Flexible automated parameterization of hydrologic models using fuzzy logic

Sudeep Samanta

Department of Forest Ecology and Management, University of Wisconsin-Madison, Madison, Wisconsin, USA

D. Scott Mackay

Department of Forest Ecology and Management and Institute for Environmental Studies, University of Wisconsin-Madison, Madison, Wisconsin, USA

Received 1 April 2002; revised 1 July 2002; accepted 8 July 2002; published 16 January 2003.

[1] Recent developments in model calibration suggest that information obtained from calibration is inherently uncertain in nature. Therefore identification of optimum parameter values is often highly nonspecific. A calibration framework using fuzzy logic is presented to deal with such uncertain information. An application of this technique to calibrate the streamflow of a hydrologic submodel embedded within an ecosystem simulation model demonstrates that objective estimates of parameter values and the range of model output associated with a failure to identify a unique solution can be obtained with suitable choices of objective functions. An iterative refinement in parameter estimates through a process of elimination was possible by incorporating multiple objective functions in calibration, thereby reducing the range of parameter values that capture the streamflow response. It is shown that objective function tradeoffs can lead to suboptimal solutions using the process of elimination without an automated procedure for reevaluation. Owing to its computational simplicity and flexibility this framework could be extended into a nonmonotonic system for automated parameter estimation. *INDEX TERMS:* 1894 Hydrology: Instruments and techniques; 1860 Hydrology: Runoff and streamflow; 1869 Hydrology: Stochastic processes; *KEYWORDS:* automated parameter estimation, hydrologic models, fuzzy logic, Monte Carlo sampling

Citation: Samanta, S., and D. S. Mackay, Flexible automated parameterization of hydrologic models using fuzzy logic, *Water Resour. Res.*, *39*(1), 1009, doi:10.1029/2002WR001349, 2003.

1. Introduction

[2] Most hydrological models are conceptual representations of ideal hydrological systems involving varying degrees of simplifications. Application of such models requires determination of appropriate parameter values defined in the model. Even for highly physically based models, it may be impossible to obtain direct measurements for all of the required parameter values due to spatial heterogeneity [*Beven*, 1989; *Binley and Beven*, 1991]. Some of the required parameters may not even be directly measurable in the form used in the model. However, it is necessary to estimate applicable values for them in order to utilize a hydrological model, and to draw useful conclusions regarding relationships that may exist between model parameters and physical watershed characteristics [*Kuczera*, 1983]. Values for such parameters are usually estimated through a calibration process that matches simulated hydrologic fluxes with a time series of observed fluxes. In an automated calibration framework, this involves selecting a model (parameters and structure) from the feasible model-parameter space. Typically, a selection is made based on the lowest degree of mismatch between simulation and observations. However, output from hydrological models with widely different parameter sets may produce nearly equal levels of measured degree of fit, making it difficult to select from among near optimal parameter sets [*Beven*, 1993]. One reason for this nonunique solution to the calibration exercise is that model parameters can compensate for each other. Furthermore, the effects of poor representation of processes within the model structure, or lack of adequate data, can sometimes be compensated for with an adjustment of parameters. For example, stream discharge data represents an accumulation of spatially variable fluxes into a single point value that cannot resolve either the individual fluxes in space or their errors. As a result, calibration using a different data set [*Beven*, 1993; *Melching*, 1995] or a different set of objective functions for evaluating model performance [*Gupta et al.*, 1998] may result in different optimal parameter sets for a model.

[3] Several calibration frameworks have been proposed for simulation models in general and hydrological models in particular, which recognize this uncertainty in calibrated parameter values, and consequently in model results [*Spear and Hornberger*, 1980; *Klepper et al.*, 1991; *Van Stratten and Keesman*, 1991; *Beven and Binley*, 1992; *Gupta et al.*, 1998; *Kuczera and Parent*, 1998]. The solutions proposed to account for this uncertainty often accept a set of models out of the model population instead of a single optimal

0043-1397/03/2002WR001349$09.00

model as the result of calibration. The idea that such a set is representative of the uncertainty in parameter values has been used in, e.g., Pareto optimal parameter sets [*Gupta et al.*, 1998]; equifinal parameter sets [*Beven and Binley*, 1992], and others. When the solution set is in the domain of a finite model population (e.g., a parameter sampling situation), then the cardinality (or number of members) of the solution set can provide a measure of uncertainty in the applicable parameter values.

[4] In this paper, we describe a method that determines this cardinality by treating the solution set of models as a fuzzy set [*Zadeh*, 1965]. This fuzzy set is transformed into a crisp set of models from which parameter variability is derived. The range of output from this crisp set of models is then used to represent the consequence of imperfect parameter identification. The approach is presented first. We then illustrate the technique with an application of the Regional Hydro-Ecological Simulation System (RHESSys) [*Mackay and Band*, 1997; *Mackay*, 2001] to a stream discharge data set from the H. J. Andrews Long Term Ecological research station.

2. Methods

2.1. Fuzzy Calibration Framework

[5] Comparison and subsequent selection of the optimum model in a traditional automated calibration system is based on the following premise:

$$x_i, x_j \in M : f(x_i) > f(x_j) \Rightarrow B(x_i, x_j) \tag{1}$$

$B(x_i, x_j)$ means x_i is a better model of the system than x_j, where x_i and x_j are two different models (i.e., structure plus parameter set) in the discrete set of feasible models M, and f is an objective function that is maximized for calibration. This premise holds if the relative ranking of x_i and x_j is not sensitive to a change in the definition of f or the calibration data. For reasons discussed in the introduction, this assumption is not strictly valid for most calibration problems and a set of models must be accepted as the solution to recognize the inherent uncertainty. Consequently, the goal of calibration may be stated as that of identifying this solution set.

[6] For a calibration process to be successful, some amount of information regarding relative suitability of models obtained from f must be independent of calibration data or specific objective function. So, the value of $f(x_i)$ would provide an estimate of the possibility that x_i is an acceptable model in relation to the other models in M. This notion is used in the generalized likelihood uncertainty estimation (GLUE) framework [*Beven and Binley*, 1992] where f values are scaled to sum to one and interpreted as the approximate likelihood of a model to be optimal. The acceptable model (or equifinal) set is then defined by a threshold value of f based on a statistical confidence limit. However, in most cases the errors from a hydrologic model do not display any fixed probabilistic properties. As a result, the probabilistic interpretation of the scaled objective function value is too restrictive and using a threshold on the objective function value to obtain the set of acceptable models under this assumption may seem arbitrary [*Gupta et al.*, 1998].

[7] An alternative to setting such a threshold is proposed by *Franks et al.* [1999] using a shaping function linked to the assumed information content of the data. However, in most cases, there is serial and cross correlation in the residuals, so the scaling factor associated with the shaping function needs to be determined empirically. In the present analysis, such a technique is not used in order to assess the information that can be obtained under a fuzzy set interpretation of objective functions without distorting the response surface.

[8] A set theoretic approach that represents parameter uncertainty is by identifying a pareto optimal set [*Yapo et al.*, 1998; *Gupta et al.*, 1998]. This pareto optimal set is associated with trade offs involved in using different objective functions used for calibration. The objective functions can differ in terms of formulation, e.g., DRMS, BIAS, and NSC [*Gupta et al.*, 1998], or in their use of independent data streams, e.g., catchment runoff and ground water levels [*Beldring*, 2002], peak flow and low flow RMSE [*Madsen*, 2000], subperiods of daily stream flow [*Boyle et al.*, 2000], and sensible heat, latent heat, ground temperature, and soil moisture [*Gupta et al.*, 1999]. Flexibility offered by this framework allows theoretically similar treatment of all such objective functions for construction of the pareto set. From theoretical considerations, the pareto optimal set grows in size with the number of objective functions used in calibration if significant tradeoffs exist among them. Consequently, a further manual step may be necessary for analysis and elimination of solutions within the tradeoff range [*Boyle et al.*, 2000] or the pareto set may need to be collapsed by using a suitably aggregated objective function [*Madsen*, 2000] depending on the application.

[9] An alternative interpretation is proposed here, in which the set of acceptable models is considered a fuzzy set. The boundary of the set of acceptable solutions to the calibration problem is considered fuzzy or uncertain. In contrast to the pareto set, the set of models representing uncertainty in the proposed approach is progressively constrained as new information is added in the form of additional objective functions. To characterize the fuzzy set, f is interpreted as a fuzzy membership grade function. In comparison, other calibration techniques have used fuzzy logic in different ways. For example, fuzzy disaggregation [*Franks and Beven*, 1997; *Franks et al.*, 1998; *Franks and Beven*, 1999; *Hankin and Beven*, 1998] uses a fuzzy measure, and the fuzzy logic based modeling approach [*See and Openshaw*, 2000] uses a fuzzy logic master model in a multimodel context.

