JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 107, NO. D20, 4443, doi:10.1029/2001JD000931, 2002

Exploring the relationship between complexity and performance in a land surface model using the multicriteria method

M. Leplastrier and A. J. Pitman
Department of Physical Geography, Macquarie University, North Ryde, New South Wales, Australia

H. Gupta
Sustainability of Semi-Arid Hydrology and Riparian Areas (SAHRA), Department of Hydrology and Water Resources, University of Arizona, Tucson, Arizona, USA

Y. Xia
Department of Physical Geography, Macquarie University, North Ryde, New South Wales, Australia

Received 6 June 2001; revised 17 April 2002; accepted 24 July 2002; published 29 October 2002.

[1] The performance of five modes of a land surface model, the Chameleon Surface Model (CHASM), was investigated after calibration via the multicriteria method to monthly totals of evaporation and runoff from the Valdai data set. The use of CHASM allows for an exploration into the relationship between surface energy balance complexity and optimal performance by isolating the impacts of different parameterizations of the surface energy balance. When compared to quantities used within the calibration process, CHASM's performance was significantly increased with calibration over default simulations regardless of calibration length or mode complexity. Within the calibration period, CHASM's performance increased with increasing complexity in the representation of the surface energy balance. Outside the calibration period there was little improvement to simulations from additional complexity in the surface energy balance representation above the simplest mode. Calibration is shown to reduce the scatter between modes suggesting that some of the differences between models in PILPS Phase 2d may be explained by the specification of parameter values. For simulations of quantities not used in calibration, performance can be reduced as a result of calibration. This implies that evaporation and runoff may not be the best quantities for calibration in order to improve model performance. It is suggested that the best quantities to calibrate may be mode and model specific. *INDEX TERMS:* 3322 Meteorology and Atmospheric Dynamics: Land/atmosphere interactions; 1878 Hydrology: Water/energy interactions; 1818 Hydrology: Evapotranspiration; *KEYWORDS:* land surface, complexity, calibration, PILPS, multicriteria

Citation: Leplastrier, M., A. J. Pitman, H. Gupta, and Y. Xia, Exploring the relationship between complexity and performance in a land surface model using the multicriteria method, *J. Geophys. Res.*, *107*(D20), 4443, doi:10.1029/2001JD000931, 2002.

1. Introduction

[2] The development of land surface models (LSMs) over the last thirty years has seen a tendency to add complexity through the replacement of implicit representations of processes [e.g., *Manabe*, 1969] with more explicit methodologies [e.g., *Deardorff*, 1978; *Sellers et al.*,1996]. Adding complexity to a LSM may improve performance but whether this comes from more physically realistic parameterizations or from changes in the effective values of parameters is unclear. *Desborough* [1999] developed a framework which enabled a single LSM to be used in a variety of configurations to explore the relationship between model performance and the complexity of the surface energy balance (SEB) parameterizations. He used this framework to investigate whether increasing the complexity of the SEB parameterization improved LSM performance.

[3] LSM performance can be related to both the parameterizations and the parameter values used within a model. It is therefore necessary to remove variations in the effective values of parameters, between models, such that parameters are effectively identical and have the same physical meaning. This cannot be achieved just by using numerically identical parameters. The Chameleon Surface Model (CHASM) provides a modeling framework whereby step-wise changes to the SEB complexity can be explored within a common modeling environment without changing the effective values of the parameters [*Desborough*, 1999]. However, CHASM uses prescribed parameter values from measurements or other estimates which can lead to erroneous conclusions as to the superiority of one LSM over another [*Bastidas et al.*, 1999] since any parameter value may be more suited to one modeling approach over another. *Bastidas et al.* [1999]

0148-0227/02/2001JD000931$09.00

Table 1. Description of Those Parameters Calibrated in CHASM With Their Maximum Feasible Ranges[a]

