Calibrating a Land Surface Model of Varying Complexity Using Multicriteria Methods and the Cabauw Dataset

Y. XIA AND A. J. PITMAN

Department of Physical Geography, Macquarie University, Sydney, Australia

H. V. GUPTA

Department of Hydrology and Water Resources, The University of Arizona, Tucson, Arizona

M. LEPLASTRIER

Department of Physical Geography, Macquarie University, Sydney, Australia

A. HENDERSON-SELLERS

Environment Division, Australian Nuclear Science and Technology Organization, Sydney, Australia

L. A. BASTIDAS

Department of Hydrology and Water Resources, The University of Arizona, Tucson, Arizona

(Manuscript received 13 November 2000, in final form 18 July 2001)

ABSTRACT

The multicriteria methodology, which provides a means to estimate optimal ranges for land surface model parameter values via calibration, is evaluated. Following calibration, differences between schemes resulting from effective parameter values can be isolated from differences resulting from scheme structure or scheme parameterizations. The method is applied to the Project for the Intercomparison of Land Surface Parameterization Schemes (PILPS) phase-2a data from the Cabauw site in the Netherlands using the Chameleon Surface Model (CHASM) as the surrogate for a range of land surface schemes. Simulations are performed calibrating six modes of CHASM, representing a range of land surface complexity, against observed net radiation and latent and sensible heat fluxes. The six modes range from a simple bucket model to a complex mosaic-type structure with separate energy balances for each mosaic tile and explicit treatment of transpiration, canopy interception, and bare-ground evaporation. Results demonstrate that the performance of CHASM depends on the complexity of the representation of the surface energy balance. If the multicriteria method is used with two observed variables, the performance of the model improves little with incremental increases in complexity until the most complex version of the model is reached. If the multicriteria method is used with three observed variables, the most complex mode is shown to calibrate more accurately and more precisely than the simple modes. In all cases, every calibrated mode performs better than simulations using the default PILPS phase-2a parameters. The performance of the most complex mode of CHASM suggests that more complex representations of the surface energy balance generally improve the calibrated performance of land surface schemes. However, all modes, when calibrated, retain a residual error that most likely is due to parameterization errors included in the scheme. Most error is contained in the simulation of the latent heat flux, which suggests that, to improve CHASM further, the representation of the surface hydrological processes should be developed. Thus, the multicriteria method provides a means to assess the performance of a single model or group of land surface models and provides guidance as to the directions scheme development should take.

1. Introduction

Developments in the parameterization of land surface processes over the last two decades have tended to add complexity to the land surface schemes or models used to represent the surface energy and water balance in climate and weather prediction simulations. The intercomparison of existing land surface models has led to the identification of large differences in the partitioning of available water between runoff and evaporation and in the partitioning of available energy between sensible and latent heat fluxes (Henderson-Sellers et al. 1995;

Corresponding author address: Professor A. J. Pitman, Dept. of Physical Geography, Macquarie University, North Ryde, NSW 2109, Australia.
E-mail: apitman@penman.es.mq.edu.au

TABLE 1. Summary of CHASM's modes (after Desborough 1999).

Surface mode	Stability correction	Surface resistance	Canopy interception	Bare-ground evaporation	Canopy resistance	Temperature differentiation
EB	√	X	X	X	X	X
RS	√	√	X	X	X	X
RS-I	√	√	√	X	X	X
RS-GI	√	√	√	√	X	X
SLAM-1T	√	√	√	√	√	X
SLAM	√	√	√	√	√	√

Shao and Henderson-Sellers 1996; Chen et al. 1997; Dirmeyer et al. 1999). Differences among models are a combination of differences in model structure and model parameters. Although intercomparison efforts have attempted to remove differences resulting from parameters by setting values in all schemes to be numerically identical, no effective mechanism existed to ensure that the parameter values were *effectively* the same. It has become clear that the effective meaning of parameters varied across land surface schemes involved in the Project for the Intercomparison of Land Surface Schemes (PILPS; Chen et al. 1997; Desborough 1999). There is therefore a need for a new methodology through which intercomparisons of this kind can be improved (cf. Henderson-Sellers 1993).

One way to ensure that parameters across a range of schemes are effectively identical and have the same physical meaning is to employ a method of parameter calibration. Carefully choosing and adjusting physical parameters for land surface models can improve the simulation of some quantities (Henderson-Sellers 1996; Dirmeyer et al. 1999), but most methods used in intercomparison exercises are subjective. The adjustment of some parameters can be difficult if they are not easily measured (e.g., surface resistance and soil hydraulic conductivity). Several objective techniques recently have been developed to select and to adjust the parameters in land surface models. For example, Sellers et al. (1989) used an iterative loop driven by a least squares reduction program and reliable micrometeorological measurements taken over the Amazonian tropical forest to estimate and to optimize physiological parameters in the Simple Biosphere Model. Their results showed that the specification of optimal parameters improved the simulation of sensible and latent heat fluxes. Franks and Beven (1997) employed the generalized likelihood uncertainty estimation technique to reduce the uncertainty in the fluxes simulated by a simple soil–vegetation–atmosphere transfer scheme. Most recent, Gupta et al. (1999) used a multicriteria parameter estimation method to estimate the ranges of optimal parameter values. They showed that the Biosphere–Atmosphere Transfer Scheme (BATS) model performance improves when its parameters were optimized using the multicriteria method.

In this study, we apply the multicriteria calibration method to six modes of the Chameleon Surface Model (CHASM; Table 1; Desborough 1999; Desborough et al. 2001) and analyze the performance of the model in these six modes in the simulation of the turbulent energy fluxes and net radiation. Our aim is to demonstrate the use of the multicriteria methodology with CHASM and to assess how much of the differences simulated by the modes of CHASM in PILPS-like experiments can be removed through calibration. The residual differences that cannot be removed by calibration should be due therefore to differences in the parameterization of the land surface. These parameterizations provide a guide to the relationship between these differences and model complexity, because the six modes of CHASM encapsulate the range of complexity used in most existing land surface models.

2. Method, data, and modeling framework

a. The multicriteria calibration methodology

Relatively little work in land surface modeling has focused on the errors caused by parameter uncertainty despite the common use of calibration in hydrology where accuracy is important for optimal design (Sorooshian and Dracup 1980) and where considerable effort has been devoted to improving forecasting (Sorooshian and Gupta 1983; Sorooshian et al. 1993; Gupta and Sorooshian 1983; Hendrickson et al. 1988). The recent development of the multicriteria calibration methodology provides an effective and efficient means of reducing the uncertainty in parameter values. The methodology was developed by Gupta et al. (1998) from a single-criteria method developed by Duan et al. (1994) that is widely used in hydrological modeling. Gupta et al. (1998) have used the multicriteria methodology to estimate the reasonable ranges of optimal parameters for the BATS land surface model. Full details of the multicriteria calibration methodology are given by Gupta et al. (1998) and Yapo et al. (1997), and a brief summary is presented below.

The first step in using the multicriteria calibration methodology is to define the feasible parameter range for each parameter to be calibrated. This range, the feasible parameter space, is then sampled. The distance between model results and observations (the model output residual) is then calculated using one or more objective functions. The objective function is generally derived from maximum likelihood or Bayesian theory

Table 2. The default, lower, and upper bounds for parameters and initial values of three state variables using in CHASM. The default values come from Beljaars and Bosveld (1997) and were used in PILPS phase 2a (Chen et al. 1997).

