WATER RESOURCES RESEARCH, VOL. 32, NO. 7, PAGES 2161–2173, JULY 1996

Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach

Jim Freer[1] and Keith Beven

Centre for Research on Environmental Systems and Statistics, Institute of Environmental and Biological Sciences, Lancaster University, Lancaster, United Kingdom

Bruno Ambroise

Centre d'Etudes et de Recherches Eco-Géographiques, Université Louis Pasteur, Strasbourg, France

Abstract. This paper addresses the problem of evaluating the predictive uncertainty of TOPMODEL using the Bayesian Generalised Likelihood Uncertainty Estimation (GLUE) methodology in an application to the small Ringelbach research catchment in the Vosges, France. The wide range of parameter sets giving acceptable simulations is demonstrated, and uncertainty bands are presented based on different likelihood measures. It is shown how the distributions of predicted discharges are non-Gaussian and vary in shape through time and with discharge. Updating of the likelihood weights using Bayes equation is demonstrated after each year of record and it is shown how the additional data can be evaluated in terms of the way they constrain the uncertainty bands.

Introduction

Traditional methods of calibration of hydrological models have been aimed at finding an optimal set of parameter values within some particular model structure to represent a catchment area (see, for example, work by *Blackie and Eeles* [1985] and the recent work of *Duan et al.* [1992], discussed below). The limitations of the optimal parameter set concept have been discussed by *Beven and Binley* [1992] and *Beven* [1989a, 1993], who suggest that there may be many parameter sets within a model structure that are equally acceptable as simulators of a catchment, and that these may often come from very different regions in the parameter space. There may be no rigorous basis for differentiating between these parameter sets, which demonstrate equifinality in simulating the observations. Other model structures may be equally viable as simulators [see *Ambroise et al.*, this issue (b)]. This equifinality arises due to the effects of error and uncertainty in the modeling process, resulting from error in the model representation of the hydrological processes and catchment characteristics and error in the boundary conditions of the simulation which, as pointed out by *Stephenson and Freeze* [1974], can never be known perfectly. A similar concept of equal acceptability of "behavioral" models underlies the recent work of *Spear et al.* [1994]. This idea can be extended to multiple model structures that are potential simulators of the catchment of interest.

In this paper it is shown how these limitations of the modeling process can be reflected in the estimation of predictive uncertainty and how additional data might be used to constrain those uncertainties. The methodology used is the Generalised Likelihood Uncertainty Estimation (GLUE) of *Beven and Binley* [1992]. This is a Bayesian Monte Carlo simulation–based technique, developed as an extension of the Generalised Sensitivity Analysis (GSA) of *Spear and Hornberger* [1980]. The method was outlined in concept by *Beven* [1989a] and other applications using different types of likelihood measure have been demonstrated by *Binley and Beven* [1991], *Beven* [1993] and *Romanowicz et al.* [1994]. The GLUE methodology is here demonstrated in an application to the simulation of the small Ringelbach research catchment in the Vosges, France, using a generalized version of TOPMODEL [see *Ambroise et al.*, this issue (a, b)].

The GLUE Methodology: Rationale

The background to the GLUE methodology has been an attempt to recognize more explicitly the fundamental limitations of hydrological models as simulators of catchment rainfall runoff processes (see discussions by *Beven* [1989b], *Grayson et al.* [1992], and *Beven* [1993]). One implication of such a recognition is that it should not be assumed that there is one "optimal" model structure or parameter set that can be found to represent a catchment (whether a lumped or distributed representation). This has been reinforced by recent detailed explorations of the parameter space for particular models [see *Duan et al.*, 1992; *Beven*, 1993] where it has been quite clearly demonstrated that there are many combinations of parameter values for a chosen model structure that may be equally good in reproducing observed discharges in terms of some quantitative objective function of goodness of fit. It also appears that the good simulations may be distributed across a wide range of values for any particular parameter, reinforcing the conclusion that it is the combined set of parameters that is important. Clearly, such conclusions can be extended to the comparison of parameter sets within different model structures and to the simulation of different types of observed variables.

One response to this undoubtedly common behavior has been to seek to develop more robust optimization algorithms [see, e.g., *Duan et al.*, 1992]. Here an approach is taken whereby it is accepted that it may not be possible to distinguish between different models or parameter sets as simulators of a

[1]Now at College of Environmental Science and Forestry, Syracuse, New York.

Paper number 95WR03723.
0043-1397/96/95WR-03723$09.00

particular catchment. Thus the GLUE methodology has been developed from an acceptance of the possible equifinality of models (where "model" will from now on be used to indicate the combination of a particular set of parameter values within some particular model structure). If there is no unique optimal model, then it is only possible to rank the model structures/ parameter sets considered on some likelihood scale as simulators of the system. Some models can certainly be rejected as "nonbehavioral" in the GSA terminology of *Spear and Hornberger* [1980] [see also *Whitehead and Hornberger*, 1984; *Hornberger et al.*, 1985]. One model, of course, will give the best fit to some period of calibration data, but there will be many others that will be almost as good, and many of these will be in very different parts of the parameter space. It is also almost certain that if a second period of data is considered, then the rankings of these possible models will change and that the best model found for the first period will not be the best for the second.

Consequently, if at least a sample of these acceptable models are retained, a range of possible behaviors will be available in making predictions with these models. In the GLUE approach these are weighted according to their calculated likelihoods from the calibration period(s), and the weights used to formulate a cumulative distribution of predictions from which uncertainty quantiles can be calculated. It is worth noting that the term "likelihood" is being used here in a wider context than the likelihood functions of classical statistics which make specific assumptions about the nature of the errors associated with the model simulations. *Romanowicz et al.* [1994] show that the classical likelihood functions can be used within the GLUE framework, although these functions, designed to find optimum parameter sets, give much greater weight to the better simulations and can result in response surfaces that are very peaked (but not necessarily with a single "optimum").

It is important to stress that the parameters are never considered independently but only ever as sets of values. The likelihood measure $L(\underline{\theta}_i|\underline{Y})$ for the ith model is associated with a particular set of parameters $\underline{\theta}_i$ conditioned on the observed data variables $\underline{Y}$. It is possible to evaluate the sensitivity of individual parameters, either by looking at the distributions of "behavioral" and "nonbehavioral" models defined by the associated likelihood weights as in the GSA proceedures of *Spear and Hornberger* [1980] (see also the hydrological example of *Beven and Binley* [1992]); or by evaluating the marginal distribution of likelihood for each parameter by integrating across the parameter space [see *Romanowicz et al.*, 1994].

The GLUE methodology also recognizes that as more data or different types of data are made available, it may be necessary to update the likelihood weights associated with different simulations. This is achieved quite easily using Bayes's equation, which allows a prior distribution of likelihood weights to be modified by a set of likelihood weights arising from the simulation of a new data set to produce an updated or posterior likelihood distribution (see below). Since each model is associated with a particular value of the chosen likelihood measure, the modification of the prior likelihood is achieved by applying Bayes equation to each retained model independently. It has been found in applying this procedure that this tends to have the effect of reducing the number of retained models with significant likelihood values; that is, as would be hoped, more data will usually reduce the feasible region of the parameter space. A corollary of this is that in order to retain a sufficient sample of models for estimation of uncertainty bounds it may be necessary to resample the parameter space as described by *Beven and Binley* [1992].

GLUE requires considerable computing resources. Thousands of model realizations may be necessary to characterize the parameter space adequately, while the results of each retained simulation must be stored to calculate the uncertainty bounds. It is, however, conceptually very simple and can be applied to models of arbitrary degrees of complexity and nonlinearity. It has been applied to both fully distributed "physically based" models [*Binley and Beven*, 1991; *Beven and Binley*, 1992] and other TOPMODEL applications [*Beven*, 1993; *Romanowicz et al.*, 1994]. The calculations presented here have been carried out on the Lancaster University 80 processor Meiko transputer system hosted by a Sun workstation. The GLUE calculations are easily implemented on parallel computing systems, especially when (as here) each individual realization can be made on a single processor.

