WATER RESOURCES RESEARCH, VOL. 36, NO. 12, PAGES 3663–3674, DECEMBER 2000

Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods

Douglas P. Boyle, Hoshin V. Gupta, and Soroosh Sorooshian
Department of Hydrology and Water Resources, University of Arizona, Tucson

Abstract. Automatic methods for model calibration seek to take advantage of the speed and power of digital computers, while being objective and relatively easy to implement. However, they do not provide parameter estimates and hydrograph simulations that are considered acceptable by the hydrologists responsible for operational forecasting and have therefore not entered into widespread use. In contrast, the manual approach which has been developed and refined over the years to result in excellent model calibrations is complicated and highly labor-intensive, and the expertise acquired by one individual with a specific model is not easily transferred to another person (or model). In this paper, we propose a hybrid approach that combines the strengths of each. A multicriteria formulation is used to "model" the evaluation techniques and strategies used in manual calibration, and the resulting optimization problem is solved by means of a computerized algorithm. The new approach provides a stronger test of model performance than methods that use a single overall statistic to aggregate model errors over a large range of hydrologic behaviors. The power of the new approach is illustrated by means of a case study using the Sacramento Soil Moisture Accounting model.

1. Introduction and Scope

Conceptual rainfall-runoff (CRR) models have become widely used for flood forecasting as the demand for timely and accurate forecasts has increased. Such models provide an approximate, lumped description of the dominant subwatershed-scale processes that contribute to the overall watershed-scale hydrologic response of the system. Most operational CRR models have of the order of 10 or more parameters, whose values must be selected so that the modeled response to rainfall closely simulates the actual behavior of the watershed of interest. For example, the Sacramento Soil Moisture Accounting (SAC-SMA) model [*Burnash et al.*, 1973] is used by the National Weather Service (NWS) for flood forecasting throughout the United States (Figure 1). The model has 17 parameters whose values must be specified (Table 1). While a few of these parameters might be estimated by relating them to observable characteristics of the watershed, most are abstract conceptual representations of nonmeasurable watershed characteristics that must be estimated through a calibration procedure.

The NWS has developed a sophisticated, highly interactive manual procedure to estimate parameter values for the SAC-SMA model [*Anderson*, 1997]. This process has been developed and refined over the years resulting in excellent model calibrations. However, the process is complicated, difficult to learn, and highly labor-intensive, requiring a substantial commitment of human resources. As a result, the expertise acquired by one individual through extensive hands-on training and experience with a specific model is not easily transferred to another person (or another model).

Over the past two and a half decades much research has been devoted to the development of automatic methods for model calibration that take advantage of the speed and power of digital computers. Such methods seek to be objective and relatively easy to implement. However, the model calibrations provided by such methods do not tend to provide parameter estimates and hydrograph simulations that are considered acceptable by the hydrologists responsible for operational forecasting. As a result, automatic methods have not yet entered into widespread use.

In this paper, we analyze the similarities and differences between the manual and automatic approaches to hydrologic model calibration and propose a hybrid approach that combines the strengths of each. The approach "models" the evaluation techniques and strategies used in manual calibration by using several objective measures to reflect different observable characteristics of watershed behavior. The resulting optimization problem is posed within the multicriterion framework presented by *Gupta et al.* [1998] and solved by means of a computerized optimization algorithm [*Yapo et al.*, 1998; *Bastidas et al.*, 1999]. Sections 2–4 describe and discuss the rationale for this methodology and illustrate the power of the approach through a case study using the SAC-SMA model.

Paper number 2000WR900207.
0043-1397/00/2000WR900207$09.00

2. Strategies for Parameter Estimation

2.1. Three-Level Classification Scheme

Model parameter estimation involves the selection of values for the parameters so that the model matches the behavior of the watershed system as closely as possible. For the purpose of this discussion we shall classify the parameter estimation process into three levels of increasing sophistication.

In level zero, approximate ranges for the parameter estimates are specified by examining lookup tables or by borrowing values from similar watersheds that have been previously calibrated. Note that these ranges reflect our uncertainty in the values of the parameters and might be termed prior estimates in the sense that they are not conditioned on any input-output

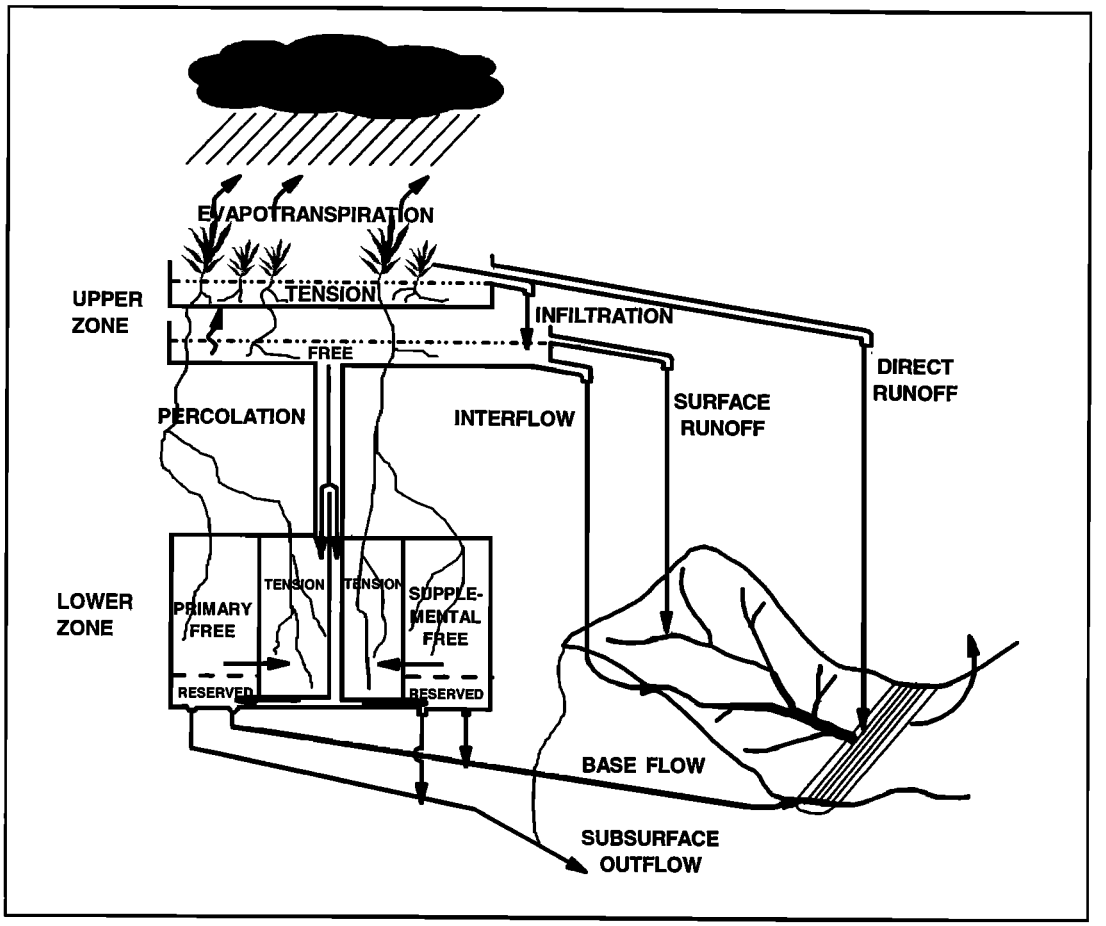

Figure 1. National Weather Service Sacramento Soil Moisture Accounting (SAC-SMA) model [from *Brazil*, 1988].

time series data collected for the watershed of interest. If a point estimate is desired, it can be selected from within this range. This approach is often applied in the case of ungaged watersheds or when historical data records are insufficient for the application of level one and two calibration techniques (see below). It is also used typically with complex "physics-based" models whose parameters are intended to represent measurable physical quantities (e.g., distributed hydrologic models such as the European hydrological system Systeme Hydrologique (SHE) and land-surface energy and water balance models such as the Biosphere-Atmosphere Transfer Scheme).

Levels one and two involve the use of progressively more sophisticated procedures for refining the parameter estimates by the use of information available in input-output time series data collected for the watershed of interest. They might therefore be termed posterior estimates. In level one the parameter ranges are refined by identifying periods in the output time series where the effects of individual parameters (or small group of parameters) are relatively dominant and can therefore be isolated. A defining characteristic of level one is that the effects of parameter interactions are essentially ignored. Such procedures are commonly described in introductory level textbooks on hydrology [e.g., *Linsley et al.*, 1982, chapter 7]. The U.S. National Weather Service follows an explicit set of procedures and guidelines for level one refinement of individual parameters (discussed further in section 2.2).