Parameter	Description	Default	Lower	Upper
Model parameters				
ALBG	Bare-ground albedo	0.2	0.1	0.4
ALBN	Snow albedo	0.75	0.7	0.9
ALBV	Vegetation albedo	0.23	0.1	0.4
ALEAFM	Leaf area index seasonality parameter	2	1	6
ALEAFS	Max leaf area index	1.5	1	3
FVEGM	Max fractional vegetation cover	0.95	0.2	0.95
FVEGS	Fractional vegetation cover seasonality	0.7	0.2	0.95
RCMIN	Canopy resistance (s m^{-1})	40	0	300
RHON	Snow density (kg m^{-3})	100	50	100
WRMAX	Available water holding capacity (kg m^{-2})	141	10	1000
ZCOL	Soil color index	5	4	6
Z0G	Bare-ground roughness length (m)	0.01	1×10^{-4}	0.01
Z0N	Snow surface roughness length (m)	4×10^{-4}	1×10^{-4}	6×10^{-4}
Z0V	Vegetation roughness length (m)	0.15	0.01	2.5
Model state variables				
TS	Aerodynamic surface temperature (K)	279	275	310
WN	Snow mass (kg m^{-2})	0	0	10
WR	Available moisture in root zone (kg m^{-2})	141	0	500

to measure a specific statistical characteristic of the output residual (Gupta et al. 1998). Because model calibration is a multiobjective problem (i.e., errors result from many sources), it is unlikely that any single objective function is best suited for model calibration; hence, the multicriteria calibration methodology allows for several objective functions to be used to measure different statistical properties of the output residual.

Once the feasible parameter space has been sampled, the multicriteria calibration methodology selects a set of parameter values that, based on the shape of the objective function space from the previous step, minimizes each of the objectives and therefore reduces the error. The methodology terminates when the process has converged to a ‘‘Pareto set.’’ Within the Pareto set, each parameter set is better than the others for at least one of the objectives but no parameter set is better than another for all of the objectives. So, within the Pareto set every parameter set is considered equal in a multiobjective sense. Because of errors (in the measurements and in the model structure), the Pareto set is not unique.

The multicriteria calibration methodology only needs one optimization run to estimate the Pareto set and is efficient because knowledge gained on the most likely position of the global optimum is retained between steps. The method does not require sampling every parameter set within the Pareto set to find the global optimum. The solutions obtained using this method have been shown experimentally to be a good representation of the Pareto set, even though not every Pareto solution is computed (Yapo et al. 1997). The methodology also finds compromise solutions, because multiple objectives are considered simultaneously in the derivation of the Pareto set. Although moving away from a parameter set that contains a local minimum for one of the objectives worsens that objective, it subsequently finds parameter sets that improve the other objectives. Other single-objective calibration procedures find parameter sets that minimize each objective function, but because the optimization runs are performed separately there is no guidance to the position of these compromise parameter sets. This is a key advantage of the multicriteria method.

b. *Observed data*

The dataset measured at Cabauw (51°58′N, 4°56′E) is a very high quality dataset described in detail by Beljaars and Bosveld (1997). The Cabauw site consists mainly of short grass divided by narrow ditches, with no obstacle or perturbation of any importance within a distance of about 200 m from the measurement site. The climate in the area is characterized as moderate maritime with prevailing westerly winds. The Cabauw data were used in PILPS phase 2a to drive a suite of land-surface schemes and then to validate these models' performance (Chen et al. 1997). The data available at Cabauw include downward shortwave radiation, downward longwave radiation, air temperature, wind at 20 m height, specific humidity at 20 m height, sensible heat flux (SH), latent heat flux (LH), ground temperature, net radiation (Rnet), and ground heat flux with a 30 min interval for 1987. This meteorological forcing is used to drive CHASM in these experiments through a single year with a 30-min time step.

The default parameters (Table 2) were provided as part of the PILPS phase-2a experiment (Chen et al. 1997), and very considerable care was taken in providing the highest-quality set of parameters possible. Beljaars and Bosveld (1997) discuss these data in more detail, but the Cabauw site was chose for PILPS phase 2a following extensive review and the recognition that it was a dataset of unusually high quality in regard to the meteorological forcing, the observed turbulent energy fluxes, and the associated parameters required by a land surface model.

c. *Description of CHASM*

CHASM was developed to explore aspects of land-surface energy balance representation (Desborough 1999). CHASM can be run in a variety of complexity modes within the same modeling environment. Each mode of CHASM utilizes the same parameterization (with the exception of the surface energy balance) and parameters, so switching between modes allows the impact of the addition or removal of a specific aspect of the surface energy balance to be explored. Six CHASM modes are used in this study, ranging from a simple Manabe-type bucket model to a complex mosaic type (i.e., Deardorff 1978) structure with separate energy balances for each mosaic tile (e.g., Koster and Suarez 1992) and explicit treatment of transpiration, canopy interception, and bare-ground evaporation. These six modes (EB, RS, RSI, RSGI, SLAM-1T, and SLAM) are shown in Table 1 and are described below.

A common hydrological module originally described by Manabe (1969) is combined with each CHASM mode. The root zone is treated as a bucket with finite water-holding capacity, and beyond this capacity runoff occurs. The use of a simple hydrological model of this kind has been shown to work well in midlatitude regions (Robock et al. 1995). Runoff also occurs when the fraction of snow cover on the ground exceeds 95%. Apart from moisture in the root zone, water can also be stored as snow or, depending on the mode, on the canopy. All modes share a common six-layer soil temperature module. Each tile is divided into area fractions of vegetation, snow, and ground. Snow cover fractions for ground and foliage surfaces are calculated as functions of the snowpack's depth and density and the vegetation's roughness length (Pitman and Desborough 1996; Desborough and Pitman 1998). The vegetation fraction is further divided into wet and dry fractions if the surface configuration mode allows for canopy interception. Each tile has a prognostic bulk temperature for the storage of energy and a diagnostic skin temperature for the calculation of surface energy fluxes.

The simplest mode of CHASM, EB, is constructed from one tile. The aerodynamic resistance to turbulent transport for heat and moisture is calculated without atmospheric stability correction. Moisture available for evaporation is stored in the root zone and on the surface as snow, resulting in two evaporation sources to which the aerodynamic resistance is applied. The RS mode is the same as EB but with a temporally invariant surface resistance added to the resistance pathway of snow-free evaporation. The aerodynamic resistance is calculated with an atmospheric stability correction. The RSI mode is the same as RS but with explicit parameterization for canopy interception of precipitation. This results in three evaporation sources because water can evaporate from the canopy store. The canopy is divided further into fractions of wet and dry areas depending on the precipitation and evaporation rates. The RSGI mode builds onto the RSI mode through the addition of bare-ground evaporation. Moisture can be stored at the surface for evaporation, and bare-ground evaporation is affected by moisture availability. SLAM-1T builds on RSGI by including a time-variable canopy resistance, which replaces the temporally invariant surface resistance. The most complex mode, SLAM, is the same as SLAM-1T but the land–atmosphere interface is divided into two tiles with the first representing a combination of bare ground and exposed snow and the other reserved for vegetation. The tiles are not necessarily the same size and are area-weighted depending on the individual fractions of the land surface type. A separate surface energy balance is calculated for each tile, which allows for temperature variations across the land–atmosphere interface, a feature not present in the less complex modes.

d. *Experimental design*

CHASM includes 14 parameters for soil and vegetation that were available for calibration (Table 2). In addition, three state variables were initialized at the beginning of the simulation, and the initial value was used as part of the calibration of the model. The ranges of the parameters and the initial values of the state variables are shown in Table 2. To explore the relative value of different observational quantities for calibrating CHASM, four calibration studies were performed. For simplicity of notation, a closed pair of curly braces will denote a calibration test; hence {SH, LH} denotes a multicriteria calibration experiment using observed sensible heat and latent heat fluxes for calibration, {SH, Rnet} used sensible heat and net radiation, {LH, Rnet} used latent heat and net radiation, and {SH, LH, Rnet} used sensible heat, latent heat, and net radiation. Root-mean-square error (rmse) was used as objective function for all experiments. We examined the sensitivity of our results to two other objective functions used by Gupta et al. (1998) (the Nash–Sutcliffe coefficient of efficiency and the mean absolute error), and the results reported in this paper are insensitive to the choice of objective function. We also examined the sensitivity of the results to different seeds and different sets of points initially sampled from the feasible parameter space. These tests indicated that the results presented in this paper were insensitive to the initial sampling point.