Requirements of GLUE

The GLUE procedure requires first that the sampling ranges be specified for each parameter to be considered. This is not normally too great a problem, since initially the ranges can be set to be as wide as considered feasible by physical argument or experience. It would normally be hoped that models at the edges of the parameter ranges will be rejected as nonbehavioral after the first period of calibration (although our experience suggests that this is not always found to be the case for hydrological models, as shown in Figure 1).

Second, a methodology for sampling the parameter space is required. In most of the applications of GLUE carried out to date this has been done by Monte Carlo simulation, using uniform random sampling across the specified parameter range. The use of uniform sampling makes the procedure simple to implement and can be retained throughout the updating of likelihoods, since the density distribution of the likelihoods is defined by the likelihood weight associated with each model. Effectively, each set of chosen parameter values is always evaluated separately as a set, and its performance implicitly reflects any interactions and insensitivities of the parameters. This thereby avoids the necessity to sample from some multivariate set of correlated distributions for the parameter values.

Third, the procedure requires a formal definition of the likelihood measure to be used and the criteria for acceptance or rejection of the models. This is a subjective choice, as with any choice of objective function. There may also be more than one objective function calculated from different types of data, and it will then be necessary to specify how these will be combined. Bayes's equation provides a consistent framework for using these likelihood measures in both combining likelihoods calculated from different types of observed data and in updating the likelihoods associated with each model as more data become available. The application to the Ringelbach presented here explores the use of different likelihood measures and of updating likelihoods as new data become available, but not of combining likelihoods from different types of data.

Application of GLUE to TOPMODEL Simulations of the Ringelbach Catchment

Two companion papers [*Ambroise et al.*, this issue (a, b)] have shown how the TOPMODEL concepts [*Kirkby*, 1975; *Beven and Kirkby*, 1979; *Beven et al.*, 1995] can be generalized

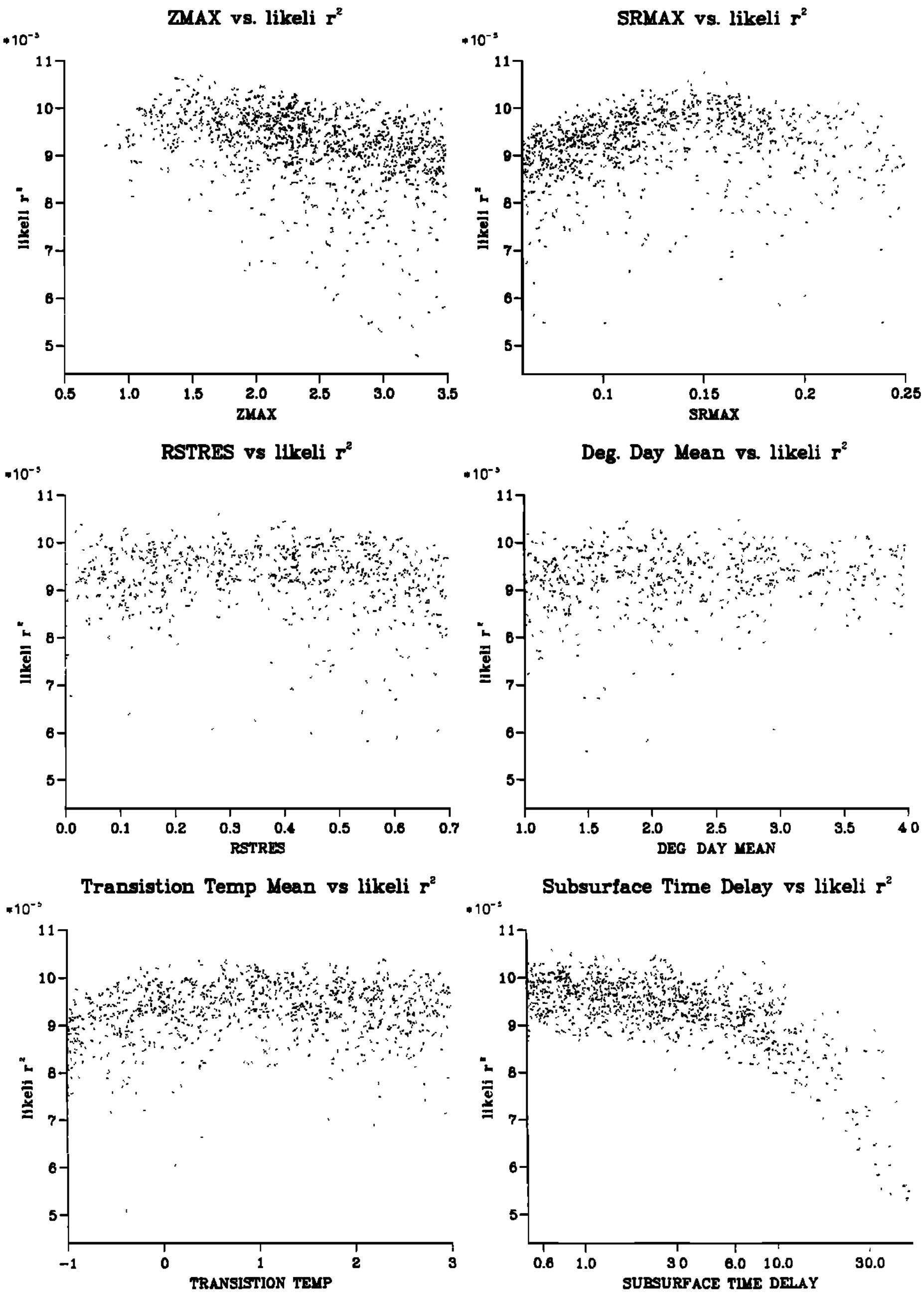


Figure 1. Scatter plot of efficiency results for each parameter in Monte Carlo simulations of the Ringelbach catchment; likelihood measure of equation (1) with $N = 1$.

to different types of soil transmissivity profile while retaining the use of the soil-topographic index of hydrological similarity as a means of distributing calculations made with catchment or subcatchment average variables back to the pattern of local responses. In this way it has been possible to construct a more realistic model of the small Ringelbach catchment in the Vosges, France. This model is based on a parabolic transmissivity profile and on a modification of the soil-topographic index distribution based on observed patterns of saturated contributing areas. Good results were obtained in simulating 4 years of detailed data using an 18-min time step with a single "optimal" parameter set [*Ambroise et al.*, this issue (b)].

The parameters of the model varied in the Monte Carlo realizations are shown in Table 1, together with the respective ranges and sampling strategy. These ranges were chosen on the basis of physical argument to be wider than the expected possible values for the catchment. To reduce the dimensionality of the parameter space, all other parameters of the model were kept at their "best guess" values given by *Ambroise et al.* [this issue (b)]. One change was made to the parameterization. Rather than specifying separate threshold temperatures for rain and snow for incoming precipitation, a mean temperature $T_{r/s}$ was allowed to vary with a constant range of 2°C between 100% snow and 100% rain.

Data Available

The GLUE methodology has been applied to the same 4 years of record used in the previous application of TOP-

Table 1. Parameter Ranges Used in Monte Carlo Simulations of the Ringelbach Catchment

Parameter	Minimum Value	Maximum Value	Sampling Strategy
m, or ZMAX, m	0.132	0.924	Uniform
S_{rmax}, m	0.06	0.25	Uniform
t_d, h/m	0.05	50.0	Uniform log values
R_{ws}, or *R*STRESS, m	0.0	0.7	Uniform
R_m, mm/°C/d	1.0	4.0	Uniform
$T_{r/s}$, °C	−1.0	3.0	Uniform

All parameters are defined by *Ambroise et al.* [this issue (b)].

MODEL to the Ringelbach catchment. Details of the hydrology for each of the 4 years are given by *Ambroise et al.* [this issue (b)]. Here the individual years are used in a sequential manner, updating the likelihood distributions associated with the model simulations after each year.

