLUE. The relationships in this section provide us with the correct descriptions of uncertainty for our simple example, and enable us to show how to generate statistically valid results using GLUE.

[16] Denote a set of streamflow observations Q_t for $t = 1,\ldots n$, to be used for model calibration by the vector Q. Using Bayes theorem the posterior density function (pdf) of the parameter vector θ based on these observations is

$$f_{\theta|Q}[\theta|Q] = c f_{\varepsilon|\theta}[\varepsilon|\theta] f_\theta[\theta], \tag{1}$$

where c is a normalization constant, θ is the parameter vector, $f_\theta[\theta]$ 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 streamflow model prediction vector which depends on the parameter vector θ. In equation (1) 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 data set Q. The pdf $f_\theta[\theta]$ is called the "prior" distribution of the model parameters, because it is based on information one has pertaining to the parameters prior to model calibration (estimation). The function $f_{\varepsilon|\theta}[\varepsilon|\theta]$ is called the likelihood function (in other words the likelihood of producing the model errors, ε); it must represent the probability of the observed flow vector Q for a given set of model parameters θ.

[17] The GLUE methodology randomly draws parameter sets θ_i, say $i = 1,\ldots T$ times, from a prior distribution. Assuming a uniform probability distribution for each of the J parameters which are independent, as has often been done in GLUE applications, yields the prior pdf,

$$f_\theta[\theta] = \frac{1}{\prod_{j=1}^{J}\left[\theta_{\max,j} - \theta_{\min,j}\right]} \text{ for } \theta_{\min,j} \le \theta_{i,j} \le \theta_{\max,j}, \tag{2}$$

where $\theta_{\min,j}$ and $\theta_{\max,j}$ are minimum and maximum values for each parameter.

[18] Application of Bayes theorem for independent errors ε_t then leads to the posterior probability for the parameter vector θ,

$$f_{\theta|Q}[\theta|Q] = c\prod_{t=1}^{n} f_{\varepsilon|\theta}[\varepsilon_t|\theta] f_\theta[\theta] = \frac{c\prod_{t=1}^{n} f_{\varepsilon|\theta}[\varepsilon_t|\theta]}{\prod_{j=1}^{J}\left[\theta_{\max,j} - \theta_{\min,j}\right]}, \tag{3}$$

where n is the number of flow observations in the calibration data set and J is the number of model parameters. The posterior probability is the probability assigned to a parameter set θ given the calibration flow vector Q used to assess the ability of those parameters to reproduce that data relative to other parameter sets. Parameter sets that produce smaller errors should have a higher likelihood function value, and thus a higher posterior probability. This is the spirit behind *Beven and Binley*'s [1992] use of goodness-of-fit criteria as likelihood measures. Nevertheless, for a given model structure, equations (1) and (3) require that the likelihood function represent the probability that the model with a given set of parameters and a prescribed model-error distribution would have generated the vector of observations Q. Equations (1) and (3) are simply an expression of Bayes Theorem which is the mathematical law describing how probabilities behave when new information is provided. The likelihood function corresponds to the pdf of the errors described by $f_{\varepsilon|\theta}[\varepsilon|\theta]$. It follows that for a given set of assumptions regarding the probabilistic structure of the errors ε, there is only one likelihood function.

[19] In keeping with our goal to keep the example simple, we assume, as did *Christensen* [2004] and many others, that the model errors, ε_t, are normal and independently distributed (NID) with zero mean and unknown variance σ_ε^2. This yields the following pdf for each ε_t conditioned upon the estimated parameters:

$$f_{\varepsilon|\theta}[\varepsilon_t|\theta] = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left[-\frac{\left(Q_t - \hat{Q}_t\right)^2}{2\sigma_\varepsilon^2}\right], \tag{4}$$

wherein Q_t are the observed flows and $\hat{Q}_t$ are the predicted flows so that $\varepsilon_t = Q_t - \hat{Q}_t$. Section 7 discusses what can be done when the model errors are neither normally distributed, independent nor homoscedastic. In the case that residual errors are correlated, a time series model can be adopted which yields a series of independent innovations [*Beven and Freer*, 2001, equation 3; *Box et al.*, 1994]; nonnormality can often be resolved by a suitable transformation.

[20] Substitution of (4) into (1) yields the correct posterior pdf,

$$f_{\theta|Q}[\theta|Q] = cf_\theta[\theta] \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left[-\frac{\left(Q_t - \hat{Q}_t\right)^2}{2\sigma_\varepsilon^2}\right] \text{ for } \theta_{\min,j} \le \theta_{i,j} \le \theta_{\max,j} \tag{5}$$

where n is the length of the climate and streamflow record used to estimate (calibrate) the parameters.

[21] To use equation (5) one 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 (MLE) $\hat{\sigma}_\varepsilon^2$ equal to

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{n}\sum_{t=1}^{n}\left(Q_t - \hat{Q}_t^{MLE}\right)^2 \tag{6}$$

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 $\hat{Q}_\varepsilon^2$ for the likelihood function in (5).

[22] After substitution of (6) into (5) and some rearranging, one obtains the posterior distribution for the model parameters θ given NID errors,

$$f_{\theta|Q}[\theta|Q] = c(\hat{\sigma}_\varepsilon)^{-n}(2\pi)^{-n/2} \exp\left[-\frac{n}{2}\frac{\sum_{t=1}^{n}\left(Q_t - \hat{Q}_t\right)^2}{\sum_{t=1}^{n}\left(Q_t - \hat{Q}_t^{MLE}\right)^2}\right] f_\theta[\theta]. \tag{7}$$

The terms $c\ (\hat{\sigma}_\varepsilon)^{-n}\ (2\ \pi)^{-\mathrm{n}/2}$ are constant in a particular application, as is the prior distribution for the parameters $f_\theta[\theta]$ using independent uniform priors on each parameter, hence they may be combined to yield

$$f_{\theta|Q}[\theta|Q] = \kappa \exp\left[-\frac{n}{2}\frac{\sum_{t=1}^{n}\left(Q_t - \hat{Q}_t\right)^2}{\sum_{t=1}^{n}\left(Q_t - \hat{Q}_t^{MLE}\right)^2}\right] \text{for } \theta_{\min,j} \le \theta_{i,j} \le \theta_{\max,j}, \tag{8}$$

where κ is simply a constant term whose value may be determined by the requirement that the integral over all the parameters of the density function should equal one. Equation (8) is equivalent to

$$f_{\theta|Q}[\theta|Q] = \kappa \exp\left[-\frac{ns_\varepsilon^2}{2\hat{\sigma}_\varepsilon^2}\right], \tag{9}$$

wherein $s_\varepsilon^2 = \sum_{t=1}^{n}(Q_t - \hat{Q}_t)^2/n$ is the mean square error for the model with a particular set of possible parameters θ and $\hat{\sigma}_\varepsilon^2$ is the MLE of the model error variance. Alternatively one can rewrite (9) as

$$f_{\theta|Q}[\theta|Q] = \kappa \exp\left\{-\frac{n}{2}\frac{[1 - R^2(\theta)]}{\left[1 - R^2\left(\hat{\theta}_{MLE}\right)\right]}\right\}, \tag{10a}$$

wherein the values of the R-squared statistics are

$$R^2(\theta) = \left[1 - \frac{s_\varepsilon^2}{s_Q^2}\right] \text{and } R^2\left(\hat{\theta}_{MLE}\right) = \left[1 - \frac{\hat{\sigma}_\varepsilon^2}{s_Q^2}\right] \tag{10b}$$

These describe how well the model with parameter set θ, in conjunction with the maximum likelihood parameter estimators fit the data. Here s_Q^2 denotes the variance of the observations Q, which is unaffected by the model parameters.

[23] Equations (8)–(10) 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 are 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. Said another way, "data provides information, and more data provides more information".