[10] To reiterate, a set of models accepted as the solution to a calibration problem can be considered an expression of the inability to precisely identify a unique solution. This uncertainty arises from the fact that it is known that the model of interest belongs to the set of alternatives but cannot be specifically identified. This type of uncertainty is called nonspecificity and can be estimated by the Hartley function [*Hartley*, 1928] in the context of crisp sets. The Hartley function is a measure of the additional information that is required to remove this nonspecificity. It is defined as

$$H(A) = \log_2 |A|, \tag{2}$$

where $H(A)$ is the Hartley function for a finite crisp set A and $|A|$ is its cardinality. The greater the cardinality of the

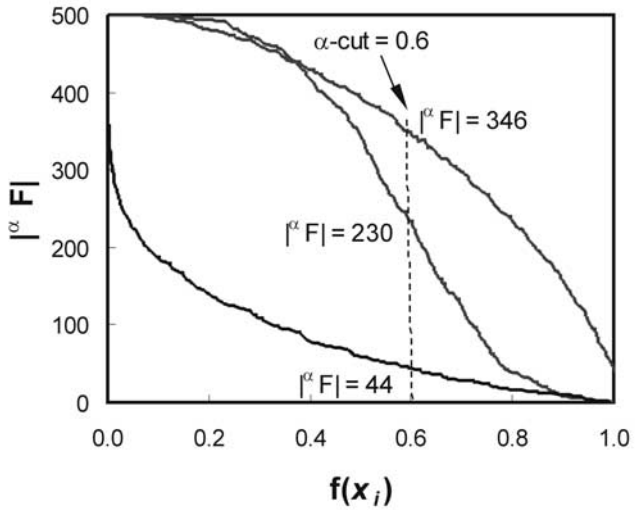


Figure 1. These hypothetical distribution functions illustrate how cardinality of a candidate set of models varies by value obtained from a measure of goodness of fit. At a given level (or α-cut), a low cardinality indicates a high level of specificity in the set of candidate models. A high cardinality indicates a high level of nonspecificity.

retained model set in proportion to the model population, the greater the nonspecificity in the model calibration. When the acceptable set of models is considered a fuzzy set, *F*, within the domain, *X,* of all feasible models, uncertainty related to the cardinality of *F* is expressed as a measure of the nonspecificity of *F*. One measure of nonspecificity in the domain of fuzzy sets and closely related to the Hartley function is the *U* uncertainty of subnormal fuzzy sets proposed by *Higashi and Klir* [1982] and refined by *Klir and Wierman* [1998]:

$$U(F) = \int_0^{h(F)} \log_2 |^{\alpha}F| d\alpha + (1 - h(F)) \log_2 |X|, \qquad (3)$$

where $U(F)$ is the *U* uncertainty associated with *F*, $|^{\alpha}F|$ is the cardinality of an α-cut of *F* (i.e., number of members that remain in the set if all members with a membership grade less than α are taken out of *F*), $h(F)$ is the height of *F* (maximum value of membership grade in *F*), and $|X|$ is the cardinality of the universal set *X* (in this case, the model population created by sampling the parameter space). An approximate solution to equation 3 is

$$U(r) = \sum_{i=2}^{n} (r_i - r_{i+1}) \log_2 i + (1 - r_1) \log_2 n, \qquad (4)$$

where *r* is the ordered possibility distribution [*Zadeh*, 1978] derived from the fuzzy set *F*, $r_1 = h(F)$, and r_{n+1} is assumed to be 0. In this context, the membership grade function defining *F* plays the role of the possibility distribution function, *r*, and the members within the set are sorted in descending order of the membership grade function. Figure 1 shows a series of hypothetical relationships between the α-cut and $|^{\alpha}F|$. At an α-cut of 0.6 the three relationships shown yield very different cardinalities. Relations that are skewed towards the low end, and thus have only a few high $f(x_i)$ models, are better than relations having too many high $f(x_i)$ values. The ideal is to have a single model with its respective $f(x_i) = 1.0$ and all other models have $f(x_i) = 0.0$. This gives a cardinality of 1.0 for the fuzzy set and a cardinality of one for the crisp set indicating that the optimal solution can be clearly identified. The more usual case is one in which the cardinality of the fuzzy set is greater than 1.0.

[11] The shape of this curve, α-cut versus $|^{\alpha}F|$, provides a basis for estimating the useful calibration information provided by the given objective function. The key is to objectively define the α-cut and the cardinality of the model set that must be retained as the solution to the calibration based on the available information. One way to do this is to use the principle of uncertainty invariance [*Klir and Wierman*, 1998], which forms a crisp set of acceptable models that approximates the respective fuzzy sets by virtue of having the same *U* uncertainty. Consider for example the ordered fuzzy set $F = \{0.9, 0.8, 0.8, 0.7, 0.6, 0.4, 0.1, 0.1\}$. The cardinality of *F* is the sum of fuzzy memberships in the set, which in this case is 4.4 and the associated *U* uncertainty is 2.2. To decide how many members must be retained given this evidence, an integer, *k*, is calculated such that the value of $|U(r) - log_2 k|$ reaches a minimum. This is obtained by equating the U uncertainty of the fuzzy set (equation 4) to the Hartley function for the desired crisp set. The value of *k* (5 in this example) is the required cardinality of the retained model set, which means the top five models in *F* are retained and the -cut is placed at 0.6. The parameter values associated with this crisp set (henceforth called the restricted set) can fill all or part of the multidimensional parameter space defined by the initial ranges or distributions used for sampling. Figure 2a shows a possible way that a restricted set might fill the parameter space for a two-parameter system. Output from this restricted set is used to set upper and lower limits on the model prediction, called a prediction envelope. This envelope can in turn be compared to an independent set of observations for verification. Characteristics of the prediction envelope are summarized by its average width and its ability to successfully contain observations during the evaluation period.

[12] Note that the models are no longer ranked within the restricted set. This follows from the argument that the objective function, once used to construct the restricted set, can provide no further information that will allow distinguishing among the member models even if some of these models may be better [*Gupta et al.*, 1998]. Further distinctions among models will need new information either in the form of additional objective functions or expert knowledge. As such, multiple objective functions can be combined by taking an intersection of the corresponding restricted sets instead of a fuzzy set intersection prior to the construction of restricted sets. The later method may result in a calibration unduly influenced by only one objective function if it is consistently poor for all models, because the operation is defined as

$$f_r(x_i) = \min\{f_1(x_i), f_2(x_i), \ldots, f_n(x_i)\}, \qquad (5)$$

where $f_r(x_i)$ is the membership grade in the intersection set for model x_i, and $f_1(x_i)$ to $f_n(x_i)$ are membership grades corresponding to *n* different objective functions. As an