The Results of Using Different Likelihood Measures

The results of the GLUE methodology depend very much on the choice of likelihood measures used. Examples of possible likelihood measures are discussed by *Beven and Binley* [1992] and *Romanowicz et al.* [1994]. There are two related aspects to this choice: the choice of a particular measure of behavior and also the choice of criteria for rejecting simulations as nonbehavioral on the basis of that measure. The definition of the ranges of parameter values to be considered (Table 1) already involves an explicit a priori likelihood evaluation of 0 for parameter values outside those ranges.

The results to be presented in this study make use of the sum of squared errors as the basic likelihood measure, in the form

$$L(\underline{\theta}_i|\underline{Y}) = (1 - \sigma_i^2/\sigma_{\text{obs}}^2)^N \qquad \sigma_i^2 < \sigma_{\text{obs}}^2 \qquad (1)$$

where $L(\underline{\theta}_i|\underline{Y})$ is the likelihood measure for the ith model conditioned on the observations, σ_i^2 is the associated error variance for the ith model, σ_{obs}^2 is the observed variance for the period under consideration, and N is a parameter. For N equal to 1, (1) is equivalent to a coefficient of determination or the *Nash and Sutcliffe* [1970] efficiency criterion. Higher N values have the effect of accentuating the weight given to the better simulations. An extreme case of this is the likelihood function used by *Romanowicz et al.* [1994], based on a first-order autocorrelated error model, which involves raising the error variance to the power of one half of the number of time steps. Results will be presented here for different values of N, and also for the function

$$L(\underline{\theta}_i|\underline{Y}) = \exp\{-N\sigma_i^2/\sigma_{\text{obs}}^2\} \qquad \sigma_i^2 < \rho_{\text{obs}}^2 \qquad (2)$$

which has the feature that in applying Bayes's equation for updating of the likelihood weights, the error variance for each period of data is given equal weight (see below).

Scatter plots for a likelihood measure based on (1) with $N = 1$ are shown in Figure 1 for each of the parameters sampled, while Figure 2 shows, for the ZMAX parameter, the effects of using different likelihood measure definitions. In all cases the likelihood measures have been evaluated on the fit to the daily discharges, accumulated from the 18-min model time steps. In Figure 2e only, the variances in (1) have been calculated using the log discharges. In Figure 1 it can be seen that for both measures, good and poor simulations are available virtually throughout the parameter ranges. It suggests that the parameter response surface is very complex and emphasizes the need to treat the parameters as sets of values: the value of a single parameter clearly has little meaning in fitting the model to the data when taken outside the context of the values of the other parameters. The change in the nature of these distributions with a high value of N is shown in Figure 2b, where the effect of N in accentuating the likelihoods associated with the better simulations is seen quite clearly. It can be inferred that a higher value of N will tend to have the effect of producing narrower confidence limits in the predictions which may or may not be acceptable in comparison with the observations. This is discussed further below.

Calculation of Likelihood Distributions and Uncertainty Bounds

The next stage involves a decision about the criterion for model rejection. In most of the studies to date, a simple threshold on the likelihood measure has been used, though with very high values of N this may not be necessary since poor simulations will, in any case, end up with very low weights. The uncertainty bounds associated with the retained simulations will depend on the choice of the likelihood measure (including the value of N) and rejection criterion. In this study the choice of a low model efficiency rejection criterion (<0.3 for $N = 1$) has been included, since this is one way of trying to ensure that the uncertainty bounds are wide enough to encompass most of the observed discharges during calibration periods.

Following rejection, the likelihood weights associated with the retained models can be rescaled to give a cumulative sum of 1.0. The rescaled weights can then be applied to their respective model discharges at each time step to give a cumulative distribution of discharges at that time step, from which the chosen discharge quantiles can be calculated to represent the model uncertainty. The results of this process for different rejection criteria and likelihood measures are shown in Figure 3 for part of the 1992 period. As noted by *Ambroise et al.* [this issue (b)], this year of data was the most difficult to model for the Ringelbach. In each case the uncertainty bounds are based on likelihood values calculated for the whole of the 1992 data set. It can be seen that the greater definition of the modal parameter region gained with high values of N has a relatively small effect on the uncertainty bounds (Figure 3b for $N = 30$). The higher value of N narrows the estimated uncertainty on the predicted discharges but this also has the result that more of the observed discharges then fall outside of the plotted 90% quantiles. This reflects the expected dominance of model structural error and input data errors over errors associated with the calibration process itself.

Figures 4 and 5 show, for selected high and low accumulated daily flows from this same period, the distribution of the predicted discharges weighted by their likelihood values for the same set of likelihood measures. The significant effect of a high value of N is seen in these distributions but also of interest is the non-Gaussian nature of the distributions and the changing shape and variance with discharge. This should not, perhaps, be unexpected with this type of nonlinear model. The very irregular distributions resulting from the use of the log discharge efficiency measure are notable (Figures 4e and 5e).

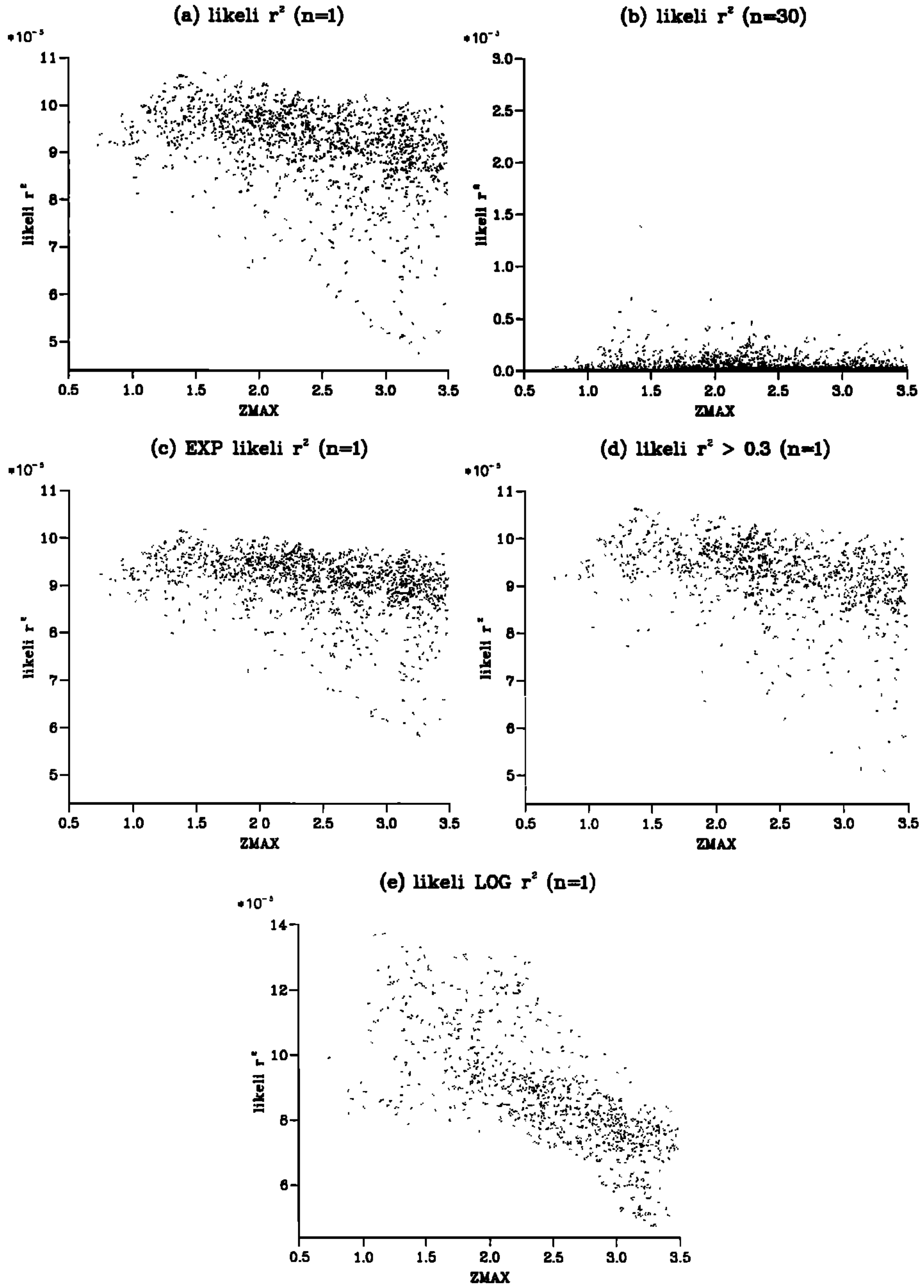


Figure 2. Scatter plot of likelihood values for different definitions of the likelihood measure and rejection criterion for simulations of the Ringelbach catchment. (a) Equation (1) with $N = 1$ and rejection of values <0.5. (b) As in Figure 2a but with $N = 30$. (c) Equation (2) with $N = 1$. (d) As in Figure 2a but with rejection of values <0.3. (e) As in Figure 2a but based on fitting log discharges.