Finally, level two consists of further refinement (narrowing) of the parameter ranges by means of a detailed analysis of parameter interactions and model performance trade-offs. This is the most difficult step in the parameter estimation process, because it requires a great deal of understanding of the model and typically involves complex decisions that weigh the multiple effects of adjusting several parameters at a time.

2.2. Approaches to Calibration

The methods employed in both levels one and two involve two components: (1) evaluation of the "closeness" between the model outputs and the corresponding measurement data and (2) adjustment of the values of the parameters to improve the closeness. Manual (sometimes called expert) and automatic calibration techniques can be compared and contrasted in terms of how each of these components is implemented.

2.2.1. Manual calibration. Manual calibration is the approach most widely used to calibrate hydrologic models. In this approach the "closeness" is typically evaluated in terms of several (more than three!) measures, and a semi-intuitive trial-and-error process is used to perform the parameter adjustment. Usually, the exact number and nature of these measures is not specified clearly. While some aspects of "closeness" may be evaluated in terms of objective measures (i.e., one or more mathematical criteria), most are, in fact, evaluated subjectively based on visual comparison of the model outputs and the data. Although the performance evaluation and parameter adjustment procedures are usually influenced by guidelines estab-

Table 1. Parameters and State Variables of the SAC-SMA Model

Parameter	Description	Level Zero	Level One	Level Two: Multiple Criteria Range	Level Two: Multiple Criteria "Best"	Level Two: Single Criterion RMSE	Level Two: Single Criterion Log RMSE
	Maximum Capacity Thresholds						
UZTWM	upper zone tension water maximum storage, mm	1.0–150.0	···	20.8–96.3	58.0	10.1	127.6
UZFWM	upper zone free water maximum storage, mm	1.0–150.0	···	26.9–49.3	45.8	31.2	21.0
LZTWM	lower zone tension water maximum storage, mm	1.0–500.0	···	196.5–250.3	247.5	257.7	210.0
LZFPM	lower zone free water primary maximum storage, mm	1.0–1000.0	120.0–140.0	122.4–128.3	124.4	120.0	120.1
LZFSM	lower zone free water supplemental maximum storage, mm	1.0–1000.0	40.0–60.0	40.0–40.9	40.5	40.1	40.1
ADIMP	additional impervious area (decimal fraction)	0.0–0.4	···	0.38–0.40	0.40	0.37	0.28
	Recession Parameters						
UZK	upper zone free water lateral depletion rate, day^{-1}	0.1–0.5	···	0.13–0.31	0.18	0.19	0.34
LZPK	lower zone primary free water depletion rate, day^{-1}	0.0001–0.025	0.006–0.01	0.006–0.008	0.006	0.010	0.007
LZSK	lower zone supplemental free water depletion rate, day^{-1}	0.01–0.25	0.15–0.20	0.17–0.19	0.18	0.20	0.15
	Percolation and Other Parameters						
ZPERC	maximum percolation rate, dimensionless	1.0–250.0	···	204.6–250.0	230.2	249.2	247.5
REXP	exponent of the percolation equation, dimensionless	0.0–5.0	···	2.5–3.8	3.2	2.5	3.5
PCTIM	impervious fraction of the watershed area (decimal fraction)	0.0–0.1	0.0–0.02	0.0033–0.0088	0.005	0.00002	0.0178
PFREE	fraction of water percolating from upper zone directly to lower zone free water storage (decimal fraction)	0.0–0.6	···	0.002–0.13	0.04	0.0002	0.25
	Parameters Not Optimized						
RIVA	riparian vegetation area (decimal fraction)	0.00					
SIDE	ratio of deep recharge to channel base flow, dimensionless	0.00					
RSERV	fraction of lower zone free water not transferable to lower zone tension water (decimal fraction)	0.30					
PXMLT	precipitation multiplication factor, dimensionless	1.00					
	State Variables						
UZTWC	upper zone tension water storage content, mm						
UZFWC	upper zone free water storage content, mm						
LZTWC	lower zone tension water storage content, mm						
LZFPC	lower zone free primary water storage content, mm						
LZFSC	lower zone free secondary water storage content, mm						
ADIMC	additional impervious area content, mm						

lished through previous experiences of model calibration, the actual sequence of parameter adjustments will vary from person to person based on their experience and training, their understanding of the model structure, the properties of the data, and the characteristics of the watershed system. In the following discussion we refer mainly to the parameter estimation procedures used by the NWS during manual calibration of the SAC-SMA flood forecast model.

The manual calibration process employed by the NWS begins with level zero parameter estimates obtained (primarily) by examining the range of parameter values for previously calibrated watersheds in the local forecast group (a group of neighboring watersheds having similar geology, hydrology, and climatology). Next, a systematic sequence of steps is followed to develop level one parameter estimates by examination of the hydrometeorological database for the watershed [*Peck*, 1976]. Periods in the observed time series data are identified where specific hydrologic processes are dominant (e.g., base flow, interflow, surface flow, evaporation, transpiration, abstraction, infiltration, etc.). For each of these periods the closeness of the model and the data is evaluated visually, and values for the relevant parameters are estimated by heuristic methods. As a trivial example, periods dominated by base flow are used to estimate the base flow recession rate parameter of the model (for each period) as the slope of the log of the hydrograph time series (Figure 2). The range of recession values identified over all the relevant periods then represents the new (refined) uncertainty in the value of the base flow recession parameter; a point estimate is typically computed by averaging the values. For the SAC-SMA model, *Peck* [1976] listed five parameters (LZFPM, LZPK, LZFSM, LZSK, and PCTIM) for which level one estimates can be easily obtained from the observed hydrograph and precipitation data. Level one estimates for another six parameters (LZTWM, UZTWM, UZK, SSOUT, UZFWM, and PFREE) are considered possible but more difficult to obtain from the data. The estimates for the remaining six parameters (SARVA, ZPERC, REXP, SIDE, ADIMP, and RSERV) cannot be refined by level one procedures.

A graphical user interface called the Interactive Calibration Program (ICP) [*National Weather Service (NWS)*, 1997] was developed by the NWS to facilitate the identification of different hydrometeorologic periods within the data for level one analysis. The ICP program will also be an integral part of the new Advanced Hydrologic Prediction System initiative recently funded by Congress starting fiscal year 2000. However, the ICP code also enables more detailed empirical and statistical analyses of the observed data, state values, and outputs simulated by the model and is therefore the primary tool used to obtain

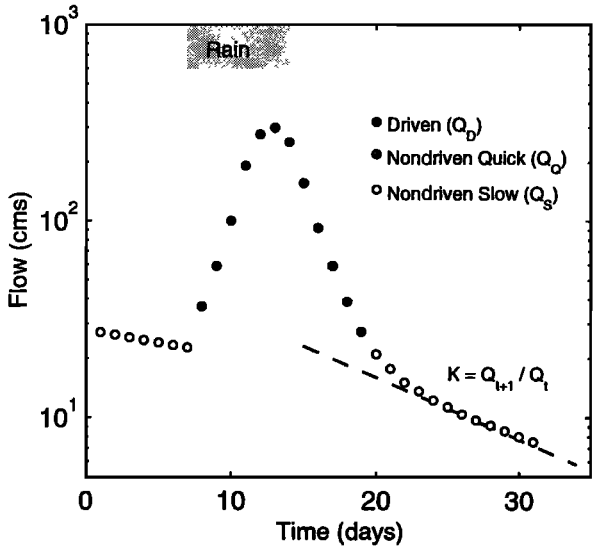

Figure 2. Partitioning of the observed hydrograph into three components: Q_D, Q_Q, and Q_S. The dashed line shows how the observed hydrograph can be used to estimate the recession constant K.

level two parameter estimates via manual calibration. The interactive graphical display allows the user to visually observe the impact of parameter adjustments on the internal behavior of the model (state variables) and on the "closeness" of the model-simulated streamflows to the observed hydrograph.

The complex effects of parameter interactions on the model responses make level two calibration a difficult process, requiring training and a great deal of practice to master. The hydrologist must simultaneously evaluate a number of subjective and objective criteria while iteratively adjusting the parameter values, some of which influence the overall behavior of the model while others have a significant impact only during certain special hydrologic events. While there are experts skilled at manual methods for level two parameter estimation, the procedures they use are not explicitly defined (some very general guidelines can be found in the ICP manual [*NWS*, 1997]).