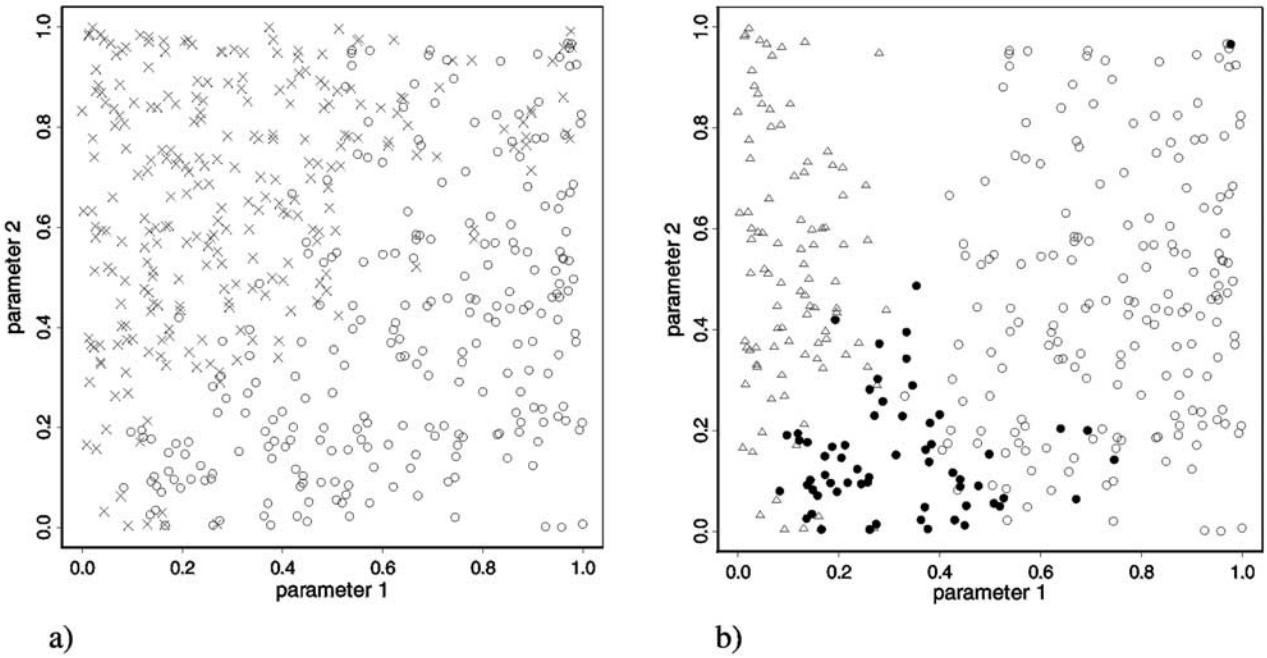


Figure 2. An example of a set of 500 simulations obtained by Monte Carlo sampling of two hypothetical model parameters 1 and 2 (scaled from 0 to 1). (a) Using the uncertainty invariance technique with a hypothetical objective function f_1 we obtain a U uncertainty value of 7.935. Based on this, 245 simulations are retained in the restricted set (open circles), the rest (crosses) are considered inadequate simulators and rejected. The retained simulations are from a more limited region in parameter space. (b) Two different objective functions f_1 and f_2 (with a restricted set cardinality of 159) are combined by intersection of the corresponding restricted sets (open circles for f_1, and triangles are for f_2) to construct a new restricted set of cardinality 62 (solid circles). The resulting set show improved parameter definition for the model as the selected parameters in the new restricted set are from limited region in the parameter space.

example of crisp set intersection, if objective functions f_1 and f_2 are used to generate the crisp sets A_1 and A_2, respectively, then the combined crisp set A_{12} is obtained by $A_1 \cap A_2$. A hypothetical outcome of combining two objective functions in a two-parameter space is shown in Figure 2b. A comparison of the cardinalities, $|A_1|$, $|A_2|$, and $|A_{12}|$, is useful in assessing additional calibration information obtained by combining f_1 and f_2 instead of using any one of these objective functions. One limitation of using crisp set intersection is that the resulting set may be too constrained because the restricted sets are only approximations of the original fuzzy sets. Therefore, caution must be exercised while combining two objective functions due to the existence of unacceptable tradeoffs or incompatibilities arising due to definitions of these functions or deficiencies in model structure.

2.2. Membership Grade Functions

[13] It is difficult to judge a priori the suitability of objective functions for use as membership grade functions (μ_F). Different objective functions evaluate different aspects of fit between simulated and observed responses, and so the process of restriction using crisp sets described above may result in an unrealistically low estimate of parameter uncertainty if deficiencies in the model structure prevent the simulation of all these aspects simultaneously. A desirable criterion is a direct mapping of f to μ_F, which is possible when f is defined so that f(x) $\rightarrow$ [0, 1]. We examined two objective function formulations widely used for hydrologic model calibration purposes. They emphasize two important aspects of the quality of fit between model output and observations. The first is the coefficient of determination, defined as

$$R^2 = \left\{ \frac{\sum_{i=1}^{N} (O_i - \bar{O})(P_i - \bar{P})}{\left[\sum_{i=1}^{N} (O_i - \bar{O})^2\right]^{0.5} \left[\sum_{i=1}^{N} (P_i - \bar{P})^2\right]^{0.5}} \right\}^2, \quad (6)$$

where N is the number of observations, P_i and are O_i are ith simulated and observed values respectively, overbar stands for the average for the calibration period. The second objective function, f_{BIAS}, is based on bias and defined as

$$BIAS = \sum_{i=1}^{N} \frac{|P_i - O_i|}{O_i}, \quad (7)$$

$$f_{BIAS} = \begin{cases} BIAS \leq 1 \Rightarrow 1 - BIAS \\ BIAS > 1 \Rightarrow 0 \end{cases}. \quad (8)$$

This transformation allows for a straightforward mapping of the information provided by bias in the simulated streamflow into membership grades. Although this technique

collapses the high bias values, it retains proportional differences in the region of interest, i.e., models with low bias values.

[14] It has been demonstrated that improved calibration of hydrologic models may be obtained by independently calibrating on subperiods of a time series [*Zhang and Lindström*, 1997; *Boyle et al.*, 2000]. Subperiods may be characterized by differences in specific processes that dominate the hydrologic behavior of the watershed. We utilized an approach similar to the one described by *Boyle et al.* [2000] to partition the hydrograph based on recorded precipitation and observed flow quantities. Each day was classified in one of three categories or classes, *F*, *P*, and *B*, following the logic described below:

$$FlowClass = \begin{cases} O_i > R_i \Rightarrow ClassF \\ O_i \leq R_i \Rightarrow ClassP \\ R_i = 0 \Rightarrow ClassB \end{cases}, \quad (9)$$

where O_i and R_i are observed streamflow and recorded precipitation for *i*th day, respectively. Observations within class *F* are during a period of rain on snow or snowmelt, class *P* observations are mainly rainfall-runoff, and stream flow during class *B* is dominated by base flow. The usefulness of these objective functions individually and in various combinations was evaluated in the context of this framework by calibrating the hydrologic submodel within RHESSys.

2.3. RHESSys Model

[15] RHESSys combines forest canopy gas exchange processes, soil moisture balance, and lateral saturated through flow within a common integrated spatial data and simulation framework [*Mackay and Band*, 1997]. RHESSys builds a hierarchical representation of watersheds. At the top level of the hierarchy, the watershed is divided into hillslope facets. At the next level, each hillslope facet is subdivided into elevation zones for adiabatic adjustment of air temperature. Each elevation zone is segmented into hydrologically uniform patches defined on intervals of the frequency distribution of the TOPMODEL topography and soil transmissivity index (*TSI*) [*Beven and Kirkby*, 1979; *Beven*, 1986; *Sivapalan et al.*, 1987; *Quinn et al.*, 1995]. Complete details on the design and implementation of RHESSys are provided in previous publications [*Band et al.*, 1993; *Mackay and Band*, 1997; *Mackay*, 2001]. For this case study, we focus on the components that directly affect basin outflow in RHESSys.

2.4. Model Parameterization

[16] The parameters that determine the behavior of the model related to its hydrologic components were obtained from previous applications of RHESSys in similar ecosystems [e.g., *Baron et al.*, 2000; *Mackay and Band*, 1997; *Watson et al.*, 1996; *White et al.*, 1998; *Mackay*, 2001]. Because a Monte Carlo sampling strategy was used to explore the parameter space, a careful selection of parameters was necessary to limit the number of simulation runs for computational reasons. Although an efficient search strategy (e.g., shuffled complex evolution [*Duan et al.*, 1992] would be able to locate acceptable parameters with less computational burden, in the current method it is necessary to approximate the entire response surface for comparing the model population. A set of initial simulations was carried out to determine the sensitivity of the stream flow output of the model to different values of hydrologically relevant parameters. Guided by this analysis the parameters, m, κ, δ, and C_{pint} were selected for calibration (Table 1) and 10,000 model realizations were simulated with random values for the selected parameters.

Table 1. Parameters of RHESSys Used for Monte Carlo Sampling During Calibration[a]

Parameter	Units	Description	Range
m		TOPMODEL parameter	0.01 to 0.20
δ	m	base flow adjustment parameter	−1.0 to 6.0
C_{pint}	m LAI^{-1} day^{-1}	rate of interception of incoming precipitation by canopy	0.0001 to 0.001
κ		scale factor used for scaling local K_0	1.0 to 10.0

[a]The values of all parameters were assumed to be uniformly distributed for sampling purposes within the indicated range.