Updating of Uncertainty Bounds

Updating of the likelihood distribution as more data becomes available may be achieved by the application of Bayes equation in the form

$$L(\underline{Y}|\underline{\theta}_i) = L(\underline{\theta}_i|\underline{Y})L_o(\underline{\theta}_i)/C \qquad (3)$$

where $L_0(\underline{\theta}_i)$ is a prior likelihood for the parameter set $\underline{\theta}$, $L(\underline{\theta}_i|\underline{Y})$ is the likelihood measure calculated with the set of observed variables $\underline{Y}$, $L(\underline{Y}|\underline{\theta}_i)$ is the posterior likelihood for the simulation of $\underline{Y}$ given $\underline{\theta}$ and C is a scaling constant calculated such that the cumulative of $L(\underline{Y}|\underline{\theta}_i)$ equals unity. The procedure outlined in the previous section can be considered as an application of (3) with a prior distribution of equal likelihoods for every model and a zero likelihood measure for those models rejected as nonbehavioral. Clearly, it is not necessary that a uniform initial prior distribution be used (this is not necessarily even the best noninformative prior for certain functions; see, for example, work by *Lee* [1992]); prior beliefs about appropriate parameter values could also be incorporated.

However, given the results from the first period of simulation, the resulting posterior distribution can be used as the

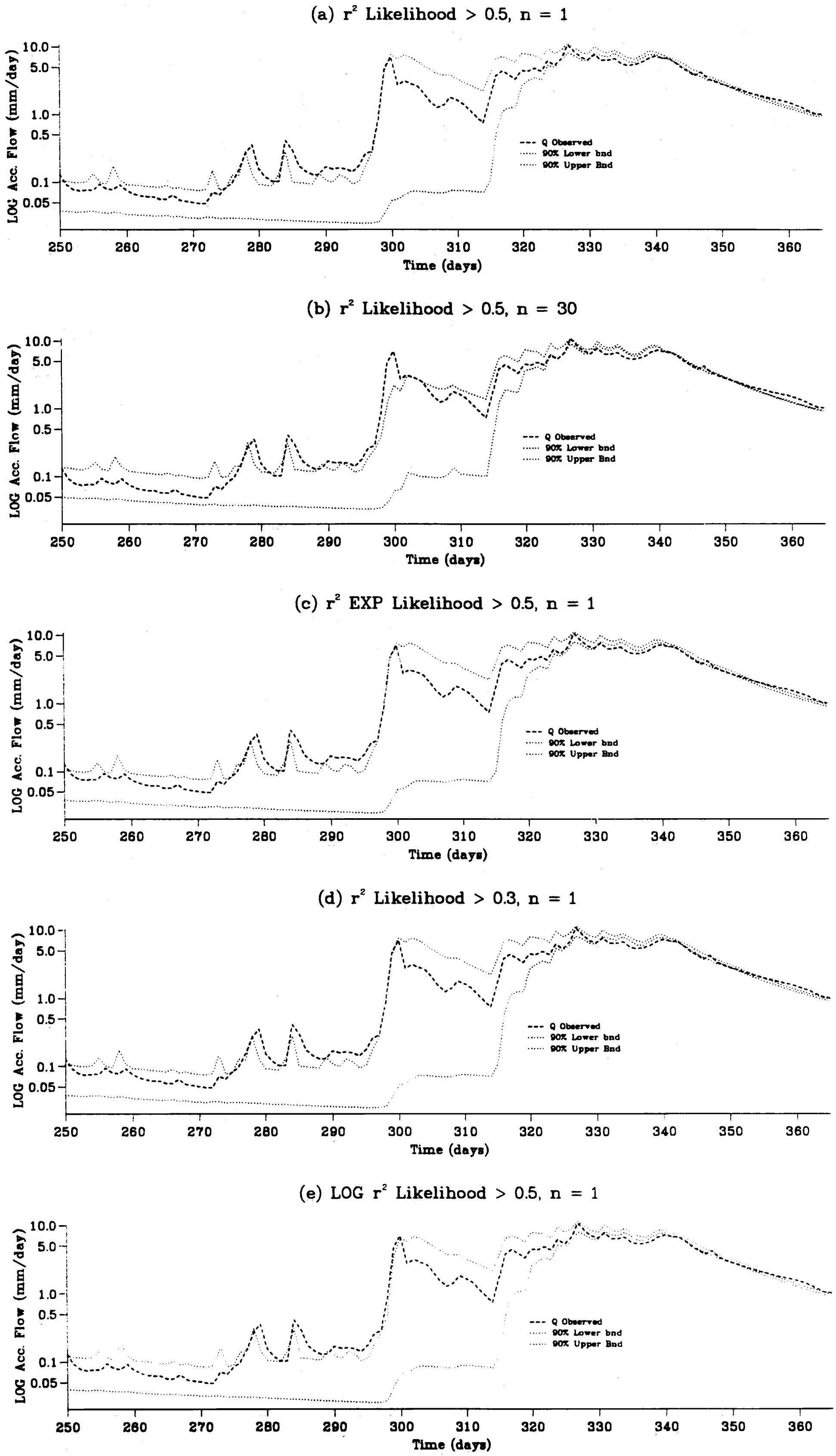


Figure 3. Ninety percent uncertainty bounds for simulations for part of 1992 in the Ringelbach catchment using same likelihood criteria defined in Figure 2.

