Large sample behaviors of the generalized likelihood uncertainty estimation (GLUE) in assessing the uncertainty of rainfall-runoff simulations

Alberto Montanari

Faculty of Engineering, University of Bologna, Bologna, Italy

Received 19 November 2004; revised 24 April 2005; accepted 10 May 2005; published 9 August 2005.

[1] Several methods have been recently proposed for quantifying the uncertainty of hydrological models. These techniques are based upon different hypotheses, are diverse in nature, and produce outputs that can significantly differ in some cases. One of the favored methods for uncertainty assessment in rainfall-runoff modeling is the generalized likelihood uncertainty estimation (GLUE). However, some fundamental questions related to its application remain unresolved. One such question is that GLUE relies on some explicit and implicit assumptions, and it is not fully clear how these may affect the uncertainty estimation when referring to large samples of data. The purpose of this study is to address this issue by assessing how GLUE performs in detecting uncertainty in the simulation of long series of synthetic river flows. The study aims to (1) discuss the hypotheses underlying GLUE and derive indications about their effects on the uncertainty estimation, and (2) compare the GLUE prediction limits with a large sample of data that is to be simulated in the presence of known sources of uncertainty. The analysis shows that the prediction limits provided by GLUE do not necessarily include a percentage close to their confidence level of the observed data. In fact, in all the experiments, GLUE underestimates the total uncertainty of the simulation provided by the hydrological model.

Citation: 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.

1. Introduction

[2] The scientific literature has recently proposed numerous contributions about uncertainty assessment in hydrological modeling. For instance, *Beven* [2002a, 2005], *Blazkova and Beven* [2002, 2004], *Krzysztofowicz* [2002], *Krzysztofowicz and Maranzano* [2004], *Vrugt et al.* [2003], *Pappenberger et al.* [2005], and *Montanari and Brath* [2004a, 2004b] are among the most recent studies. Uncertainty assessment is also one of the main goals of the Prediction in Ungauged Basins (PUB) initiative promoted by the International Association of Hydrological Sciences. This intense scientific activity about the uncertainty analysis has resulted in the proposal of many different approaches for quantifying the reliability of hydrological models. However, the results of these various techniques are conditioned by the assumptions underlying each method.

[3] One of the most known and used uncertainty assessment methods in hydrological modeling is the generalized likelihood uncertainty estimation (GLUE), proposed by *Beven and Binley* [1992], that is briefly described in section 2 of the present article. The reasons for the popularity of GLUE are manifold. First, it has been already used in numerous real-world applications [see, e.g., *Freer et al.*, 1996; *Cameron et al.*, 1999, 2000; *Blazkova and Beven*, 2002, 2004; *Aronica et al.*, 2002]. Second, GLUE can in principle be applied to ungauged basins. Third, GLUE is easy to use and can account for all causes of uncertainty in hydrological modeling, either explicitly or implicitly. Finally, GLUE gives the possibility to consider different competing modeling approaches that are individually evaluated. This includes the possible rejection of some of them as nonbehavioral (see section 2). The acceptance of the existence of multiple behavioral parameter sets has been called equifinality [*Beven*, 1993]. It constitutes the basis of the alternative blueprint for hydrological modeling recently proposed by *Beven* [2001, 2002a, 2002b, 2005].

[4] However, a number of questions related to the application of GLUE still remain unresolved. These interrogatives are mainly originated by the necessity to make some subjective decisions when applying GLUE, and it is not fully clear how the resulting prediction limits are conditioned by these assumptions.

[5] The purpose of this article is to present the results of a series of experiments, described in section 3, in which GLUE is applied for assessing the uncertainty in the simulation of synthetic river flow data. In each experiment, an attempt to assess the effect on the uncertainty estimation of a given assumption within GLUE is carried out. In detail, hourly river flows are preliminarily generated (synthetic river flows) by using a first rainfall-runoff model that is supposed to represent reality. A second rainfall-runoff model of reduced complexity is then used to approximate the simulation provided by the first model. As a consequence, uncertainty in the second simulation is given both by the model structural uncertainty, because the second model is of reduced complexity, and by the parameter

0043-1397/05/2004WR003826$09.00

uncertainty, because the presence of model structural uncertainty may induce identifiability problems with regard to model parameters. Afterward, GLUE is applied to assess the uncertainty of the output of the second model. Hence, with this approach, it is possible to assess to what extent the parameter set may compensate for model structure in practical applications. It is worth noting that the synthetic variables are, of course, not affected by measurement errors or other kinds of uncertainty. In order to inspect how GLUE performs in the presence of multiple sources of error, therefore emulating a real-world application, the uncertainty assessment is repeated after having corrupted some of the input (rainfall) and output (river flow) variables of the second rainfall-runoff model (corrupted data analysis). Finally, the use within GLUE of the likelihood measure based on the *Whittle* [1953] approximation to the Gaussian maximum likelihood function is proposed and discussed.

[6] The purposes of the experiments are twofold: (1) to discuss the hypotheses underlying GLUE and to derive indications about their effects on the uncertainty estimation, and (2) to apply GLUE in a context characterized by the availability of a large sample of data in order to be able to compare the GLUE output with what one would expect from standard statistical inference (i.e., about 95% of the data should lie within the 95% confidence limits).

[7] It is important to remark that even in the ideal context this study refers to, one should not expect that a specified proportion of synthetic observations will lie within the GLUE prediction limits [*Beven and Freer*, 2001]. This is not the primary aim within GLUE, whose main purpose is to assess the uncertainty induced by the presence of (possible) competing modeling solutions (see section 5 for additional comments about this issue on the basis of the results obtained in the experiments). However, the comparison described in point 2 listed above may help develop a better understanding of the key features of the GLUE output.

[8] Finally, section 5 of this article outlines some possible conclusions that may give useful indications for evaluating the reliability of hydrological simulation studies in the presence of uncertainty and equifinality.

2. Brief Description of GLUE

[9] GLUE rejects the concept of an optimum model and parameter set and assumes that prior to input of data into a model, all model structures and parameter sets have an equal likelihood of being acceptable. Once different candidate models are identified, GLUE is performed by first identifying for each model the parameters which most affect the output. Then, a high number of parameter sets is generated via uniform sampling, or incorporating prior knowledge about the distribution of parameters. The hydrological models are then run for each of the sets and the model output is compared to a record of observed data (e.g., observed hydrographs or annual maximum peak flows [see *Cameron et al.*, 1999]). The performance of each trial is assessed through likelihood measures. As examples, *Cameron et al.* [1999] used the *Nash and Sutcliffe* [1970] efficiency to evaluate the likelihood of the simulation of a continuous hydrograph, and *Beven and Freer* [2001] present a list of likelihood measures that can be used in rainfall-runoff modeling. Performance evaluation includes rejecting some models and/or parameter sets as nonbehavioral. It is also possible that all models are rejected in a way that a statistical method would not allow [*Freer et al.*, 2003].

[10] All models and their corresponding parameter sets that provide a likelihood measure that reaches a minimum threshold are retained. The likelihood weighted uncertainty bounds can be calculated using the following standard procedure [*Freer et al.*, 1996]. Let us suppose that the river flow at time t is simulated by n alternative modeling solutions, that provide the simulations $Q^i_{\mathrm{sim}}(t)$, $i = 1, \ldots n$, each one characterized by its own likelihood measure. First, the calculated likelihoods are rescaled to produce a cumulative sum of 1.0. Therefore one associates to each $Q^i_{\mathrm{sim}}(t)$ its rescaled likelihood weight. The $Q^i_{\mathrm{sim}}(t)$ are subsequently ranked in ascending order and a probability of not exceedance is subsequently assigned to each of them, that is equal to the cumulated sum of the rescaled likelihood weights up to the considered $Q^i_{\mathrm{sim}}(t)$. Therefore a cumulative distribution function of simulated discharges is then constructed, with the highest $Q^i_{\mathrm{sim}}(t)$ associated to a probability of not exceedance equal to 1. This allows uncertainty bounds corresponding to an assigned confidence level to be derived, in addition to a median simulation. There is an implicit assumption in GLUE that the errors in the prediction will be similar to those in the evaluation period [*Beven*, 2005].

[11] GLUE requires the user to take some subjective decisions by thinking about all the different sources of uncertainty. By considering different parameter sets, one is in fact explicitly evaluating the effects of parameter uncertainty. By considering either different model input and output or different model structures, one is explicitly considering the effect of observable uncertainty and model structural uncertainty, respectively. It is important to note that the effects of the different sources of approximation cannot be evaluated separately in an additive way. For instance, parameter uncertainty is strictly related to both model structural uncertainty and observable uncertainty because either an imperfect model or an imperfect input may induce identifiability problems, and therefore possible equifinality in the estimation of model parameters. Thus one should not expect to always be able to treat each source of uncertainty individually with GLUE but, rather, to implicitly deal with different causes of approximation.