A comprehensive understanding of the model, the real watershed system, and the data is required to produce consistent and reliable results with manual calibration. Unfortunately, this knowledge is not easily transferred from person to person or model to model and must come from extensive hands-on training and experience with the specific model. Even when performed by an expert, the process is highly labor-intensive and requires a substantial commitment of human resources. Manual calibration of the SAC-SMA model to a single watershed may require several hundred hours of effort. Further, the subjectivity of the evaluation procedure and the complex tradeoffs in the model's abilities make it very likely that many different solutions (sets of parameter values) may seem to produce equally "good" results. The NWS conducts several workshops annually to train hydrologists in calibration procedures for their models. Development of the skills needed to conduct a competent calibration for the SAC-SMA model can require several months of practice (M. Smith, Hydrologic Research Laboratory, NWS, personal communication, 1999).

2.2.2. Automatic calibration. Automatic calibration procedures for hydrologic models have been under development for at least three and a half decades, with the degree of sophistication generally paralleling the increases in computing power. The goal has been to develop an objective strategy for parameter estimation that provides consistent performance by eliminating the kinds of subjective human judgements involved in the manual approach. In the classical approach, borrowed from systems and operations research theory, the "closeness" is typically evaluated in terms of a single objective measure (a mathematical criterion) of the overall difference between the simulated and observed hydrographs, and the parameter adjustment is performed using an optimization algorithm. Considerable research has been devoted to trying to identify a "best" criterion [e.g., *Sorooshian and Dracup*, 1980; *James and Burges*, 1982; *Kuczera*, 1983a, 1983b] and the "best" optimization algorithms [e.g., *Brazil and Krajewski*, 1987; *Brazil*, 1988; *Wang*, 1991; *Duan et al.*, 1992, 1993; *Sorooshian et al.*, 1993]. Recent work has suggested that automatic calibration procedures also can be extended to handle multiple criteria [*Gupta et al.*, 1998, 1999; *Yapo et al.*, 1998].

In the automatic approach the performance evaluation and parameter adjustment procedures are objective in the sense that they establish explicit rules by which the actual sequence of parameter adjustments is made. The power of the procedure therefore depends on how well it has been designed to reflect the factors important to a successful calibration, and much effort has been devoted to establishing what those factors are. The following discussion is based primarily on our own experiences with the development and testing of automatic parameter estimation procedures for the SAC-SMA flood forecast model. In this discussion we explicitly distinguish between the classical single-criterion method and the new multicriteria approach.

In general, implementation of an automatic calibration process requires the user to specify a region of the parameter space which is considered to contain feasible values for the parameters; this is given typically as upper and lower bounds on each parameter. It is common practice for these bounds to be established via the level zero and level one parameter estimation procedures described above for manual calibration. The single-criterion and multicriteria automatic calibration methods therefore differ only in their approach to level two calibration.

The single-criterion approach searches the feasible parameter space for a single point that optimizes the mathematical criterion selected to measure the closeness of the model output and the data. The OPT3 automatic calibration code used by the NWS allows the user to select from among several choices for the criterion and the optimization algorithm. In general, the criterion most commonly used in the literature has been the root-mean-squared error (RMSE) evaluated on either the streamflows or the log of the streamflows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{t=1}^{N}(z_t^{\mathrm{sim}} - z_t^{\mathrm{obs}})^2}, \qquad (1)$$

where z_t^{sim} is either the simulated streamflow or log streamflow and z_t^{obs} is the corresponding observed value. Of the several optimization algorithms that have been tested, the Shuffled Complex Evolution-University of Arizona (SCE-UA) method has generally proved to be both robust and efficient [e.g., *Duan et al.*, 1992; *Gan and Biftu*, 1996; *Kuzcera*, 1997; *Thyer et al.*, 1999].

It has been suggested that although the single-criterion automatic approach is able to provide good parameter estimates, it can also "degenerate into pure curve fitting and produce a set of parameters that fit the calibration reasonably well but are hydrologically unrealistic" [*Peck*, 1976]. For this reason and

because automatic approaches require some mathematical background and familiarization with more complicated computational tools, the approach has not become popular for routine calibration of hydrologic models. It is not difficult to see that the approach has both a major strength and a major weakness. Its strength is its use of a robust optimization procedure to rapidly make successful parameter adjustments in the presence of strong parameter interaction. Its weakness, however, is its "complete dependency on one error function" [*Peck*, 1976].

The main problem with the single-criterion automatic approach is that the criterion must be chosen carefully to measure the difference between the simulated and observed hydrographs in a manner that does not place undue emphasis on matching one aspect of the hydrograph at the expense of another. While an expert can give due consideration to this fact in the course of manual calibration, the selection of a single criterion before initiating the process of calibration can predispose the automatic procedure toward an inappropriate (even hydrologically unrealistic) solution. For example, it has been our experience that use of the RMSE criterion evaluated on the streamflows tends to overemphasize fitting of the flood peaks and leads to strongly biased simulations of the recessions (and hence incorrect tracking of soil-moisture storages).

The multicriteria approach addresses this problem through a two-step process. In the first step an automatic search of the feasible parameter space is used to find the set of solutions (the so-called "Pareto optimal" region) which simultaneously optimizes several user-selected criteria that measure different aspects of the closeness of the model output and the data. This quickly results in several viable solutions, reflecting the range of different ways in which the hydrograph can be simulated with different kinds of "minimal" error [*Yapo et al.*, 1997; *Gupta et al.*, 1998]. In the second step the solutions having unacceptable trade-offs in fitting of the different parts of the hydrograph are rejected, and additional criteria (both objective and subjective) are employed to narrow the solution space. The major objective of this paper is to demonstrate how the two-step multicriteria automatic approach can be used to develop a hybrid strategy that combines the strengths of the manual and automatic calibration, resulting in efficient, yet acceptable, estimates for the parameters of a conceptual hydrologic model.

3. Combining the Strengths of Manual and Automatic Calibration

Manual calibration uses a subjective process of visual inspection and comparison of the model output and the observed data to implicitly evaluate (measure) the ability of the model to simulate specific aspects of the hydrologic behavior. Examples of these behaviors include the magnitude and timing of the peak flows and the shapes of the rising and falling limbs of the hydrograph. In addition, statistical criteria such as monthly and seasonal biases are typically used to provide an objective evaluation of the overall (long-term) behavior of the model. In principle, it could be said that the hydrologists doing manual calibration keep track of a number of different "criteria," most of which are not explicitly defined, while adjusting the model parameters. This allows them to balance the trade-offs in the ability of the model to simulate various aspects of the hydrograph (i.e., to consider the characteristics of the model structural error) with recognition of any potential errors in the observed data. Emulation of this process via automatic calibration requires the use of explicit mathematical criteria to approximate (model) the implicit measures used in the manual approach. This development involves identification of several characteristic features of the observed streamflow hydrograph, each representing a distinct (preferably unique) aspect of the behavior of the watershed. The "closeness" of the model output to the observed data for each of these features is measured objectively by means of a mathematical criterion.

There have been several recent reports on the exploration of hybrid methodologies that emulate manual procedures through stepwise application of optimization techniques. *Sugawara et al.* [1984] used a rule-based procedure to identify subperiods in the simulated hydrograph for automated adjustment of parameter subgroups of the Tank model. *Brazil* [1988] used an interactive procedure to perform level one estimates of parameter values of the SAC-SMA model, followed by automated global random search and iterative parameter optimization steps. A postsimulation multicriteria evaluation (based on up to 10 criteria) was used to obtain level two parameter estimates. *Harlin* [1991] and later *Zhang and Lindstrom* [1997] developed automated approaches that emulate the manual calibration procedures for the Hydrologiska Byrans Vattenbalansavdelning (HVB) model [*Bergstrom*, 1976]. Their procedures involve partitioning the runoff time series into several subperiods associated with specific, dominant hydrologic processes (e.g., rain flood, snowmelt flood, and base flow). Each subperiod is used to calibrate a different parameter (or subgroup). An iterative automated procedure cycles through the different periods to search for an "optimal" parameter set.