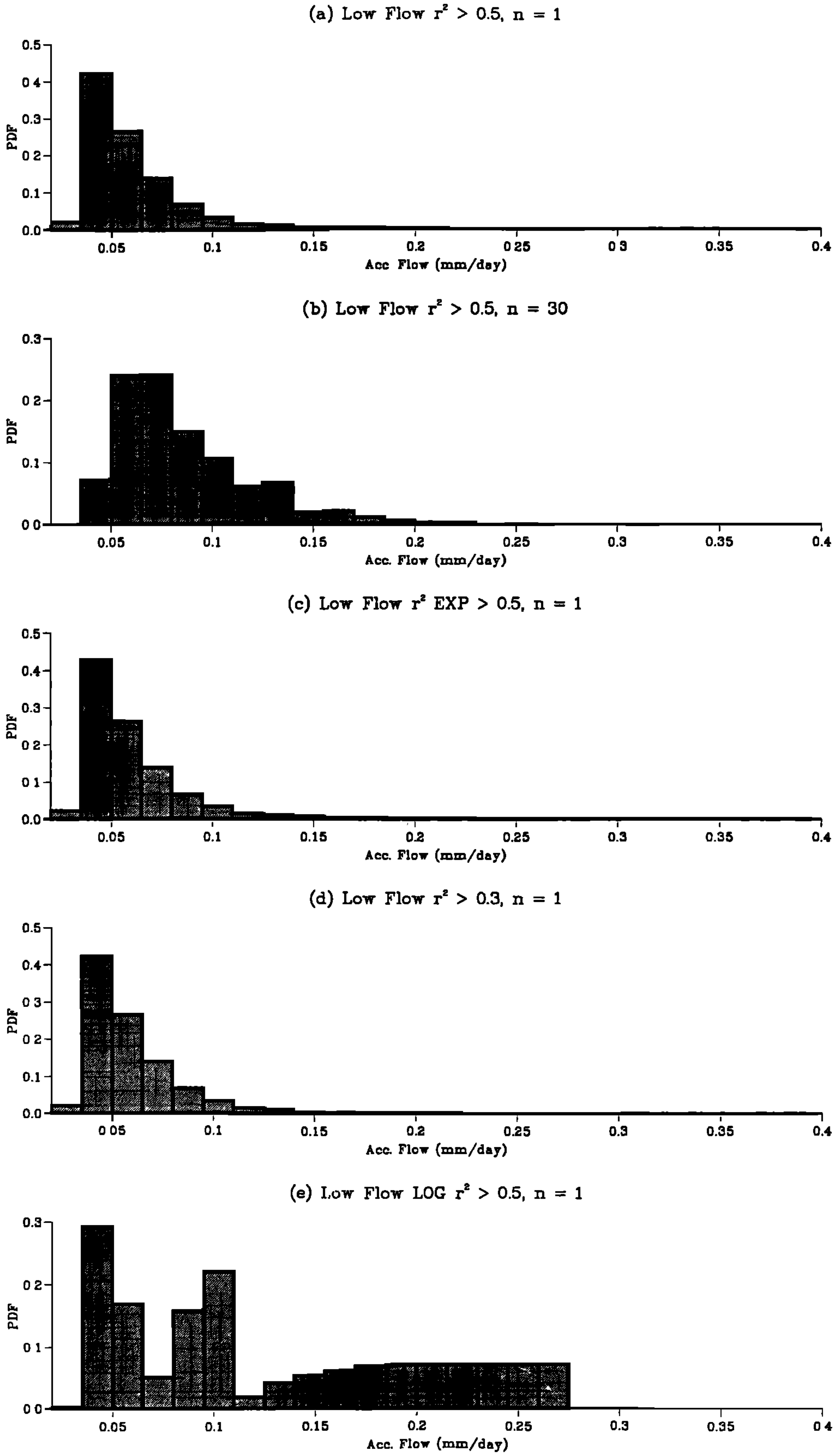


Figure 4. Ringelbach catchment simulations: likelihood weighted distributions of predicted accumulated discharges for day 259 of 1992 (observed flow equals 0.093 mm/day). Likelihood measures as defined in Figure 2.

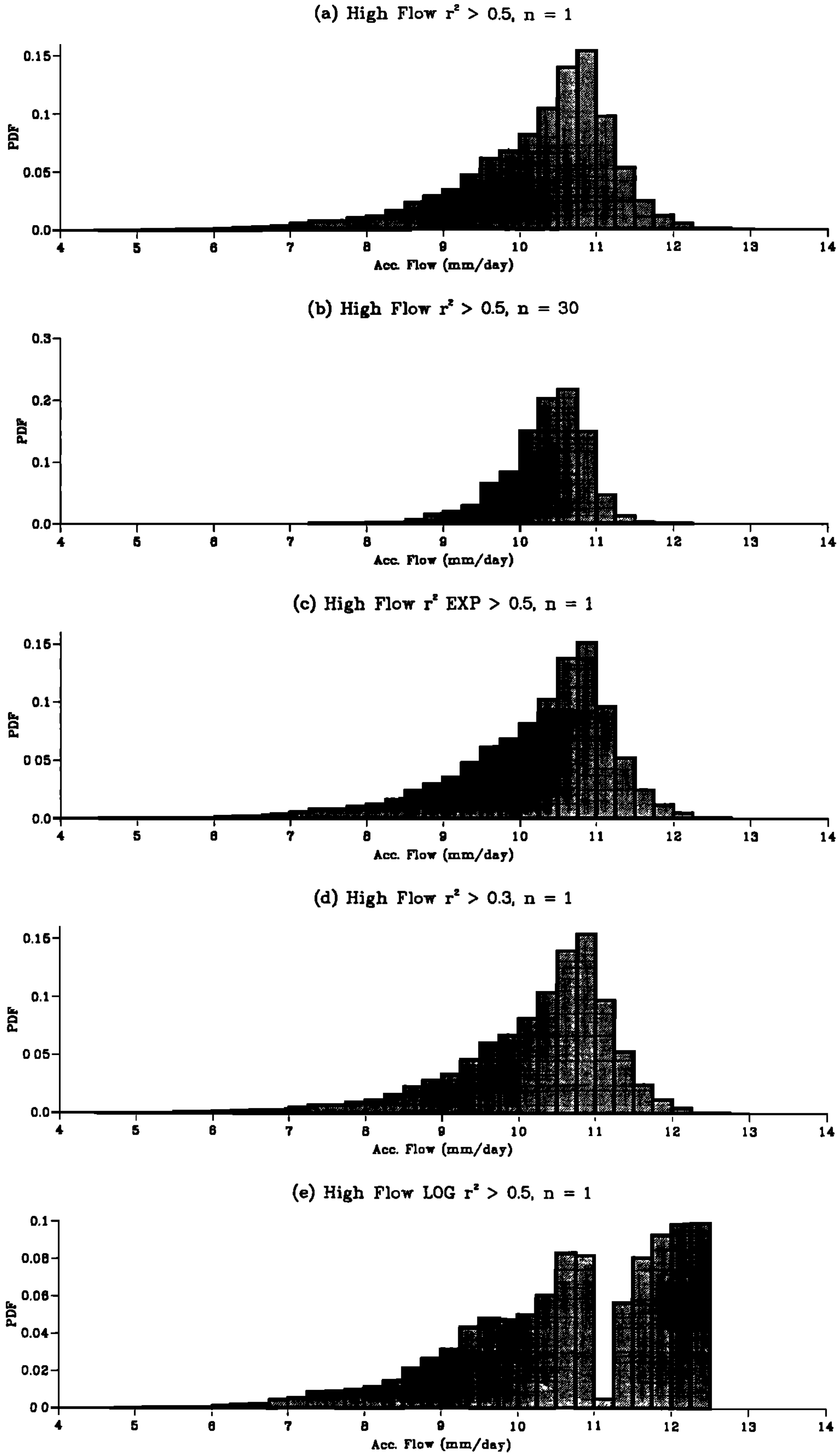


Figure 5. Ringelbach catchment simulations: likelihood weighted distributions of predicted accumulated discharges for day 327 of 1992 (observed flow equals 10.744 mm/day). Likelihood measures as defined in Figure 2.