2.5. Study Site

[17] A data set available for H.J. Andrews Experimental forest, Oregon, a long-term ecological research (LTER) site was used for this study. A small catchment, WS 2 [*U.S. Department of Agriculture*, 1986], within this basin was selected for Monte Carlo simulations. Predominant land cover for WS 2 is old growth conifer forest. WS 2 is a first order catchment with an area of 60.3 hectares. The elevation ranges from 548 m to 1070 m with a mean slope of 27.1 degrees. A 30-meter digital elevation model (USGS level 2) was used to divide the catchment into six hillslope partitions. Soils data was obtained from a 1964 survey. Daily stream flow has been continuously recorded for this catchment for over 40 years. There are various recommendations regarding the length of data required for calibration. *Yapo et al.* [1996] recommend using approximately eight years of representative data, while *Sorooshian and Gupta* [1995] suggest using a data set at least 20 times the number of parameters to be estimated noting that the marginal improvements may become small after 500 to 1000 data points. In order to keep the problem simpler, the later suggestion was adopted for the simulation experiment using data from 1959 to 1965. The first 638 days, up to the beginning of a water year, were regarded as the time necessary for model spin up. Stream flow observations for the next two water years, with 109, 232, and 389 days in classes F, P and B, respectively, were used for calibration purposes. The same classes had 105, 231, and 394 days, respectively, for the two subsequent years used for validation. Daily precipitation, minimum and maximum temperatures were obtained from a meteorological station (CS2MET) located within WS 2. The mean annual precipitation recorded at this station over the past 40 years is 224 cm.

3. Results and Discussion

3.1. Coefficient of Determination (R^2) as Membership Grade Function

[18] The first objective functions considered were Coefficients of Determination (R^2) (equation 4) taking the entire

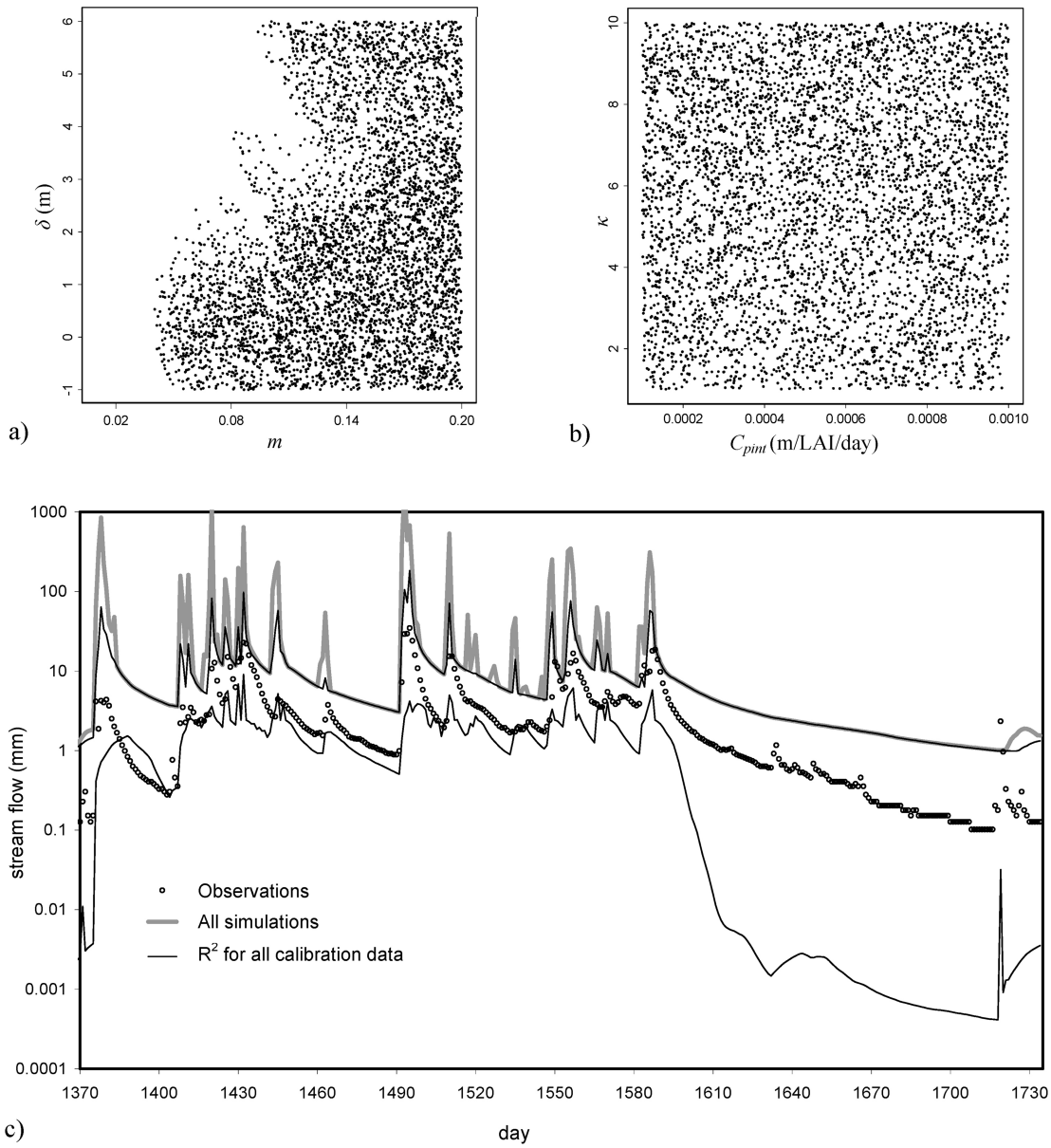


Figure 3. Parameter values corresponding to the restricted set using R^2 for the calibration time series as a whole are shown in Figures 3a and 3b. The values of parameters m and δ (a) are more important in determining R^2 values compared to C_{pint} and κ (b). The prediction envelope obtained using R^2 values (c) set a tighter limit on the predicted stream flow (solid lines) compared to the model population (shaded line). These limits are shown for 1 year in the test period. The lower limit is essentially zero for the model population and not depicted in the log scale. It can be seen that using R^2 as objective function can obtain a considerably narrower prediction envelope, particularly at the high and low ends of observations.

calibration time series as a whole, and for each class of observations (equation 9) considered separately. For the whole time series data, the restricted set has a cardinality of 4,955 and a corresponding α-cut value of 0.4807. The resulting parameter spaces are shown in Figures 3a and 3b, taking two parameters at a time. Parameters κ and C_{pint} do not show any pronounced influence on the R^2 values, most likely because of insensitivity [*Yapo et al.*, 1996]. However, R^2 values are influenced by the selected value of the m parameter because in RHESSys this parameter controls both the sensitivity of the spatial distribution of the saturated zone to topography and soil characteristics, and the response of base flow to mean saturation deficit. The parameter δ allows for a direct adjustment of base flow, which strongly influences R^2 values at low values of m. The prediction envelope (Figure 3c) based on the restricted set is considerably narrower than the prediction envelope generated by the model population and is successful in rejecting some of the models that generate extreme upper and lower limits in the prediction envelope for the model population. An overall improvement is achieved in the prediction envelope using the restricted set. A comparison of the two

Table 2. Summary of Properties of Prediction Envelopes for Different Objective Functions[a]

Membership Grade Function	Cardinality Of Restricted Set	Characteristics of Prediction Envelope for Test Period							
		Percent Observations Contained				Average Width, mm			
		All Data	Class F	Class P	Class B	All Data	Class F	Class P	Class B
None	10,000	95.3	87.5	97.0	96.4	34.7	23.7	82.8	9.5
R^2 - all data	4,955	83.2	73.3	83.5	85.5	7.9	9.5	13.4	4.3
R^2 - class F	4,232	73.3	62.9	75.3	74.9	7.4	8.7	13.0	3.9
R^2 - class P	5,472	88.9	79.0	88.7	91.6	8.8	10.9	14.8	4.7
R^2 - class B	6,418	83.6	73.3	86.1	84.8	9.2	10.1	17.3	4.3
R^2 - Intersect of class F, P, and B	3,586	72.7	61.0	72.7	75.9	6.5	8.0	10.7	3.7
BIAS - all data	1,843	94.5	84.8	96.1	96.2	28.9	20.2	68.5	8.1
BIAS - class F	5,376	94.9	85.7	96.5	96.4	30.4	20.4	72.0	8.6
BIAS - class P	1,245	88.2	76.2	88.7	91.1	12.8	10.1	27.8	4.7
BIAS - class B	4,764	87.9	75.2	88.3	91.1	14.4	12.0	32.0	4.7
R^2 - all data $\cap$ BIAS- all data	440	78.8	60.0	78.4	84.0	6.3	6.8	10.1	3.8
R^2 - all data $\cap$ BIAS- all data *lo*δ cluster	380	43.8	22.9	35.9	54.1	2.2	2.9	4.0	1.0
R^2 - all data $\cap$ BIAS- all data *hi*δ cluster	60	6.2	7.6	9.5	3.8	1.3	1.8	1.4	1.1

[a]The considered interval is for 2 years of test period as a whole and also the classes F, P and B individually. Note that the performance of the calibrated model set in validation period varies considerably among classes. Cardinality provides some idea of the uncertainty in parameter values associated with each objective function. However, a significantly lower cardinality does not always result in a significant or consistent improvement in prediction.