[24] This whole idea plays a very important role in the analysis of the precision of maximum likelihood estimation. In general, asymptotically, the sampling variance matrix Σ for the model parameter estimators is equal to the negative of the inverse of the expected second derivative of the log likelihood function [*Bickel and Doksum*, 2001]. Thus,

$$\Sigma^{-1} = -E\left[\frac{\partial^2 \ln L(Q|\theta)}{\partial \theta^2}\right] = -nE\left[\frac{\partial^2 \ln f(Q_i|\theta)}{\partial \theta^2}\right], \qquad (11)$$

wherein $L(Q|\theta)$ is the likelihood function for the entire data set Q, assuming independent observations, and $f(Q_i|\theta)$ is the probability density function for a single observation. We mention this general theory here because it makes so clear the value of each observation in a correct analysis, and how the precision of the estimated parameters depends on the number of observations n.

[25] The power of appropriate use of the correct likelihood function for the data is truly impressive. As illustrated by the classical analysis presented by *Kendall and Stuart* [1961], subject to several regularity conditions, the maximum likelihood estimators asymptotically achieve the minimum variance among all asymptotically unbiased estimators, which is the Cramér-Rao inequality. However, this analysis requires that one use the likelihood function that describes the distribution of the data that would be available for inference in repeated sampling.

[26] As a practical matter, we are interested in a Monte Carlo procedure that will generate a set of discrete parameter sets $\{\theta_i\}$ and their associated probabilities p_i which jointly provide a consistent and unbiased representation of $f_{\theta|Q}[\theta|Q]$. This could be done in several ways, but the simplest is that employed by GLUE. If values of θ_i are randomly drawn using the prior pdf for the parameter vector $f_\theta[\theta]$ in equations (1) and (2), and those parameter vectors are then assigned probabilities proportional to the likelihood function in equations (1) and (3), one obtains

$$p_i = \frac{f_{\theta|Q}[\theta_i|Q]}{\sum_{i=1}^{m} f_{\theta|Q}[\theta_i|Q]}, \qquad (12)$$

for $i = 1,...m$. Consequently $\{\theta_i, p_i\}$ jointly provide the needed representation of $f_{\theta|Q}[\theta_i|Q]$. Note that the constant c in equations (1) and (3) cancels out in equation (12).

[27] Now a critical issue that should not be missed is that equation (12) describes the probability distribution for the unknown parameter vector θ. It is absolutely **not** the predictive distribution describing what might be the value of a future observation, which has been assumed in many previous GLUE applications. This fact is discussed by *Christensen* [2004] and *Blasone et al.* [2008a, 2008b]. *Mantovan and Todini* [2006] and *Mantovan et al.* [2007] emphasize this point.

[28] To generate an uncertainty distribution for what might be a future observation, one needs to consider the uncertainty in the parameters described by $f_{\theta|Q}[\hat{\theta}_i|Q]$ 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, 2006a, 2006b], *Vrugt et al.* [2005], *Moradkhani et al.* [2005a, 2005b], *Kuczera et al.* [2006], *Clark and Vrugt* [2006], and *Huard and Mailhot* [2006].

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

$$f_{Q_f|Q}[Q_f|Q] = \int f_{Q_f|\theta}[Q_f|\theta] f_{\theta|Q}[\theta|Q] d\theta, \qquad (13)$$

with $\varepsilon_t = Q_t - \hat{Q}_t$ and $\hat{Q}_t = M(\theta)$. Thus in developing a predictive distribution for a future observation, one needs to consider the uncertainty in the parameters, and also, what is generally more important, the deviations of the observed flows from even the best prediction [*Mantovan and Todini*, 2006, p. 373].

[30] 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.

3. Likelihood Measures Used with GLUE

[31] The GLUE idea is to combine a priori knowledge of the model parameters captured by the prior pdf, with new information (i.e., observed data) represented by the likelihood 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 for a likelihood measure arguing that "the choice of a likelihood measure will be inherently subjective". They require that "It should be zero for all simulations that are considered to exhibit behavior dissimilar to the system in question, and that 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 measure 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].

[32] Several likelihood measures have been proposed and used in previous applications of the GLUE methodology.

Table 1 of *Beven et al.* [2000] provides a summary. A popular likelihood measure, inverse error variance, introduced by both *Beven* [1989] and *Beven and Binley* [1992] is

$$L_{IV} = \left(s_\varepsilon^2\right)^{-N}, \tag{14}$$

where s_ε is the standard deviation of the model errors, and N is called the "shaping factor". *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 lead to different uncertainty intervals [*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 likelihood weights.

[33] The likelihood measure applied most frequently is based on the efficiency index introduced by *Nash and Sutcliffe* [1970],

$$L_{NS} = \left[1 - \frac{s_\varepsilon^2}{s_Q^2}\right]^N, \tag{15}$$

where s_ε is the standard deviation of the errors, s_Q is the standard deviation of the observations and again, N is again a "shaping factor". For examples, see *Kinner and Stallard* [2004], *Uhlenbrook and Sieber* [2005] and other studies summarized in Table 1 of *Beven et al.* [2000]. *Freer et al.* [1996] used (15) with $N = 1$ and 30, as well as

$$L_{EXP} = \exp\left[-Ns_\varepsilon^2/s_Q^2\right]. \tag{16}$$

which is included in Table 1 of *Beven and Freer* [2001] without the factor s_Q^2. Without s_Q^2 the value of L_{EXP} would depend on the units employed to measure flows, which makes no sense. 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 $s_\varepsilon^2 \to s_Q^2$. 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].

[34] In our experiments, we employ the likelihood function in equation (14) introduced by *Beven and Binley* [1992] to see how closely it resembles a statistically correct likelihood function for NID model errors. Here the sample variance of the errors ε_i is computed assuming the true mean is zero, which provides a penalty for bias. Although the most common value for the shaping factor is $N = 1$ in previous studies [see *Beven et al.*, 2000, Table 1], we consider a range of values for the shaping factor N. Using standard notation, the likelihood in equation (14) can be rewritten as

$$f_{IV}[\theta|Q] = \left[\frac{\sum_{t=1}^{n}\left(Q_t - \hat{Q}\right)^2}{n}\right]^{-N}, \tag{17}$$

which can be further rewritten to resemble (9) and (10) as follows:

$$f_{IV}[\theta|Q] = \kappa\left[s_\varepsilon^2\right]^{-N} = \kappa\left[\left(1 - R^2(\theta)s_Q^2\right]^{-N}, \tag{18}$$

where κ is a constant term chosen to make $\sum_{i=1}^{T} f_{IV}(\theta_i|Q) = 1$ across all T parameter sets, s_Q^2 describes the variance of the observed streamflows and $R^2(\theta)$ is defined in (10b).

[35] The form of the Bayesian likelihood functions in (8)–(10) are quite different from the informal likelihood function in (18), and this would be the case for any of the informal likelihood functions or likelihood measures suggested by *Beven and Binley* [1992] and many others. Only by chance will the subjectively selected informal likelihood function in (18) be representative of the likelihood function in (8)–(10) as is shown below in our example. The correct likelihood function depends critically upon $R^2(\hat{\theta}_{MLE})$ which reflects how well the model really can fit the data. It also depends 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 (18). *Mantovan and Todini* [2006, pp. 373–374] use the term incoherence to describe the failure of informal likelihood functions to account for the value of n.

[36] The GLUE likelihood measure commonly used in the past and given in (18) 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 (10). 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 in equations (14), (15) and (16) that are so often used with GLUE fail to recognize this critical message.

[37] *Vick* [2002, pp. 37–38] calls Bayes Theorem the "crowning achievement of the classical probabilists". "By melding evidence from frequency-based observations with prior belief based on other kinds of knowledge and judgment, it rationalized for us how they could be so apparently adaptable to both". However, if we replace the statistically correct interpretation of the evidence represented by our best description of the probability of seeing the observed sample (as a function of the parameters), with an arbitrary and subjective likelihood function, then the correct and appropriate link to the real data is lost. Bayesian analysis is an algebra for probabilistic statements. Knowingly choosing to forego this algebra is like saying, as a reviewer pointed out, "We elect not to use the formal Navier-Stokes equation because the boundary conditions are unknown. Instead, we will use a subjective equation that does not preserve mass continuity since our assumed boundary conditions would be inaccurate anyway".

4. An Illustration

4.1. A Linear Watershed Model

[38] Here a simple linear regression model is adopted because it enables us to compare the uncertainty intervals

generated using commonly adopted GLUE likelihood measures, with exact confidence and prediction intervals based on classical statistical theory which are known to generate intervals that in repeated trials contain the target parameters with the required frequency.