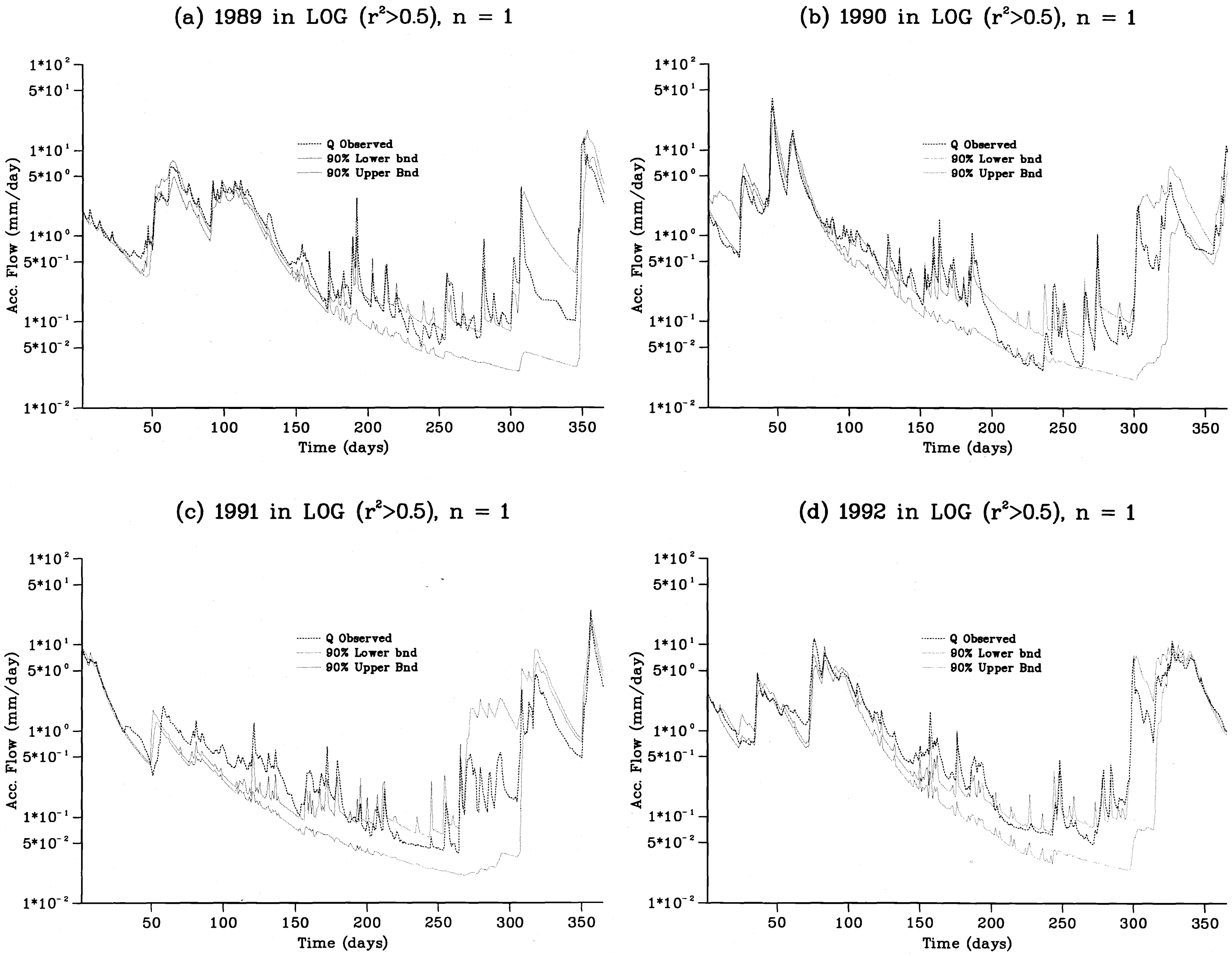


Figure 6. Ringelbach catchment simulations: 90% uncertainty bounds for 4 years of record with updating after each year, using likelihood measure of equation (1) with $N = 1$ and rejection criterion of <0.5.

Table 2. Ringelbach Catchment Simulations

Year(s)	$N = 1$	$N = 30$
1989	13.46	12.19
1989–1990	13.46	11.71
1989–1991	13.45	11.01
1989–1992	13.44	10.79

Calculated Shannon entropy values for posterior likelihood measures based on equation (1) with $N = 1$ and $N = 30$ after updating with the addition of each additional year of data.

prior distribution for the next period. If, in fact, further observations from the next period are available, the resulting likelihood values can be used with (3) to update the posterior likelihood distribution for any further predictions. An important difference between the likelihood measures defined by (2) and (3) becomes apparent here. Because of the multiplicative nature of (3), the use of (1) as a likelihood measure means that after a number of updatings, the error variance associated with earlier periods carries less and less weight in the calculation of the posterior likelihoods. The use of (2), however, allows the error variances from additional periods to contribute linearly within the exponential. Thus only with (2) will the same likelihood distributions be obtained from using a period of data as a single entity compared with splitting it down into smaller periods and applying Bayes's equation after evaluating the likelihoods after each period. It is worth adding, however, that this feature of (1) might be desirable when a catchment (and by inference the model parameters) is undergoing gradual change.

Figure 6 shows the results of using likelihood measure (1) in predicting the hydrographs for the complete 4 years of record using posterior distributions updated using the year of observations. Computing constraints meant that all the retained (behavioral) models were reinitialized using the observed discharge at the start of each year. It can be noted from Figure 6 that the observed discharge is generally contained within the calculated uncertainty bounds, but not always, and for some periods departs quite markedly from the range of model predictions, particularly in 1991. This is a reminder that this type of analysis is not a complete panacea for errors arising either from the specification of the input data, or for model structural errors. Some of the most prominent departures, for example, were associated with snowmelt events, which are known to be difficult to predict correctly in both amount and timing.

There is a tendency for the uncertainty bounds to become wider with increasing flow, but one interesting feature of the Figure 6 plots is the wide uncertainty associated with the events occurring after the long summer dry period. These are also notoriously difficult to predict but in all 4 years (with the exception of one peak in 1991) the observed discharges are contained within the uncertainty bounds.

On the Value of Additional Data

Beven and Binley [1992] have shown how the GLUE methodology can be used to evaluate the value of additional data in constraining the predictive uncertainty by the explicit definition of uncertainty measures. The Shannon entropy measure, H, given the set of scaled likelihood weights $L_i = L(\underline{Y}|\underline{\theta}_i)$, is defined as

$$H = -\sum_i L_i \log_2 L_i \tag{4}$$

Table 2 shows the results of using the Shannon entropy to evaluate the reduction in uncertainty as each year of data is added for likelihood values based on (1) with $N = 1$ and $N = 30$. The results show that for each likelihood measure, the addition of each year of data results in a fall in uncertainty and consequent narrowing of the uncertainty bounds. This fall is, however, relatively small, which is consistent with our finding that relatively few sets of simulations become "nonbehavioral" as each new year of data is added. It will be interesting to see if this trend continues as further data become available, perhaps with the inclusion of more extreme conditions.

On the Sensitivity to Individual Parameters

In the application of the GLUE methodology we have stressed that the behavior of the model is assessed in terms of the performance of sets of parameters. It is the combination of parameter values that produces behavioral or nonbehavioral simulations within the chosen model structure. The interaction between parameter values results in the broad regions of acceptable simulations when individual parameters are considered, as in Figures 1 and 2 [see also *Duan et al.*, 1992; *Beven*, 1993]. Such plots reveal little, however, about the sensitivity of the model predictions to the individual parameters, except where some strong change in the likelihood measure is observed in some range of a particular parameter.

More is revealed about sensitivity by an extension of the Generalised Sensitivity Analysis of *Spear and Hornberger* [1980]. They constructed distributions for each parameter conditioned on a classification of the Monte Carlo simulations into two classes; behavioral and nonbehavioral. The criterion for differentiating between the two classes was a subjectively chosen value of a goodness of fit measure, as we have used here. Similarity between the cumulative distributions for the two classes suggests insensitivity to that parameter; strong differences between the distributions reveals a sensitive parameter.

In fact, a more detailed analysis can reveal more subtle sensitivities within the model. Figure 7 shows the cumulative distributions for 10 sets of behavioral simulations (likelihood value >0.5 in (1) with $N = 1$) in the application of TOPMODEL to the Ringelbach. The 10 sets represent equal numbers of behavioral simulations (760 in each set), here equally weighted, with set 1 representing the highest likelihood values, and set 10 the lowest (but still >0.5). Four parameters are shown: m, $S_{\max}$, t_d, and R_{ws}. The last, R_{ws}, shows no sensitivity at all, whereas all the others show changing distributions as the likelihood values increase. There tends to be some refinement of the feasible range as the likelihood threshold increases from set 10 to set 1. However, even in the very best set of simulations these figures show that the distribution is spread across most of the range of values considered. Clearly, the complex parameter response surface does not result from insensitivity of the parameter values alone.