3. Applying the multicriteria method to CHASM

a. *Parameter estimation for the two-variable calibration*

Figure 1 shows the results for the three experiments in which CHASM was calibrated against two observed

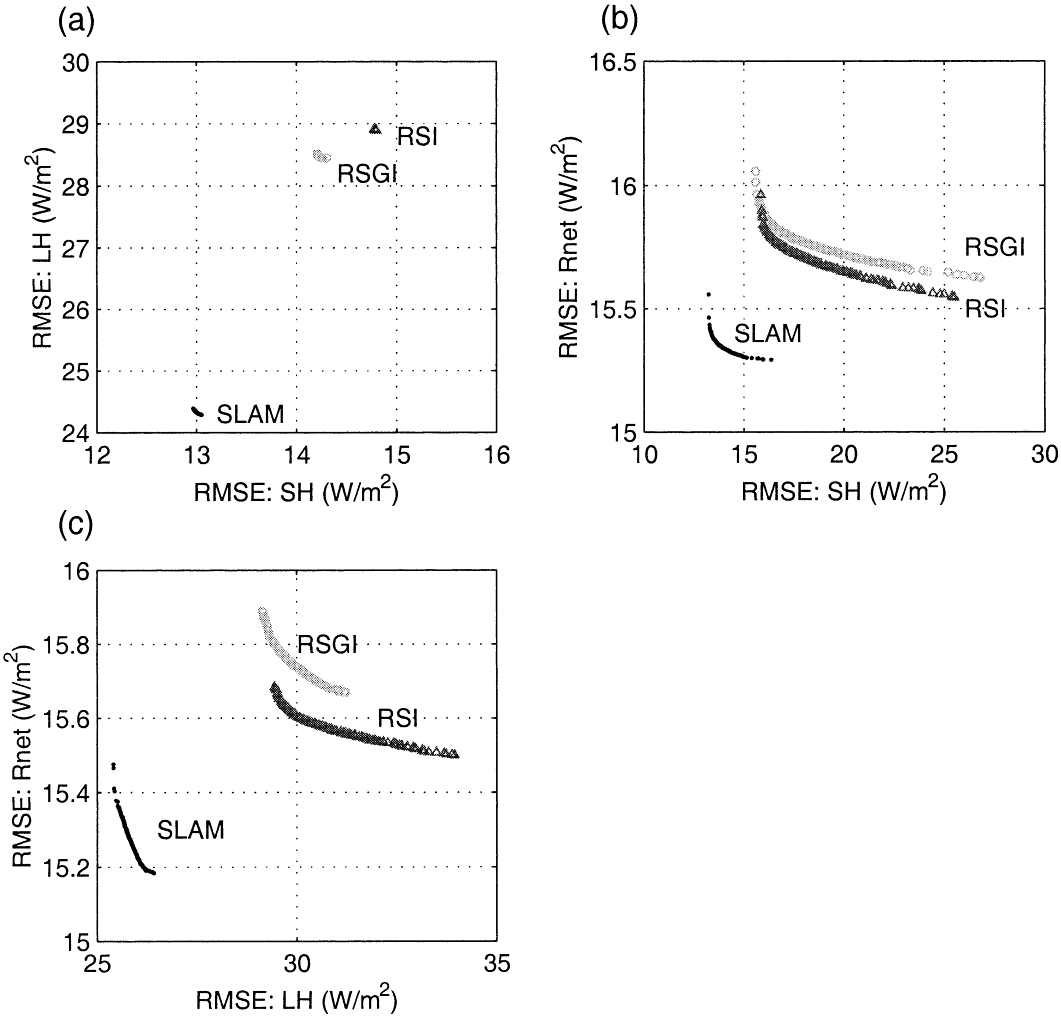

FIG. 1. Pareto fronts simulated for the three experiments: (a) {LH, SH}, (b) {RNET, SH}, and (c) {RNET, LH}, for three modes of CHASM (see Table 1).

quantities ({SH, LH}, {SH, Rnet}, or {LH, Rnet}). Results are plotted for modes RSI, RSGI, and SLAM, because these modes represent the key range of complexity available with CHASM. Each simulation shows a Pareto front as a curve, where each point on the curve represents a parameter set that provides a solution that is equally good if both of the observed quantities are taken into account.

There are two key results in Fig. 1. First, the length of the Pareto front varies between the modes of CHASM, that is, with simpler modes (e.g., RSGI and RSI), more extreme values of rmse are simulated for both the quantities plotted. Second, the position of Pareto set varies among modes, with the most complex mode (SLAM) always lying closer to a zero rmse. This implies that a more complex model can be calibrated better than a simpler model and indicates that a more realistic representation of the surface energy balance provides a greater opportunity to calibrate a model against observed data. The improved calibration is not related to the number of parameters being calibrated because these do not vary between CHASM modes.

Figure 2 shows the results from all six modes for {SH, Rnet} (the results from the other experiments give similar results). The results from the simulations for each mode using the default parameter set [as used in Chen et al. (1997) and see Table 2] are identified in Fig. 2. In all cases, the Pareto set obtained through calibration is a major improvement over the results obtained using the default parameter set, which implies that calibration is useful irrespective of the complexity of the model. The improvement in rmse is mainly in reducing errors associated with Rnet (i.e., calibration improves SH relatively little) except with SLAM, for which calibration improves both fluxes significantly. Figure 2 also shows that the results simulated by CHASM tend to improve little as more complex modes are employed