[12] A key issue when applying GLUE therefore is to decide which sources of uncertainty should be treated explicitly after identifying the purposes of the analysis, the hydrological model, and the expected level of uncertainty in model input, output, and structure.

[13] From the description given above, one can see that GLUE prediction limits are always conditional on some subjective decisions. For instance, when different modeling approaches and parameter sets are considered, it is still an open question how to identify nonbehavioral solutions (see *Beven* [2005] for a comprehensive discussion about this problem, that will not be dealt with in this article). Moreover, some sources of uncertainty are accounted for by GLUE implicitly, and therefore it might be not straightforward to decide which uncertainty sources are to be explicitly taken into account. Finally, since uncertainty is estimated in GLUE by weighting multiple behavioral solutions, it is not yet clear how the derived prediction

limits behave with respect to the confidence limits estimated by using statistical inference (when applicable).

[14] Initial answers to the above questions might be found by comparing the output of GLUE with a large sample of the observed variables to be simulated. However, since the observed data are often available with a limited sample size, this is rarely, if ever, possible in real-world applications. Moreover, real-world data, even if measured, are affected by uncertainties that cannot be easily quantified in some cases. Therefore it is not clear to what extent the indications derived by comparing the observed variables with the GLUE output will be reliable.

[15] In this study, the *Nash and Sutcliffe* [1970] efficiency in the simulation of the hourly river flow data was used as the main performance measure from which model likelihood values were derived. The Nash efficiency has been widely used in the past for both model optimization and GLUE studies. It is essentially a transformed version of the sum of squared errors, and therefore it is not an unbiased estimator when the rainfall-runoff model errors are correlated or heteroscedastic, which is frequently the case in practical applications. Therefore the parameter estimates may result biased for the presence of correlation (for instance, due to timing errors) and heteroscedasticity in residuals. This latter heteroscedasticity translates into emphasis given to large errors that tend to occur at higher flows.

[16] In the fifth experiment (see section 3.10), the effect of using different likelihood measures was inspected.

3. Description of the Experiments

[17] The experiments consist of applying GLUE for assessing the uncertainty in the river flow data simulated by the HYMOD rainfall-runoff model that is described below. A first series of experiments (the five experiments described in sections 3.6–3.10) was carried out by using as input to HYMOD synthetic rainfall and river flow data that were generated as described in sections 3.2 and 3.3 (uncorrupted data analysis). These synthetic data are not affected by measurement errors or other kinds of uncertainty. Afterward, in order to emulate a situation were the observed input and output variables used to calibrate the rainfall-runoff model are affected by uncertainty (observable uncertainty), as it is always the case in real-world applications, a second series of experiments was carried out by using corrupted hourly synthetic rainfall and river flow data. The data corruption was carried out as described in section 3.4 (corrupted data analysis).

3.1. The River Basin

[18] The experiments refer to the Secchia River basin that is located in northern Italy. The Secchia River flows northward across the Apennine Mountains and is a right tributary to the Po River. The contributing area is 1214 km^2 at the Bacchello Bridge river cross section that is located about 62 km upstream of the confluence in the Po River. The maximum altitude is 2121 m above sea level (a.s.l.) at Mount Cusna. The main stream length up to Bacchello Bridge is about 98 km, and the basin concentration time is about 12 hours. The mean annual rainfall depth ranges between 700 and more than 2000 mm/yr over the basin area. The maximum peak discharge observed at Bacchello Bridge in the period 1923–1981 was 823 m^3/s (20 April 1960).

3.2. Generation of Synthetic Rainfall and Temperature Data

[19] The synthetic rainfall data were generated by using the generalized multivariate Neyman-Scott rectangular pulses model [*Cowpertwait*, 1995]. This stochastic model represents the total rainfall intensity at time t as the sum of the intensities given by a random sequence of rain cells active at time t.

[20] In detail, the model represents the storm origins as occurrences of a Poisson process with rate λ and with the arrival times being the same at every point in the catchment. Each storm origin generates a random number of circular rain cells according to a Poisson process with rate ν. The spatial position of these cells is given by a two-dimensional Poisson process with rate δ, and their radius is an independent exponential random variable with parameter γ. For the starting time of a rain cell, the waiting time after a storm origin is an independent exponential random variable with parameter β. Each rain cell has a random duration and a random intensity, where the intensity itself is held constant over the circular area covered by the cell and throughout the cell duration. Therefore a rectangular pulse of rain is associated with each rain cell. The cell intensity and duration are both distributed exponentially with parameters μ and η.

[21] The total rainfall intensity at an arbitrary time t at a point m is the summation of the intensities of all cells active at time t that overlap point m. This total rainfall intensity is scaled by a factor that can be a function of the altitude of the point to account for the effects of orography. In order to account for seasonality, the model parameters can assume different values in each calendar month.

[22] The rainfall model was applied to the Secchia River basin by generating data for five locations where rain gauges are present. By using the method of moments, historical hourly rainfall data observed in the 2-year period 1972–1973 were used to calibrate the Neyman-Scott model parameters. In detail, calibration was performed by optimizing the fit of the mean, variance, and proportion of dry days of the hourly rainfall data in each rain gauge, as well as the spatial cross correlation at lag zero among the five locations. Fifty years of hourly rainfall data were subsequently generated for the five rain gauges. The depth duration frequency curves for rainfall in each of the rain gauges were well reproduced by the simulated data; in particular, the percentage error in the simulation of the 12-hour (a time span comparable to the concentration time of the basin) cumulated rainfall with return period of 100 years was always lower than 6% in all the rain gauges. The mean areal hourly rainfall over the basin was then estimated through a weighted average of the hourly rainfall data in the five locations; that is,

$$P_m(t) = \sum_{j=1}^{5} w_j p_j(t), \tag{1}$$

where $P_m(t)$ is the mean areal hourly rainfall over the basin at time t, the w_j are weights, and $p_j(t)$ is the hourly rainfall at

time t in rain gauge j. The weights w_j are estimated with the Thiessen polygons method. Therefore a 50-year record of lumped hourly rainfall $P_m(t)$ over the basin was obtained that was used as input to the rainfall-runoff simulation models.

[23] The generation of the synthetic hourly temperature record was performed by referring to a location where historical data are available and was carried out by applying a fractionally differenced autoregressive integrated moving average (FARIMA) model. On many occasions, this class of models turned out to be able to fit the autocorrelation structure of temperature series, which is very often affected by a slow decay, which may suggest the presence of long-term persistence, implying this way the presence of the Hurst effect [*Montanari*, 2003]. More details on FARIMA models and the simulation procedure herein applied are given by *Montanari et al.* [1997]. A mean areal value of the hourly temperature data was obtained by rescaling the synthetic observations to the mean altitude of the basin area, by adopting a standard temperature gradient (0.6°C per 100 m of altitude shift).

3.3. Generation of Synthetic River Flow Data

[24] Synthetic river flow data were obtained by using the previously generated synthetic rainfall and temperature records as input to the lumped version of the rainfall-runoff model ADM (a distributed model [*Franchini*, 1996]). The ADM model is derived from the Xinanjiang model [*Zhao et al.*, 1980] and is based upon the same concept of probability distributed soil moisture storage capacity. The model is divided into two main blocks. The first block represents the water balance at soil level, that is, the balance between the moisture content and the incoming (precipitation) and outgoing (evapotranspiration, surface runoff, interflow, and base flow) water flows. The second block represents the transfer of runoff production to the basin outlet. The soil is divided into two zones: The upper zone produces surface and subsurface runoff (interflow), and the lower zone produces base flow runoff. The transfer of these components to the outlet section takes place in two distinct stages. The first represents the flow along the hillslopes toward the channel network, while the second represents the flow along the channel network toward the basin outlet. Surface runoff and interflow are summed and transferred along the hillslopes, with a transfer function obtained as the solution of the convective diffusive flow equation when a lateral, uniformly distributed input is considered. An analogous function is used for the transfer of the total runoff along the river network to the closure section. Evapotranspiration is computed via the radiation method [*Doorembos and Pruitt*, 1977], which yields estimates of the hourly evapotranspirated water.

[25] Using a generic algorithm [*Duan et al.*, 1992], the nine parameters of the ADM model were estimated by automatically optimizing the simulation of historical hourly river flow data that were observed at Bacchello Bridge in 1972 and part of 1973. The model was validated by simulating the fraction of the 1973 flows that was not used for calibration. The resulting efficiency was 0.81 for the validation period. By using as input data the synthetic rainfall and temperature series, the ADM model was subsequently applied to generate a 50-year-long record of hourly synthetic river flows $Q_{obs}(t)$ of the Secchia River at the Bacchello Bridge cross-river section.