Conclusions

This paper has presented the results of applying the GLUE methodology to the TOPMODEL simulation of the small Ringelbach catchment in the Vosges, France. Some aspects of the methods are explored in further detail than before, in particular the comparison of different model rejection criteria and different likelihood measure definitions. It has been shown how both uncertainty bounds and the distributions of predicted discharges are directly affected by the choice of likelihood

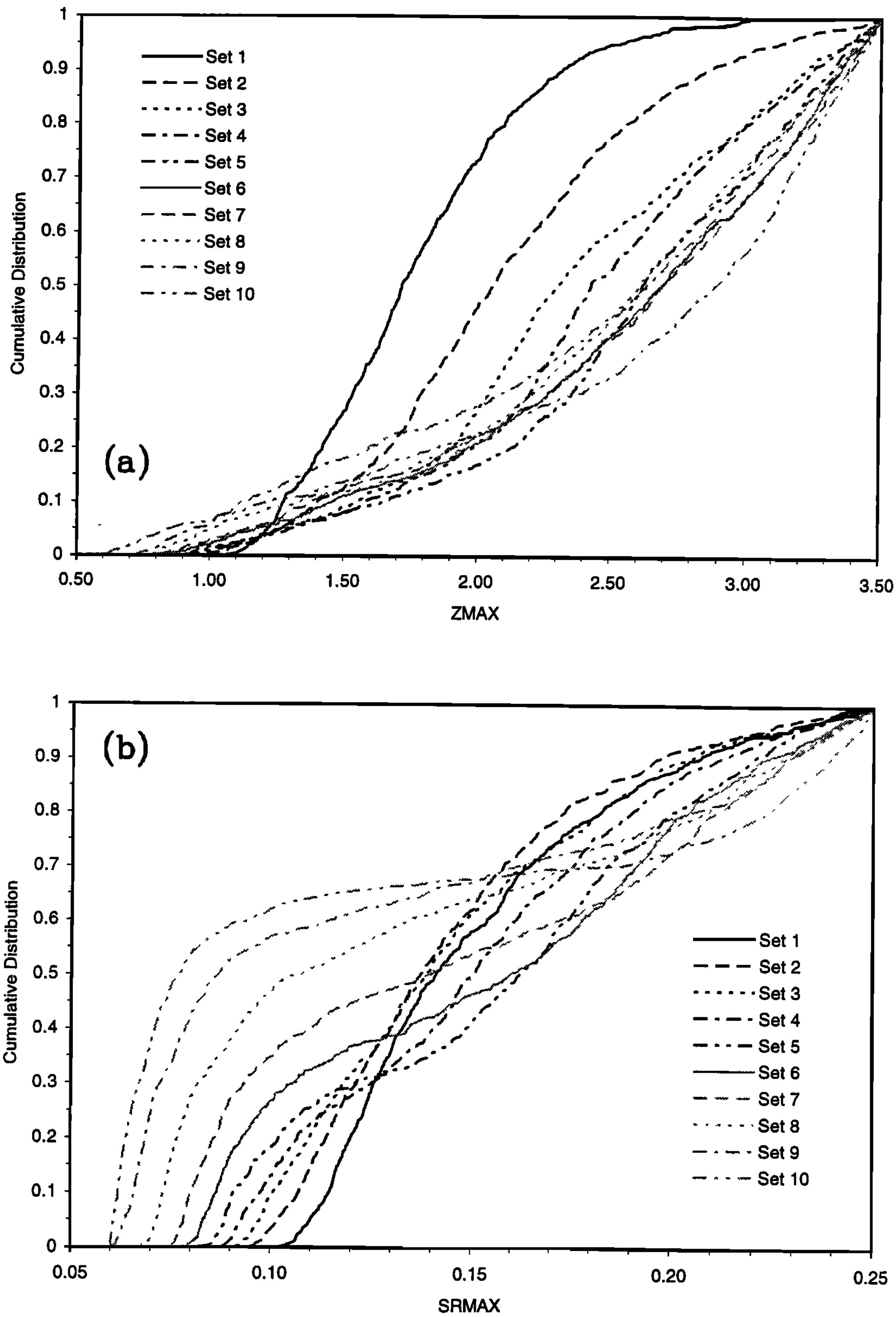
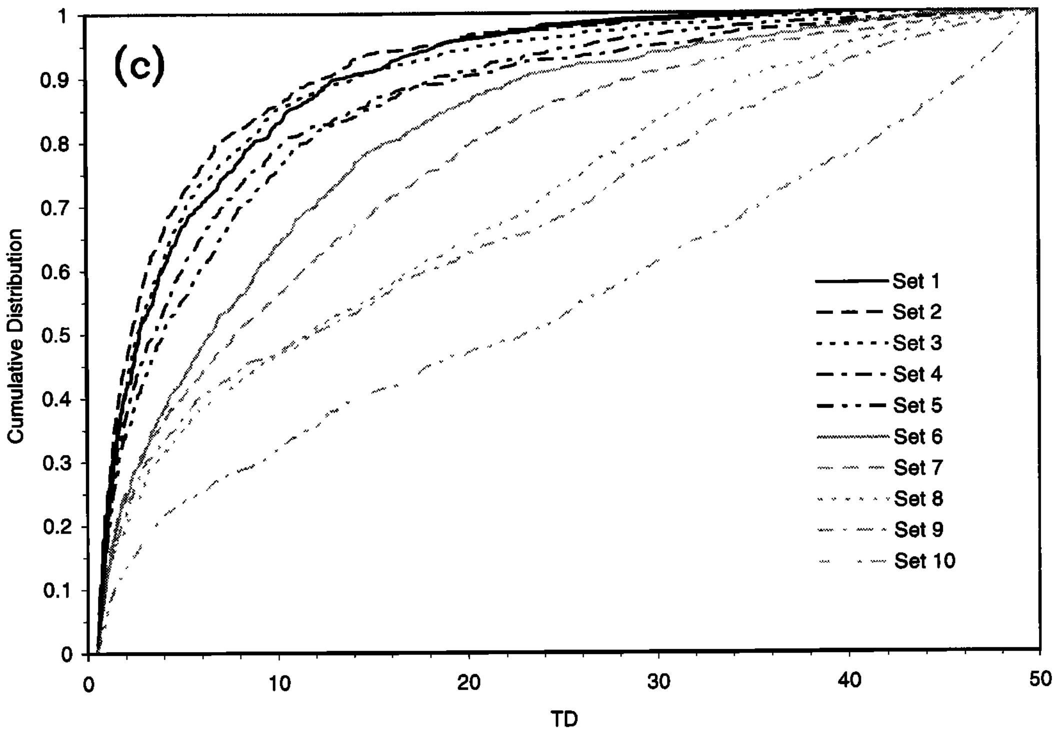


Figure 7. Sensitivity of individual parameters expressed as cumulative distributions of values in 10 equal sets (by number) of the retained "behavioral" simulations (likelihood based on equation (1) with $N = 1$; rejection of values <0.5) differentiated by likelihood value (set 1, highest values; set 10, lowest values). (a) ZMAX or m parameter; (b) $S_{r\max}$ parameter; (c) t_d parameter; and (d) RSTRESS or R_{ws} parameter.

measure. Those distributions of predicted discharges are inherently non-Gaussian and vary in shape through the simulated period, suggesting that evaluation of predictive uncertainty for the type of nonlinear model used here requires GLUE's type of Monte Carlo approach. Wide uncertainty bounds are predicted during some peak discharge periods and particularly for the first storms after the long dry summer period.