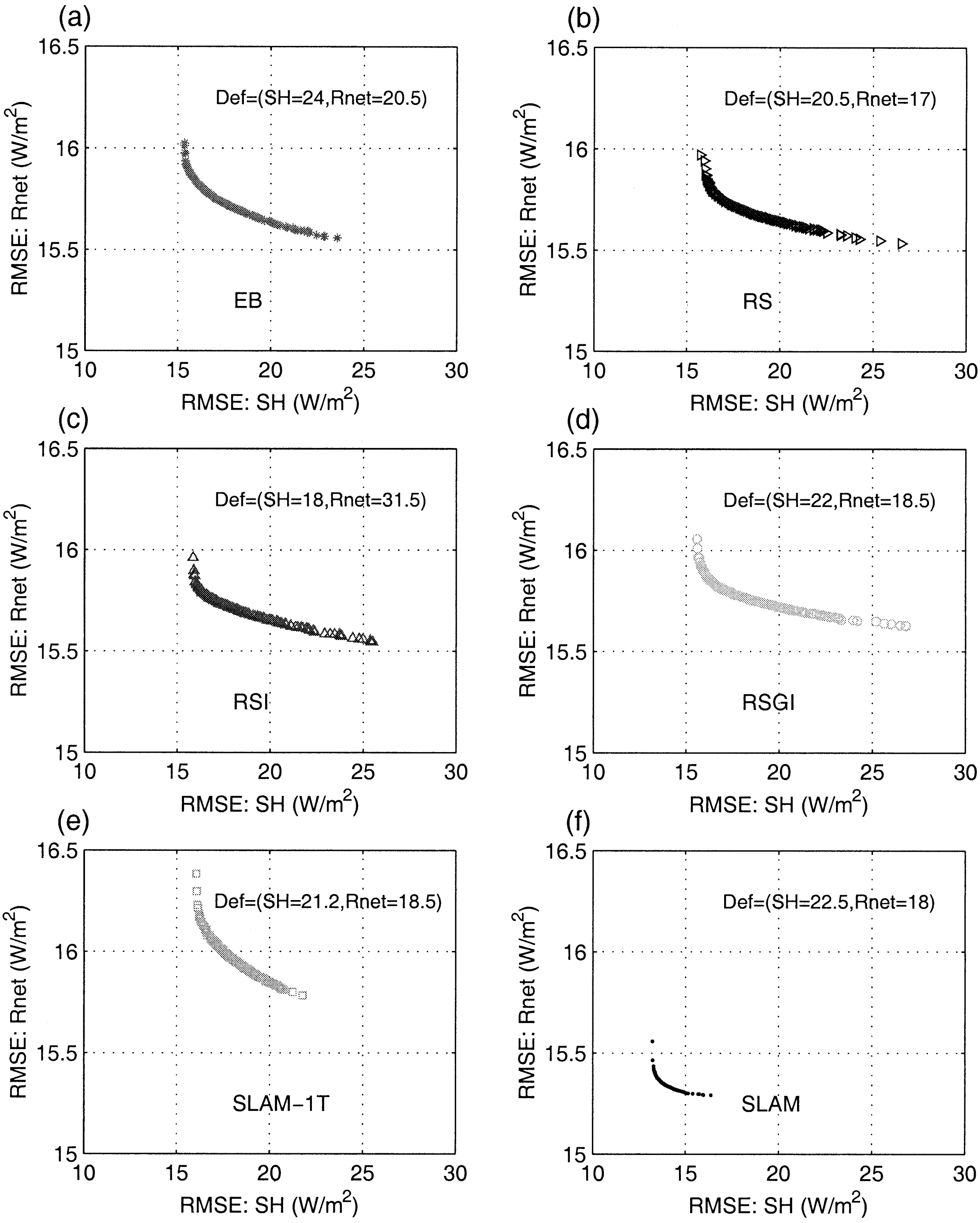

FIG. 2. Results from six modes of CHASM from the {RNET, SH} experiment: (a) EB, (b) RS, (c) RSI, (d) RSGI, (e) SLAM-1T, and (f) SLAM. The results from simulations using the default parameter set (Table 2) are also noted.

until the complexity of SLAM is reached, at which there is a marked improvement in both the length of the Pareto curve (implying more precise calibration) and in the position of the Pareto curve (implying a more accurate simulation).

The results from Figs. 1 and 2 show that the lowest rmse values obtained using CHASM were about 25 W m^{-2} for LH and 13 W m^{-2} for SH (Fig. 1). This rmse error is the "residual" error that cannot be explained by a failure to specify parameter values correctly. This

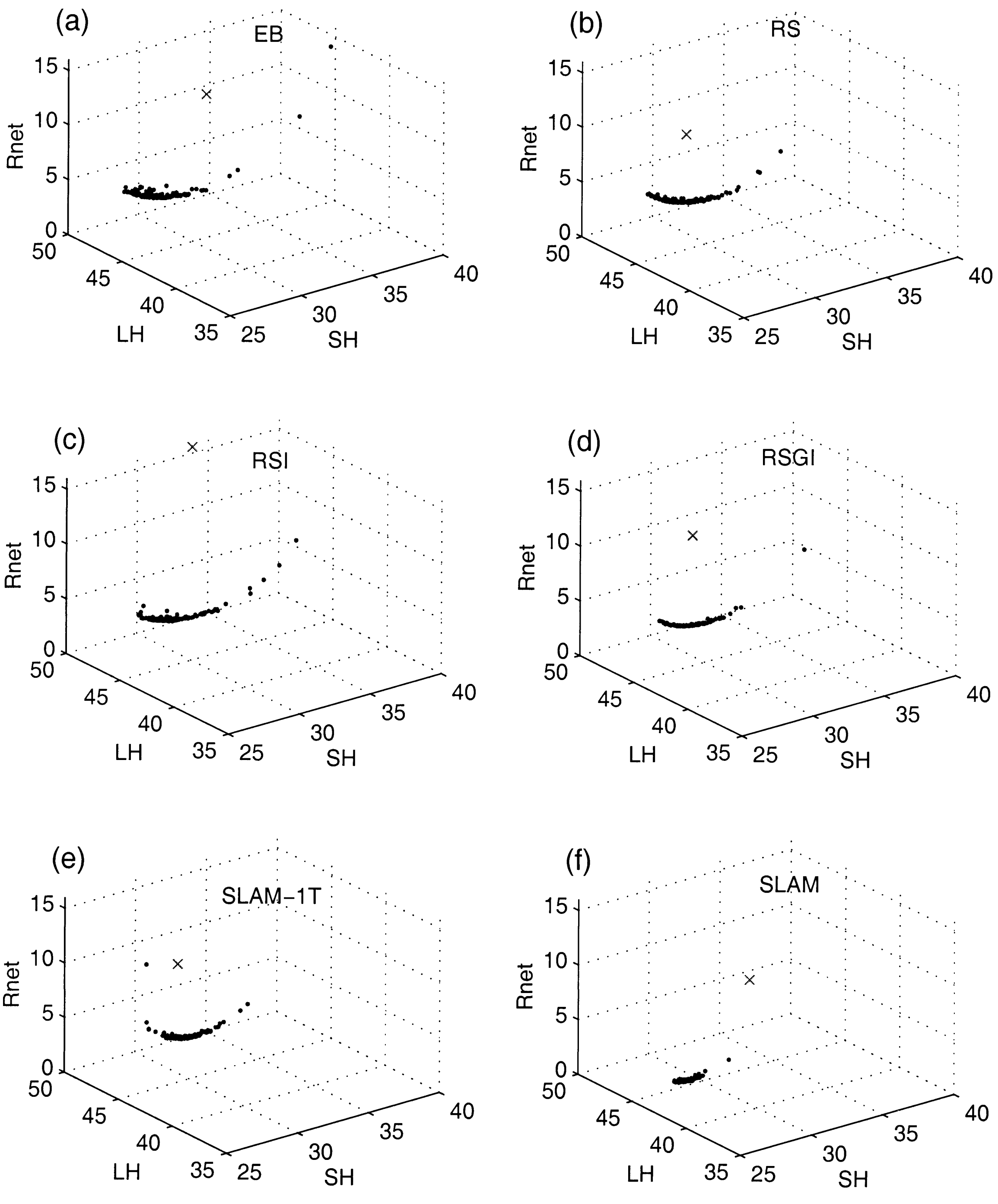

FIG. 3. Results from six modes of CHASM from the {RNET, LH, SH} experiment: (a) EB, (b) RS, (c) RSI, (d) RSGI, (e) SLAM-1T, and (f) SLAM. The results from simulations using the default parameter set (Table 2) are also shown ("×").

residual therefore stems from either errors in the observational dataset or in model parameterization. Given the quality of the Cabauw dataset, the errors probably are related mostly to model error, and, given that a large fraction of the error is in LH rather than SH, this suggests that the modeling of hydrological processes should be the focus for future attempts to improve the model.

b. Parameter estimation in the three-criteria calibration

The multicriteria calibration method can be used to calibrate a model using three observed quantities simultaneously. The final experiment conducted was therefore {SH, LH, Rnet}. Figure 3 shows the rmse ranges calculated for each mode using the default parameters (Table 2) and shows the Pareto front simulated by each mode.

The smallest rmse values are obtained for SLAM, with the other modes showing similar rmse values (Fig. 3). The three-criteria calibration shows a long Pareto front with outliers for all modes. However, in the case of SLAM, the absence of significant outliers means that the extremes of rmse values are lower than the result from the default simulation. With SLAM, the calibration of the model improves the simulation of sensible heat (by 20%–30%), latent heat (by 9%–14%), and net radiation (by 11%–15%) over using default parameters. For simpler modes, although the minimum rmse values simulated for each mode are an improvement over the default, the maximum values are either equivalent to, or worse than, the default. However, this generally is caused either by one or two outliers at the extreme of the Pareto front (see for example EB or RSI in Fig. 3) or by a small fraction of the Pareto front (e.g., RSGI in Fig. 3). The vast majority of calibrated solutions for all modes represent a substantial improvement over simulations using the default parameters.