Calibrated Parameter Descriptions	
Model Parameters	Feasible Parameter Range
Bare ground albedo	0.05–0.45
Snow albedo	0.4–0.95
Vegetation albedo	0.16–0.26
LAI seasonality parameter	3–5
Maximum LAI	0–5
Maximum fractional vegetation cover	0.6–0.95
Fractional vegetation cover seasonality	0–0.95
Canopy resistance (s/m)	0–250
Snow density (kg/m3)	50–450
Available water holding capacity (kg/m3)	100–300
Soil color index	0–3
Bare ground roughness length (m)	0.0003–0.01
Snow surface roughness length (m)	0.0001–0.0007
Vegetation roughness length (m)	0.08–0.06
Model state variables	
Aerodynamic surface temperature (K)	260–300
Snow mass (kg/m3)	20–165
Available moisture in root zone (kg/m^2)	35–235

[a] The state variables are included here since they are included in the calibration to avoid initialization problems [following *Gupta et al.*, 1999].

Table 2. Summary of the Features Which Identify Each of the Different SEB Configurations (Surface Modes) of CHASM

Surface Mode	Stability Correction	Surface Resistance	Canopy Interception	Bare Ground Evaporation	Temperature Differentiation
EB	-	-	-	-	-
RS	✓	✓	-	-	-
RSI	✓	✓	✓	-	-
RSGI	✓	✓	✓	✓	-
SLAM	✓	✓	✓	✓	✓

suggested that a comparison of the optimal performances of a model through objective calibration is the preferred method to base conclusions. There are several objective techniques available to select and adjust the parameters in LSMs [e.g., *Sellers et al.*, 1989; *Franks and Beven*, 1997]. Recently, parameter estimation using multicriteria methods (MCM) [*Gupta et al.*, 1998] and in particular the multiobjective complex evolution algorithm [*Yapo et al.*, 1998] have been shown to provide an effective way to reduce the errors associated with parameter uncertainty by calibrating a LSM to several quantities simultaneously, thus enabling a LSM to achieve optimal performance [*Gupta et al.*, 1999].

[4] The purpose of this paper is to explore the relationship between the complexity of the SEB representation and model performance. By calibrating each mode of CHASM, the best performance at each level of complexity can be achieved. This permits conclusions to be reached regarding the relationship between model SEB complexity and performance. A secondary aim is to estimate the amount of scatter in default simulations which may be attributed to the effective value of parameters. This would provide a step towards reducing scatter in PILPS-like land surface model intercomparison exercises.

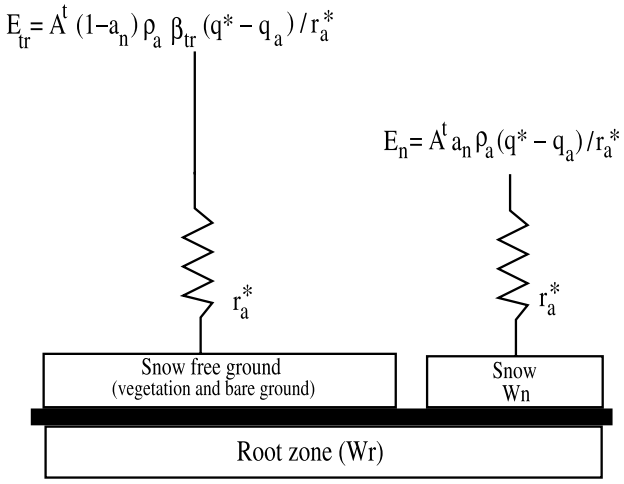
2. Chameleon Surface Model (CHASM)

[5] CHASM was designed to explore the impact of SEB complexity on model behavior [*Desborough*, 1999]. The model utilizes the same parameterizations and parameters (Table 1) for most of the components of the model over a wide complexity range in the SEB configuration. CHASM's SEB configurations range from a simple homogenous surface (as in the *Manabe* [1969] bucket) to a complex *Deardorff* [1978] type 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. To resolve the SEB, CHASM combines similar elements throughout a grid square to form tiles (called a "grouped mosaic approach" [e.g., *Koster and Suarez*, 1992]). Each tile is divided into fractions of vegetation, snow and ground. Snow cover fractions for ground and foliage surfaces are calculated as functions of the snow pack depth and density and the vegetation roughness length. The vegetation fraction is divided into wet and dry fractions if the SEB mode allows for canopy interception. Each tile has a prognostic bulk temperature and a diagnostic skin temperature. Parameters are included for albedo and roughness length for each type of cover and seasonality parameters for leaf area index and vegetation fraction.