[39] Let the annual streamflow observations Q_t be related to annual rainfall observations P_t by

$$Q_t = \alpha + \beta \cdot P_t + \varepsilon_t, \tag{19}$$

wherein α and β are model parameters, and ε_t are NID model errors with zero mean and constant variance σ_ε^2. Here ε_t represents the failure of the observed precipitation P_t to capture the true watershed average, and the inability of the linear model to perfectly predict runoff even if it had as input the true watershed average precipitation. For simplicity, we assume that the marginal distributions of Q and P are normally distributed. Suppose that mean values of P and Q are μ_p = 100 cm and μ_Q = 60 cm respectively, and their respective standard deviations are σ_P = 20 cm and σ_Q = 15 cm. Thus coefficients of variation of P and Q are 0.20 and 0.25. 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. We describe model precision by R^2_{Model}, so as a result

$$\sigma_\varepsilon^2 = \left(1 - R^2_{Model}\right)\sigma_Q^2, \tag{20}$$

where the corresponding values of the parameters are

$$\beta = R \cdot \left(\frac{\sigma_Q}{\sigma_P}\right) = 0.75 \cdot R_{Model} \tag{21}$$

$$\alpha = \mu_Q - \beta \cdot \mu_P = 60 - 0.75 \cdot R_{Model} \cdot 100. \tag{22}$$

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 point rain gauge may not reflect the average rainfall for the basin hence we consider as our prior on β a uniform distribution on the interval (0, 2).

[40] 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 60 cm; 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 $\alpha = -100$ cm. 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 tried to follow the logic GLUE applications adopt to provide a reasonable implementation of that method.

[41] For the purposes of streamflow simulations with both good and poor parameter estimates, the estimate of the flow based on the linear model is taken to be

$$\hat{Q} = \max[0, a + b \cdot P], \tag{23}$$

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

4.2. Confidence and Prediction Intervals for A Linear Regression

[42] Most introductory statistics textbooks provide the derivation of confidence and prediction intervals for the linear regression analysis that would be associated with calibration of the linear watershed model in (19). Substitution of OLS estimates a and b of the regression model parameters into (19) yields a model prediction of Q, denoted $\hat{Q}$, for a particular value P_o of precipitation,

$$\hat{Q} = a + bP_o. \tag{24}$$

The variance of a prediction $\hat{Q}$ describes the likely difference between that prediction and a possible future flow Q_f for a particular precipitation value P_o, and is given by

$$\sigma^2_{\text{Pred}} = E\left[\left(Q_f - \hat{Q}\right)^2\right] = \sigma_\varepsilon^2\left(1 + \frac{1}{n} + \frac{\left(P_o - \overline{P}\right)^2}{\sum_{t=1}^{n}\left(P_t - \overline{P}\right)^2}\right), \tag{25}$$

where we estimate σ_ε^2 by s^2 the classic unbiased estimator of the residual variance. Note that for large sample sizes, the uncertainty in the two model parameter estimates a and b slowly vanishes like $1/n$. Thus σ^2_{Pred} reduces to just s_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.

[43] Similarly if one's interest is in a mean prediction based on the regression $\hat{Q} = a + bP_o$, then the variance of concern will be less than in (25) and is given by

$$E\left[\left(\hat{Q} - a - bP\right)^2\right] = \sigma_\varepsilon^2\left(\frac{1}{n} + \frac{\left(P_o - \overline{P}\right)^2}{\sum_{t=1}^{n}\left(P_t - \overline{P}\right)^2}\right), \tag{26}$$

where we estimate σ_ε^2 by s_ε^2.

[44] In ordinary statistical parlance, equations (25) and (26) would be used to generate prediction and confidence intervals, respectively, given the linear model in (19). On the basis of (25) and (26) and using a Students t distribution with n - 2 degrees of freedom (because s_ε^2 is an estimator), Figure 1 displays 95% prediction intervals for a future observation and 95% confidence intervals for the model mean for different P values. Figure 1 is based on a single sample of length n = 40 generated from (19). Clearly the GLUE methodology should yield similar intervals when applied to this sample with the linear model because equations (25) and (26) are the correct answer. Using a correct Bayesian analysis with a noninformative prior yields essentially the same result, though the Bayesian interpretation would be different [*Zellner*, 1971].

4.3. Uncertainty Intervals Using GLUE

[45] Here we evaluate the ability of GLUE to reproduce the correct prediction and confidence intervals provided by equations (25) and (26), respectively. That exercise will demonstrate the importance of using the Bayesian likelihood function derived in (7)–(10). Our experiment proceeded 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 (20)–(22). (2) Generate a single sample of precipitation P_t, model errors ε_t and

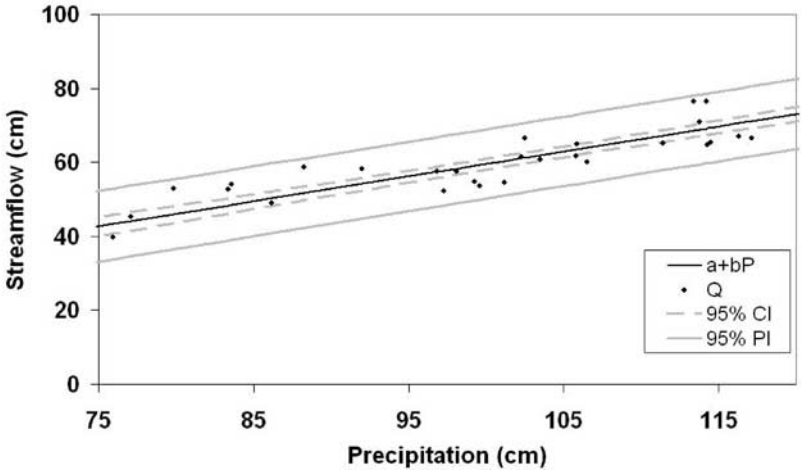

Figure 1. Example of confidence (uncertainty) intervals (CI) for the mean flow associated with each precipitation value and prediction intervals (PI) for an observed flow given each precipitation value for a simple linear regression based on $n = 40$ observations computed using equations (25) and (26) with the Students t distribution with $n - 2$ degrees of freedom to reflect uncertainty in the sample variance.

corresponding streamflow Q_t observations $t = 1,\ldots n$ ($n = 40$) to mimic the problem faced by a hydrologist who typically only has 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 MLEs. (4) For Figure 1, compute the exact 95% confidence intervals for the mean flow associated with precipitation using (26), and the 95% prediction intervals for an observed flow using (25), for a simple linear regression based on the $n = 40$ observation. Confidence intervals and prediction intervals are similarly computed for a specific value P_o (the 90th percentile of P) for comparison with GLUE results described below. (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 $n = 1, \ldots 40$ 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 $i = 1,\ldots,m$. (8) Using the m values of $R^2(\theta_i)$ for $i = 1,\ldots m$ along with n, $R(\hat{Q}_{MLE})$ and s_Q, compute the Bayesian likelihood function (10), 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 nonbehavioral parameter sets. (10) For each set of behavioral parameters compute probabilities $p(a_i, b_i)$, $i = 1,\ldots,m$, using the different likelihood functions. (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. See Figure 2.

5. Results

5.1. Impact of Likelihood Functions on Uncertainty Intervals

[46] This section compares the uncertainty intervals generated by GLUE with different likelihood functions, with confidence intervals based on classical regression theory. Figure 2 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 functions: The likelihood function for normal and independent distributed (NID) model errors in (4), the Inverse Variance (IV) likelihood function in (14) with $N = 1$, and the Nash-Sutcliffe (NS) efficiency index in (15) with $N = 1$ and 30. Some 90% of the generated parameter sets had negative values of $R^2(\theta)$. The inverse variance likelihood function assigned to these parameter sets a probability of over 40%. This was judged to be unreasonable, so for the IV likelihood behavioral thresholds of $R^2(\theta) = 0$ and 50% were adopted resulting in cases IV00 (N = 1) and IV50 (N = 1).