When default values are compared with the optimized parameter ranges, many of the calibrated parameter values fall well outside of the range of the default parameter range. Figure 4 shows the ranges of five key parameters and the default value prescribed by Chen et al. (1997), and it shows that SLAM always calibrates to a tighter range for each parameter. This result is not related to number of parameters being calibrated given that this number does not vary among CHASM modes. For albedo, the default value specified for the PILPS phase-2a experiments is similar to the range predicted by the multicriteria calibration method for each mode. In contrast, all modes calibrate to a higher value of ALEAFM and a lower value of FVEGM than the default value, although in both cases SLAM is the closest to the default. The calibrated value of Z0V is very much higher in most modes, with the notable exception of SLAM, which predicts a value close to the default. Overall, the default parameter values (which are the best estimates for the Cabauw site) are most closely approximated by SLAM. Anomalous parameter values are obtained for the other modes. This result suggests that the multicriteria method finds anomalous parameter values to compensate for poor model design in the intermediate modes, whereas in the most complex mode, with the highest percentage of explicit and physically realistic parameterizations, the multicriteria method does find generally reasonable parameter values with which to optimize model performance. Thus, a default parameter value that is very different from the calibrated range may indicate that the multicriteria method finds parameter values that minimize the total error in the system, not just errors associated with parameter uncertainty. Although the feasible parameter space restricts the multicriteria method from choosing values that are nonphysical (the feasible parameter space is shown in Table 2), the values derived using the multicriteria method may not be correct because of the presence of errors in model structure and errors in observations.

Despite this caveat, the calibrated modes of CHASM perform much better than when using the default parameters. To show the improvement more clearly, Figs. 5 and 6 show two 5-day time series of sensible heat, latent heat, and net radiation for 1–5 July and 27–31 December for SLAM. The net radiation is simulated very well in both December (Fig. 5) and July (Fig. 6). The default simulations for sensible and latent heat in December show some anomalous behavior (e.g., toward the end of day 3), but overall the calibrated performance is slightly better. In July, SLAM using default parameters clearly overestimates sensible heat and underestimates latent heat every day by about 150 W m^{-2}. The time series generated with the parameter sets that produced the maximum and minimum rmse values in experiment {RNET, SH, LH} are clear improvements over the default. Both sensible and latent heat are simulated very well in both December and July, and the systematic errors in the simulation of these fluxes obtained using the default parameter set are almost completely removed. If daily modeled and observed fluxes are compared for the whole of July and January, the improvement in CHASM's performance resulting from calibration is clear. Figure 7 shows the line of best fit (and accompanying variance r^2 value for sensible heat, latent heat, and net radiation). Net radiation is simulated very well (Figs. 7a,b). In July, a significant improvement is clear in the simulation of sensible and latent heat from the default in terms of the trend line. The r^2 value increases from 0.88 (default) to 0.91 (calibrated) for the sensible heat (Figs. 7c,e). In January, the r^2 values are also improved by calibration for both sensible and latent heat. Figure 8 shows r^2 and the statistic that describes the slope m for sensible and latent heat. In the case of the sensible heat flux (Fig. 8a) there is a clear improvement in the r^2 values in the calibrated experiments but little improvement in the slope. In July, there is a small improvement in the r^2 values, but the slope is considerably improved in the calibrated experiments. For the latent heat flux (Fig. 8b), the improvement in January is in both the slope and the r^2 values. In July, there is no improvement in the r^2 value, but the slope is improved. Overall, calibration moves the solutions toward the 1-to-1 interception point in Fig. 8, indicating that the calibration has improved both the r^2 and slope statistics.

4. Discussion and conclusions

In this paper, we investigated the ability of the multicriteria method to estimate optimized parameters for six CHASM modes that vary in terms of the surface energy balance complexity. The aim of this paper was to examine the multicriteria method using CHASM and to explore the relationship between calibration and model complexity. In phase 2a of PILPS (Chen et al. 1997),

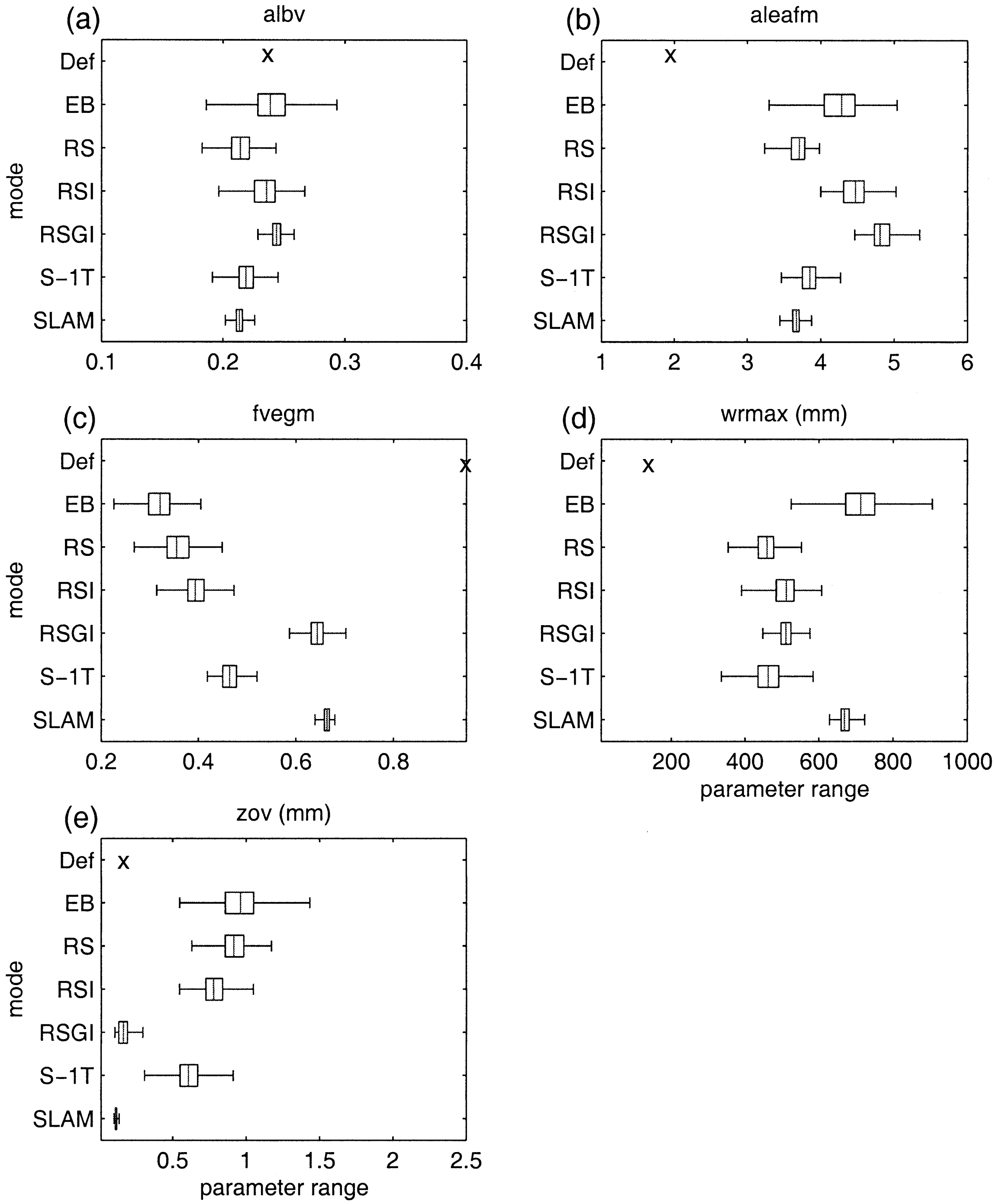

FIG. 4. Calibrated ranges for five parameters for six modes of CHASM: (a) vegetation albedo, (b) maximum fractional vegetation cover, (c) maximum leaf area index, (d) water holding capacity, and (e) vegetation roughness length. The median is in the center of the box, the edge of the box shows the interquartile range, and the bars show the size of a single standard deviation.

errors were reported in the simulation of sensible and latent heat by a range of models. Our hope was to use the multicriteria method to see whether the potential existed to use the methodology to isolate errors resulting from the specification of parameter values given that the effective value of parameters varies across land surface schemes (Chen et al. 1997; Desborough 1999).