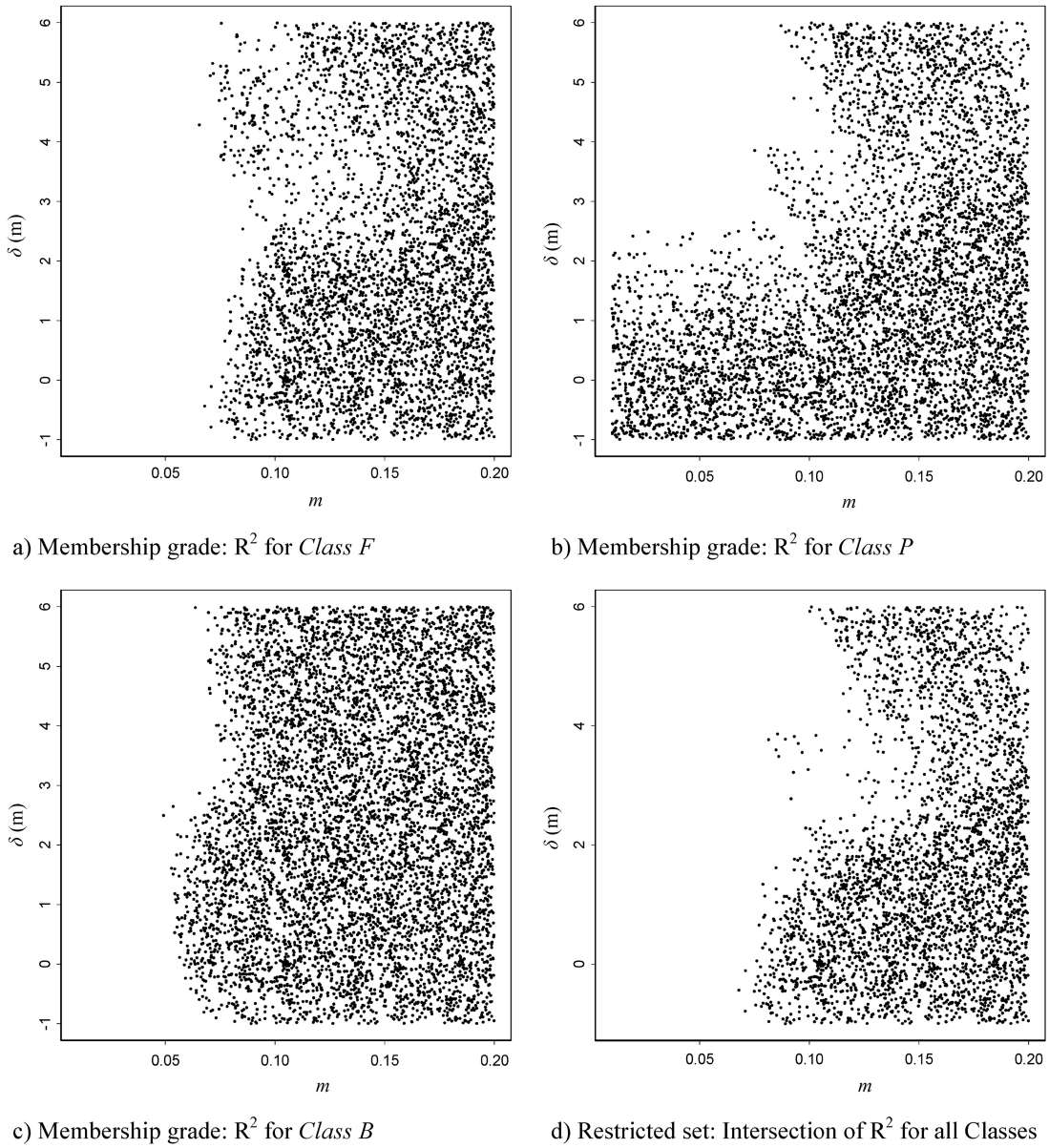
prediction envelopes shows the greatest improvement for peak flows. The average width of the modified prediction envelope is considerably less for all classes (Table 2) although a somewhat higher number of observations now lie outside this interval. A 77% reduction in the average width of the interval is achieved at the expense of losing only 12% of observations from the prediction envelope. However, the observations that lie outside tend to occur together (for example between day 1386 and 1404 in Figure 3c indicating trends that may be due to a lack of flexibility allowed by the four parameters selected for random sampling. It may be possible to calibrate RHESSys to achieve a better fit to these observations alone disregarding the others, but the resulting parameters values may not be representative of long-term system behavior. Including such simulations in the restricted set would result in a prediction envelope that is apparently better in containing observations at the cost of widening the prediction envelope. Moreover, the parameters in the restricted set would include some parameter combinations that are difficult to interpret. Another possible reason for the loss of observations from the prediction envelope could be that R^2 for the time series as a whole is not providing sufficient information regarding parameter values at certain periods, as it may be disproportionately influenced by extreme values in the time series [*Legates and McCabe*, 1999]. It is also noted that the average width of the prediction envelope differs among classes. For example, the average width of the prediction envelope for class P observations (13.4 mm) is much higher than that for class B (4.3mm). This is due to that fact that stream discharge during low flow periods requires fewer model components, leading to smaller errors. In addition, the variability within class B is considerably less.

[19] This variability among classes prompted the use of restricted sets for class-based R^2 values as a possible way of obtaining additional information from the same set of observations. Differences among restricted sets for classes F, P and B, taken individually, are visible in the resulting parameter clusters (Figures 4a, 4b, and 4c, respectively) as well as the prediction envelope characteristics (Table 2). The restricted sets had cardinalities of 4232, 5472, and 6418 respectively. For class B many simulations provided high R^2 values. A lack of precipitation for long periods at the study site results in temporally autocorrelated observations in class B, which reduces the number of independent observations in this class. Consequently, the prediction envelopes for other classes are poor when observations solely from this class are used for calibration. These differences indicate that parameter estimates change depending on the class of data used for calibration. Differences are also apparent in the constructed prediction envelopes. Poor performance of the prediction envelope for class F is possibly because parameters influencing snowmelt were not adjusted, as these did not show a large influence during initial simulations (Section 2.3).

[20] Individually the restricted sets for class-based R^2 values do not provide consistent improvement in the characteristics of the prediction envelope. In spite of this, the restricted set for combined class-based R^2 values provides some improvement over the overall R^2 based restricted set. The new restricted set retained about a third (cardinality 3586) of the model population. The corresponding parameter values are similar, but more clustered than parameters retained by the overall R^2 measure (Figure 4d). At this step of refinement, the width of the uncertainty interval is reduced by 17% at the cost of about 10% of observations in the evaluation period. A comparison of the two prediction envelopes (Figure 5) shows that, due to the rejection of many models with lower *m* values, some of the spikes at the upper bound of the envelope have been removed along with a consistent removal of very low simulated stream flow values. Most of the observations that now lie outside this envelope are at the low end but still lie close to the lower limit of predictions (e.g., between days 1459 and 1489). This indicates that separate time subseries calibration can result in tangible, although marginal improvement in the current context.