[47] For the uncertainty distribution of the mean flow associated with $P_o = 125.6$ cm, illustrated in Figure 2, the NID likelihood is the correct description of the error distribution, so those results provide the correct posterior distribution for the mean flow. One observes that the posterior distribution generated by the inverse-variance likelihood IV 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 function and the shaping factor N matter.

[48] Figure 3 further illustrates the effect of the shaping factor N. Figure 3 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 (26) 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. With NS, the GLUE values match the correct intervals for N values in the 20–30 range; with IV, the GLUE values would only match the correct result for N

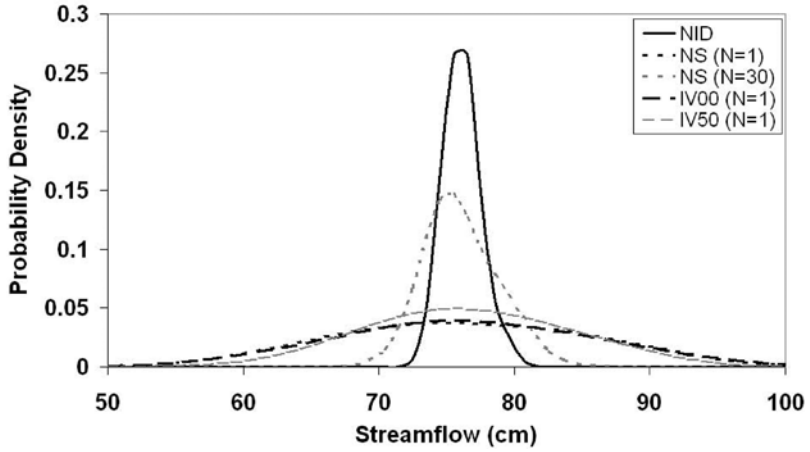

Figure 2. Posterior probability density functions for the mean flow associated with precipitation value P_o = 125.6 cm generated by GLUE with several likelihood functions (n = 40).

greater than 100 for either threshold. For $N < 10$, NS and IV uncertainty intervals diverge and are much too large.

[49] As noted above, sample size is very important in determining the precision of parameter estimators. Figure 4 is the same as Figure 3, except the sample size changes from 10 to 100. While NS with $N = 30$ and $n < 20$ comes close to the correct result, it is far away for $n > 60$. NS and IV results with $N = 1$ are absurd.

[50] A concern with subjective GLUE likelihood measures is their lack of dependence on sample size n. One expects the uncertainty intervals for a regression model to become narrower as more observations are available for calibration [*Mantovan and Todini*, 2006]. As Figure 4 shows, the uncertainty intervals for the mean flow associated with P_o = 125.6 cm generated using GLUE likelihood measures as likelihood functions remain relatively constant regardless of sample size. This makes no sense: more data tell us something and thereby increase the precision of parameter estimators; the problem here is that the NS and IV likelihood measures fail to reflect the value of sample information. *Beven et al.* [2007] have suggested that this incoherence can be resolved by using informal likelihood measures for separate blocks which could be combined with Bayes theorem; whether this would produce the correct result in the end depends upon whether a correct likelihood function is employed for each block, and if blocks are effectively independent so that it is correct to multiply their likelihood functions.

[51] Figure 5 looks at this issue in yet another way by reporting the values of the probabilities assigned to different parameter sets. Each parameter set can be described by the goodness-of-fit value $R^2(\theta)$ computed using equation (10b). Equation (10a) for NID, equation (18) for IV, and $L_{NS} = [R^2(\theta)]^N$ for NS express the three likelihood functions in terms of $R^2(\theta)$. Figure 5 displays the different probabilities that result. IV50 is only employed for $R^2(\theta) > 0.50$. 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.

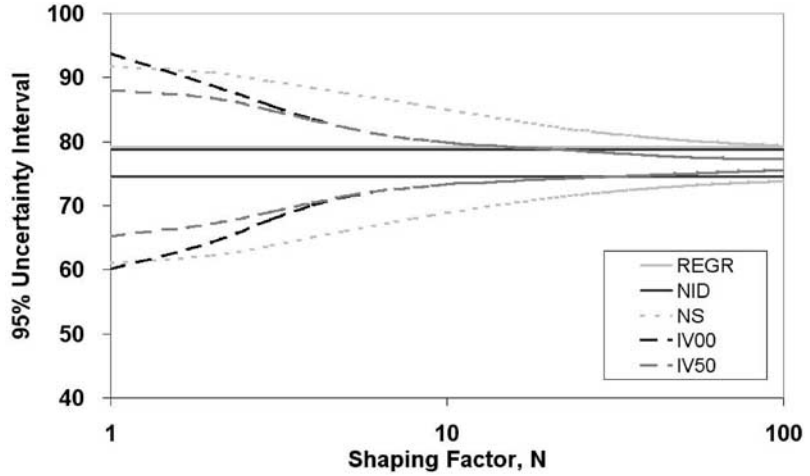

Figure 3. Effect of shaping factor N for NS and IV likelihoods on generated uncertainty intervals for the mean flow associated with precipitation value P_o = 125.6 cm, relative to the confidence intervals computed for the simple linear regression using equation (26) with the Students t distribution (n = 40).

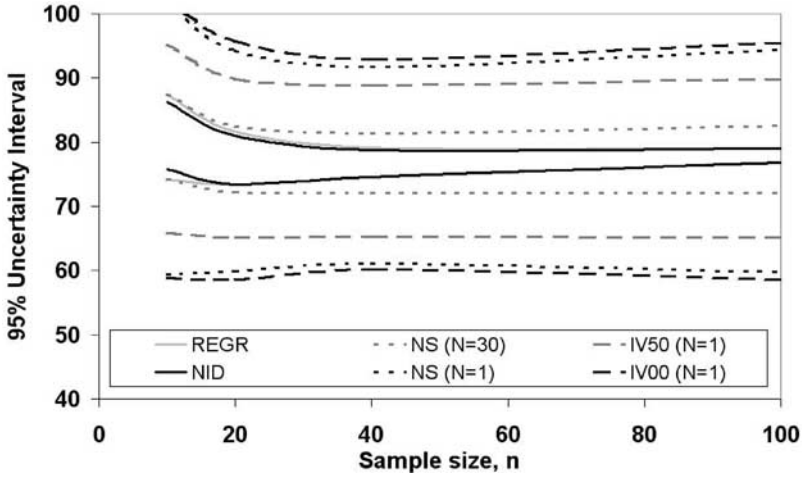

Figure 4. Effect of sample size n on uncertainty intervals for the mean flow associated with precipitation value P_o = 125.6 cm for simple regression using equation (26) with the Students t distribution compared with the corresponding uncertainty intervals obtained using various likelihood functions.

[52] Figure 5 illustrates very important differences among the probabilities generated with different likelihood functions. With 10,000 randomly generated parameter pairs (2 parameters per set), 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 Figure 5. A correct statistical analysis using GLUE with 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 nearly certain that such parameter sets do not represent the true parameter values. However, GLUE with the NS and IV informal likelihood functions find many parameter sets to be plausible when they are actually beyond the realm of credibility. This explains the need in many GLUE analyses to impose a behavioral constraint on parameters sets. No such constraint was needed when applying GLUE with the correct likelihood function for NID data.

5.2. Effect of R^2_{Model}

[53] Figures 1–5 consider the case where the true model had an R^2_{Model} of 0.90. Figure 6 is like Figures 3–4 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 6 describes the effect of the value of R^2_{Model} on 95% uncertainty intervals computed with different likelihood functions. In Figure 6 one sees that if the correct NID likelihood function is adopted, the 95% confidence intervals for the mean flow associated with P_o collapse to a point as R^2_{Model} goes to 1, as is expected from equation (26). If the model is perfect, it only takes a few data points to determine the values of the coefficients exactly. However, that truth is only honored with REGR and NID.