The results show that the multicriteria method works in a consistent and robust manner with all six modes of CHASM, and thus the method works with a range of

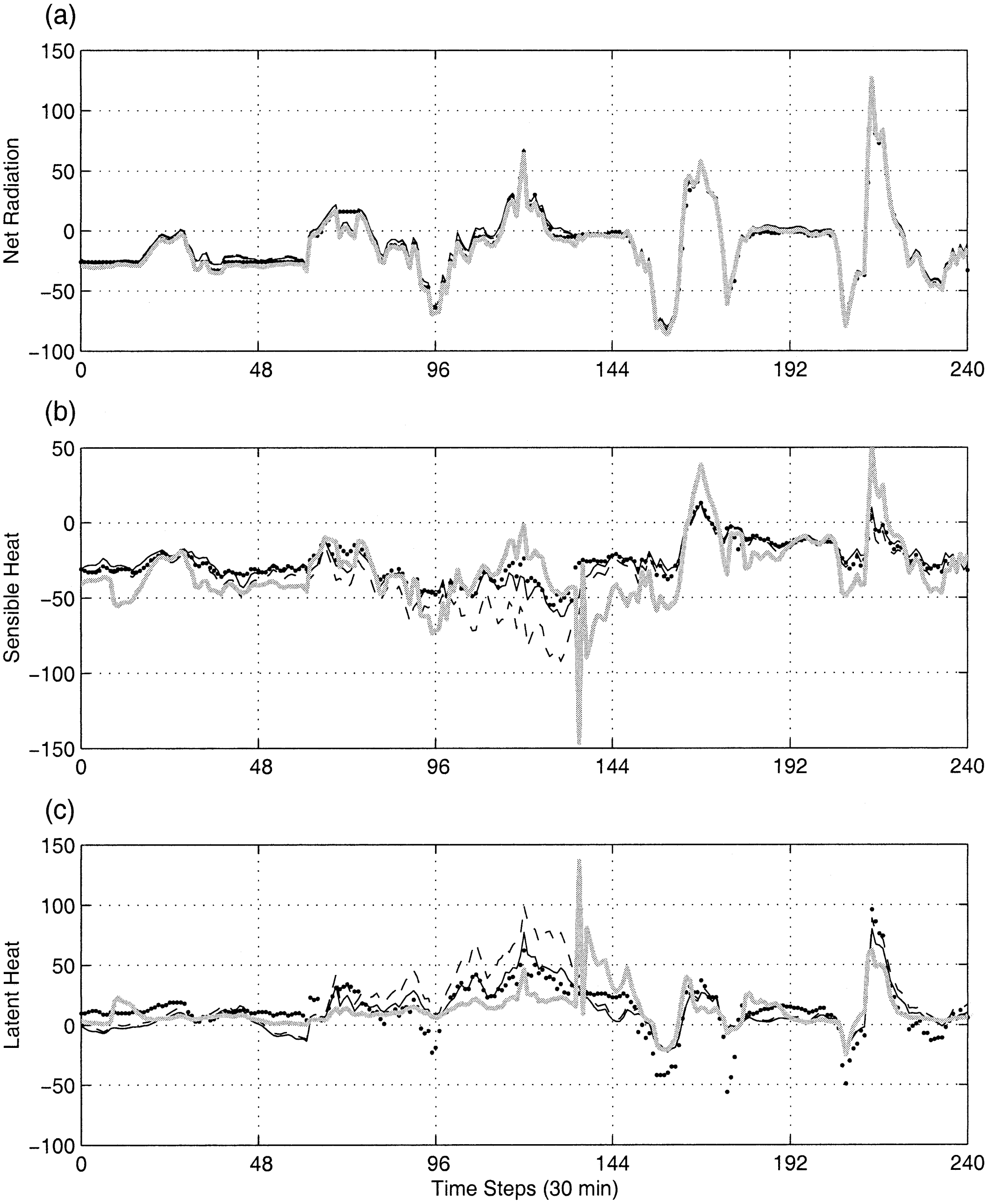

FIG. 5. Simulation of (a) net radiation (W m^{-2}), (b) sensible heat (W m^{-2}), and (c) latent heat (W m^{-2}) for SLAM during 27–31 Dec obtained using the maximum and minimum parameter sets from the Pareto front shown in Fig. 3 for CHASM. The dots represent observed quantities, and the gray line is the result using default parameter values. The solid black line and dashed black line represent the simulations using the parameter sets from the extreme points of the Pareto solution set (see Table 2).

levels of complexity. Results were shown generally to vary relatively little as a function of complexity until the most complex mode, SLAM, was used, which calibrated most accurately (minimizing the rmse) and most precisely (with the shortest Pareto front). When three criteria were used to calibrate the model ({LH, SH, Rnet}), SLAM was shown to perform better than the other modes. However, although the most complex mode calibrates better than the simpler modes, all modes show significant improvement in performance following the calibration of parameter values in comparison with the simulations obtained using the default parameters. SLAM performed best but still retained some residual error. The largest error was in evaporation, providing a guide to where further model development should take place. Further, many of the default parameters, provided