3.2. Bias as Membership Grade Function

[21] We used f_{BIAS} to evaluate the information that can be gained by looking at the long-term bias in simulated discharge. Values of f_{BIAS} are reasonably well distributed over the range [0,1], particularly for overall bias and class P,



a) Membership grade: R^2 for *Class F*

b) Membership grade: R^2 for *Class P*

c) Membership grade: R^2 for *Class B*

d) Restricted set: Intersection of R^2 for all Classes

Figure 4. R^2 restricted sets for the three classes (a) F, (b) P, and (c) B, show different response to the values of parameters *m* and δ. Combining the three classes by set intersection yields (d) a restricted set similar to the restricted set based on over all R^2 (Figure 3a) but shows somewhat better defined parameter values.

which result in low cardinalities for the corresponding restricted sets (Table 2). However, a low cardinality does not always translate into a better prediction envelope. For example, using overall f_{BIAS} results in an average prediction envelope width of 28.9 mm, primarily due to the existence of widely separated parameter values that yield similar levels of f_{BIAS}. Taken individually and considering only the performance of the prediction envelope, the best results are obtained with f_{BIAS} for class P. However, the selected parameter values for class P are very different from those that satisfy the overall f_{BIAS}. In general, models selected based on the four different objectives (Figures 6a, 6b, 6c, and 6d) tend to be different. In this instance, class F shows some overlap with both class P and B, but the three classes taken together show no appreciable region of intersection. When all the classes are combined together with overall bias (Figure 6e), only six models remain at the edge of the parameter space. In appearance, this set has very well defined parameter values in a restricted region of the parameter space. However, comparing the output from these models with the observations reveals that high values of f_{BIAS} for this set are achieved by compensation, overestimating the discharge at some points and underestimating the same at others (figure not shown as these models belong to the *hi*δ cluster shown in Figure 7b and provide similar outputs). Models retained by this criterion include those that achieved the goal of low bias by compensation, possibly in addition to the models that simulate the behavior of the system adequately. This illustrates how a process of elimination can progress too far in the presence of considerable

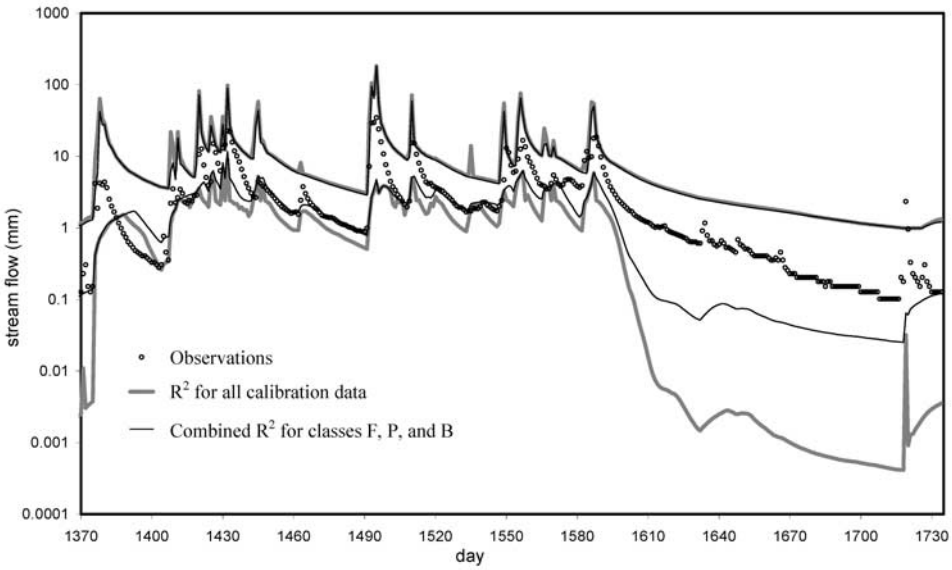


Figure 5. A comparison of prediction envelopes for over all R^2 (shaded line) and combination of R^2 for classes F, P, and B (solid line) shows that the later envelope is consistently narrower and does not have some of the spikes in the first prediction envelope (e.g., at day 1535). Most observations that moved outside of the new prediction envelope are low observed stream flow values that remain close to the lower limit (e.g., between days 1470 to 1490).

tradeoffs between objective functions without a reliable automated technique for expanding the restricted set when necessary, and the importance of selecting noncommensurable measures of information [*Gupta et al.*, 1998]. In this case, where models can achieve a consistently better f_{BIAS} value by compensation; f_{BIAS} should be combined with other objective function(s) that provide additional information required to distinguish between the two cases. To avoid the above problem, the restricted set for overall f_{BIAS} was considered appropriate for combining with R^2.

3.3. Combining Coefficient of Determination and Bias

[22] R^2 and f_{BIAS} result in crisp sets of models with different shortcomings (refer to Table 2). On the one hand, R^2 produces a set with a high cardinality and a wide parameter range for the restricted set, but a reasonable prediction envelope in terms of containing observations. On the other hand, the crisp set obtained by f_{BIAS} is better defined in terms of cardinality, but has a wider prediction envelope due to the presence of widely separated parameter clusters. An examination of the parameter space (Figures 3a and 6a) reveals that the parameters for the two calibration objective functions differ considerably. The restricted set obtained by combining these two objective functions (Figure 6f) has a much lower cardinality of 440. The prediction envelope is similar or better in terms of its ability to successfully contain observations within a narrow range for all classes compared to those for individual objective functions or a combination of R^2 for classes (Table 2). An examination of the prediction envelope (Figure 7a) shows it to be similar to the one obtained by overall R^2 (see Figure 5), but with a lower average width for higher values of observations (classes F and P). However, the parameter values for this combined objective function tend to occur in two distinct clusters widely separated in the parameter space (Figure 6f). One of these clusters corresponds to lower values of m and δ and the other to high values of δ (referred to as *lo*δ and *hi*δ clusters respectively). The existence of such widely separate clusters indicates a contradiction in parameter values that must be resolved to achieve satisfactory calibration of the model. The existence of such contradictory sets of parameter values was not detected with any single objective function. When the predictions from each of the clusters are viewed separately (Figure 7b), the prediction envelope from the combined f_{BIAS} and R^2 can be seen as a composite of two disjoint intervals corresponding to two parameter clusters. The simulations associated with the *lo*δ cluster show a better visual match to the dynamics of the observed streamflow. Clearly, the models associated with the *hi*δ cluster underpredict high flows and over-predicts low flows. The prediction envelope associated with the *lo*δ cluster is only 2.2mm wide on the average, but contains about 44% of observations (Table 2).

[23] The restricted set obtained by the intersection of individual class based R^2 sets with that for overall f_{BIAS} show characteristics similar to the restricted set described above. It has a cardinality of 66 and displays two distinct clusters at similar locations with 40 models in the *lo*δ cluster. Prediction envelopes associated with these clusters also show similar characteristics. The envelope constructed using only the models in the *lo*δ cluster has an average width of 0.7 mm and contains 20% of the observations over all.

[24] In both of the above cases, the parameters that yield better simulators of the dynamics by matching the timing of peak flows also tend to overestimate the magnitude of the peaks. It is important to realize that a four-parameter selection system limits the degrees of freedom. As such, the parameter combinations that provide a better match between observed and simulated peak flows (e.g., the *hi*δ cluster) do not necessarily match low flows and the dynamics. We hypothesize that these limitations are due to (1)