While the previous methods have been shown to be effective, they avoid dealing explicitly with the effects of interactions among the model parameters and, in particular, provide little insight into the performance trade-offs associated with model structural errors. In addition, they are largely model-dependent and must be adapted for use with different hydrologic models. The automatic multicriteria calibration strategy, however, provides a more versatile approach to emulating the evaluation procedures used in the manual approach to level two calibration, while taking advantage of the power and efficiency of an optimization algorithm to search the parameter space. In this approach a multicriteria optimization problem is defined in terms of simultaneous minimization of all the model performance criteria representing the "closeness" of the model output to the observed streamflow hydrograph. A multicriteria optimization methodology, such as the multi-objective complex evolution (MOCOM-UA) algorithm [*Yapo et al.*, 1998], is used to search for the Pareto solution space containing the "good" solutions to the problem. Finally, one or more "acceptable" parameter estimates are selected from within the Pareto solution space by rejecting solutions with unacceptable trade-offs and/or poor overall (long-term) statistical characteristics (e.g., annual, monthly, and flow group biases).

The MOCOM-UA is a general purpose global multiobjective optimization algorithm that provides an effective and efficient estimate of the Pareto space with a single optimization run and is based on an extension of the SCE-UA population evolution method [*Duan et al.*, 1993]. A detailed description and explanation of the method are given by *Yapo et al.* [1997, 1998], and so they will not be repeated here. In brief, the MOCOM-UA method involves the initial selection of a population of p points distributed randomly throughout the s-

dimensional feasible parameter space. In the absence of prior information about the location of the Pareto optimum a uniform sampling distribution is used. For each point the multiobjective vector **F** is computed, and the population is ranked and sorted using a Parteo-ranking procedure suggested by *Goldberg* [1989]. Simplexes of $s + 1$ points are then selected from the population according to a robust rank-based selection method [*Whitley*, 1989]. A multiobjective extension of the downhill simplex method is used to evolve each simplex in a multiobjective improvement direction. Iterative application of the ranking and evolution procedures causes the entire population to converge toward the Pareto optimum. The procedure terminates automatically when all points in the population become nondominated. Experiments conducted using standard synthetic multiobjective test problems have shown that the final population provides a fairly uniform approximation of the Pareto solution space [*Yapo et al.*, 1997, 1998].

In this paper, we investigate an automated multicriteria scheme that emulates the manual approach with a simple two-step procedure. In the first step the hydrograph is partitioned into three components based on the reasonable assumption that the behavior of the watershed is inherently different during periods "driven" by rainfall and periods without rain. Further, the periods immediately following the cessation of rainfall and dominated by interflow can be distinguished from the later periods that are dominated by base flow. The streamflow hydrograph can therefore be partitioned into three components (Figure 2), which we call "driven" (Q_D), "nondriven quick" (Q_Q), and "nondriven slow" (Q_S). The time steps corresponding to each of these components are identified through an analysis of the precipitation data and the time of concentration for the watershed. The time steps with nonzero rainfalls, lagged by the time of concentration for the watershed, are classified as driven. Of the remaining (nondriven) time steps those with streamflows lower than a certain threshold value (e.g., mean of the logarithms of the flows) are classified as nondriven slow, and the rest are classified as nondriven quick. In this paper, a simple empirical method is used for partitioning the nondriven flows; note that there are a wide range of different techniques based on physical reasoning [e.g., *Rutledge*, 1993; *Nathan and McMahon*, 1990] available for base flow separation that could be used to improve the partitioning of the nondriven component. For each of the components the closeness between the model outputs and the corresponding observed values is estimated separately with one or more statistical criteria. Finally, MOCOM-UA is used to search for the Pareto solution space containing the "good" solutions to the problem.

In the second step, one or more "acceptable" parameter estimates are selected from within the Pareto solution space by rejecting solutions with unacceptable trade-offs and/or poor overall (long-term) statistical characteristics (e.g., annual, monthly, and flow group biases). This approach is illustrated in section 4 with a case study involving parameter estimation for the SAC-SMA model.

4. Case Study

4.1. Introduction

In this study, the multicriteria approach outlined above was used to estimate values for the parameters of the SAC-SMA flood forecast model using historical data from the Leaf River watershed (1950 km^2) located north of Collins, Mississippi. Forty consecutive years of data (water years (WY) 1948–1988) are available for this watershed, representing a wide variety of hydrologic conditions. The details of the SAC-SMA model and the Leaf River data have been discussed previously in the literature [e.g., *Burnash et al.*, 1973; *Peck*, 1976; *Brazil and Hudlow*, 1981; *Sorooshian and Gupta*, 1983]. The level zero and level one estimates of parameter uncertainty (Table 1), the unit hydrograph ordinates, and potential evapotranspiration demand curve for this watershed have been determined previously by *Brazil* [1988]. Of the 17 model parameters, only 13 are usually estimated via level two parameter estimation procedures. An 11-year period (WY 1952–1962 inclusive) was selected to be used for this purpose.

4.2. Level Two Parameter Estimation

As described in section 3, the hydrograph was partitioned into three components: driven (Q_D), nondriven quick (Q_Q), and nondriven slow (Q_S). For each of the components the simulation error was measured using the RMSE statistic, resulting in three evaluation criteria, designated as FD (driven), FQ (nondriven quick), and FS (nondriven slow). The Pareto optimal solution space for the three criteria was estimated using the MOCOM-UA multicriteria optimization algorithm [*Yapo et al.*, 1997]; the algorithm used 15,533 trials and approximately 1 hour of computer time (on a Sun workstation) to converge to a set of 500 Pareto optimal solutions.

The results of the multicriteria automatic calibration run are shown in Figures 3 and 4. Figures 3a–3c present two-dimensional projections of the three-criterion trade-off surface represented by the 500 Pareto optimal parameter sets (indicated by the shaded dots). Figures 3a–3c clearly illustrate the inability of the model to simultaneously match all three aspects of the hydrograph. For example, Figure 3b indicates a smoothly varying trade-off between the models' ability to match the driven (Q_D) and the nondriven slow (Q_S) portions of the hydrograph (similarly, see Figure 3c). Figure 3a, however, shows that the FD and FQ criteria are very highly correlated, indicating that these two portions of the hydrograph contain very similar information about the parameters of this model and that one of these criteria can be considered to be redundant in this case study. As a result, equivalent solutions can be expected (for this model structure and data set) using only two criteria, either a combination of FD and FQ with FS, or FD with FS, or FQ with FS. The behavior of the model within FD and FQ is strongly influenced by the shape of the unit hydrograph, which was fixed (not optimized) in this case study. Optimization of the unit hydrograph shape, or replacement and optimization of a different routing model, may increase the model's ability to fit the FD and FQ components and, as a result, reduce some of the correlation. Note also that the solutions corresponding to the best fits to the Q_D, Q_Q, and Q_S portions of the hydrograph (i.e., solutions corresponding to minimal FD, FQ, and FS, respectively) correspond to the extreme points of the set of shaded dots on each plot.

Figures 3a–3c also reveal that the trade-offs in fitting the three hydrograph components are quite significant. For example, the RMSE error in fitting the slow-flow component Q_S increases from 2.5 to 4 m^3/s as the RMSE error in fitting the driven (rising limb) component Q_D decreases from 25 to 22 m^3/s. The 1.5-m^3/s variation in Q_S error is a fairly large fraction (33%) of the average slow-flow level (4.5 m^3/s), while the 3-m^3/s variation in Q_D error is only 9% of the average rising limb flow level (35 m^3/s).