[54] Use of the subjective likelihood functions NS and IV results in intervals that do not collapse to a point as R^2_{Model} goes to 1. The width of their probability intervals

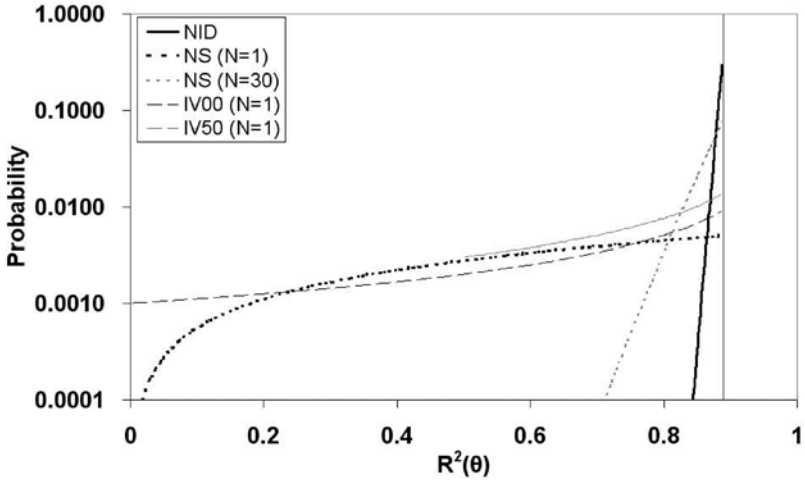

Figure 5. Comparison of the probabilities assigned to different parameter sets employing GLUE likelihood functions as a function of model goodness of fit $R^2(\theta)$, for the proposed parameter vector θ for n = 40 when R_{Model}^2 = 0.90.

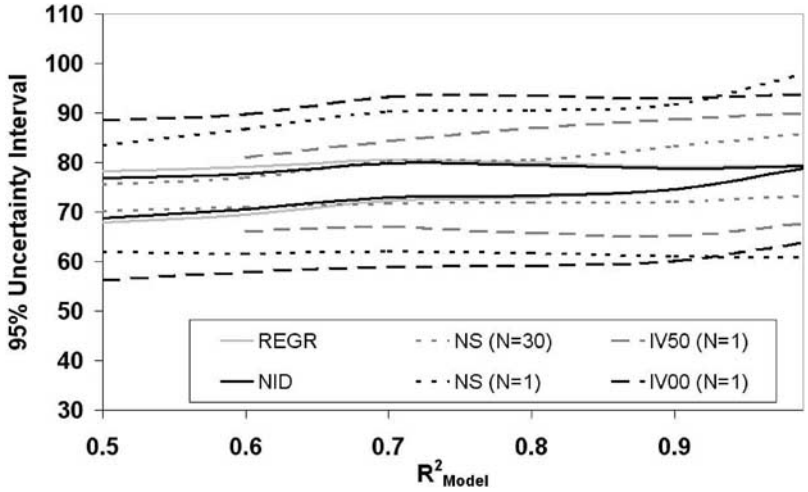

Figure 6. Effect of R_{Model}^2 on 95% uncertainty intervals for the mean flow when P_o = 125.6 cm obtained with different likelihood functions.

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 nonbehavioral threshold which is imposed on the analysis. How can it be true that the precision of the estimated parameters does not depend on the precision of the model?

5.3. A Reappraisal of the Concept of "Equifinality"

[55] The understanding one gains from Figures 5 and 6, and GLUE's acceptance of statistically implausible parameter sets, suggest a reappraisal of the repeated claim for equifinality: "that there are . . .many different parameter sets within a chosen model structure that may be behavioral or acceptable in reproducing the observed behavior of that system" [*Beven and Freer*, 2001, abstract]. While alternative optima and redundancy in models certainly are a reality and a concern, much of the "equifinality" that has been reported is most likely a result of the use of informal likelihood measures that do not distinguish a statistically valid alternative parameter set from just a bad fit. In reference to the dotty plot in Figures 1a–1c of *Beven and Freer* [2001, p. 15] showing likelihoods versus the value of three parameters, the paper observes that, "There are many simulation . . .that, on the basis of the error variance alone, are virtually indistinguishable from one another". Figure 1 of *Beven* [2006a], provides many examples to support his manifesto. What we see in these plots of informal likelihoods versus the values of different parameters are many poor solutions which could have been identified as such, if a statistically realistic likelihood function were employed. Often it only takes one parameter in the set of parameters to have a bad value for a model to perform poorly.

[56] Rather than the performance of the model outputs corresponding to different parameter sets being indistinguishable, we assert that a problem is that GLUE with NS and IV fails to distinguish good solutions with statistically valid parameters from poor model fits resulting from parameters that are statistically invalid. Also of critical importance we would ask that, having created a few thousand alternative behavioral parameter sets by independently and randomly drawing tens of thousands of values for each parameter, why should we believe with complex models that even one truly good set of parameters has been examined to reveal what is possible?

5.4. Prediction Intervals for Future Observations

[57] 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 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 a simple subjective likelihood measure 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 equation 19) that describes the likely difference 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 as illustrated in equation (12). Here we illustrate how this can be done using Monte Carlo simulation.

[58] 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 (19). 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 (25), so we include those prediction intervals as our standard.

[59] 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 param-

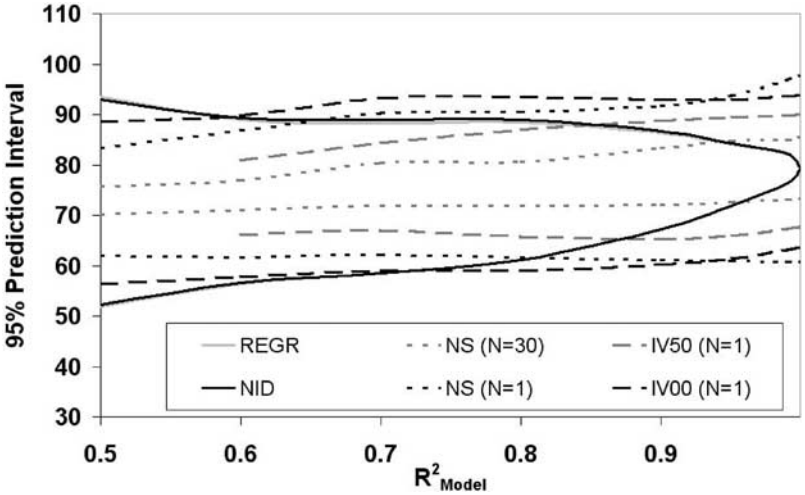

Figure 7. Effect of R_{Model}^2 on 95% prediction intervals for the future flow when P_o = 125.6 cm obtained with different likelihood functions for n = 40.

eters 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 and independent 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.

[60] Combining the 10,000 different possible sets of parameters that provide predictions of Q for each observed P, with 50 normal replicates, yields 500,000 possible values of the Q that go with P_0. These were used to construct the 95% prediction intervals in Figure 7. Figure 7 illustrates the effect of R^2_{Model} on 95% prediction intervals for a future flow Q for P_o = 125.6 cm obtained with different likelihood functions. One can see that the results with the NID likelihood function are almost identical with the correct statistical result obtained with equation (25), denoted REGR. Also shown are the GLUE prediction 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 all by GLUE with the NS and IV likelihood functions. 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.

[61] Table 1 provides another way to represent the results in Figure 7. 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 outside of the 90% prediction intervals for each observation in each record. With 90% prediction intervals and n = 40, on average four observations should fall outside the intervals, E[K] = 4, neglecting the fact that the prediction intervals are based on an analysis of the 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 IV00 (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 $R^2_{Model} > 0.90$.

[62] Figure 7 and Table 1 illustrate 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 and subjective likelihood function will lead to arbitrary intervals without statistical validity, and which fail to include future observations with the target frequency.

5.5. Calibration of Prediction Intervals

[63] GLUE users are resourceful and the suggestion has been made to adjust the behavioral threshold for R^2 (θ), and perhaps the shaping factor N, so that generated uncertainty intervals contain observations across the calibration period with the desired frequency. *Blasone et al.* [2008a, p. 639–640, 646] suggest this "tuning" resolves the subjectivity of the choice of a threshold because one now has an objective

Table 1. Number of Observations K in the Three n = 40 Period Records That Fall Outside the 90% Prediction Intervals for Each Record[a]

	R^2_{Model}		
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

[a]Anticipate $K \approx 4$.

selection criterion for the threshold; however, they then observed that when sufficient parameter sets were retained to achieve the average coverage desired, the accuracy of the median forecast was significantly compromised. Furthermore, they correctly observe that the calibrated prediction intervals misrepresent the real uncertainty over different flow ranges.