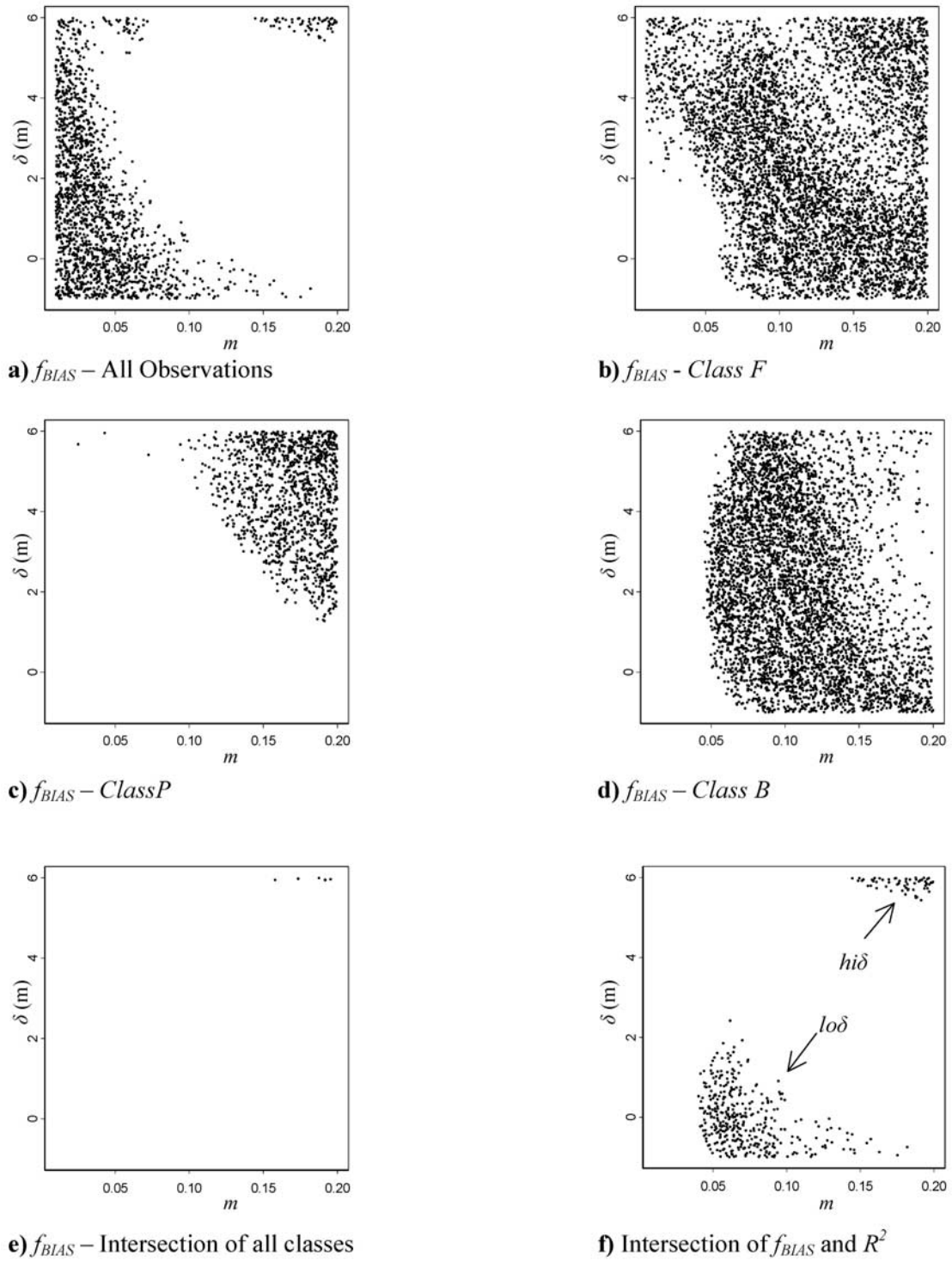


Figure 6. Restricted sets for f_{BIAS} as membership grade function are shown for (a) all observations, (b) class F, (c) class P, and (d) class B. These sets show more distinct differences among classes than those using class based R^2. When the three class based f_{BIAS} are combined, only six simulations are retained in the resulting restricted set, and all are restricted to only one parameter region obtained by overall f_{BIAS}. In order to explore the parameter space more thoroughly, over all f_{BIAS} and R^2 are to form (f) a restricted set, which has a cardinality of 440. Parameter values associated with this restricted set occur in two distinct clusters in parameter space. Separation between these clusters (identified as *hi*δ and *lo*δ, depending on the associated values of δ) provides further opportunity to evaluate and refine the calibration.

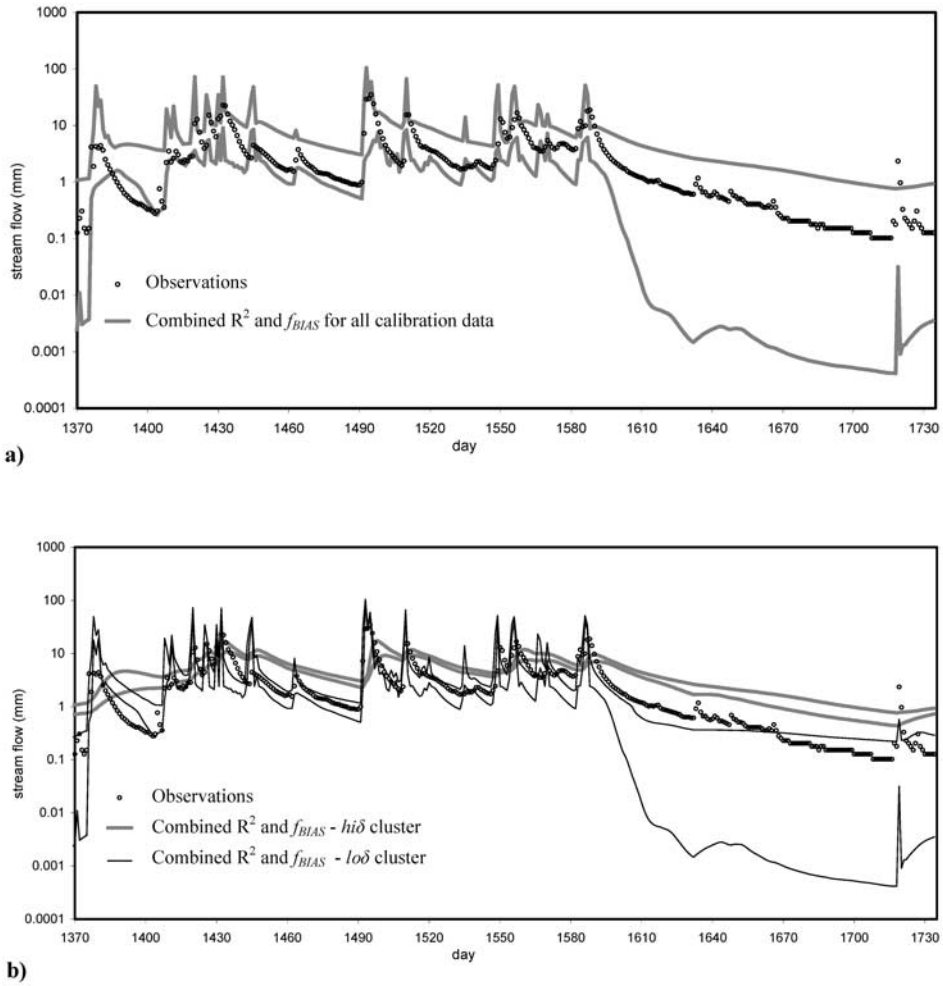


Figure 7. Prediction envelope from the restricted set for the combined over all f_{BIAS} and R^2 (a) performs reasonably well during the test period. However, considering the *hi*δ and *lo*δ parameter clusters seperately show that this prediction envelope is compsed of two prediction envelopes (b) of very different caharacteristics. Prediction envelope from simulations in the *lo*δ parameter cluster resemble the system behavior more closely.

restrictions imposed by using only four parameters for calibration or (2) some structural inadequacies in the model that need to be addressed. These issues would have to be resolved for any model that is to be used for predictive purposes.

[25] From the above results, it is observed that predictions from deterministic models are usually not randomly distributed within the predicted interval, particularly when there are multiple optima in the parameter space (e.g., the clusters obtained by combining R^2 and f_{BIAS}). The resulting predicted interval might be composed of more than one disjoint interval making it difficult to assign a probabilistic interpretation to this interval as in GLUE [*Beven*, 1993]. While the problem of equifinality [*Beven and Binley*, 1992] remains, it can be considerably reduced by incremental refinement using combinations of objective functions under the fuzzy interpretation. The concept of pareto optimality [*Yapo et al.*, 1998; *Gupta et al.*, 1998] provides a powerful framework for combining multiple objective functions and assessing parameter uncertainty arising out of tradeoffs among objective functions. However, an assumption of equal significance to all objective functions in the Pareto set of solutions (refer to equation 2 of *Gupta et al.* [1998]) may retain parameter sets that ''fit the data'' but are unacceptable to manual calibration experts [*Boyle et al.*, 2000] making it difficult to incorporate an automated rejection scheme strictly within this framework. However, satisfactory compromise solutions can be obtained by selection based on expert knowledge. As is shown with f_{BIAS}, and R^2 here, existence of significant tradeoffs can prevent satisfactory parameterization of a model such as RHESSys. It may be easier to formulate automatic rejection schemes for such models using rule-based criteria in addition to strictly calculated objective functions. The ability to separate out clusters of parameter values using a fuzzy logic framework would make it easier to formulate relevant knowledge-based rules useful for this purpose. However, the robustness of our framework depends on the sampling frequency of the parameter space and an efficient search procedure e.g., multiobjective complex evolution (MOCOM-UA [*Yapo et al.*, 1998]) or shuffled complex evolution (SCE-UA [*Duan et al.*, 1992]) may be employed to ensure that the promising

parameter regions are well represented among the sampled parameters.

4. Conclusions

[26] The calibration and uncertainty estimation framework described in this paper provides a basis for making an objective estimate of parameters and the range of model output associated with a failure to identify a unique solution. A fuzzy ranking of models, instead of a strict one as implied by equation 1, results in a nonspecific solution to the calibration problem. However, the iterative approach proposed here supports a nonlinear but continual refinement in selection of parameter values while recognizing the uncertainties inherent in the calibration process. The method can be used to identify contradictory parameter clusters that are subsequently evaluated by different objective functions or other decision criteria. However, as seen from the results, careful choice and evaluation of the membership grade function or combinations thereof is still necessary to obtain reliable parameter estimates. This is a monotonic refinement technique based solely on the elimination of models considered unacceptable at any calibration step. For an automated calibration and uncertainty estimation framework to be effective, we suggest that this method needs to be extended to incorporate methods for accepting as well as rejecting candidate models (i.e., a nonmonotonic system). As the described method and pareto optimality are both grounded in set theory, it may be possible to construct such a framework by combining the two approaches. Flexibility provided by this approach makes it suitable for implementation as a rejection method in a more complete automated calibration framework or an expert system for model calibration integrating rule-based techniques along with calibration based on objective function evaluation.