3.4. Synthetic Rainfall and River Flow Data Corruption

[26] The rainfall data corruption was carried out by varying the weights w_j used to compute the mean areal rainfall over the basin in equation (1). The weights were perturbed by randomly generating, at each time step, each w_j accordingly to a uniform distribution in the range ±20% of the value that was used when computing the synthetic mean areal rainfall data with equation (1). Then, the w_j obtained at each time step are rescaled so that their cumulative sum is equal to one.

[27] The rainfall data corruption introduced an uncertainty that can be quantified by the coefficient of determination of the linear regression of corrupted versus uncorrupted rainfall depths. In this case, it resulted equal to 0.76. The type of uncertainty introduced when corrupting the rainfall data is not unlikely in practical applications. In fact, the weights attributed to each rain gauge for the computation of the mean areal rainfall over the basin can change in time, depending on the spatial variability of the rainfall field.

[28] The river flow data corruption was carried out by multiplying each observation by a coefficient that assumed different values at each hourly time step. The time series of the coefficient values was obtained by generating outcomes from a uniform distribution in the range 0.8–1.2. The coefficient of determination of the linear regression of corrupted versus uncorrupted river flow data is equal to 0.86. Clearly, there are many possible ways for corrupting a river flow series. The procedure used here is representative of a situation where each river flow measurement is affected by a random inaccuracy. For instance, this latter can be caused by measurement errors or by the presence of unsteady flow conditions in the case the river discharge is measured through a steady state rating curve, that in many real-world situations is used to convert river stage observations.

3.5. Brief Description of the HYMOD Model

[29] HYMOD is a five-parameter conceptual rainfall-runoff model that was introduced by *Boyle* [2000] and recently used by *Wagener et al.* [2001] and *Vrugt et al.* [2003]. HYMOD consists of a relatively simple rainfall excess model, described in detail by *Moore* [1985], that is connected with two series of linear reservoirs: three identical reservoirs for the quick response and a single reservoir for the slow response. This model requires the optimization of five parameters on the basis of observed streamflow data: the maximum storage capacity in the catchment, C [L], the degree of spatial variability of the soil moisture capacity within the catchment, b [−], the factor distributing the flow between the two series of reservoirs, α [−], and the residence time of the linear quick and slow reservoirs, K_s [T] and K_l [T], respectively. Evapotranspiration is accounted for in the same way as in ADM, i.e., by using the hourly evapotranspiration estimates of the radiation method as input data for HYMOD.

[30] With a total of only five parameters, HYMOD can be considered an approach of reduced complexity with respect to ADM that was previously used for generating the synthetic river flow data $Q_{obs}(t)$. Thus it is anticipated that

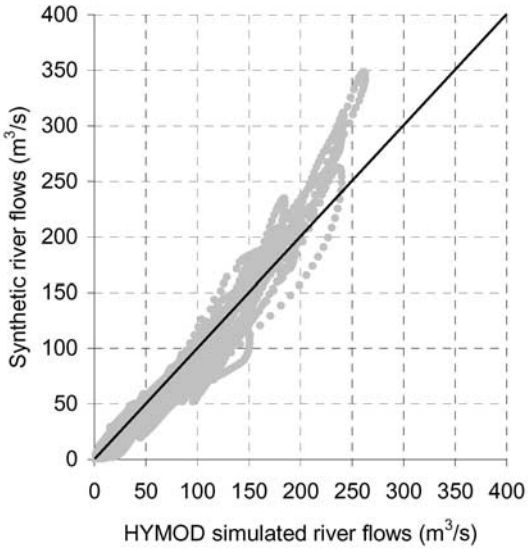

Figure 1. Scatterplot of synthetic versus HYMOD simulated hourly river flow data (calibration mode, uncorrupted data). The plot refers to years 1–10 of the simulated record.

HYMOD will provide a simulation $Q_{sim}(t)$ that does not perfectly reproduce the series $Q_{obs}(t)$ and therefore give rise to the presence of simulation uncertainty that is caused by both model structural uncertainty and HYMOD parameters uncertainty.

[31] By using a 30-year-long record of uncorrupted synthetic data (years from 1 to 30 of the synthetic records), HYMOD was first calibrated in an optimality context by minimizing the squared difference between $Q_{obs}(t)$ and $Q_{sim}(t)$ through a generic algorithm [*Duan et al.*, 1992]. This operation allowed an optimal parameter set to be obtained. Model validation was done with respect to a 20-year-long record of synthetic data that had not been used in the calibration phase (years from 31 to 50). The Nash efficiency computed when using uncorrupted data in validation is 0.91. Figure 1 shows a scatterplot of synthetic versus HYMOD simulated hourly uncorrupted river flows in calibration mode. The plot refers to the years from 1 to 10 of the synthetic river flow record.

[32] HYMOD was also calibrated with the same procedure described above, but using the corrupted rainfall and river flow series. The Nash efficiency in validation is 0.71; this is a feasible value in real-world applications.

3.6. First Experiment

[33] In order to assess the uncertainty of the HYMOD simulation, the first experiment was performed by applying GLUE and allowing multiple parameter sets to be considered. In detail, GLUE was implemented in calibration mode by running HYMOD using a 30-year-long record (from years 1 to 30) of hourly rainfall data (a total of 262,800 synthetic data). Ten thousand simulations were performed by using parameter sets obtained by randomly generating uniformly distributed values of the five HYMOD parameters, in the range ±50% of their optimal values estimated as described in section 3.5. The uniform distribution was adopted because a prior information on the parameter distributions was not available. The parameters were generated by using a variant of the multiplicative congruential algorithm known as the Wichmann-Hill algorithm [*Maclaren*, 1989] and were allowed to vary all simultaneously from set to set. A number n of the parameters sets that provided the simulations characterized by the highest values of the Nash efficiency (see also section 3.10) were retained as behavioral and were used to provide n GLUE runs of 20 years of synthetic data (years from 31 to 50) that were not used in the calibration phase (validation mode). In the uncorrupted data analysis, the results were also computed in each experiment for a 10-year-long validation in order to analyze the sensitivity to the length of the model run.

[34] It is worth noting that some applications presented by the literature identify the behavioral modeling solutions by fixing a threshold value of the Nash efficiency. All models that do not reach such threshold efficiency are rejected. This is equivalent to what is done here, where the threshold value is determined by fixing the number n of behavioral parameter sets.

[35] We allowed n to vary in order to inspect its effects on the uncertainty estimation. In detail, values of n equal to 1000, 2000, 3000, 4000, 6000, and 8000 were considered. The obtained median GLUE outputs, along with the respective prediction limits, were subsequently compared with the corresponding synthetic observations to be simulated.

[36] The GLUE implementation as described above is used in most applications presented in the literature and allows the explicit evaluation of the parameter uncertainty. However, as it was noted in section 2, it is expected that other sources of uncertainty (in this case, model structural uncertainty) might be treated implicitly. In particular, in the uncorrupted data analysis the use of error-free input data allows the examination of the compensation effect between model structural uncertainty and parameter values that is found in real applications.

3.7. Second Experiment

[37] The second experiment was performed by assuming that the rainfall input provided to the model is uncertain. It is well known that in real-world applications the input data are never certain because they contain measurement errors. For example, rainfall uncertainty was explicitly considered in the application of GLUE presented by *Cameron et al.* [1999]. Rainfall uncertainty is also taken into account by other uncertainty assessment techniques. For instance, in the Bayesian approach to total error analysis (BATEA [*Kavetski et al.*, 2003]), an uncorrelated error is assumed to affect the rainfall observations.

[38] Hence, in contrast with the results of the first experiment where only the parameter uncertainty was explicitly considered, the purpose of the second experiment is to assess the response of GLUE when input uncertainty, as well as parameter uncertainty, is explicitly taken into account.

[39] There are many different methods to account for rainfall uncertainty within GLUE. For this investigation, multiple likely rainfall inputs were obtained by perturbing the weights w_j that were used to compute the mean areal rainfall over the basin (see equation (1)). In detail, 10,000 HYMOD runs were performed by using randomly generated parameter sets. For each run, the rainfall forcing was varied by randomly generating the weights w_j according to a

uniform distribution in the range ±20% of the value that was used when computing the synthetic rainfall data. Then, the obtained w_j are rescaled so that their cumulative sum is equal to one. The weights were kept constant through each model simulation, and a number $n = 1000$ of runs were retained as behavioral.

[40] Of course, this is only one possible solution for evaluating input uncertainty, and, in general, the GLUE results will vary depending on the method chosen.