[64] Basically this is a terrible idea. By comparing equations (25) and (26) whose results are displayed in Figure 1, one sees that the uncertainty intervals which describe parameter uncertainty in (26) flair at the extremes, and are different than those representing prediction uncertainty which have almost fixed width. Thus if one inflates the parameter uncertainty intervals so they have the correct average coverage when used as prediction intervals, one will over estimate the uncertainty associated with large and small flows, while underestimating the uncertainty associated with average flows. For his groundwater example, *Christensen* [2004, p. 56] observes that a tuned threshold that generates the correct coverage for the head at one location failed to solve the problem elsewhere. And hopefully, no one would use such calibrated results to construct uncertainty intervals for parameters. Clearly if one wants prediction intervals, they should use the calibration-period errors to construct a valid description of prediction errors [*Mantovan et al.*, 2007], including the period-to-period error correlations if such descriptions of prediction uncertainty are to be used in an operational setting.

6. Manifesto or Misguided Epistle?

[65] *Beven* [2006a, p. 18] in his manifesto for GLUE notes with some amazement that other authors have fixated on finding the optimal parameter set which yields the minimum observed forecast error. Instead the GLUE focus appears to be the generation of a large number of solutions that exceed some nominal threshold that defines behavioral solutions. Beven admits the threshold selected is arbitrary, and inspection reveals that even many of the behavioral solutions provide mediocre if not just poor performance. *Beven*'s [2006a] manifesto which advocates the notion of equifinality fails to reflect the intent of most rainfall-runoff studies which is to develop an operational model that allows us to make the best predictions that we can, as well as to describe the precision of those predictions and the precision with which the operationally best parameters can be resolved. It would be naive to believe that all the data are accurate and that we have the correct model structure, or that other mathematical structures could not yield input-output relationships that are operationally indistinguishable across the range of the data and anticipated predictions. Our data have limited precision and our models are very crude approximations of reality, and mathematically such approximations can be expressed in an infinite number of ways.

[66] *Beven* [2006a] provides a long discussion of the problems of model identification and calibration to justify the manifesto of "equifinality". We think that discussion fundamentally misunderstands the points above and the purpose and aims of uncertainty analysis in operational rainfall-runoff modeling. The goals are to find the best model for predicting future events given the observed data that are available for model calibration, to quantify potential errors associated with future predictions, and to quantify watershed model parameter uncertainty to guide model development. GLUE as generally applied to date with informal likelihood functions such as NS and IV fails in all three aspects.

6.1. Model Identifiability

[67] As often applied, GLUE does not identify the best model for use in operational studies. In fact no explicit effort is made to identify the best parameters or to demonstrate that parameters that have been generated include the best that are possible.

6.2. Quantification of Model Prediction Error

[68] As discussed above, GLUE (as generally applied in the past) does not provide a correct uncertainty analysis that yields prediction intervals which are consistent with correct classical or Bayesian statistical theory, and thus they have been observed to perform poorly. Most applications of GLUE have only included the influence of parameter uncertainty and have neglected the importance of model errors as well as measurement errors associated with both inputs and outputs, all of which often dominate the difference between model predictions and observed values. It is important to acknowledge that even if the values of the parameters are known, most models would still exhibit significant forecast errors [see *Kuczera et al.*, 2006], and GLUE, as generally implemented, fails to recognize this reality.

6.3. Quantification of Model Parameter Uncertainty

[69] Watershed modeling studies need to have descriptions of the statistical precision of watershed model parameter estimators. These are needed to determine the appropriate complexity of a model given one's ability to resolve parameters describing different hydrologic processes. But as documented here, GLUE as often applied does not correctly quantify the precision of the model parameters because it does not use a correct likelihood function. Its analysis is not statistically valid and the assumed relationship between data and parameter precision is grossly misrepresented.

6.4. What Do We Have?

[70] Overall, GLUE fails to provide the discriminating statistical analysis needed to support the model development processes. Thus we find the manifesto articulated by *Beven* [2006a] based upon GLUE analyses with subjective likelihood measures to be a misguided and unwise call for a modeling revolution.

7. Recommendations for Improving GLUE

[71] In practice, environmental simulation models are far more complex than the simple linear model with NID errors used in our analysis. Thus error models should be adopted that address nonnormal, heteroscedastic and serially correlated residuals as well as other complexities. We believe that improvements in our ability to characterize complex model error structures provide a fruitful and richly rewarding path for improving GLUE and other procedures used for uncertainty analysis and/or calibration.

[72] A common challenge in hydrologic modeling involves both the heteroscedasticity and nonnormality of model residuals, owing to a variety of different processes

whose importance varies over time with groundwater and soil moisture values, and the distribution of rainfall. Heteroscedasticity and nonnormality are often related and sometimes, a single transformation of the residuals solves both of these problems simultaneously. For example, *Romanowicz et al.* [1994] used a logarithmic transformation of the residuals to deal with both nonnormality and heteroscedasticity. Square root and other power transformations have also been employed by *Kuczera and Parent* [1998] and by *Sorooshian and Dracup* [1980] who developed a heteroscedastic maximum likelihood estimator of the model parameters. *Schaefli et al.* [2007] recommends the use of a mixture distribution. Approaches to ensure homoscedasticity of the residuals include (1) the use of a logarithmic or other transformation of the residuals, (2) modeling portions of the hydrologic record separately (i.e., high flows, non-snow periods) and (3) use of weighted least squares where weights on the residuals are inversely proportional to the standard deviation of the residuals at a given time and flow level.

[73] Another common concern is that residuals exhibit temporal persistence. Again 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, or (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]. 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.

[74] Another critical challenge associated with uncertainty analysis for highly nonlinear models with multidimensional parameter spaces is to generate truly good sets of parameters, as illustrated by papers in the volume edited by *Duan et al.* [2003]. Attractive algorithms are presented by *Kavetski et al.* [2002], *Marshall et al.* [2004], *Vrugt et al.* [2005], *Mugunthan and Shoemaker* [2006], *Tolson and Shoemaker* [2008] and *Blasone et al.* [2008a].

[75] Most environmental 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. Much of the observed correlation in model predictions is due to errors in estimated model inputs such as rainfall working their way though a watershed which has memory. *Thiemann et al.* [2001], *Kavetski et al.* [2002, 2006a, 2006b], *Kuczera et al.* [2006], *Vrugt et al.* [2005], *Moradkhani et al.* [2005a, 2005b], *Clark and Vrugt* [2006], *Huard and Mailhot* [2006] and *Ajami et al.* [2007] discuss very 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.

[76] Many researchers believe, and claims have been made that the basic GLUE methodology with the traditional likelihood measures, equations (14)–(16) [*Beven and Freer*, 2001, Table 1 equation (1a)–(1c)], addresses concerns with input-data errors, model errors, and corresponding correlations in the observed calibration errors. (For example, see quote in section 5.4). This is fundamentally not true.

[77] All three of these likelihood measures are monotonic functions of s_ε^2, the residual mean square error for a given parameter set over the calibration period. 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 functions will do nothing to fundamentally reflect in the analysis nonnormal and heteroscedastic errors, or 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. No adjustment is made to individual errors to account for differences in variances (heteroscedasticity), distribution (nonnormality) or autocorrelation (persistence).

[78] For GLUE studies to be effective at identifying reasonable parameter sets and providing a valid sensitivity analysis, given input-data errors, model errors, and output-variable measurement errors, resulting in correlated heteroscedastic residual errors, effort needs to be invested to develop goodness-of-fit criteria that reflect these error distributions, just as it is required to develop realistic statistical likelihood functions. In fact, it is the same challenge, because the goodness-of-fit criterion has to realistically measure how well a model fits the data in the dimensions of concern and, to describe if differences between observed data and model predictions can reasonably be explained by chance.