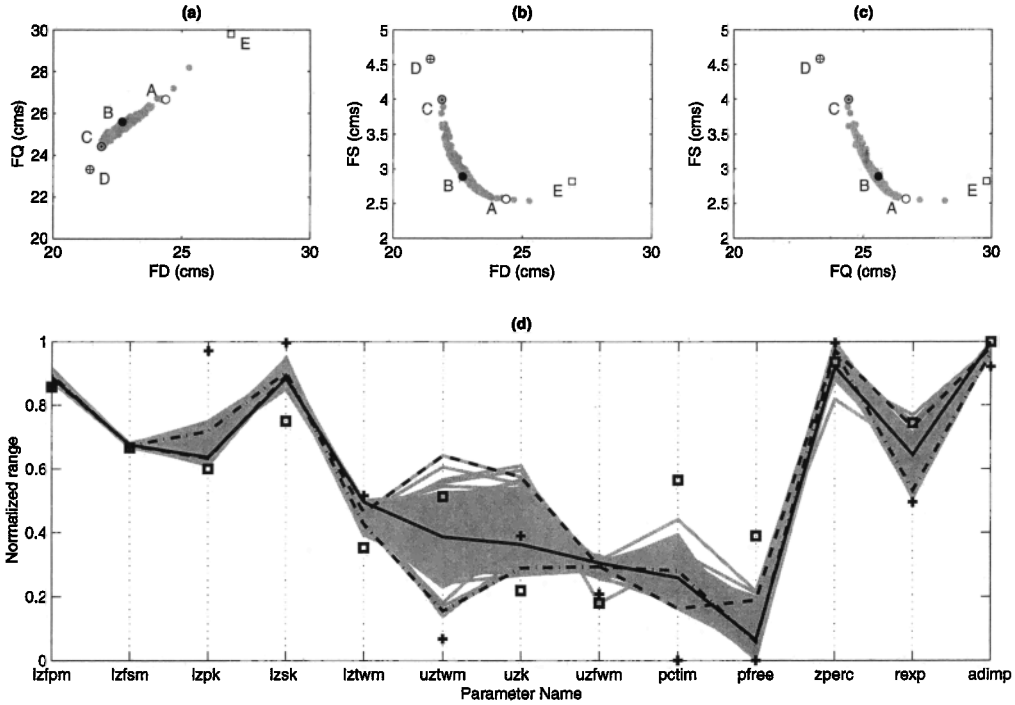

Figure 3. Pareto solutions obtained with the automatic multicriteria approach to calibrate the SAC-SMA model: (a–c) two-dimensional projections of objective space and (d) normalized parameter space. Marked points correspond to 500 Pareto solutions (shaded points), extreme points in RMSE and bias trade-off analysis (A and C), "best" point selected from trade-off analysis (B), Shuffled Complex Evolution (SCE) RMSE solution (D), and SCE log RMSE solution (E). Curves correspond to 500 Pareto solutions (shaded), solution A (dashed), solution B (solid), and solution C (dashed-dotted).

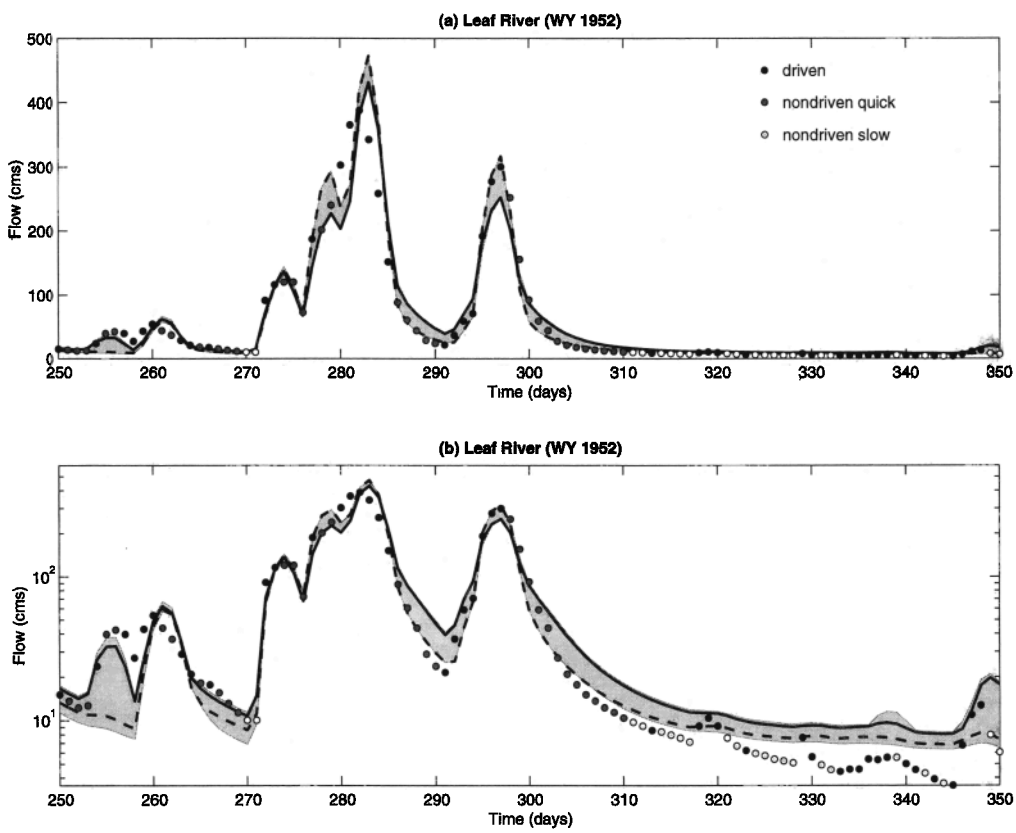

Figure 4. Hydrograph range associated with the Pareto solution set for (a) arithmetic scale and (b) logarithmic scale. The solid circles correspond to the observed streamflow data, the shaded region corresponds to the 500 Pareto solutions, the solid curve corresponds to the minimal FD solution, and the dashed curve corresponds to the minimal FS solution.

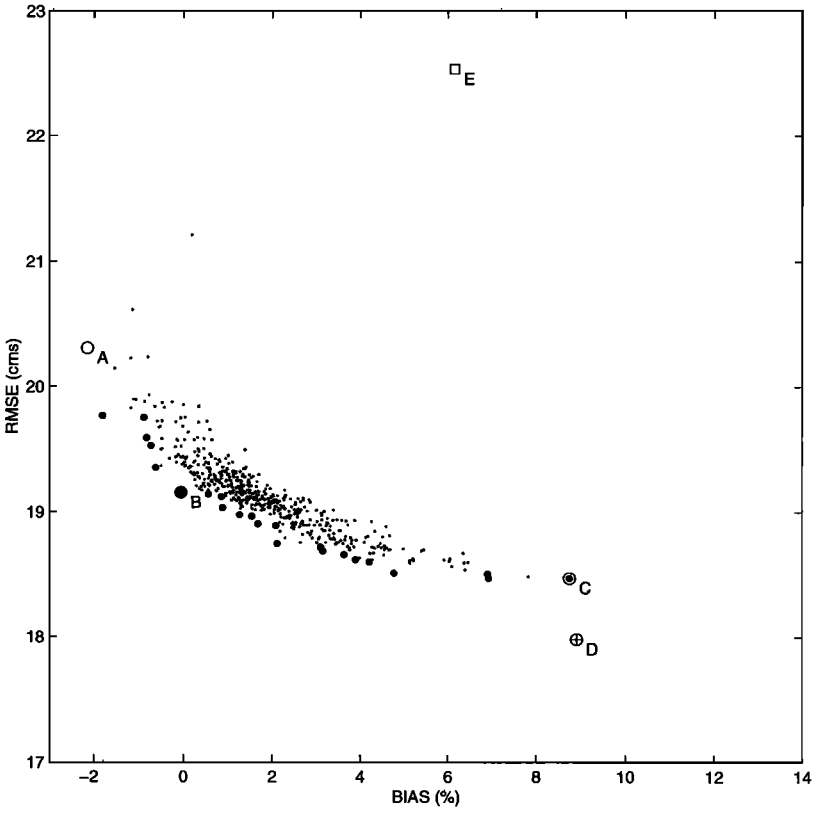

Figure 5. Trade-off analysis of 500 Pareto solutions evaluated over entire calibration period. Marked points correspond to 24 Pareto frontier solutions (solid points), extreme points in trade-off analysis (A and C), "best" point selected from trade-off analysis (B), SCE RMSE solution (D), and SCE log RMSE solution (E).

The normalized parameter plot presented in Figure 3d shows the variability in the parameter values for the 500 Pareto optimal solutions (indicated by the shaded lines). Each line across the graph represents one of the parameter sets. The maximum range for each parameter is the parameter uncertainty remaining after the level one analysis (Table 1). Notice that the parameter uncertainty has been reduced significantly by the multicriteria optimization. In particular, the parameters LZFPM and LZFSM show virtually no sensitivity to the trade-off analysis, indicating that these values are very precisely determined. In addition, the parameters LZPK and LZSK show only small amounts of variability. These results provide strong support for the claim by *Peck* [1976] that estimates for these four lower zone parameters (which are primarily associated with the shape of the slow-flow component of the hydrograph) can be easily obtained from the observed data (see section 2.2).