3.8. Third Experiment

[41] The third experiment was performed in order to inspect the effect on the uncertainty estimation of the length of the calibration period. In detail, while the validation period was kept at 20 years long, from years 31 to 50, five different lengths of the calibration period were selected: 2, 5, 10, 20, and 30 years. As in the first experiment, 10,000 randomly generated HYMOD parameter sets were considered and a value of $n = 1000$ was assumed. On the basis of intuitive considerations, one would expect that the uncertainty detected by GLUE decreases for an increasing length of the calibration period.

[42] The necessity to calibrate GLUE on periods of increasing length is common in practical applications. For instance, one may perform a first application of GLUE when only a short record of data is available. After then, one may obtain more data that allows calibrating GLUE on an extended period. The question then arises about how to update the GLUE confidence bands with the new information. In fact, there are different possible alternatives in computing the likelihood of a simulation of increasing length. Recalibrating GLUE over the whole period, therefore, by using both the former data and the new ones, is the choice that has been adopted here. Another choice would be to update, as the new data become available, the probability distribution of model parameters in a Bayesian framework. According to this choice, one recalibrates GLUE by considering only the new data and using as prior distribution of the rainfall-runoff model parameters the probability distribution resulting from the former GLUE calibration. This second choice gives much more weight to the new data. The identification of the proper choice is a critical decision that may significantly affect the uncertainty estimation [*Beven*, 2005]. Methods for combining information from different periods, including Bayesian combination, were previously considered by *Freer et al.* [1996] and *Beven and Freer* [2001].

[43] The behavior of the Whittle's estimates with respect to observation periods of increasing length are also inspected within the fifth experiment (see section 3.10) so that a comparison of the results obtained via different likelihood measures is possible.

3.9. Fourth Experiment

[44] The fourth experiment was performed by considering multiple parameter sets and multiple rainfall-runoff model structures. The same data set of the first experiment was used. Besides the HYMOD model described above, a second and third model were also considered. The second is a *Nash* [1958] model composed by a cascade of three linear reservoirs. The Nash model is a linear approach that in this case, was applied to synthetic data that were generated by using a nonlinear rainfall-runoff transformation. The Nash model derives the simulated river flow $Q_{\mathrm{sim}}(t)$ accordingly to the relationship

$$Q_{sim}(t) = \int_0^t A\phi P_m(t-\tau)\frac{1}{k(Ns-1)}\left(\frac{\tau}{k}\right)^{Ns-1} e^{-\tau/k}d\tau, \qquad (2)$$

where Ns [−] is the number of reservoirs ($Ns = 3$ in the present case); $P_m(t)$ [L/T] is mean areal rainfall intensity at time t; $\phi P_m(t)$ [L/T] is the net rainfall; A [L^2] is the basin area; and k [T] is the proportionality constant between the discharge $q(t)$ [L^3/T] from each reservoir and its water content $w_s(t)$ [L^3],

$$q(t) = \frac{w_s(t)}{k}. \qquad (3)$$

Here k and ϕ are parameters to be calibrated. The third considered model is ADM, which in this case represents the "perfect" model, i.e., the model used to generate the uncorrupted synthetic data. Since the true parameters are included among the candidate parameter sets of ADM, this latter model is expected to provide the "perfect" simulation when using the uncorrupted data. Therefore, in the case of the perfect data, the fourth experiment should give important indications about the response of GLUE in an ideal situation where simulation uncertainty is not present.

[45] For each model, 4000 parameter sets were randomly generated in the range ±50% of their optimal values and $n = 3000$ of runs were retained as behavioral.

3.10. Fifth Experiment

[46] The fifth experiment was performed by considering multiple parameter sets and different likelihood measures. The same data set of the first experiment was used. Besides the Nash efficiency E, the mean absolute relative error M_{Re}, the sum of squared errors S_{Sq}, and the likelihood function proposed by *Whittle* [1953], $W(\theta)$, were considered.

[47] E, M_{Re}, and S_{Sq} are performance measures that are frequently used when calibrating rainfall-runoff models. They are computed through the relationships

$$E = 1 - \frac{\sum_{t=1}^{N}[Q_{obs}(t) - Q_{sim}(t)]^2}{\sum_{t=1}^{n}[Q_{obs}(t) - Q_m]^2}, \qquad (4)$$

$$M_{Re} = \frac{1}{N}\sum_{t=1}^{N}\frac{|Q_{obs}(t) - Q_{sim}(t)|}{Q_{obs}(t)}, \qquad (5)$$

$$S_{Sq} = \sum_{t=1}^{N}[Q_{obs}(t) - Q_{sim}(t)]^2, \qquad (6)$$

where $Q_{\mathrm{obs}}(t)$ and $Q_{\mathrm{sim}}(t)$ [L^3/T] are observed and simulated hourly river discharge at time t, respectively, Q_m is the mean value of the observed sample, and N [−] is the sample size. M_{Re} is expected to give more weight to the simulation of the lower river flows.

[48] The likelihood measure proposed by *Whittle* [1953] for a stationary time series is computed on the spectral density [see, e.g., *Beran*, 1994; *Montanari et al.*, 2000;

Chouduri et al., 2004]. Whittle's likelihood has been widely used in the time series literature for constructing estimators. For the purpose of computing $W(\theta)$, the rainfall-runoff transformation can be written as (a similar representation was used by *Beven and Freer* [2001])

$$Q_{obs}(t) = M[\theta, \mathrm{I}(t)] + e(t). \tag{7}$$

Here $M[\theta, \mathrm{I}(t)]$ is the transformation operated by the hydrological model in which θ is the parameter vector and $\mathrm{I}(t)$ is the input vector (for instance, rainfall and temperature at time t); and $e(t)$ are the hydrological model residuals whose mean value is supposed to be equal to zero; since $e(t)$ is usually dependent, it was modeled here by using an autoregressive model of order one,

$$e(t) = \phi e(t-1) + \varepsilon(t), \tag{8}$$

where ϕ is an autoregressive parameter and $\varepsilon(t)$ represents a zero mean, independent and identically distributed (i.i.d.) random variable. In general, the autoregressive process is conditional on the hydrological model $M[\theta, \mathrm{I}(t)]$ that is used, and higher-order models could also be considered. However, an extensive study carried out by the *World Meteorological Organization* [1992] proved that a first-order autoregressive process is sufficient to account for dependence in the residual series in many practical applications of rainfall-runoff models.

[49] The Whittle's likelihood for the model given by (7) can be computed through the relationship

$$L(\theta) = \exp\left[-\sum_{j=1}^{N/2}\left\{\log\left[f_M(\lambda_j,\theta) + f_e(\lambda_j,\phi)\right] + \frac{J(\lambda_j)}{f_M(\lambda_j,\theta) + f_e(\lambda_j,\phi)}\right\}\right], \tag{9}$$

where λ_j are the Fourier frequencies; J is the spectral density of the observed sample; f_M is the spectral density of the hydrological model that depends on the parameter vector θ; and f_e is the spectral density of the first-order autoregressive operator that depends on the autoregressive parameters ϕ. Thus model calibration can be carried out by minimizing [*Beran*, 1994]

$$W(\theta) = \sum_{j=1}^{N/2}\left\{\log\left[f_M(\lambda_j,\theta) + f_e(\lambda_j,\phi)\right] + \frac{J(\lambda_j)}{f_M(\lambda_j,\theta) + f_e(\lambda_j,\phi)}\right\}. \tag{10}$$

In the above expression, the spectral density function of the model, $f_M(\lambda_j, \theta)$, has been estimated by computing the smoothed periodogram of a long model run. Smoothing was conducted by using the loess smoother that operates through a local regression model [*Cleveland and Devlin*, 1988]. A smoothing span equal to 50 was used.

[50] The spectral density function of the first-order autoregressive process is given by [*Beran*, 1994]

$$f_e(\lambda_j,\phi) = \frac{\sigma_\varepsilon^2}{2\pi}\left(1 - 2\phi\cos\lambda + \phi^2\right)^{-1}, \tag{11}$$

where σ_ε is the standard deviation of $\varepsilon(t)$.

[51] Under the assumption of zero mean and i.i.d. $\varepsilon(t)$, Whittle's likelihood provides asymptotically consistent and normally distributed estimates in the case of non-Gaussian and linear models [*Giraitis and Surgailis*, 1990]. Asymptotical normality is no more guaranteed for the case of nonlinear models, as it is in the present case.