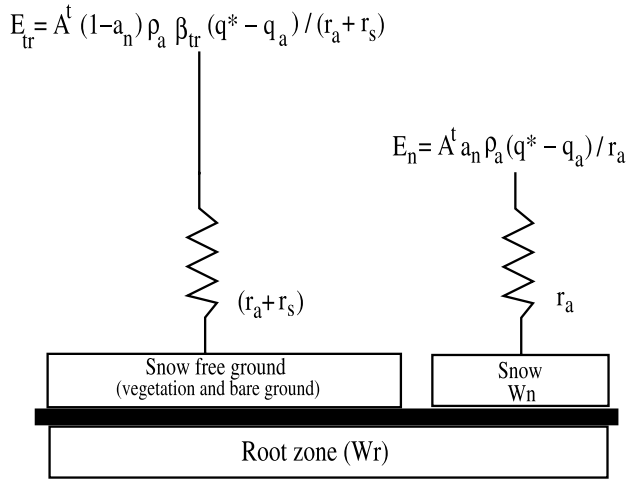
[6] CHASM's hydrology follows *Manabe* [1969] in that the root zone is treated as a bucket with finite water holding capacity and beyond this capacity runoff occurs. Water can also be stored as snow or depending on the mode, on the canopy following interception or on the surface for bare ground evaporation. The use of a simple hydrology model has been shown to work well in midlatitude regions [*Robock et al.*, 1995]. Soil temperature is simulated using four layers using a finite difference method and a zero-flux boundary condition at the base of the profile. Each tile, depending on the mode, can have up to four evaporation sources for canopy evaporation, transpiration, bare ground evaporation and snow sublimation. Again depending on the mode, resistances may be applied to reduce evaporation and transpiration rates.

[7] Table 2 lists the different modes of CHASM used in this study in increasing order of complexity, and Figure 1 shows resistance diagrams for each mode. EB, the simplest mode of CHASM, is constructed from one tile. The aerodynamic resistance to turbulent transport for heat and

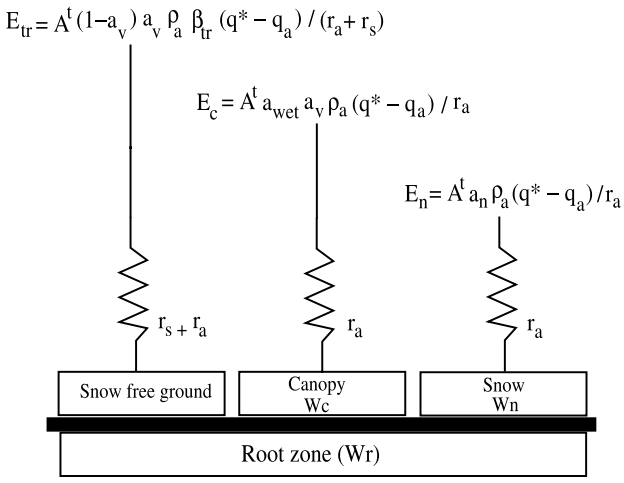
Figure 1. (opposite) Illustration of CHASM's modes. (a) mode EB with two evaporation sources: from the vegetation and soil (E_{tr}) and from the snow (E_n) and two moisture storage terms: root zone (W_r) and snow (W_n) (b) the RS-mode. Where the aerodynamic resistance (r_a) is calculated with an atmospheric stability constant (r_a^* is calculated without stability correction) and canopy resistance (r_s) is added to the resistance pathway of E_{tr}; (c) RSI which includes canopy interception (W_c). RSI has three evaporation sources, E_{tr}, evaporation of intercepted water (E_c) and E_n, and three moisture storage terms W_r, W_c and W_n; (d) RSGI which includes a bare ground parameterization, an extra soil moisture storage term (W_g) and an extra evaporative source from the soil (E_g); and (e) SLAM which includes two tiles with separate SEBs plus a variable canopy resistance (r_c), which is applied to the evaporation pathway, E_{tr}. Other terms include air density (ρ_a), A^t which is the area extent of the tile, a_{wet} is the fraction of the canopy that is wet. q* is the saturated vapor pressure of the surface, q_s is the vapor pressure of the air, a_g is the fraction of soil, a_n is the fraction of snow and a_v is the vegetation fraction.