In general, there is larger variability in the estimates for PCTIM, UZTWM, UZK, ZPERC, REXP, and PFREE. These parameters are associated with partitioning of the hyetograph into quick flow (overland and interflow) and infiltration (called percolation in the SAC-SMA model). The variability in these parameters is consistent with the intuitive notion that the structural error in the model is associated primarily with difficulties in correctly modeling the nonlinear and spatially variable infiltration processes at the surface.

Taken together, these results suggest (not surprisingly) that the lumped conceptual representation of lower zone recession processes in the SAC-SMA model is actually quite adequate and that further attempts to improve the model would be most productive if focused on upper zone processes. These results also imply that the long sequences of systematic errors commonly observed in the slow-flow portions of the simulated hydrograph are caused primarily by incorrect estimation of the volume and timing of percolation.

Figure 4a presents the results in the hydrograph space for a 100-day portion of the calibration period. Figure 4b shows the same information in the log-transformed flow space so the slow-flow behavior can be clearly seen. The solid circles correspond to the observed data, the shaded region corresponds to the hydrograph trade-off uncertainty associated with the 500 Pareto optimal solutions, the solid line corresponds to the minimal FD (driven) solution, and the dashed line corresponds to the minimal FS (nondriven slow) solution. Notice that the minimal FD solution tends to fit the peaks better at the expense of overestimating the recessions, while the minimal FS solution fits the recessions better at the expense of overestimating the peaks (the minimal FQ solution is almost identical to the minimal FD solution and is not plotted). More interesting, however, is that the uncertainty region as a whole tends to overestimate the hydrograph recessions, again supporting the inference that the percolation estimates are incorrect (biased).

4.3. Selecting a "Best" Parameter Set

The analysis presented above illustrates how the automatic multicriteria approach generates a set of Pareto optimal solutions which provide useful information about the characteristic behaviors of the model and its strengths and limitations. However, to use the model for on-line streamflow forecasting, it is desirable to select a single representative parameter set that provides an acceptable trade-off in fitting of the different parts

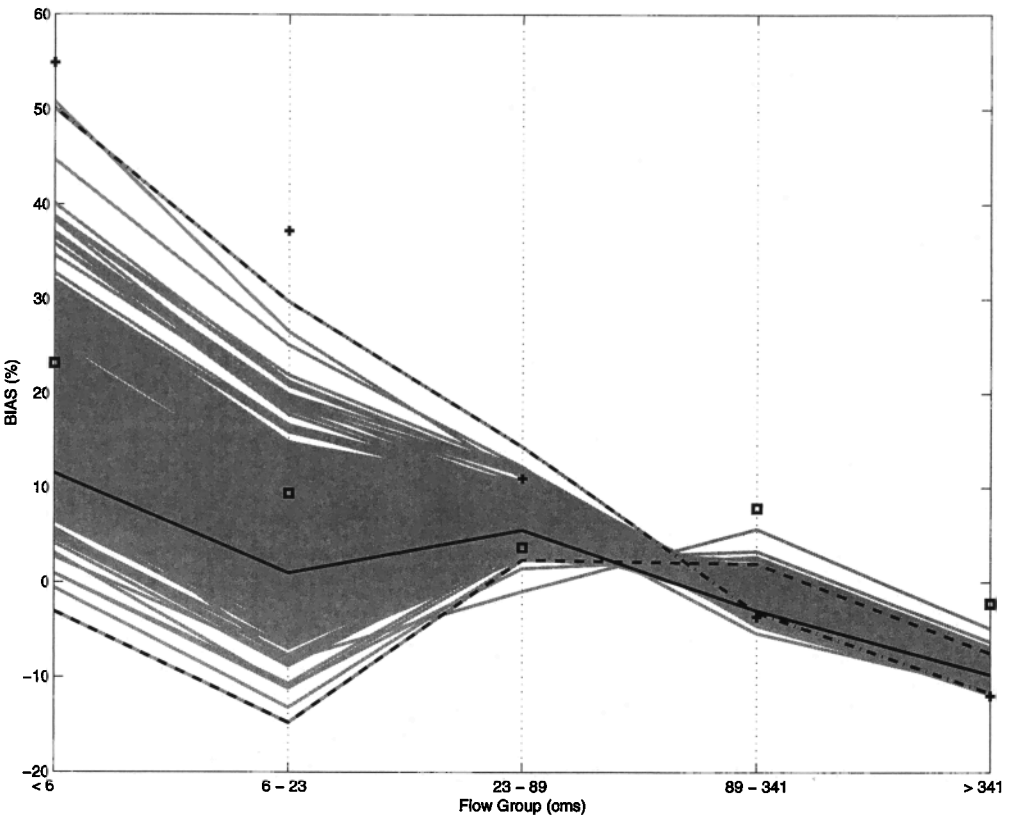

Figure 6. Evaluation of model performance over calibration period in terms of flow groups. Marked points and lines correspond to 500 Pareto solutions (shaded lines), multicriteria extreme point solution A (dashed line), best multicriteria solution B (solid line), multicriteria extreme point solution C (dashed-dotted line), SCE RMSE solution D (plus signs), and SCE log RMSE solution E (squares).

of the hydrograph. For this purpose, we draw again on the standard methods employed in manual calibration by NWS hydrologists. Two important considerations are that the long-term bias of the model simulations is as close to zero as possible (i.e., the model preserves the observed water balance) and the overall residual variance is relatively small. A third important consideration is that the bias by flow level also tends to be close to zero. Figure 5 shows a bicriterion plot of the overall 11-year percent bias against residual variance for the 500 potential solutions generated by the multicriteria optimization run. The pattern of the points clearly indicates a trade-off between the bias and variance: The points with minimal variance tend to have strong positive bias, whereas the points having close to zero bias have somewhat larger error variances. This plot shows that the classical approach, which uses a single overall RMSE criterion for model calibration, may give small error variances but at the expense of significant model bias (leading to parameter estimates that "fit the data" but are unacceptable to manual calibration experts). *Lindstrom* [1997] also acknowledged this fact and proposed a single-criterion approach to correct for it based on a weighted combination of variance and bias measures.

Figure 5 also shows that only 24 of the original 500 points fall on the Pareto frontier (i.e., these 24 points are superior to the remaining 476 points in a multicriteria sense). The three points marked A, B, and C are of special interest; A and C represent the extreme points in the trade-off analysis, while B has the smallest overall bias (essentially zero). Figure 6 shows the percent bias by flow group for each of the 500 solutions, with points A, B, and C highlighted. Notice that the minimum variance point C (dashed-dotted curve) tends to severely overestimate the low and medium flow portions of the hydrograph, while point A (dashed curve) tends to underestimate the low-flow region. However, point B (solid curve) seems to be relatively balanced across the full range of flows. On the basis of this analysis we would select point B as the parameter set to be used for streamflow forecasting.

The relative locations of the points A, B, and C in the original multicriteria (FD, FQ, and FS) space are shown in Figure 3. Points A and C fall near opposite ends of the trade-off surface. However, point B, selected to have minimal overall bias, seems to represent a reasonable compromise between matching of the three criteria (FD(B) = 22.7, FQ(B) = 25.6, and FS(B) = 2.9). For completeness, Figure 3d shows the relative position of the three points in the parameter space.

4.4. Comparison With Conventional Single-Criterion Calibration

To demonstrate the advantages of the level two multicriteria approach, we show a comparison with the results obtained by conventional single-criterion automatic calibration [e.g., *Sorooshian and Gupta*, 1993]. The SAC-SMA model was calibrated by separately fitting to the RMSE criterion and the log RMSE criterion (RMSE evaluated after log transformation of the flows) using the SCE-UA global optimization algorithm [*Duan et al.*, 1993]. The SCE RMSE and SCE log RMSE calibration results have been plotted as points D (pluses) and E (squares) in Figures 3, 5, and 6 for easy comparison with the

Table 2. Overall Statistics for Selected Parameter Sets

	Bias, %		RMSE, m^3/s	
	Calibration	Evaluation	Calibration	Evaluation
Multicriteria (B)	−0.05	6.5	19.2	19.7
Single Criterion (D)	8.9	14.4	18.0	18.8
Single Criterion (E)	6.1	11.7	22.5	24.6

multicriteria results. From Figures 3a–3c and Figure 6 we see (as expected) that the SCE RMSE solution is biased toward fitting the high-flow portions of the hydrograph (particularly the nondriven quick (FQ) portion) at the expense of poor fitting of the nondriven slow portion, while the SCE log RMSE solution is biased toward fitting the low-flow portions of the hydrograph at the expense of poor fitting of the driven and nondriven quick portions. In general, Figures 3, 5, and 6 show that the SCE RMSE (point D) parameter values are similar (and provide similar performance) to those for point C (Pareto set solution having minimum overall variance) and tend to provide hydrographs with larger overall bias.