[52] It is worth pointing out that some of the underlying assumptions of the estimator given by (10) are indeed likely to be violated in hydrological modeling. First of all, the $\varepsilon(t)$ are often heteroscedastic because of the presence of higher hydrological model errors during peak flow periods. Consequently, a bias can be induced in the parameter estimates and prediction limits. In this study, we found that the variance of $\varepsilon(t)$ is small, and therefore we assumed that the effects of heteroscedasticity are negligible. Indeed, when dealing with uncorrupted rainfall and river flow data, the hydrological model explained 90%, on average, of the variance $\sigma^2_{Q\mathrm{obs}}$ of the synthetic data, while the combination of the hydrological model and autoregressive operator explained 99%, on average, of $\sigma^2_{Q\mathrm{obs}}$. Therefore the bias induced by eteroscedasticity and nonstationarity in $\varepsilon(t)$ is likely to have a marginal effect on hydrological model estimation.

[53] The Whittle's likelihood was used within this study by rejecting those parameter sets that lead to residuals $e(t)$ of the hydrological model whose mean value is significantly different from zero at the 95% confidence level. A large sample of experiments performed by using the HYMOD model confirmed empirically the consistency of the Whittle's estimator as expressed by (10). The distribution of the estimates resulted in general non-Gaussian. Other types of likelihood measure were considered by *Romanowicz et al.* [1994], *Freer et al.* [1996], and *Beven and Freer* [2001]. However, the evaluation of these results did not refer to a full validation mode.

[54] In this study, 10,000 runs were performed by using the same parameter sets of the first experiment. For each trial the values of E, M_{Re}, S_{Sq}, and $W(q)$ were computed, and the best performing 1000 parameter sets were retained for each performance measure.

[55] The Whittle estimation was repeated by considering different lengths of the calibration period, 10 and 2 years. The purpose was to check whether an alternative likelihood measure behaves differently with respect to Nash efficiency when information from periods of increasing length is evaluated.

[56] Finally, a combination of the Whittle's likelihood and sum of squares, $W(q)$ and S_{Sq}, respectively, was also tested. This was done by first applying GLUE using the Whittle likelihood. This allows a probability distribution for the HYMOD parameters to be derived. Subsequently, GLUE was applied again by using S_{Sq} as the performance measure, by sampling the candidate parameter sets from the probability distributions previously estimated. The aim is to apply GLUE by including a prior knowledge about the distribution of parameters, which is derived through a preceding GLUE application by using a different likelihood measure. Each GLUE application was developed by generating 5000 parameter sets, 1000 of which were retained as behavioral.

4. Results of the Experiments

[57] The results of the experiments are assessed by comparing the output of the 20-year-long validations to the synthetic data that is to be simulated. First, the Nash

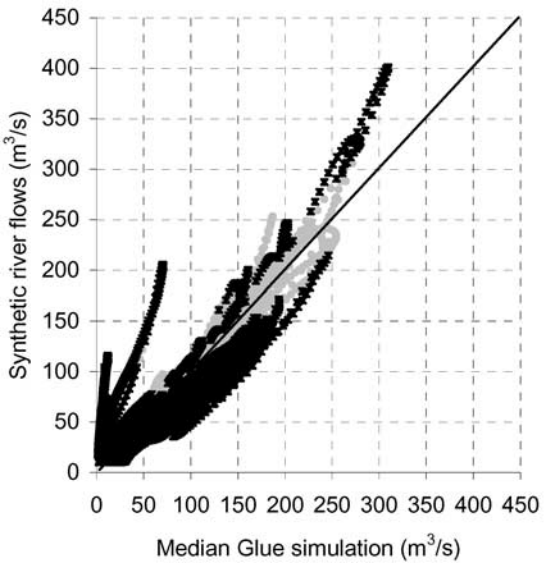

Figure 2. First experiment, calibration period of 30 years, $n = 1000$. Synthetic data to be simulated (hourly river flows, uncorrupted data) versus median GLUE output for a validation period of 10 years. Crosses indicate the points lying outside the 95% GLUE prediction limits. Ten thousand randomly selected parameter sets were considered. The plot refers to years 31–40 of the simulated record.

efficiency of the median GLUE output was computed to provide an indication of the reliability of the given modeling solution in the evaluation of the best river flow estimate. Second, an appraisal of the GLUE capability to assess the simulation uncertainty in validation mode is carried out by counting the number of observed data lying inside the 95% prediction bands of the simulation. Figure 2 illustrates the results of the first experiment, uncorrupted data analysis, with $n = 1000$ and validation length of 10 years. The scatterplot displays the synthetic river flow data versus the median GLUE output. Crosses indicate the points outside the 95% prediction envelope. Therefore, if the total uncertainty is completely captured by GLUE, the crosses should be 5% of the total points and uniformly distributed along the range of the simulated river flows, though this result is not expected for the reasons mentioned at the end of section 1. Moreover, for the first, second, and third experiments, only the parameter (and input, for the second experiment) uncertainty is explicitly considered by GLUE. However, because part of the uncertainty is potentially treated implicitly, it is interesting to verify how much of total uncertainty is actually captured.

4.1. Uncorrupted Data Analysis

[58] Table 1 shows a summary of all GLUE applications in each experiment in validation mode. In particular, the Nash efficiency of the median GLUE output and the percentage of simulated data lying outside the 95% prediction limits (crosses in Figures 2–5) are indicated. In order to show the sensitivity of the results to the run length, the same statistics referring to a 10-year-long trial are also reported, in parentheses. Figure 3, 4, and 5 are scatterplots (similar to the one shown in Figure 2) that respectively refer to the first experiment with $n = 8000$, the fourth experiment, and the fifth experiment with the Whittle likelihood.

[59] Before looking at the results, it is worth commenting about the performances of the three rainfall-runoff models involved in the fourth experiment. As it was already mentioned in section 3.9, 4000 runs were performed for each one of them, and the 3000 best ones were retained as behavioral. Among them, 2363 were provided by ADM and 637 were provided by HYMOD. Because none of the runs provided by the Nash model was among the 3000 behavioral trials, the Nash model was rejected as nonbehavioral.

[60] It is also worth noting that the lowest efficiencies in all of the experiments are very high; this situation is quite

Table 1. Results of the Experiments With Uncorrupted Rainfall and River Flow Data[a]

Experiment	n	Calibration Period, years	Likelihood Measure	Highest Likelihood (Calibration)	Lowest Likelihood (Calibration)	K	E_v (Validation)
1	1000	30	E	0.92	0.89	40% (39%)	0.88 (0.85)
1	2000	30	E	0.92	0.89	33% (32%)	0.88 (0.85)
1	3000	30	E	0.92	0.88	29% (28%)	0.88 (0.85)
1	4000	30	E	0.92	0.87	26% (25%)	0.88 (0.85)
1	6000	30	E	0.92	0.85	19% (18%)	0.88 (0.86)
1	8000	30	E	0.92	0.82	17% (16%)	0.88 (0.85)
2	1000	30	E	0.92	0.90	40% (39%)	0.88 (0.85)
3	1000	20	E	0.91	0.89	39% (39%)	0.88 (0.86)
3	1000	10	E	0.91	0.88	31% (29%)	0.90 (0.88)
3	1000	5	E	0.89	0.86	30% (27%)	0.90 (0.88)
3	1000	2	E	0.81	0.76	27% (24%)	0.90 (0.88)
4	3000	30	E	1.00	0.89	0% (0%)	0.98 (0.98)
5	1000	30	M_{Re}	0.56	0.83	35% (34%)	0.72 (0.69)
5	1000	30	S_{Sq}	102.82	125.86	40% (40%)	0.88 (0.85)
5	1000	30	$W(\theta)$	0.0003	3.2539	14% (13%)	0.89 (0.86)
5	1000	10	$W(\theta)$	0.0002	3.1714	11% (10%)	0.90 (0.89)
5	1000	2	$W(\theta)$	0.0080	4.0024	10% (8%)	0.88 (0.86)
5	1000	30	$W(\theta) + S_{Sq}$	103.15	135.93	45% (44%)	0.88 (0.84)

[a] E_c is the Nash-Sutcliffe efficiency of the median GLUE output in validation; K indicates the percentage of simulated data lying outside the 95% prediction limits; n is the number of behavioral parameter sets that have been retained; E indicates the Nash-Sutcliffe efficiency computed on the hourly river flows; M_{Re} is the mean relative error; S_{Sq} is the sum of squares; and $W(\theta)$ is the Whittle's likelihood measure. The length of the validation period is 20 years. Values of K and E_v computed in 10-year-long runs are reported in parentheses, in order to show the sensitivity of the results to the length of the model run.