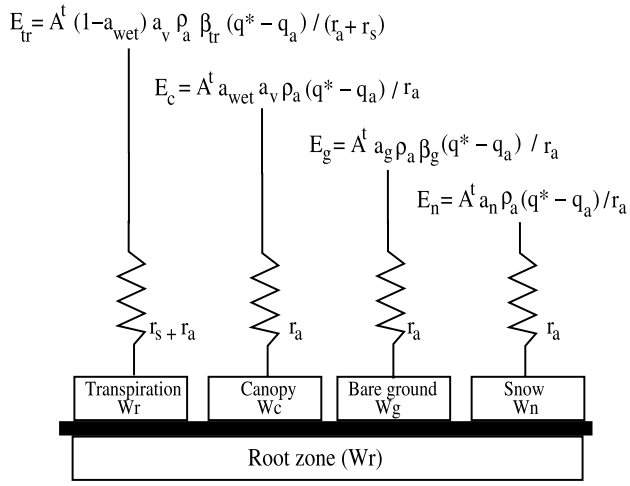
(a) EB



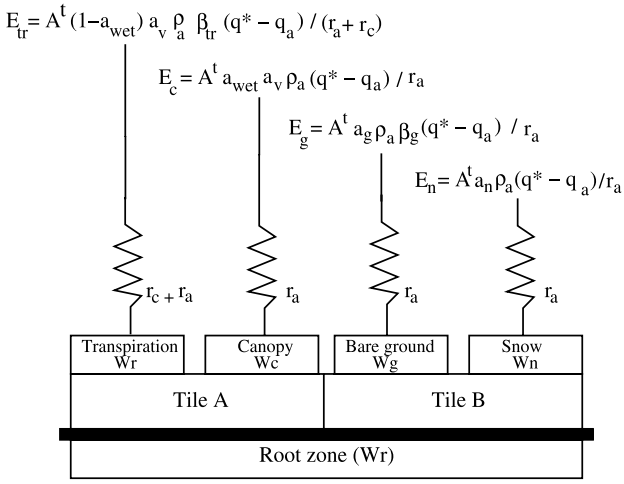
(b) RS



(c) RSI



(d) RSGI



(e) SLAM

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 parameterizations for canopy interception of precipitation. This results in three evaporation sources since water can evaporate from the canopy store. The canopy is further divided 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 builds on RSGI by including a time variable canopy resistance which replaces the temporally invariant surface resistance. 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 surface type. A separate SEB is calculated for each tile which allows for temperature variations across the land atmosphere interface.

[8] Overall, CHASM has proven useful for explaining some of the results obtained by PILPS [e.g., *Desborough*, 1999] by identifying the role of differences in effective parameters in explaining interscheme variations. This is due to maintaining a common modeling environment by ensuring that most parameters retain the same effective value when switching between modes. Further, the number of parameters being calibrated in each mode of CHASM does not vary. These factors allow the effect of increasing the complexity of the SEB configuration on model performance to be explored by sequentially adding explicit parameterizations.

3. Forcing and Calibration Data Sets

[9] The observational data from Valdai (57.6°N, 33.1°E) have been used to explore land surface processes before [e.g., *Robock et al.*, 1995; *Vinnikov et al.*, 1996; *Schlosser et al.*, 1997, 2000] (PILPS Phase 2d). The data span 18 years, permitting an examination of seasonal and interannual model performance. The vegetation cover is mainly grassland meadow. Near surface air temperatures rise above 15°C in summer and fall below −10°C in winter providing almost continuous snow cover between November and April. The majority of precipitation falls in the summer and autumn months with an annual average of 730 mm.