To verify the consistency and reliability of the results, the performance of the multicriteria estimate B and the single-criterion estimates D and E were evaluated over a 36-water-year period of the available data. Table 2 shows that the overall percent bias and variance statistics for the parameter estimates change (increase) in a similar way from calibration to evaluation period. However, the multicriteria estimate B provides smaller overall bias. Figure 7 shows the percent bias and variance statistics broken out for each of the 36 individual water years. The three columns show the statistics plotted against the mean annual flow (as a measure of "wetness") for the three parameter sets B (Figures 7a and 7b), D (Figures 7c and 7d), and E (Figures 7e and 7f). Figure 8 shows the frequency plots for each statistic. On the average, the percent bias for wet years tends to be similar for all three parameter sets, while the percent bias for dry and average years tends to be larger (more positively biased) for the single-criterion estimate D (Figures 7a, 7c, and 7e). This difference in the statistical tendencies can be seen clearly in the percent bias frequency plot (Figure 8). Further, the annual RMSE statistic tends to increase with annual wetness for all three parameter sets (Figures 7b, 7d, and 7f), but the statistical tendency is similar (on average) for the B and D parameter sets and significantly better than the E parameter set. In general, we can conclude that the multicriteria estimate B provides statistically similar distributions of annual error variance as estimate D but with smaller overall bias.

4.5. Comments About "Equifinality"

It has been suggested [e.g., see *Beven and Binley*, 1992] that many models are overparameterized resulting in "equifinality" of model performance associated with widely different values for the model parameters (where we understand equifinality to mean essentially indistinguishable model behaviors). Figure 9 shows a frequency plot of the overall calibration period RMSE for all 500 parameter estimates belonging to the Pareto solution set. The plot shows that about 90% of the points have very similar overall RMSE values, seemingly supporting the argument for equifinality. However, Figures 3a–3c clearly show that these same points do not appear to be similar when examined in terms of their abilities to simulate different portions of the hydrograph, clearly indicating that the points cannot be considered to be equifinal.

We argue therefore that any single overall statistic which aggregates model performance errors over a large range of hydrologic behaviors is a relatively weak test of model performance. One cannot therefore conclude equifinality by recourse to such a test. Furthermore, one should be particularly careful not to infer erroneous conclusions about parameter identifiability without recourse to examination of a number of different measures, each emphasizing a different important aspect of model behavior.

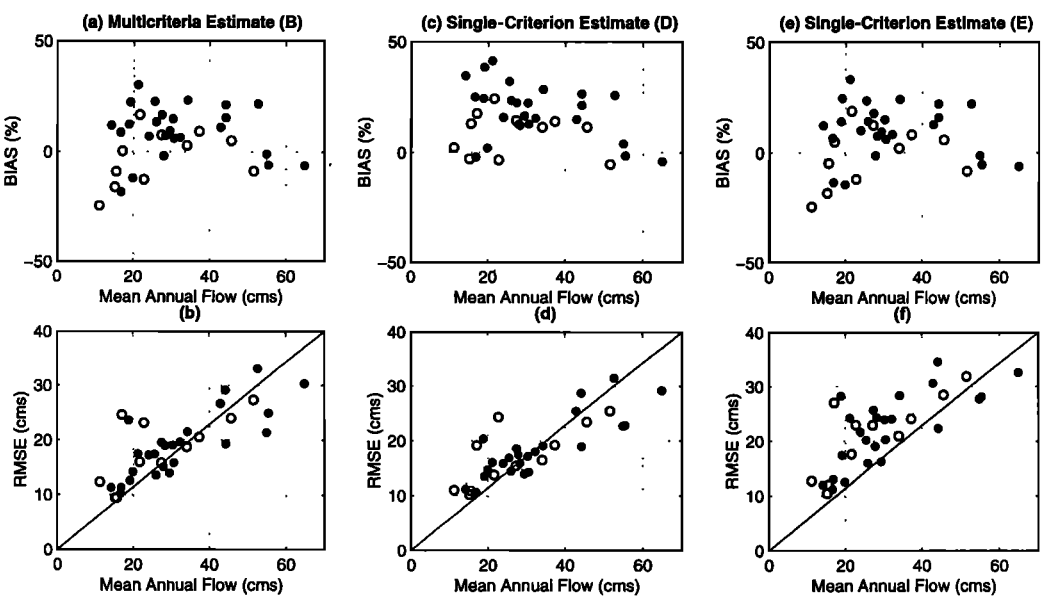

Figure 7. Evaluation of model performance over each of the 36 water years: (a and b) best multicriteria solution B, (c and d) SCE RMSE solution D, and (e and f) SCE log RMSE solution E. Marked points correspond to years used for calibration (solid circles) and years not used for calibration (open circles).

5. Conclusions

There is an increase in the level of effort involved in obtaining progressively more refined estimates for the parameters of hydrologic models. Level two parameter estimates (that account for parameter interactions and model performance trade-offs) are particularly difficult to obtain by manual methods, requiring extensive training as well as considerable experience with the specific model. This expertise is difficult to transfer from person to person and model to model.

This paper explores the relationship between the manual and automatic procedures for model calibration and presents a two-step automatic multicriteria approach for obtaining level two parameter estimates. The approach combines the strengths of manual and automatic methods by emulating the evaluation techniques and strategies used in manual calibration. By solving the resulting multicriteria optimization problem with a computerized global search procedure algorithm, the time and effort required to estimate the parameter range representing the trade-offs in the performance of the model are dramatically reduced. As a result, the attention of the hydrologist can be redirected from the tedious effort of manually searching for the "good" region to the more productive task of evaluating solutions from within the region found with the use of the automatic search algorithm. Finally, a simple strategy, modeled on manual procedures, is suggested as an aid to the hydrologist in analysis and elimination of the solutions within the trade-off range.

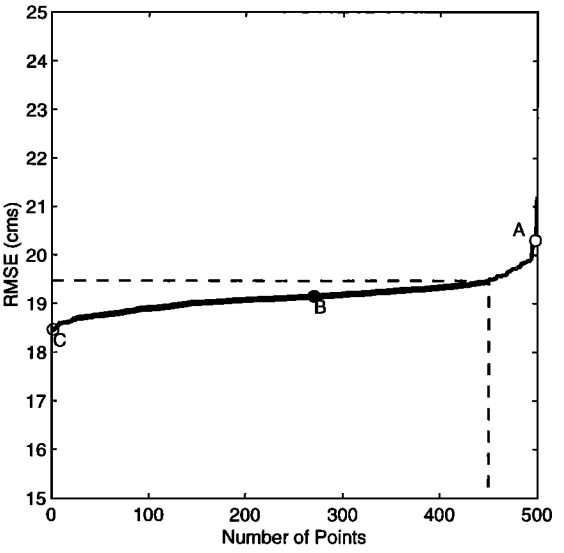

Figure 9. Frequency plot of RMSE over calibration period for all 500 Pareto solutions. Marked points correspond to extreme points in trade-off analysis (solutions A and C) and the "best" multicriteria solution B. The dashed line highlights the 90% of the 500 Pareto solutions which have very similar RMSE values.

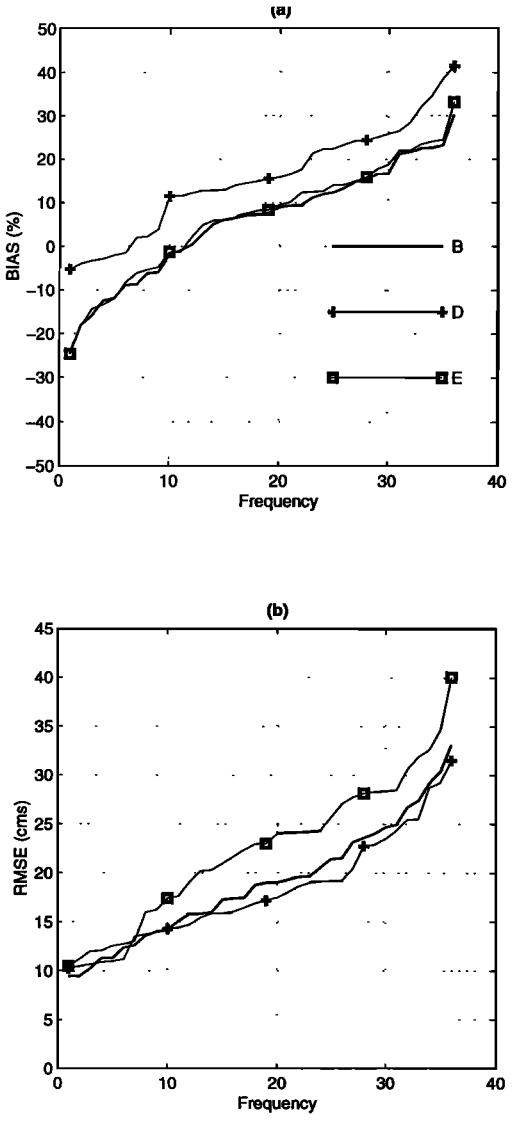

Figure 8. Frequency plots for each statistic shown in Figure 7: (a) percent bias and (b) RMSE. Lines correspond to best multicriteria solution B (solid line), SCE RMSE solution D (solid line with plus symbols), and SCE log RMSE solution E (solid line with squares).