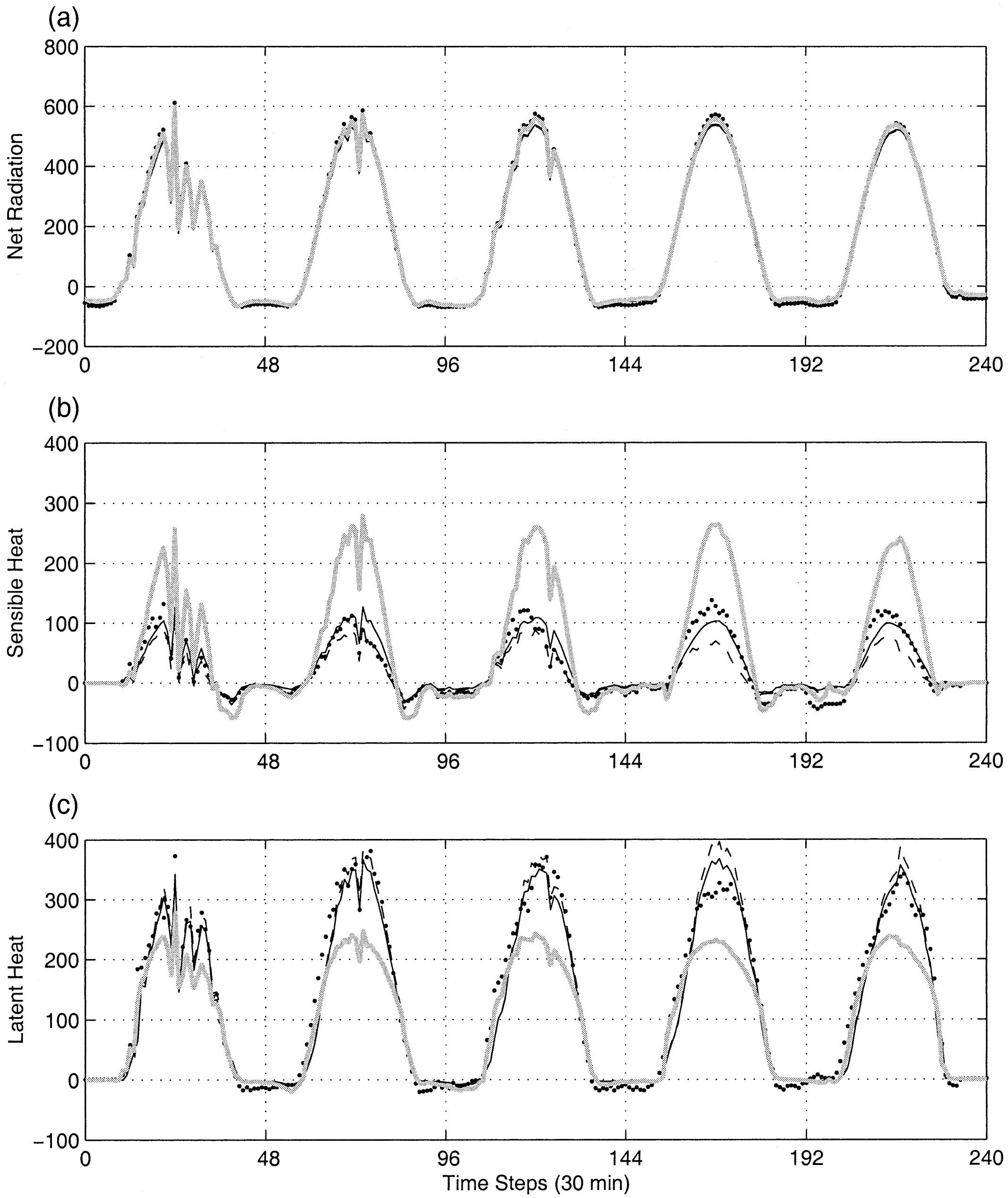

FIG. 6. As in Fig. 5 but for 1–5 Jul.

by Beljaars and Bosveld (1997), fall outside the range obtained via calibration. One example is the maximum leaf area index (ALEAFM), specified as 2.0, for which the best simulations actually were achieved with values of about 3.5. This result may be a peculiarity of CHASM or may be a problem with the observed value for this parameter. The use of the multicriteria methodology provides a way of identifying peculiarities in the parameter data and an avenue for further investigation. However, it is a concern that CHASM calibrates to fairly high values of vegetation roughness length for all modes bar SLAM and RSGI. This result suggests that optimal performance may be achieved via anomalous parameter values that help to compensate for poor model design. This issue needs to be investigated further with a wide range of meteorological forcing data, a variety of surface vegetation types, and examination of the results against data not used in calibration.

One of the key reasons for pursuing this research was the wish to identify the cause of the scatter generally found in the PILPS results (e.g., phase 2a, Chen et al. 1997). Errors in the observations, in the model parameterization, and in the parameter values specified for the model can all contribute to the scatter identified by

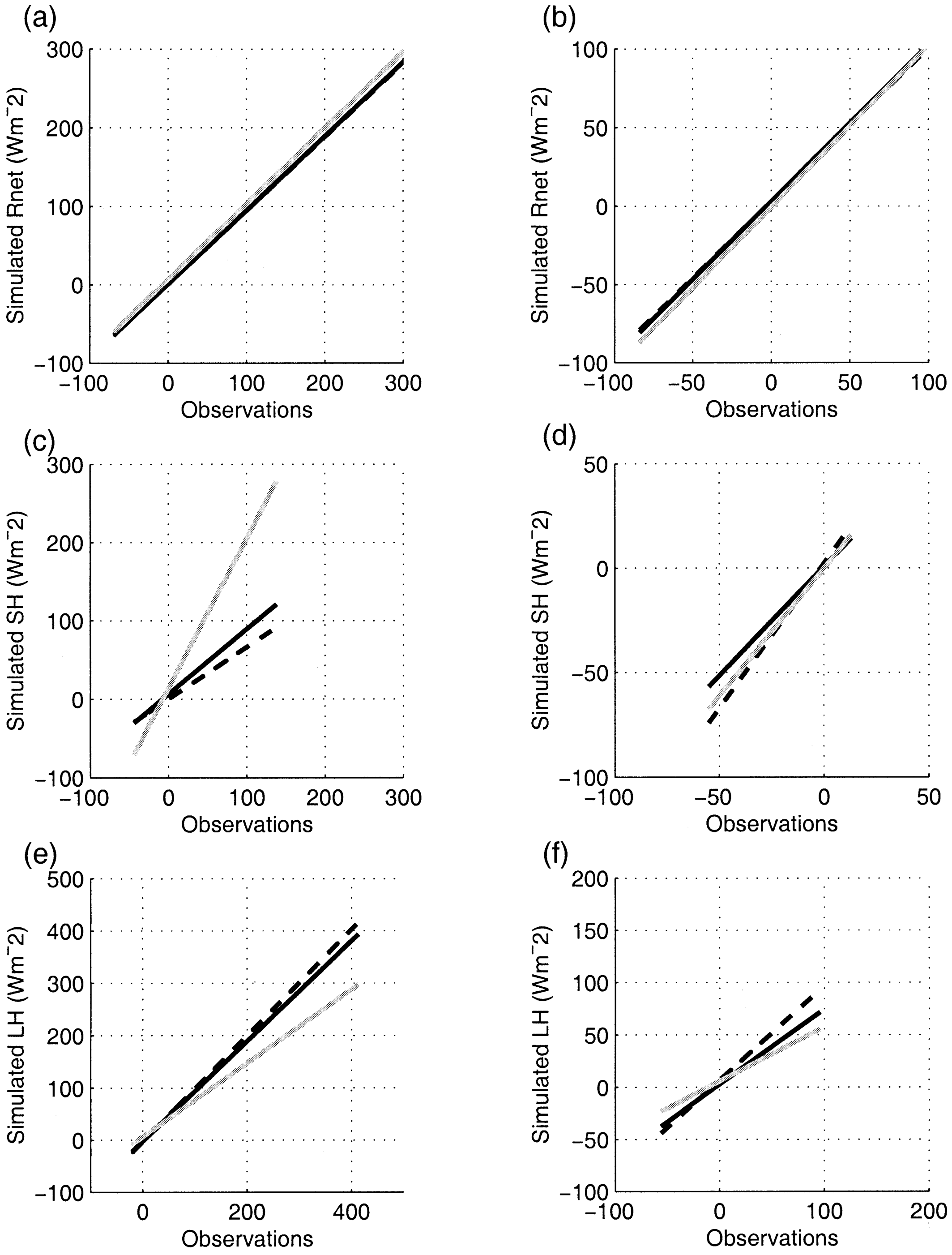

FIG. 7. Lines of best fit for time-step data for (a) net radiation, (c) sensible heat, and (e) and latent heat for Jul. (b), (d), (f). The same, respectively, for Jan. The gray line is for the simulation using the default parameters, and the solid black line and dashed black line represent the simulations using the parameter sets from the extreme points of the Pareto solution set (see Table 2). The r^2 values are given for each simulation.

PILPS. Figure 4 shows that the effective parameter values that maximize the performance of a land surface scheme vary according to the nature of the scheme and that forcing all schemes to use the same parameter data will have different implications for different models. The multicriteria methodology provides a means to remove differences resulting from the specification of parameter values. Allowing all schemes to calibrate to the same data, in an objective way, should maximize the schemes' performance. Because the same observational