[27] **Acknowledgments.** Funding for this research was provided by a NASA Land Surface Hydrology grant (NAG5-8554). Additional funding from the Wisconsin Alumni Research Foundation (WARF) is also acknowledged. Computing facilities of the UW-Madison Integrated Remote Sensing Resources Center, a NASA Center of Excellence in Remote Sensing (grant NAG5-6535), were used in this research. We thank Alan Vonderohe of the Department of Civil and Environmental Engineering for providing access to additional computing resources. The helpful comments and constructive suggestions from two anonymous reviewers and the associate editor are gratefully acknowledged.

References

Band, L. E., P. Patterson, R. Nemani, and S. W. Running, Forest ecosystem processes at the watershed scale: Incorporating hillslope hydrology, *Agric. For. Meteorol.*, *63*, 93–126, 1993.

Baron, J. S., M. D. Hartman, L. E. Band, and R. B. Lammers, Sensitivity of a high elevation Rocky Mountain watershed to altered climate and CO_2, *Water Resour. Res.*, *36*(1), 89–99, 2000.

Beldring, S., Multi-criteria validation of a precipitation-runoff model, *J. Hydrol.*, *257*, 189–211, 2002.

Beven, K. J., Runoff production and flood frequency in catchments of order *n*: An alternative approach, in *Scale Problems in Hydrology*, edited by V. K. Gupta, I. Rodriguez-Iturbe, and E. F. Wood, pp. 107–131, D. Reidel, Norwell, Mass., 1986.

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

Beven, K. J., Prophecy, reality and uncertainty in distributed hydrological modeling, *Adv. Water Resour.*, *16*, 41–55, 1993.

Beven, K. J., and A. M. Binley, The future of distributed models: Model calibration and uncertainty prediction, *Hydrol. Processes*, *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.

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

Boyle, D. P., H. V. Gupta, and S. Sorooshian, Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods, *Water Resour. Res.*, *36*(12), 3663–3674, 2000.

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

Franks, S. W., and K. J. Beven, Estimation of evapotranspiration at the landscape scale: A fuzzy disaggregation approach, *Water Resour. Res.*, *33*(12), 2929–2938, 1997.

Franks, S. W., and K. J. Beven, Conditioning a multiple-patch SVAT model using uncertain time-space estimates of latent heat fluxes as inferred from remotely sensed data, *Water Resour. Res.*, *35*(9), 2751–2761, 1999.

Franks, S. W., P. Gineste, K. J. Beven, and P. Merot, On constraining the predictions of a distributed model: The incorporation of fuzzy estimates of saturated areas into calibration process, *Water Resour. Res.*, *34*(4), 787–797, 1998.

Franks, S. W., K. J. Beven, and J. H. C. Gash, Multi-objective conditioning of a simple SVAT model, *Hydrol. Earth Syst. Sci.*, *3*(4), 477–489, 1999.

Gupta, H. V., S. Sorooshian, and P. O. Yapo, Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, *Water Resour. Res.*, *34*(4), 751–763, 1998.

Gupta, H. V., L. A. Bastidas, S. Sorooshian, W. J. Shuttleworth, and Z. L. Yang, Parameter estimation of a land surface scheme using multicriteria methods, *J. Geophys. Res.*, *104*(D16), 19,491–19,503, 1999.

Hankin, B. G., and K. J. Beven, Modelling dispersion in complex open channel flows: Fuzzy calibration (2), *Stochastic Hydrol. Hydraul.*, *12*(6), 397–412, 1998.

Hartley, R. V. L., Transmission of information, *Bell Syst. Tech. J.*, *7*(3), 535–563, 1928.

Higashi, M., and G. J. Klir, Measures of uncertainty and information based on possibility distributions, *Int. J. Gen. Syst.*, *9*, 43–58, 1982.

Klepper, O., H. Scholten, and J. P. G. Van De Kamer, Prediction uncertainty in an ecological model of the Oosterschelde estuary, *J. Forecasting*, *10*, 191–209, 1991.

Klir, G. J., and M. J. Wierman, *Uncertainty-Based Information: Elements of Generalized Information Theory, Studies in Fuzziness and Soft Computing*, vol. 15, edited by J. Kacprzyk, Physica, New York, 1998.

Kuczera, G., Improved parameter inference in catchment models, 1, Evaluating parameter uncertainty, *Water Resour. Res.*, *19*(5), 1151–1162, 1983.

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

Legates, D. R., and G. J. McCabe, Jr., Evaluating the use of "goodness of fit" measures in hydrologic and hydroclimatic model validation, *Water Resour. Res.*, *35*(1), 233–241, 1999.

Mackay, D. S., Evaluation of hydrologic equilibrium in a mountainous watershed: Incorporating forest canopy spatial adjustment to soil biogeochemical processes, *Adv. Water Resour.*, *24*, 1211–1227, 2001.

Mackay, D. S., and L. E. Band, Forest ecosystem processes at the watershed scale: Dynamic coupling of distributed hydrology and canopy growth, *Hydrol. Processes*, *11*, 1197–1217, 1997.

Madsen, H., Automatic calibration of a conceptual rainfall-runoff model using multiple objectives, *J. Hydrol.*, *235*, 276–288, 2000.

Melching, C. S., Reliability estimation, in *Computer Models of Watershed Hydrology*, edited by V. P. Singh, pp. 69–118, Water Resour. Publ., Highlands Ranch, Colo., 1995.

Quinn, P. F., K. J. Beven, and R. Lamb, The ln(a/tanβ) index: How to calculate it and how to use it within the TOPMODEL framework, *Hydrol. Processes*, *9*, 161–182, 1995.

See, L., and S. Openshaw, A hybrid multi-model approach to river level forecasting, *Hydrol. Sci. J.*, *45*(4), 523–536, 2000.

Sivapalan, M., K. J. Beven, and E. F. Wood, On hydrologic similarity, 2, A scaled model of storm water runoff production, *Water Resour. Res.*, *23*(12), 2266–2278, 1987.

Sorooshian, S., and V. K. Gupta, Model calibration, in *Computer Models of Watershed Hydrology*, edited by V. P. Singh, pp. 23–68, Water Resour. Publ., Highlands Ranch, Colo., 1995.

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

U.S. Department of Agriculture, Map of H. J. Andrews Experimental Forest, Blue River Ranger District, Willamatte Natl. For., Oreg., 1986.

Van Stratten, G., and K. J. Keesman, Uncertainty propagation and speculation in projective forecasts of environmental change: A lake-eutrophication example, *J. Forecasting*, *10*, 163–190, 1991.

Watson, G. R., R. A. Vertessy, and L. E. Band, Distributed parameterization of a large scale water balance model for an Australian forested region, in *Hydrogis 96: Application of Geographic Information Systems in Hydrology and Water Resources Management (Proceedings of Vienna Conference)*, *IAHS Publ.*, *235*, 157–166, 1996.

White, J. D., S. W. Running, P. E. Thornton, R. E. Keane, K. C. Ryan, D. B. Farge, and C. H. Key, Assessing simulated ecosystem processes for climate variability research at Glacier National Park, USA, *Ecol. Appl.*, *8*(3), 805–823, 1998.

Yapo, P. O., H. V. Gupta, and S. Sorooshian, Automatic calibration of conceptual rainfall-runoff models: Sensitivity to calibration data, *J. Hydrol.*, *181*, 23–48, 1996.

Yapo, P. O., H. V. Gupta, and S. Sorooshian, Multi-objective global optimization for hydrologic models, *J. Hydrol.*, *204*, 83–97, 1998.

Zadeh, L. A., Fuzzy sets, *Inf. Control*, *8*(3), 338–353, 1965.

Zadeh, L. A., Fuzzy sets as a basis for a theory of possibility, *Fuzzy Sets Syst.*, *1*, 3–28, 1978.

Zhang, X., and G. Lindström, Development of an automatic calibration scheme for the HBV hydrological model, *Hydrol. Processes*, *11*, 1671–1682, 1997.

S. Samanta and D. S. Mackay, Department of Forest Ecology and Management, University of Wisconsin-Madison, Madison, WI 53715, USA. (ssamanta@students.wisc.edu)