8. Conclusions

[79] *Christensen* [2004], *Montanari* [2005], *Mantovan and Todini* [2006], and this study document that the widely accepted GLUE approach to describing the impact of parameter uncertainty for environmental models produces prediction intervals which fail to agree with results based on proven methods of uncertainty analysis. Here we documented this phenomenon using a simple linear regression model. That choice enables us to compare uncertainty intervals generated by GLUE using several recommended informal "likelihood" measures with exact analytic uncertainty intervals available in most introductory statistics textbooks, which are known in repeated trials to generate uncertainty and prediction intervals that contain selected parameters with the anticipated frequency.

[80] We discussed why the choice of a likelihood function is critical and needs to be a reasonable description of the distribution of the model errors for the statistical inference and resulting uncertainty and prediction intervals to be valid. This warning is equally valid if GLUE with an informal likelihood measure is to be used for model calibration and sensitivity analysis: one needs goodness-

of-fit criteria that reflect the nonnormality, heteroscedasticity, and correlation among the residual errors if one is to evaluate the reasonableness of different model-parameter sets: the traditional subjective GLUE likelihood functions fail to do that.

[81] 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 we employed the likelihood function for NID errors, a behavioral threshold was unnecessary and 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-based/informal likelihood functions recommended by *Beven and Binley* [1992] and others are used with GLUE.

[82] *Beven and Binley* [1992], *Beven and Freer* [2001], and others have suggested that the choice of the likelihood measure used in GLUE is 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, p. 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.

[83] Although the experiments performed in this paper assume NID model errors, these assumptions are not a necessary assumption to employ the general form of the likelihood function introduced in (3). Equation (3) is the general form of the likelihood function which should be used, regardless of the assumed structure of the model errors. The general likelihood function in (3) may be adopted for situations when one has nonnormal, heteroscedastic and autocorrelated model errors. The fundamental message here is that for every assumed model structure, there is a corresponding likelihood function appropriate for use with the GLUE methodology which will lead to uncertainty analyses which remain consistent with both classical and Bayesian statistics.

[84] Earlier studies by *Christensen* [2004], *Montanari* [2005] and *Mantovan and Todini* [2006] have illustrated and discussed problems with GLUE analyses. We document how the GLUE methodology can be applied to generate realistic uncertainty and prediction intervals using model simulations which are consistent with Bayesian and classical statistical theory. 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 the additional needed extensions are made to describe prediction uncertainty, then the GLUE methodology should be a valuable tool for both model calibration and for estimating model uncertainty with the advantages that have made it so popular. If an arbitrary likelihood measure 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.

[85] **Acknowledgments.** An early version of this research was presented at the 2005 Spring AGU meeting in New Orleans for which Rebecca Batchelder won an ''Outstanding Student Paper Award''. The authors are indebted to Shafiqul Islam, Linfield Brown, Dennis Lettenmaier, Keith Beven, Thomas E. Adams, Richard P. Hooper, Charles N. Kroll, Ezio Todini, Alberto Montanari, David Huard, Steen Christensen, and four anonymous reviewers for their helpful suggestions on early versions of the manuscript. We appreciate some support provided by the U.S. Department of Agriculture, through the grant: Integrating data and models from the Cannonsville NY Watershed, USDA/CSREES 2005-51130-03338.

References

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

Beven, K. J. (1989), Interflow, in *Unsaturated Flow in Hydrological Modeling*, edited by H. J. Morel-Seytoux, pp. 191–219, D. Reidel, Dordrecht, Netherlands.

Beven, K. J. (2001), *Rainfall-Runoff Modelling–The Primer*, 360 pp., John Wiley, Hoboken, N. J.

Beven, K. J. (2006a), A manifesto for the equifinality thesis, *J. Hydrol.*, *320*(1–2), 18–36.

Beven, K. J. (2006b), On undermining the science?, *Hydrol. Processes*, *20*, 3141–3146, doi:10.1002/hyp.6396.

Beven, K. J., and A. M. Binley (1992), The future of distributed models: Model calibration and uncertainty prediction, *Hydrol. Processes*, *6*, 279–298, doi:10.1002/hyp.3360060305.

Beven, K. J., and J. Freer (2001), Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, *J. Hydrol.*, *249*, 11–29, doi:10.1016/S0022-1694(01)00421-8.

Beven, K., J. Freer, B. Hankin, and K. Schulz (2000), The use of generalized likelihood measures for uncertainty estimation in high-order models of environmental systems, in *Nonlinear and Nonstationary Signal Processing*, edited by W. J. Fitzgerald et al., pp. 115–151, Cambridge Univ. Press, Cambridge, U. K.

Beven, K, P. Smith, and J. Freer (2007), Comment on ''Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology'', by P. Mantovan and E. Todini, *J. Hydrol.*, *338*, 315–318, doi:10.1016/j.jhydrol.2007.02.023.

Beven, K., P. Smith, and J. Freer (2008), So just why would a modeler choose to be incoherent?, *J. Hydrol.*, *354*, 15–32, doi:10.1016/j.hydrol.2008.02.007.

Bianchini, G., A. Cortes, T. Margalef, and E. Luque (2006), Improved prediction methods for wildfires using high performance computing: A comparison, *Lect. Notes Comput. Sci.*, *3991*, 539–546, doi:10.1007/11758501_73.

Bickel, P. J., and K. A. Doksum (2001), *Mathematical Statistics: Basic Ideas and Selected Topics*, vol. 1, 2nd ed., Prentice Hall, Upper Saddle River, N. J.

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

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

Box, G. E. P., G. M. Jenkins, and G. C. Reinsel (1994), *Time Series Analysis: Forecasting and Control*, 3rd ed., Prentice Hall, Upper Saddle River, N. J.

Carrera, J., A. Alcolea, J. Hidalgo, and L. J. Slooten (2005), The inverse problem in hydrology, *J. Hydrol.*, *13*, 206–222.

Christensen, S. (2004), A synthetic groundwater modeling study of the accuracy of GLUE uncertainty intervals, *Nord. Hydrol.*, *35*(1), 45–59.

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

Duan, D., S. Sorooshian, and R. P. Ibbitt (1988), A maximum likelihood criterion for use with data collected at unequal time intervals, *Water Resour. Res.*, *24*(7), 1163–1173, doi:10.1029/WR024i007p01163.

Duan, D., H. V. Gupta, S. Sorooshian, A. N. Rousseau, and R. Turcotte (Eds.) (2003), *Calibration of Watershed Models*, *Water Sci. Appl.*, vol. 6, edited by Q. Duan et al., 653 pp., AGU, Washington, D. C.

Engeland, K., and L. Gottschalk (2002), Bayesian estimation of parameters in a regional hydrological model, *Hydrol. Earth Syst. Sci.*, *6*, 883–898.

Freer, J., J. K. Beven, and B. Ambroise (1996), 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.

Hellweger, F. L., and U. Lall (2004), Modeling the effect of algal dynamics on arsenic speciation in Lake Biwa, *Environ. Sci. Technol.*, *38*, 6716–6723, doi:10.1021/es049660k.

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

International Association of Hydrological Sciences (2003), PUB science and implementation plan, version 4, report, Gentbrugge, Belgium.

Jia, Y. B., and T. B. Culver (2008), Uncertainty analysis for watershed modeling using generalized likelihood uncertainty estimation with multiple calibration measures, *J. Water Resour. Plann. Manage*, *134*(2), doi:10.1061/(ASCE)0733-9496(2008)134:2(97).

Kavetski, D. N., S. W. Franks, and G. Kuczera (2002) Confronting input uncertainty in environmental modelling, in *Calibration of Watershed Models*, *Water Sci. Appl.*, vol. 6, edited by Q. Duan et al., pp. 49–68, AGU, Washington, D. C.

Kavetski, D. N., G. Kuczera, and S. W. Franks (2006a), Bayesian analysis of input uncertainty in hydrological modeling: Theory, I., *Water Resour. Res.*, *42*, W03407, doi:10.1029/2005WR004368.

Kavetski, D. N., G. Kuczera, and S. W. Franks (2006b), Bayesian analysis of input uncertainty in hydrological modeling: IApplication, I., *Water Resour. Res.*, *42*, W03408, doi:10.1029/2005WR004376.