[10] The atmospheric data were measured at a grassland plot [*Schlosser et al.*, 1997]. Atmospheric pressure, air temperature and humidity were recorded at a height of 2 m and wind speed at a height of 10 m. Shortwave radiation and longwave radiation were derived following *Schlosser et al.* [1997].

[11] Soil moisture samples were taken from 14 representative sites close to the end of each month. Soil moisture was calculated using the thermostat-weight (gravimetric) technique described by *Robock et al.* [1995]. The errors in soil moisture measurements in the top 1 m of soil are less than 1 cm and the seasonal variations of total soil moisture in the top 1 m between the 14 sites are very small [*Vinnikov et al.*, 1996]. Following *Schlosser et al.* [1997] the soil moisture from 11 of the sites were averaged.

[12] *Schlosser et al.* [1997] suggests that a high degree of confidence can be placed on the runoff measurements obtained in the warmer months. However, when the stream is frozen in winter or when the stream overflows in springtime the runoff measurements are less accurate. Modifications to the observed runoff were made by *Schlosser et al.* [1997] according to variations in the observed averaged water table depth.

[13] Monthly measurements of evaporation from the catchment were described by *Federov* [1977] during May to October using weighing lysimeters (1960–1973). Estimations of evaporation for the remaining months (November to April) were calculated using the *Budyko* [1956] algorithm for potential evaporation. *Schlosser et al.* [1997] compared the monthly evaporation calculated from the residual of the water balance from the top 1 m of soil with the lysimeter measurements and found that their seasonal cycles were in good agreement.

4. Multicriteria Calibration Methodology

[14] 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 and forecasting skill is important. The multicriteria calibration methodology was developed by *Gupta et al.* [1998] from a single-criteria method [*Duan et al.*, 1994]. *Gupta et al.* [1998] have used the multicriterial methodology to estimate reasonable ranges of optimal parameters for the BATS LSM.

[15] The first step in using the multicriteria calibration methodology is to define the feasible parameter range for each parameter to be calibrated (see Table 1 for a list of parameters 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 to measure a specific statistical characteristic of the output residual [*Gupta et al.*, 1998]. Since model calibration is a multiobjective problem 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.

[16] Once the feasible parameter space (Table 1) has been sampled, the multicriteria calibration methodology selects a set of parameter values which, based on the shape of the objective function space from the previous step, minimizes each of the objectives and therefore reduces the error (model output residual). The methodology terminates when the process has converged to a "Pareto set" or "solution set". A Pareto set contains a prespecified number of parameter sets where each parameter set gives solutions that are better than all others for at least one of the objectives. Thus, no parameter set is better than another for all of the objectives,

and within the Pareto set every parameter set is considered equal in a multiobjective sense. Due to measurement and modeling errors the Pareto set consists of several solutions. The Pareto set obtained using the multicriteria calibration methodology has been shown to be a good representation of the Pareto Set, even though not every Pareto solution is computed [*Yapo et al.*, 1998]. 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.

5. Experimental Methodology

[17] Each of the five modes of CHASM were calibrated to monthly totals of evaporation and runoff. Calibrating to an observational frequency (e.g. monthly totals) that differs from forcing data frequency (e.g. 30-minutes) had not been attempted before, but proved to be possible. Since there are no observations for the parameters required by CHASM, all 14 parameters values were derived via the MC method, as well as initial values of the 3 state variables listed in Table 1. The initial boundaries for each parameter (the feasible parameter space) were chosen from the literature. Calibration lengths one to eight years using data over the period 1966–1973 (i.e. 1966, 1966–67, 1966–68,. . . ., 1966–73 and 1967–73, 1968–73, 1969–73,. . . ., 1972–73, 1973) were conducted as well as the remaining single years: 1967, 1968, 1969, 1970, 1971, 1972. No spin-up period was used for the calibration runs whereas for default simulations, each mode was allowed to equilibrate. The conclusions were found to be insensitive to the use of three different objective functions (e.g. Root Mean Squared Error (RMSE), mean absolute error and Nash-Sutcliffe error [*Gupta et al.*, 1998]) and so only the results from using RMSE are reported in this paper. The results were also examined using six different initial sampling points (the first step in the MC-method) and different termination points (i.e. the number of solutions in the Pareto set was varied from 150 to 1000) and results were found to be insensitive.