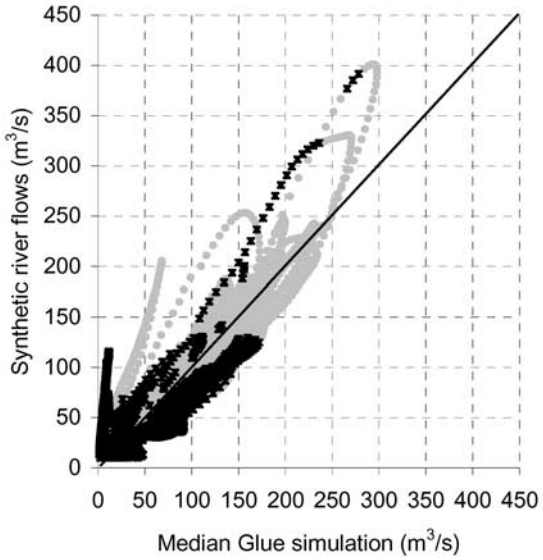

Figure 3. First experiment, calibration period of 30 years, $n = 8000$. Synthetic data to be simulated (hourly river flows, uncorrupted data) versus median GLUE output for a validation period of 10 years. Crosses indicate the points lying outside the 95% GLUE prediction limits. Ten thousand randomly selected parameter sets were considered.

rare in real-world applications and might indeed affect the results of the uncertainty estimation. In fact, the GLUE prediction band will certainly be narrower than what is expected for practical implementations. However, this is just a consequence of the good performances of HYMOD (for instance, input and output uncertainty are not present when analyzing the uncorrupted data) and is not expected to have much influence on the indications that can be derived about the reliability of the uncertainty estimation.

[61] Briefly, the results of the experiments can be summarized by the following points.

[62] 1. The first experiment shows that the subjective choice of the number of behavioral parameter sets has a strong effect on the uncertainty estimation. In fact, the percentage K of points lying outside the 95% prediction limits varies from 40% to 17% for $n = 1000$ and $n = 8000$, respectively. One may also note that even when choosing $n = 8000$, the uncertainty as estimated by GLUE is lower than the total simulation uncertainty that would be completely captured for values of K around 5%. Therefore one may conclude that by considering multiple parameters, one does not completely capture the effects induced by model structural uncertainty. This result is evident even in the present case in which the HYMOD model provides quite good performances (Nash efficiency is 0.91 in validation). It should also be remarked that n does not significantly affect the efficiency of the median GLUE output.

[63] 2. The second experiment shows that the uncertainty estimated by GLUE does not increase when multiple inputs are considered. This proves the efficiency of GLUE in explicitly detecting the effects of input uncertainty in the present case. Indeed, since input uncertainty is not present, GLUE is correct in detecting its null contribution. Thus it can be concluded that either the model parameters do compensate when running GLUE with multiple inputs or only trials operated with slightly corrupted rainfalls were retained as behavioral.

[64] 3. The third experiment shows, on the one hand, that GLUE indicates increasing uncertainty for calibration periods of decreasing length. This result is coherent with what one would expect on the basis of intuition. On the other hand, the Nash efficiency of the median GLUE output does not appear to be affected much by the length of the calibration period. However, this result is to be evaluated in view of the fact that uncorrupted and strictly stationary data

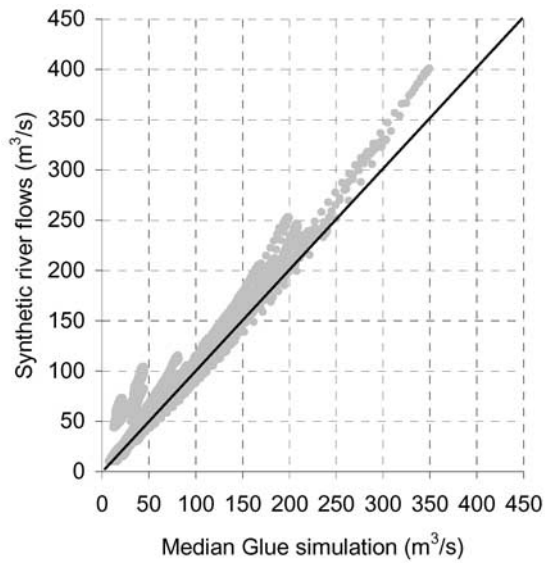

Figure 4. Fourth experiment, calibration period of 30 years, $n = 3000$. Synthetic data to be simulated (hourly river flows, uncorrupted data) versus median GLUE output for a validation period of 10 years. There are no points lying outside the 95% GLUE prediction limits. Three rainfall-runoff models were considered.

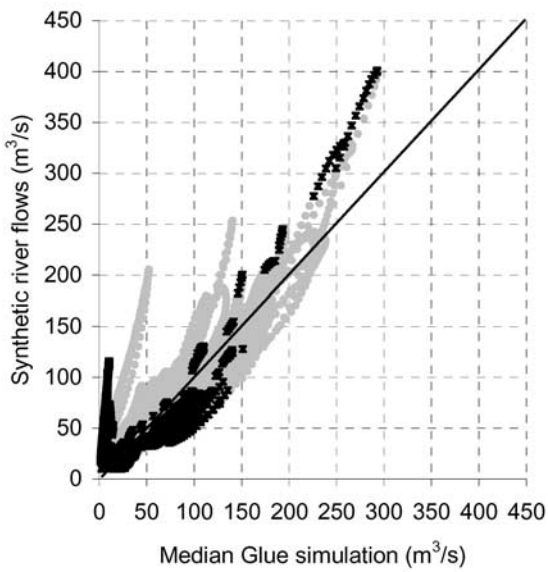

Figure 5. Fifth experiment, calibration period of 30 years, $n = 1000$, Whittle likelihood. Synthetic data to be simulated (hourly river flows, uncorrupted data) versus median GLUE output for a validation period of 10 years. Crosses indicate the points lying outside the 95% GLUE prediction limits.

were analyzed here. When dealing with real-world observations, the sample size required for performing a stable simulation, in terms of efficiency of the results, might be much larger than the minimum calibration size of 2 years that was considered in this study.

[65] 4. The fourth experiment shows that GLUE correctly includes all data inside the 95% prediction limits when the perfect model is considered. However, it is interesting to note that the median GLUE output is still characterized by the presence of uncertainty (Nash efficiency is 0.98). This point is illustrated by Figure 4 in which the data of the scatterplot are still scattered around the straight line that indicates equality between synthetic and simulated data. Therefore GLUE indicates the presence of uncertainty even when the perfect model is considered, though in a reduced amount with respect to the previous experiments. It is expected that varying the amount of information used to calibrate GLUE, by either changing the length of the calibration period or using different likelihood measures, would lead to a change in the efficiency of the GLUE output. For instance, the Nash efficiency is equal to 0.93 when using a calibration period of only 2 years. By contrast, an increase of the calibration period length did not lead to a significant change in the results. As a matter of fact, when dealing with the perfect model and perfect (stationary) data, a calibration period that covers 30 years (that is, 262,800 hourly data) is extended enough in order to provide a stable efficiency of the median GLUE output. In real situations this result is not expected because the perfect model and perfect data are not available. It is important to note that the uncertainty detected by GLUE is lower with respect to the first experiment in which only one rainfall-runoff model was considered. Therefore including multiple model structures, while leading to a possible increase in equifinality because a wider range of models is considered, does not always lead to an increase in uncertainty.

[66] 5. The fifth experiment shows some significant differences among the considered likelihood measures. First of all, one notes that using S_{Sq} gives results similar to using E. This was to be expected since the two measures are different ways of scaling the same type of information and both rely implicitly on the assumption of the absence of dependence in the hydrological model errors. Using the mean relative error M_{Re} gives rise to a lower efficiency of the median GLUE output, 0.72 versus 0.88 of E and S_{Sq}. This result is reasonable because using M_{Re} leads to a better simulation of lower river flows while the efficiency of the median GLUE output is much more affected by errors in the higher data. An inspection of the percentage of points lying inside the GLUE prediction bands reveals that M_{Re} is more efficient in the uncertainty assessment phase than E (35% versus 40% of outside points). This result is probably because the response surface defined through M_{Re} is less peaky than the one defined through E and S_{Sq}.

[67] Interesting remarks can also be drawn by looking at the results obtained with the Whittle likelihood. The efficiency of the median GLUE output is comparable to using E and S_{Sq} (0.89, versus 0.88 of E and S_{Sq}) while including much more observed points within the prediction bands with respect to the other likelihood measures (14% for $W(\theta)$, versus 40% of E and 35% of M_{Re}). Similar results are obtained when operating with a shorter calibration period. Therefore Whittle's likelihood appears to provide a less peaked response surface with respect to E while preserving the capability of providing a efficient median output. The shape of the response surface of Whittle's likelihood is not surprising; in fact, its less peaked appearance is probably influenced by the smoothing operation conducted when estimating the spectral density function $f_M(\lambda, \theta)$ of the model (see section 3.10). This smoothing procedure is needed in order to perform a robust estimation of $f_M(\lambda, \theta)$, but it undoubtedly leads to increasing equifinality and therefore to widening of the GLUE prediction bands. In summary, Whittle's likelihood appears to be a potentially valuable performances measure that evaluates the goodness of the fit provided by the model over the whole range of the Fourier frequencies of the data.