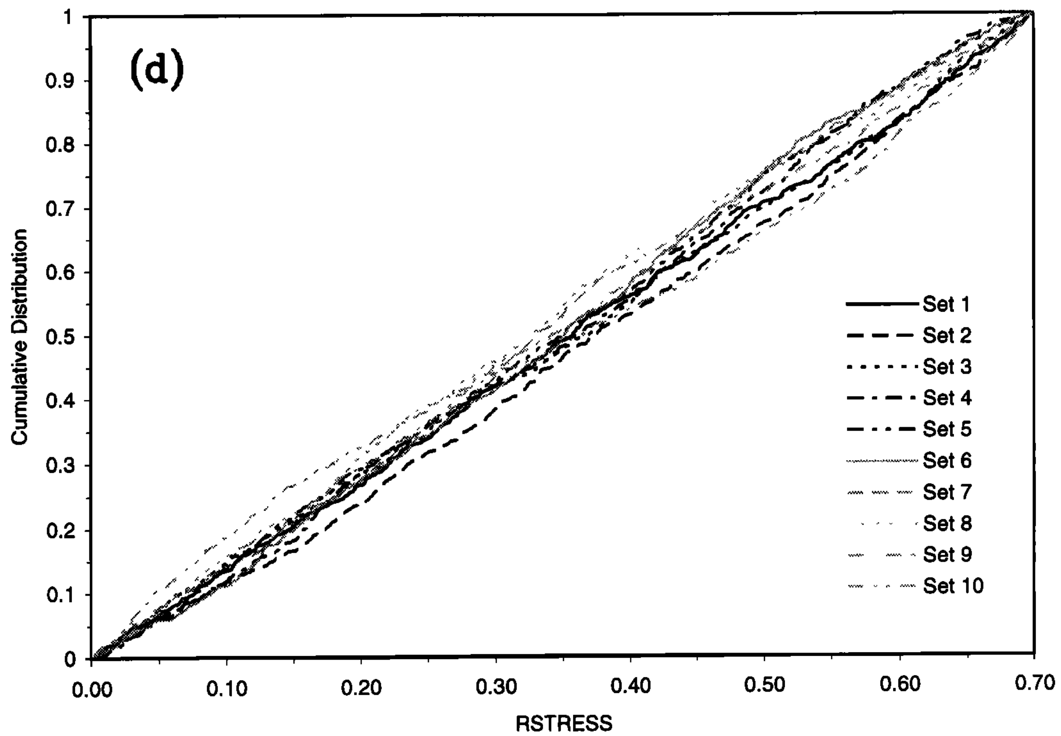


Figure 7. (continued)

The results presented have been shown to depend on the subjective choice of likelihood measure. Given the complex nature of the parameter response surface and modeling errors it would appear to be difficult to avoid some subjectivity in this choice, but those tried here have resulted in qualitatively reasonable uncertainty bounds. Marked departures of the observed discharges from the predicted uncertainty bounds are seen for some short periods of the record. This procedure cannot compensate completely for errors in either the input data or model structure. It can serve as a pointer to periods when either source of error may dominate the normal performance of the model.

It is interesting to note that the change in the estimated uncertainty bounds with the chosen likelihood measure is perhaps less than might be expected. This is because the procedure retains only those simulations that it is reasonable to consider "behavioral," as defined in terms of the likelihood measure used (all of which are here based on the sum of squared errors). Thus the different likelihood measures have essentially common sets of simulations; it is only the weighting of the tails that varies between them. If a very different criterion of behavior was used, a more significant effect might have been seen, as shown in Figures 4e and 5e. In this study, model performance has only been evaluated with respect to the predicted discharges. It is hoped in the future to extend the analysis to the comparison of observed and predicted internal state variables, using a Bayesian approach to combining different measures of performance.

It is worth noting that qualitative evaluations may also be included within this framework. *Beven* [1993] gives the example that if a model simulates a runoff response by an overland flow mechanism, but that this is considered to be an unimportant process in the catchment under study, then a binary likelihood measure can be used to distinguish models predicting overland flow responses (nonbehavioral) from those predicting subsurface flows (behavioral), even though those models predicting overland flow might yield good fits to the observed discharges. *Ambroise et al.* [this issue (b)] note that the use of internal state data is not easy, since the availability of such data will suggest spatial anomalies that will be better predicted if heterogeneities in parameter values (such as transmissivities) were allowed. This, however, will also have the effect of increasing the dimensionality of the parameter space. The use of spatial data for the calibration of distributed catchment models remains largely to be explored.

Acknowledgments. This research was funded by the STEP-ENCORE Programme of the EEC, the Programme PRH (CNRS-INSU), and by grant GR3/8950 of the UK Natural Environment Research Council. Field studies in the Ringelbach research catchment were funded by PIREN-Eau/Alsace Programme (CNRS, Ministère de l'Environnement, Région Alsace), the ATP "Bilan Hydrique" (CNRS-INRA-ORSTOM-CIRAD), the "Eau-Sol-Plante" Action (Programme Environnement du CNRS), and the "Bassins de Recherche Vosgiens" Programme (Ministère de la Recherche et de la Technologie).

References

Ambroise, B., K. J. Beven, and J. Freer, Toward a generalization of the TOPMODEL concepts: Topographic indices of hydrological similarity, *Water Resour. Res.*, this issue (a).

Ambroise, B., J. Freer, and K. J. Beven, Application of a generalized TOPMODEL to the small Ringelbach catchment, Vosges, France, *Water Resour. Res.*, this issue (b).

Beven, K. J., Interflow, in *Unsaturated Flow in Hydrologic Modelling*, edited by H. J. Morel-Seytoux, pp. 191–219, D. Reidel, Norwell, Mass., 1989a.

Beven, K. J., Changing ideas in hydrology: The case of physically based models, *J. Hydrol.*, *105*, 157–172, 1989b.

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

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

Beven, K. J., and M. J. Kirkby, A physically based variable contributing area model of basin hydrology, *Hydrol. Sci. Bull.*, *24*(1), 43–69, 1979.

Beven, K. J., R. Lamb, P. F. Quinn, R. Romanowicz, and J. Freer, TOPMODEL, in *Computer Models of Watershed Hydrology*, edited by V. P. Singh, pp. 627–668, Water Resour. Publ., Highlands Ranch, Colo., 1995.

Binley, A. M., and K. J. Beven, Physically-based modelling of catchment hydrology: A likelihood approach to reducing predictive uncertainty, in *Computer Modelling in the Environmental Sciences*, edited by D. G. Farmer and M. J. Rycroft, pp. 75–88, Clarendon, Oxford, 1991.

Blackie, J. R., and C. W. O. Eeles, Lumped catchment models, in *Hydrological Forecasting*, edited by M. G. Anderson and T. P. Burt, John Wiley, New York, 1985.

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

Grayson, R. B., I. D. Moore, and T. A. McMahon, Physically based hydrologic modeling, 2, Is the concept realistic?, *Water Resour. Res.*, *28*, 2659–2666, 1992.

Hornberger, G. M., K. J. Beven, B. J. Cosby, and D. E. Sappington, Shenandoah watershed study: Calibration of a topography-based, variable contributing area hydrological model to a small forested catchment, *Water Resour. Res.*, *21*, 1841–1850, 1985.

Kirkby, M. J., Hydrograph modelling strategies, in *Process in Physical and Human Geography*, edited by R. Peel, M. Chisholm, and P. Haggett, pp. 69–90, Heinemann, London, 1975.

Lee, P. M., *Bayesian Statistics: An Introduction*, Edward Arnold, London, 1989.

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

Romanowicz, R., K. J. Beven, and J. Tawn, 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, New York, 1994.

Spear, R. C., and G. M. Hornberger, Eutrophication in Peel Inlet, II, Identification of critical uncertainties via Generalised Sensitivity Analysis, *Water Res.*, *14*, 43–49, 1980.

Spear, R. C., T. M. Grieb, and N. Shang, Parameter uncertainty and interaction in complex environmental models, *Water Resour. Res.*, *30*(11), 3159–3170, 1994.

Stephenson, G. R., and R. A. Freeze, Mathematical simulation of subsurface flow contributions to snowmelt runoff, Reynolds Creek watershed, Idaho, *Water Resour. Res.*, *10*, 284–298, 1974.

Whitehead, P. G., and G. M. Hornberger, Modeling algal behavior in the River Thames, *Water Res.*, *18*, 945–953, 1984.

B. Ambroise, Centre d'Etudes et de Recherches Eco-Géographiques (CEREG, URA 95 CNRS), 3 rue de l'Argonne, F-67083 Strasbourg Cedex, France. (e-mail: ambroise@geographie.u-strasbg.fr)

K. Beven, Centre for Research on Environmental Systems and Statistics, Institute of Environmental and Biological Sciences, Lancaster University, Lancaster LA1 4YQ, UK. (e-mail: k.beven@lancaster.ac.uk)

J. Freer, College of Environmental Science and Forestry, Marshall, Syracuse, NY 13210. (e-mail: jefreer@mailbox.syr.edu)

(Received November 21, 1994; revised December 4, 1995; accepted December 6, 1995.)