Kendall, M. A., and A. Stuart (1961). *The Advanced Theory of Statistics*, vol. 2, *Inference and Relationship*, pp. 17.13–17.29, 18.22–18.34, Hafner, New York.

Kinner, D. A., and R. F. Stallard (2004), Identifying storm flow pathways in a rainforest catchment using hydrological and geochemical modelling, *Hydrol. Processes*, *18*, 2851–2875, doi:10.1002/hyp.1498.

Kuczera, G., and E. Parent (1998), Monte Carlo assessment of parameter uncertainty in conceptual catchment models: The Metropolis algorithm, *J. Hydrol.*, *211*, 69–85, doi:10.1016/S0022-1694(98)00198-X.

Kuczera, G., D. Kavetski, S. Franks, and M. Thyer (2006), Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterizing model error using storm-dependent parameters, *J. Hydrol.*, *331*, 161–177, doi:10.1016/j.jhydrol.2006.05.010.

Liang, S., R. C. Spear, and E. Seto, et al. (2005), A multi-group model of *Schistosoma japonicum* transmission dynamics and control: Model calibration and control prediction, *Trop. Med. Int. Health*, *10*, 263–278, doi:10.1111/j.1365-3156.2005.01386.x.

Mantovan, P., and E. Todini (2006), Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology, *J. Hydrol.*, *330*, 368–381, doi:10.1016/j.jhydrol.2006.04.046.

Mantovan, P., E. Todini, and M. L. V. Martina (2007), Reply to comment by Keith Beven, Paul Smith, and Jim Freer on "Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology", *J. Hydrol.*, *338*, 319–324, doi:10.1016/j.jhydrol.2007.02.029.

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

Mertens, J., et al. (2004), Including prior information in the estimation of effective soil parameters in unsaturated zone modeling, *J. Hydrol.*, *294*, 251–269, doi:10.1016/j.jhydrol.2004.02.011.

Mo, X. G., and K. J. Beven (2004), Multi-objective parameter conditioning of a three-source wheat canopy model, *Agric. For. Meteorol.*, *122*, 39–63, doi:10.1016/j.agrformet.2003.09.009.

Montanari, A. (2005), Large sample behaviors of the generalized likelihood uncertainty estimation (GLUE) in assessing the uncertainty of rainfall-runoff simulations, *Water Resour. Res.*, *41*, W08406, doi:10.1029/2004WR003826.

Montanari, A. (2007), What do we mean by uncertainty? The need for a consistent wording about uncertainty assessment in hydrology, *Hydrol. Processes*, *21*, 841–845, doi:10.1002/hyp.6623.

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

Moradkhani, H., S. Sorooshian, H. V. Gupta, and P. Houser (2005b), Dual state-parameter estimation of hydrological models using ensemble Kalman filter, *Adv. Water Resour.*, *28*, 135–147, doi:10.1016/j.advwatres.2004.09.002.

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

Nash, J., and J. Sutcliffe (1970), River flow forecasting through conceptual models, 1. A discussion of principles, *J. Hydrol.*, *10*, 282–290, doi:10.1016/0022-1694(70)90255-6.

Page, T., et al. (2004), Uncertainty in modelled estimates of acid deposition across Wales: A GLUE approach, *Atmos. Environ.*, *38*, 2079–2090, doi:10.1016/j.atmosenv.2004.01.029..

Pappenberger, F., et al. (2005), Uncertainty in the calibration of effective roughness parameters in HEC-RAS using inundation and downstream level observations, *J. Hydrol.*, *302*, 46–69, doi:10.1016/j.jhydrol.2004.06.036.

Pinol, J., K. J. Beven, and D. Viegas (2005), Modelling the effect of fire-exclusion and prescribed fire on wildfire size in Mediterranean ecosystems, *Ecol. Modell.*, *183*, 397–409, doi:10.1016/j.ecolmodel.2004.09.001.

Ratto, M., S. Tarantola, and A. Saltelli (2001), Sensitivity analysis in model calibration: GSA-GLUE approach, *Comput. Phys. Commun.*, *136*, 212–224, doi:10.1016/S0010-4655(01)00159-X.

Romanowicz, R., and K. J. Beven (1998), Dynamic real-time prediction of flood inundation probabilities, *Hydrol. Sci. J.*, *43*, 181–196.

Romanowicz, R., K. J. Beven, and J. Tawn (1994), Evaluation of predictive uncertainty in non-linear hydrological models using a Bayesian approach, in *Statistics for the Environment II: Water Related Issues*, edited by V. Barnett and K. F. Turkman, pp. 297–317, John Wiley, Hoboken, N. J.

Romanowicz, R., K. J. Beven, and J. Tawn (1996). Bayesian calibration of flood inundation models, in *Floodplain Processes*, edited by M. G. Anderson, D. E. Walling, and P. D. Bates, pp. 333–360, John Wiley, Chichester, U. K.

Ruessink, B. G. (2005), Predictive uncertainty of a nearshore bed evolution model, *Cont. Shelf Res.*, *25*(9), 1053–1069, doi:10.1016/j.csr.2004.12.007.

Salas, J. D., O. G. Sveinsson, W. L. Lane, and D. K. Frevert (2006), Stochastic streamflow simulation using SAMS-2003, *J. Irrig. Drain. Engr.*, *132*(2), 112–122.

Schaefli, B., D. B. Talamba, and A. Busy (2007), Quantifying hydrological modeling errors through a mixture of normal distributions, *J. Hydrol*, *332*, 303–315, doi:10.1016/j.jhydrol.2006.07.005.

Smith, R. M. S., D. J. Evans, and H. S. Wheater (2005), Evaluation of two hybrid metric-conceptual models, for simulating phosphorus transfer from agricultural land in the river enborne, a lowland UK catchment, *J. Hydrol.*, *304*, 366–380, doi:10.1016/j.jhydrol.2004.07.046.

Sorooshian, S., and J. Dracup (1980), Stochastic parameter estimation procedures for hydrologic rainfall-runoff models: Correlated and heteroscedastic error cases, *Water Resour. Res.*, *16*(2), 430–442, doi:10.1029/WR016i002p00430.

Spear, R. C., and G. M. Hornberger (1980), Eutrophication in Peel Inlet. II. Identification of critical uncertainties via generalized sensitivity analysis, *Water Res.*, *14*, 43–49, doi:10.1016/0043-1354(80)90040-8.

Thiemann, M., M. Trosset, H. Gupta, and S. Sorooshian (2001), Bayesian recursive parameter estimation for hydrological models, *Water Resour. Res.*, *37*(10), 2521–2535, doi:10.1029/2000WR900405.

Tolson, B. A., and C. A. Shoemaker (2007), Dynamically dimensioned search algorithm for computationally efficient watershed model calibration, *Water Resour. Res.*, *43*, W01413, doi:10.1029/2005WR004723.

Tolson, B. A., and C. A. Shoemaker (2008), Efficient prediction uncertainty approximation in the calibration of environmental simulation models, *Water Resour. Res.*, *44*, W04411, doi:10.1029/2007WR005869.

Tremblay, M., and D. Wallach (2004), Comparison of parameter estimation methods for crop models, *Agronomie*, *24*, 351–365, doi:10.1051/agro:2004033.

Uhlenbrook, S., and A. Sieber (2005), On the value of experimental data to reduce the prediction uncertainty of a process-oriented catchment model, *Environ. Model. Softw.*, *20*, 19–32, doi:10.1016/j.envsoft.2003.12.006.

Vick, S. G. (2002), *Degrees of Belief*, ASCE Press, Reston, Va.

Vrugt, J. A., C. G. H. Diks, H. V. Gupta, W. Bouten, and J. M. Verstraten (2005), 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.

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

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, John Wiley, Hoboken, N. J.

R. Batchelder and R. M. Vogel, Department of Civil and Environmental Engineering, Tufts University, Medford, MA 02155, USA. (rbatch@alumni.tufts.edu; richard.vogel@tufts.edu)

S. U. Lee and J. R. Stedinger, School of Civil and Environmental Engineering, Cornell University, Ithaca, NY 14853, USA. (sl673@cornell.edu; jrs5@cornell.edu)