[68] The combination of Whittle's likelihood and S_{Sq} provides consistent results. Because the prior distribution of model parameters given by Whittle's likelihood is used, the equifinality is reduced and thereby the prediction bands narrowed and the identifiability problems limited. In general, the results of the combination, including the median output efficiency, will depend both on which likelihood measure is used first and the number of total and behavioral parameter sets considered during the GLUE application.

[69] Finally, one can see that the results presented in Table 1 do not change significantly when increasing the length of the validation period from 10 to 20 years.

4.2. Corrupted Data Analysis

[70] Table 2 shows a summary of all GLUE applications to corrupted data, for each experiment in a 20-year-long validation. One can see that in this case the Nash-Sutcliffe efficiencies are closer to values of real-world case studies.

[71] The results of the corrupted data analysis largely confirm the conclusions drawn in section 4.1. However, some interesting differences can be noted. In the first experiment, one can see that the percentage of points lying outside the confidence bands is much greater (see Figure 6). As it was expected, the newly introduced observable uncertainty is largely not accounted for by GLUE when multiple parameters are considered. Moreover, one can see that the Nash efficiency of the median GLUE output in the different runs is more fluctuating with respect to the uncorrupted data analysis. This can be explained by considering that the increased simulation uncertainty, which was introduced through a random perturbation of the rainfall and river flow, introduces additional instability in the HYMOD performances.

[72] The second experiment shows that the uncertainty estimated by GLUE increases when multiple inputs are considered in the corrupted data analysis. This outcome confirms the efficiency of GLUE in explicitly detecting at least part of the uncertainty introduced in the corrupted rainfall data.

[73] The results of the third experiment partly disagree with what was found in the uncorrupted data analysis. In detail, one can see that the uncertainty detected by GLUE does not increase for calibration periods of decreasing length. However, one should consider that the data corruption has introduced a large amount of uncertainty in the simulation. Therefore it is possible that the uncertainty

Table 2. Results of the Experiments With Corrupted Rainfall and River Flow Data[a]

Experiment	n	Calibration Period, years	Likelihood Measure	Highest Likelihood (Calibration)	Lowest Likelihood (Calibration)	K	E_v (Validation)
1	1000	30	E	0.79	0.76	61%	0.66
1	2000	30	E	0.79	0.74	56%	0.63
1	3000	30	E	0.79	0.72	51%	0.65
1	4000	30	E	0.79	0.71	49%	0.59
1	6000	30	E	0.79	0.68	43%	0.63
1	8000	30	E	0.79	0.63	40%	0.61
2	1000	30	E	0.80	0.75	55%	0.67
3	1000	20	E	0.78	0.75	62%	0.65
3	1000	10	E	0.78	0.75	62%	0.66
3	1000	5	E	0.78	0.76	61%	0.64
3	1000	2	E	0.81	0.78	60%	0.60
4	3000	30	E	0.84	0.75	38%	0.70
5	1000	30	M_{Re}	0.58	0.79	49%	0.28
5	1000	30	S_{Sq}	206.15	233.55	61%	0.66
5	1000	30	$W(\theta)$	0.0131	0.1086	52%	0.61
5	1000	10	$W(\theta)$	0.0002	3.1714	54%	0.63
5	1000	2	$W(\theta)$	0.1560	0.1755	50%	0.50
5	1000	30	$W(\theta) + S_{Sq}$	103.15	135.93	66%	0.65

[a]Symbols are as in Table 1. The length of the validation period is 20 years.

induced by the reduced length of the calibration period is partly hidden, because of compensation effects among different sources of error.

[74] The fourth experiment stimulates an interesting consideration. In fact, one can see that in contrast to what was found in the corrupted data analysis, the results of the uncertainty estimation do not change significantly when ADM (the ''perfect model'') is included among the competing modeling solutions. The reason of this outcome is that ADM does not significantly outperform HYMOD anymore when corrupted data are used. In fact, the errors in the input and output variables partly compensate for the uncertainty in the model structure. This a typical case where a model (HYMOD) might result behavioral for the presence of data errors and not for its effective capability of well simulating the hydrological processes. However, it is remarkable to note that even in this case, the median GLUE output is more efficient when considering multiple rainfall-runoff models. Therefore it is confirmed that including multiple model structures may lead to a decreased uncertainty in real-world applications.

[75] The fifth experiment confirms that the use of Whittle's likelihood still allows us to include much more observed points within the prediction bands.

5. Concluding Remarks

[76] The results of the experiments summarized above highlight that the subjective choices to be taken within GLUE are highly effective on the uncertainty estimation. Indeed, the prediction intervals given by GLUE are dependent upon which models are used, which parameter ranges are sampled, which likelihood measures are evaluated, and how they are combined. They will also be implicitly dependent upon the unknown errors in the input data and observations that are compared with the model simulations. In summary, the GLUE results appear to be remarkably conditioned on the assumptions made.

[77] The analysis confirms that the prediction limits provided by GLUE do not necessarily include a specified proportion of the observed data. In fact, in all the experiments herein developed the GLUE prediction bands appear to underestimate the total uncertainty of the simulation provided by the hydrological model. In detail, when corrupted rainfall and river flow data are used, and a real-world application is therefore emulated, the 95% prediction limits capture no more than 62% of the observed data.

[78] On the other hand, one may note that as mentioned in section 2, it might be not meaningful to compare uncertainty assessed by GLUE with total simulation uncertainty (see *Beven* [2005] and *Beven and Freer* [2001] for a comprehensive discussion about this point). In fact, GLUE provides

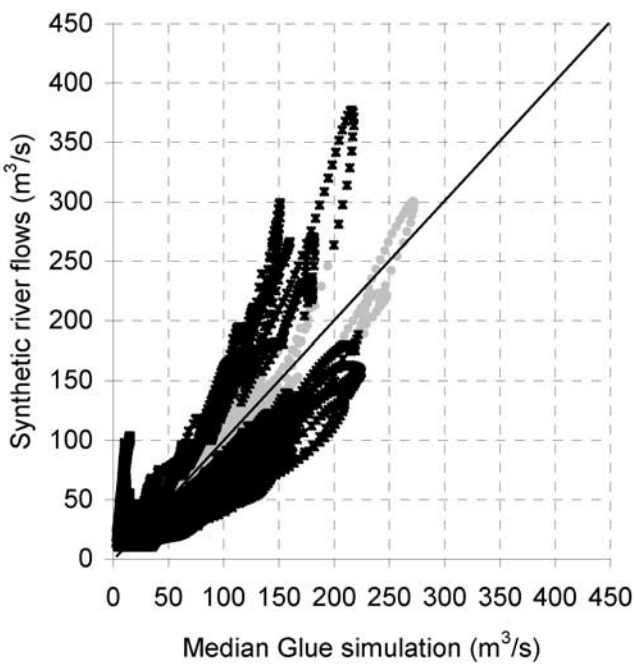

Figure 6. First experiment, calibration period of 30 years, $n = 1000$. Synthetic data to be simulated (hourly river flows, corrupted data) versus median GLUE output for a validation period of 10 years. Crosses indicate the points lying outside the 95% GLUE prediction limits. Ten thousand randomly selected parameter sets were considered. The plot refers to years 31–40 of the simulated record.

an estimate of the uncertainty originating from the insecurity of the user who might be dubious about which model structure, model parameter, and model input to use. This is a kind of uncertainty induced by equifinality, or identifiability problems, that we may call uncertainty in the modeling process. This is distinctly different from the total simulation uncertainty that gives indications about the actual magnitude of the simulation error. In some cases, the two types of uncertainty may be comparable, as the results shown in Tables 1 and 2 demonstrate. However, some important differences between them can be highlighted.

[79] Total simulation uncertainty can be reliably quantified only with respect to ideal situations such as the one considered in this study. However, it is objective, in principle, once a given modeling approach is selected. Moreover, provided that long and reliable historical records are available, it can be assessed by comparing observed and simulated data, though generally this is never the case in real-world applications. Finally, it is expected to vanish in an ideal context if the perfect modeling solution is used.