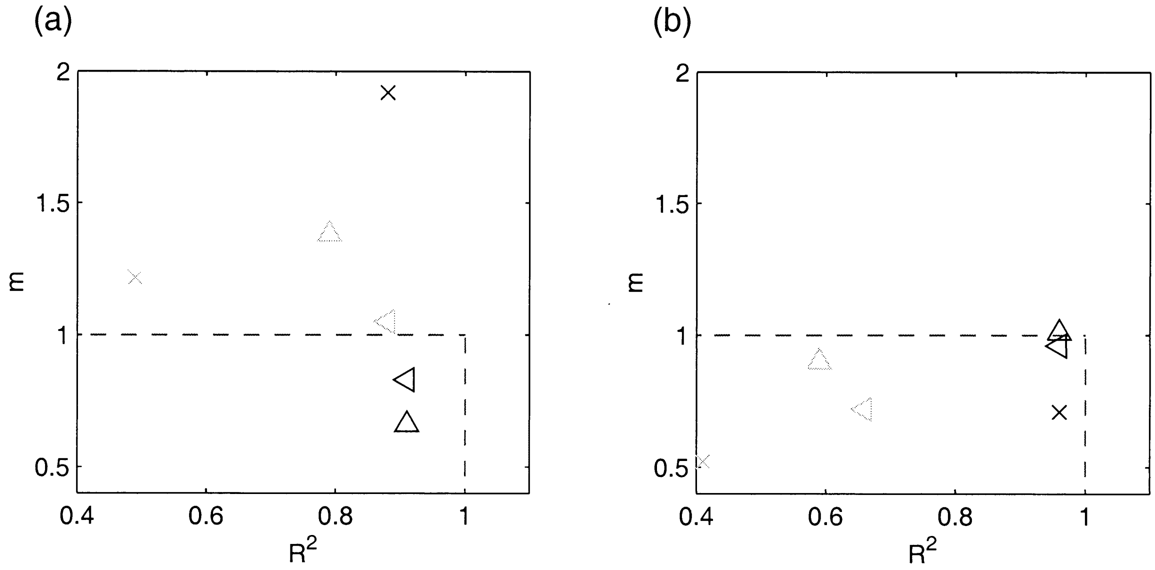

FIG. 8. Values for r^2 and for the slope m for (a) sensible heat flux and (b) latent heat flux. The gray symbols are for Jan, and the black symbols are for Jul for the minimum (△) and maximum (◁) solutions from the three-calibration experiment and the default experiment (×).

data would be used for each land surface scheme, a comparison of residual errors among a range of models would permit the identification of the relative merits of the schemes, independent of implications resulting from choices of parameter values. According to the results of this paper, climate models should include a parameterization of the surface energy balance that is relatively complex. Intermediate levels of complexity, such as adding canopy resistance to a Manabe-type parameterization with a constant surface resistance, would not appear to be effective. This conclusion needs to be explored with a large range of climate-forcing datasets. It is hoped that such an exploration will lead to specific guidance on how to improve land surface models.

The results reported in this paper were derived using point-based data and are not easily extrapolated into the climate-model environment. However, Sen et al. (2001) have explored the impact of calibrating a model in the same way as reported here and then using the calibrated parameters within a climate model. They found that point-based calibration can be used to derive parameters that then can be used in a climate model to improve the simulated climate. Our results, which show that calibration improves all modes of CHASM, suggest that we would find improvements in the simulated climate if we used calibration-derived parameter datasets.

In conclusion, an exploration of the multicriteria method with CHASM opens up a range of useful directions for exploring the performance of land surface schemes without the difficulty of choosing parameter values in ways that do not bias results to one particular model (i.e., the effective parameter dataset that best suits one model disadvantages another in any objective intercomparison exercise). The methodology can also provide a means to examine the large differences apparent in PILPS and other land surface intercomparison results, and we hope to examine this area in the future.

Acknowledgments. This work was funded, in part, by the Australian Research Council. We thank an anonymous reviewer for insightful comments.

REFERENCES

Beljaars, A. C. M., and F. C. Bosveld, 1997: Cabauw data for the validation of land surface parameterization schemes. *J. Climate,* **10,** 1172–1193.

Chen, T. H., and Coauthors, 1997: Cabauw experimental results from the Project for Intercomparison of Land-Surface Parameterization Schemes. *J. Climate,* **10,** 1194–1215.

Deardorff, J. W., 1978: Efficient prediction of ground surface temperature and moisture with inclusion of a layer of vegetation. *J. Geophys. Res.,* **83,** 1889–1903.

Desborough, C. E., 1999: Surface energy balance complexity in GCM land surface models. *Climate Dyn.,* **15,** 389–403.

——, and A. J. Pitman, 1998: The BASE land surface model. *Global Planet. Change,* **19,** 3–18.

——, ——, and B. McAvaney, 2001: Surface energy balance complexity in GCM land surface models. Part II: Coupled simulations. *Climate Dyn.,* **17,** 615–626.

Dirmeyer, P. A., A. J. Dolman, and N. Sato, 1999: The pilot phase of the Global Soil Wetness Project. *Bull. Amer. Meteor. Soc.,* **80,** 851–878.

Duan, Q., S. Sorooshian, and V. K. Gupta, 1994: Optimal use of the SCE-UA global optimisation method for calibrating watershed models. *J. Hydrol.,* **158,** 265–284.

Franks, S. W., and K. J. Beven, 1997: Bayesian estimation of uncertainty in land-surface–atmosphere flux predictions. *J. Geophys. Res.,* **102,** 23 991–23 999.

Gupta, V. K., and S. Sorooshian, 1983: The calibration of conceptual catchment models using derivative-based optimization algorithms. *Water Resour. Res.,* **19,** 269–276.

——, ——, and P. O. Yapo, 1998: Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resour. Res.,* **34,** 751–761.

——, L. A. Bastidas, S. Sorooshian, W. J. Shuttleworth, and Z.-L. Yang, 1999: Parameter estimation of a land surface scheme using multicriteria methods. *J. Geophys. Res.,* **104,** 19 491–19 503.

Henderson-Sellers, A., 1993: A factorial assessment of the sensitivity of the BATS land-surface parameterization scheme. *J. Climate,* **6,** 227–247.

——, 1996: Soil moisture simulation: Achievements of the RICE and PILPS intercomparison workshop and future directions. *Global Planet. Change,* **13,** 99–115.

——, A. J. Pitman, P. K. Love, P. Irannejad, and T. H. Chen, 1995: The Project for Intercomparison of Land Surface Parameterization Schemes (PILPS): Phases 2 and 3. *Bull. Amer. Meteor. Soc.,* **76,** 489–503.

Hendrickson, J. D., S. Sorooshian, and L. Brazil, 1988: Comparison of Newton-type and direct search algorithms for calibration of conceptual rainfall–runoff models. *Water Resour. Res.,* **24,** 691–700.

Koster, R. D., and M. J. Suarez, 1992: Modeling the land-surface boundary in climate models as a composite of independent vegetation stands. *J. Geophys. Res.,* **97,** 2697–2715.

Manabe, S., 1969: Climate and the ocean circulation: 1. The atmospheric circulation and the hydrology of the earth's surface. *Mon. Wea. Rev.,* **97,** 739–805.

Pitman, A. J., and C. E. Desborough, 1996: Brief description of Bare Essentials of Surface Transfer and results from simulations with the HAPEX-MOBILHY data. *Global Planet. Change,* **13,** 135–143.

Robock, A., K. Ya. Vinnikov, C. A. Schlosser, N. A. Speranskaya, and Y. Xue, 1995: Use of midlatitude soil moisture and meteorological observations to validate soil moisture simulations with biosphere and bucket models. *J. Climate,* **8,** 15–35.

Sellers, P. J., W. J. Shuttleworth, J. L. Dorman, A. Dalcher, and J. M. Roberts, 1989: Calibrating the Simple Biosphere Model for Amazonian tropical forest using field and remote sensing data. Part I: Average calibration with field data. *J. Appl. Meteor.,* **28,** 727–759.

Sen, O. L., L. A. Bastidas, W. J. Shuttleworth, Z. L. Yang, H. V. Gupta, and S. Sorooshian, 2001: Impact of field calibrated vegetation parameters on GCM climate simulations. *Quart. J. Roy. Meteor. Soc.,* **127,** 1199–1224.

Shao, Y., and A. Henderson-Sellers, 1996: Modeling soil moisture: A Project for Intercomparison of Land-Surface Parameterization Schemes, phase 2(b). *J. Geophys. Res.,* **101,** 7461–7475.

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

——, and V. K. Gupta, 1983: Automatic calibration of conceptual rainfall–runoff models: The question of parameter observability and uniqueness. *Water Resour. Res.,* **19,** 260–268.

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

Yapo, P. O., H. V. Gupta, and S. Sorooshian, 1997: Multi-objective global optimization for hydrologic models. *J. Hydrol.,* **204,** 83–97.