A case study using the SAC-SMA model was used to demonstrate the performance of the new approach. The results indicate that parameter sets selected from within the Pareto region tend to provide consistent and reliable model forecasts. Further, the properties of the Pareto region provide information useful for evaluating the limitations of the various components of the watershed model, thereby pointing toward potential structural improvements. The approach provides a stronger test of model performance than methods that use a single overall statistic to aggregate model errors over a large range of hydrologic behaviors.

Research aimed at further development of the approach is ongoing. This includes more sophisticated partitioning schemes (selection of criteria), evaluation of the sensitivity of the results to the number of criteria, methods to improve selection of preferred solutions from the Pareto region, and uses of the approach for evaluating the appropriate levels of model structural complexity. The results of this work will be reported in due course. As always, we invite dialog with others interested in these topics.

Acknowledgments. Partial financial support for this research was provided by the National Science Foundation (EAR-9418147), by the Hydrologic Research Laboratory of the National Weather Service (grants NA47WG0408 and NA77WH0425), and by the National Aeronautics and Space Administration (NASA-EOS grant NAGW2425).

References

Anderson, E. A., Hydrologic model calibration using the Interactive Calibration Program (ICP), report, Hydrol. Res. Lab., U.S. Natl. Weather Serv., Silver Spring, Md., 1997.

Bastidas, L. A., H. V. Gupta, and S. Sorooshian, The Multi-Objective COMplex evolution algorithm, MOCOM-UA, User's Guide, report, Dep. of Hydrol. and Water Resour., Univ. of Ariz., Tucson, 1999.

Bergstrom, L., Development and application of a conceptual runoff model for Scandinavian catchments, *SMHI Rep. RHO 7*, Swed. Meteorol. and Hydrol. Inst., Norrkoping, Sweden, 1976.

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

Brazil, L. E., Multilevel calibration strategy for complex hydrologic simulation models, Ph.D. dissertation, 217 pp., Colo. State Univ., Fort Collins, 1988.

Brazil, L. E., and M. D. Hudlow, Calibration procedures used with the National Weather Service Forecast System, in *Water and Related Land Resource Systems*, edited by Y. Y. Haimes and J. Kindler, pp. 457–466, Pergamon, Tarrytown, N. Y., 1981.

Brazil, L. E., and W. F. Krajewski, Optimization of complex hydrologic models using random search methods, paper presented at Conference on Engineering Hydrology, Hydraul. Div., Am. Soc. of Civ. Eng., Williamsburg, Va., Aug. 3–7, 1987.

Burnash, R. J. C., R. L. Ferral, and R. A. McGuire, A generalized streamflow simulation system: Conceptual modeling for digital computers, Dep. of Water Resour., State of Calif., Sacramento, 1973.

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

Duan, Q., V. K. Gupta, and S. Sorooshian, A shuffled complex evolution approach for effective and efficient global minimization, *J. Optim. Theory Appl.*, *76*(3), 501–521, 1993.

Gan, T. Y., and G. F. Biftu, Automatic calibration of conceptual rainfall-runoff models: Optimization algorithms, catchment conditions, and model structure, *Water Resour. Res.*, *32*(12), 3513–3524, 1996.

Goldberg, D. E., *Genetic Algorithms in Search, Optimization, and Machine Learning*, 412 pp., Addison-Wesley-Longman, Reading, Mass., 1989.

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., S. Sorooshian, and P. O. Yapo, Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration, *J. Hydrol. Eng.*, *4*(2), 135–143, 1999.

Harlin, J., Development of a process oriented calibration scheme for the HBV hydrologic model, *Nord. Hydrol.*, *22*, 15–36, 1991.

James, L. D., and S. J. Burges, Selection, calibration, and testing of hydrologic models, in *Hydrologic Testing of Small Watersheds*, edited by C. T. Haan, H. P. Johnson, and D. L. Brakensiek, pp. 435–472, Am. Soc. of Agric. Eng., St. Joseph, Mich., 1982.

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

Kuczera, G., Improved parameter inference in catchment models, 2, Combining different kinds of hydrologic data and testing their compatibility, *Water Resour. Res.*, *19*(5), 1163–1172, 1983b.

Kuczera, G., Efficient subspace probabilistic parameter optimization for catchment models, *Water Resour. Res.*, *33*(1), 177–185, 1997.

Lindstrom, G., A simple automatic calibration routine for the HBV model, *Nord. Hydrol.*, *28*, 153–168, 1997.

Linsley, R. K., M. A. Kohler, and J. L. H. Paulhus, *Hydrology for Engineers*, 3rd ed., 508 pp., McGraw-Hill, New York, 1982.

Nathan, R. J., and T. A. McMahon, Evaluation of automated techniques for base flow and recession analyses, *Water Resour. Res.*, *26*(7), 1465–1473, 1990.

National Weather Service (NWS), *National Weather Service River Forecast System User's Manual*, Part IV.4.1, Natl. Weather Serv., Silver Spring, Md., 1997.

Peck, E. L., Catchment modeling and initial parameter estimation for the National Weather Service river forecast system, *NOAA Tech. Memo. NWS HYDRO-31*, Natl. Weather Serv., Silver Spring, Md., 1976.

Rutledge, A. T., Computer programs for describing the recession of ground-water discharge and for estimating mean ground-water recharge and discharge from streamflow records, *U.S. Geol. Surv. Water Resour. Invest. Rep.*, *93-4121*, 45 pp., 1993.

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

Sorooshian, S., and V. K. Gupta, Automatic calibration of conceptual rainfall-runoff models: The question of parameter observability and uniqueness, *Water Resour. Res.*, *19*(1), 251–259, 1983.

Sorooshian, S., Q. Duan, and V. K. Gupta, Calibration of rainfall-runoff models: Application of global optimization to the Sacramento soil moisture accounting model, *Water Resour. Res.*, *29*(4), 1185–1194, 1993.

Sugawara, M., I. Watanabe, E. Ozaki, and Y. Katsuyama, Tank model with snow component, research notes, 293 pp., Natl. Res. Cent. for Disaster Prev., Ibaraki-ken, Japan, 1984.

Thyer, M., G. Kuczera, and B. C. Bates, Probabilistic optimization for conceptual rainfall-runoff models: A comparison of the shuffled complex evolution and simulated annealing algorithms, *Water Resour. Res.*, *35*(3), 767–773, 1999.

Wang, Q. J., The genetic algorithm and its application to calibrating conceptual rainfall-runoff models, *Water Resour. Res.*, *27*(9), 2467–2471, 1991.

Whitley, D., The genitor algorithm and selection pressure: Why rank-based allocation of reproductive trials is best, paper presented at Third International Conference on Genetic Algorithms, 1989.

Yapo, P. O., H. V. Gupta, and S. Sorooshian, A multiobjective global optimization algorithm with application to calibration of hydrologic models, *HWR Tech. Rep. 97-050*, Dep. of Hydrol. and Water Resour., Univ. of Ariz., Tucson, 1997.

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

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

D. P. Boyle, H. V. Gupta, and S. Sorooshian, Department of Hydrology and Water Resources, University of Arizona, P.O. Box 210011, Tucson, AZ 87521-0011. (boyle@hwr.arizona.edu; gupta@hwr.arizona.edu; sorooshian@hwr.arizona.edu)

(Received October 8, 1999; revised June 13, 2000; accepted July 5, 2000.)