[80] Uncertainty in the modeling process is dependent on the choices done by the user, for instance, the number of model structures, the number of behavioral parameter sets, etc. This is reasonable because the decisions taken by the user reflect his feeling about equifinality. For instance, if one considered that the input uncertainty is negligible, then this uncertainty source would not be considered. Last, uncertainty in the modeling process does not vanish if the perfect model is employed, because the user does not know that he has the perfect solution available and therefore he might nevertheless feel trapped in equifinality [see *Beven*, 2002a, 2005]. Distinguishing between the two types of uncertainty might be advisable in order to correctly evaluate the response of GLUE or any other uncertainty assessment technique. Uncertainty detected by GLUE might coincide with total simulation uncertainty under a proper selection of the subjective decisions to be made within GLUE, even though one may note that a similar solution is not necessarily an aim in the GLUE framework.

[81] Given that any uncertainty assessment method is conditioned on some assumptions, the choice of the most reasonable and robust technique should be made on the basis of the knowledge of the system, the input and output data, and the available modeling approaches. In general, depending on the scopes of the analysis, different techniques for uncertainty estimation may be required. However, regardless of which technique is applied, it is extremely important that all of the underlying assumptions are stated explicitly and the consequent limitations discussed in full detail.

[82] **Acknowledgments.** The author thanks Keith Beven for providing many and very useful comments and advice. Three anonymous referees are also acknowledged for providing helpful and constructive reviews. The work presented here has been carried out in the framework of the activity of the Working Group at the University of Bologna of the Prediction in Ungauged Basins (PUB) initiative of the International Association of Hydrological Sciences.

References

Aronica, G., P. D. Bates, and M. S. Horritt (2002), Assessing the uncertainty in distributed model predictions using observed binary pattern information within GLUE, *Hydrol. Processes*, *16*, 2001–2016.

Beran, J. (1994), *Statistics for Long-Memory Processes*, CRC Press, Boca Raton, Fla.

Beven, K. J. (1993), Prophesy, reality and uncertainty in distributed hydrological modeling, *Adv. Water Resour.*, *16*, 41–51.

Beven, K. J. (2001), How far can we go in distributed hydrological modeling, *Hydrol. Earth Syst. Sci.*, *5*, 1–12.

Beven, K. J. (2002a), Towards an alternative blueprint for a physically based digitally simulated hydrologic response modeling system, *Hydrol. Processes*, *16*, 189–206.

Beven, K. J. (2002b), Towards a coherent philosophy for environmental modelling, *Proc. R. Soc. London A*, *460*(458), 2465–2484.

Beven, K. J. (2005), A manifesto for the equifinality thesis, *J. Hydrol.*, in press.

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

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.

Blazkova, S., and K. J. Beven (2002), Flood frequency estimation by continuous simulation for a catchment treated as ungauged (with uncertainty), *Water Resour. Res.*, *38*(8), 1139, doi:10.1029/2001WR000500.

Blazkova, S., and K. J. Beven (2004), Flood frequency estimation by continuous simulation of subcatchment rainfalls and discharges with the aim of improving dam safety assessment in a large basin in the Czech Republic, *J. Hydrol.*, *292*, 153–172.

Boyle, D. P. (2000), Multicriteria calibration of hydrological models, Ph.D. dissertation, Dep. of Hydrol. and Water Resour., Univ. of Ariz., Tucson.

Cameron, D. S., K. J. Beven, J. Tawn, S. Blazkova, and P. Naden (1999), Flood frequency estimation by continuous simulation for a gauged upland catchment (with uncertainty), *J. Hydrol.*, *219*, 169–187.

Cameron, D. S., K. J. Beven, and P. Naden (2000), Flood frequency estimation by continuous simulation under climate change (with uncertainty), *Hydrol. Earth Syst. Sci.*, *4*, 393–405.

Chouduri, N., S. Ghosal, and A. Roy (2004), Contiguity of the Whittle measure for a Gaussian time series, *Biometrika*, *91*, 211–218.

Cleveland, W. S., and S. J. Devlin (1988), Locally-weighted regression: An approach to regression analysis by local fitting, *J. Am. Stat. Assoc.*, *83*, 596–610.

Cowpertwait, P. S. P. (1995), A generalized spatial-temporal model of rainfall based on a clustered point process, *Proc. R. Soc. London A*, *450*, 163–175.

Doorembos, J., and W. O. Pruitt (1977), Guidelines for predicting crop water requirements, *U.N. Food and Agric. Organ. Irrig. Drain. Pap.*, *24*, 180 pp., Rome.

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

Franchini, M. (1996), Use of a genetic algorithm combined with a local search method for the automatic calibration of conceptual rainfall runoff models, *Hydrol. Sci. J.*, *41*, 21–39.

Freer, J., K. J. 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*, 2163–2173.

Freer, J., K. J. Beven, and N. Peters (2003), Multivariate seasonal period model rejection within the generalised likelihood uncertainty estimation procedure, in *Calibration of Watershed Models*, *Water Sci. Appl. Ser.*, vol. 6, edited by Q. Duan et al., pp. 69–87, AGU, Washington, D. C.

Giraitis, L., and D. Surgailis (1990), A central limit theorem for quadratic forms in strongly dependent linear variables and application to asymptotical normality of Whittle's estimate, *Probab. Theory Related Fields*, *86*, 87–104.

Kavetski, D., S. W. Franks, and G. Kuczera (2003), Confronting input uncertainty in environmental modelling, in *Calibration of Watershed Models*, *Water Sci. Appl. Ser.*, vol. 6, edited by Q. Duan et al., pp. 49–68, AGU, Washington, D. C.

Krzysztofowicz, R. (2002), Bayesian system for probabilistic river stage forecasting, *J. Hydrol.*, *268*, 16–40.

Krzysztofowicz, R., and C. J. Maranzano (2004), Bayesian system for probabilistic stage transition forecasting, *J. Hydrol.*, *293*, 57–73.

Maclaren, N. M. (1989), The generation of multiple independent sequences of pseudorandom numbers, *Appl. Stat.*, *38*, 351–359.

Montanari, A. (2003), Long-range dependence in hydrology, in *Theory and Applications of Long-Range Dependence*, edited by P. Doukhan et al., pp. 461–472, Springer, New York.

Montanari, A., and A. Brath (2004a), A stochastic approach for assessing the uncertainty of rainfall-runoff simulations, *Water Resour. Res.*, *40*, W01106, doi:10.1029/2003WR002540.

Montanari, A., and A. Brath (2004b), Assessing the uncertainty of rainfall-runoff simulations through a meta-Gaussian approach, in *Recent*

Advances in Peak River Flow Modelling, Prediction and Real-Time Forecasting—Assessment of the Impacts of Land-Use and Climate Changes, edited by A. Brath et al., pp. 79–104, BIOS, Cosenza, Italy.

Montanari, A., R. Rosso, and M. S. Taqqu (1997), Fractionally differenced ARIMA models applied to hydrologic time series: Identification, estimation, and simulation, *Water Resour. Res.*, *33*, 1035–1044.

Montanari, A., R. Rosso, and M. S. Taqqu (2000), A seasonal fractional ARIMA model applied to the Nile River monthly flows at Aswan, *Water Resour. Res.*, *36*, 1249–1259.

Moore, R. J. (1985), The probability-distributed principle and runoff production at point and basin scales, *Hydrol. Sci. J.*, *30*, 273–297.

Nash, J. E. (1958), The form of the instantaneous unit hydrograph, *IAHS Publ.*, *42*, 114–118.

Nash, J. E., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models: 1. A discussion of principles, *J. Hydrol.*, *10*, 282–290.

Pappenberger, F., K. J. Beven, M. Horritt, and S. Blazkova (2005), Uncertainty in the calibration of effective roughness parameters in HEC-RAS using inundation and downstream level observations, *J. Hydrol.*, *302*(1–4), 46–69.

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*, vol. 2, *Water Related Issues*, edited by V. Barnett and K. F. Turkman, pp. 297–317, John Wiley, Hoboken, N. J.

Vrugt, J. A., H. V. Gupta, W. Bouten, and S. Sorooshian (2003), A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters, *Water Resour. Res.*, *39*(8), 1201, doi:10.1029/2002WR001642.

Wagener, T., D. P. Boyle, M. J. Lees, H. S. Wheater, H. V. Gupta, and S. Sorooshian (2001), A framework for development and application of hydrological models, *Hydrol. Earth Syst. Sci.*, *5*, 13–26.

Whittle, P. (1953), Estimation and information in stationary time series, *Ark. Mat.*, *2*, 423–434.

World Meteorological Organization (1992), Simulated real-time intercomparison of hydrological models, *Oper. Hydrol. Rep. 38*, 241 pp., Geneva.

Zhao, R. J., Y. L. Zhuang, L. R. Fang, X. R. Liu, and Q. S. Zhang (1980), The Xinanjiang model, *IAHS Publ.*, *129*, 351–356.

A. Montanari, Faculty of Engineering, University of Bologna, Via del Risorgimento 2, I-40136 Bologna, Italy. (alberto.montanari@unibo.it)