6. Results

[18] The performance of an LSM should be assessed against quantities used within the calibration period, both within and outside the calibration period as well as against quantities not used in calibration. This section is therefore divided into three parts: first, simulations of runoff and evaporation over the calibration period; secondly simulations of evaporation and runoff outside the calibration period and finally simulations of soil moisture and sensible heat.

6.1. Simulations of Runoff and Evaporation Over the Calibration Period

[19] Figure 2 shows the Pareto fronts for five modes of CHASM. They are formed by plotting the RMSE for both objective functions using each of the 150 different parameter sets that form the Pareto set using CHASM over the full 8-years of the observational record. The corresponding simulations for evaporation RMSE and runoff RMSE for each parameter set then forms the Pareto front. Figure 2 also shows the single solution generated by each mode of CHASM using the default parameter set (which is the best estimate of the model parameters available). Figure 2 shows three key findings. First, in all cases the calibration of CHASM leads to a large number of solutions which are superior in both the simulation of runoff and evaporation compared to the default solution. Some of the tails of the Pareto front have a worse RMSE for runoff than the default solution, but these are a minority of solutions and the RMSE for evaporation is almost always superior to the default simulation. This demonstrates that calibration of each mode of CHASM improves default model performance when compared to quantities within the calibration period.

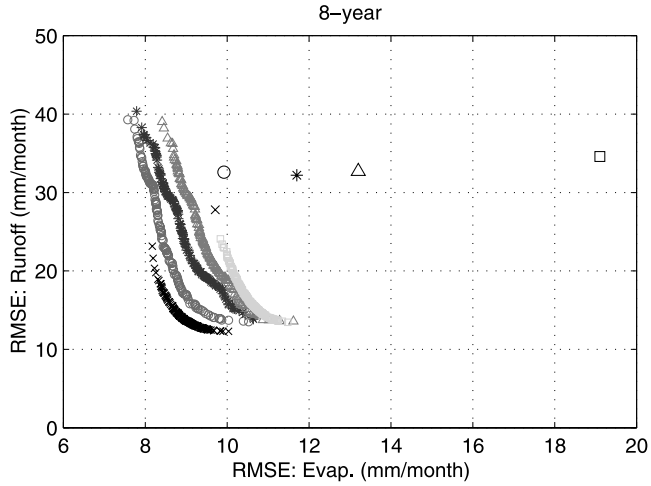


Figure 2. The Pareto fronts for all five modes obtained from calibration using the multicriteria method over the full 8 years of observational record. SLAM = cross, RSGI = circle, RSI = star, RS = triangle, EB = square. Default simulations have the same but larger symbol.

[20] A second result shown in Figure 2 is that the performance of more complex modes of CHASM is superior to more simple modes. This is true in the case of default parameters as well as calibrated parameters. The position of the Pareto fronts are sequential in terms of mode complexity with SLAM performing best and EB performing worst. It is noteworthy, however, that the length of the Pareto front in the increasing runoff RMSE direction is large for the intermediate modes. *Desborough* [1999] found that these modes were sensitive to small changes to the constant surface resistance parameter which is unique to these modes. Therefore small changes in parameters via calibration may give these model structures too many degrees of freedom which results in small improvements to evaporation at the expense of a large degradation to runoff simulations. EB and SLAM, which do not include the constant surface resistance are therefore less likely to produce anomalous runoff simulations. In the calibrated simulations, there is a greater improvement in the simulations of evaporation compared to runoff as complexity is added. This is due to changes in

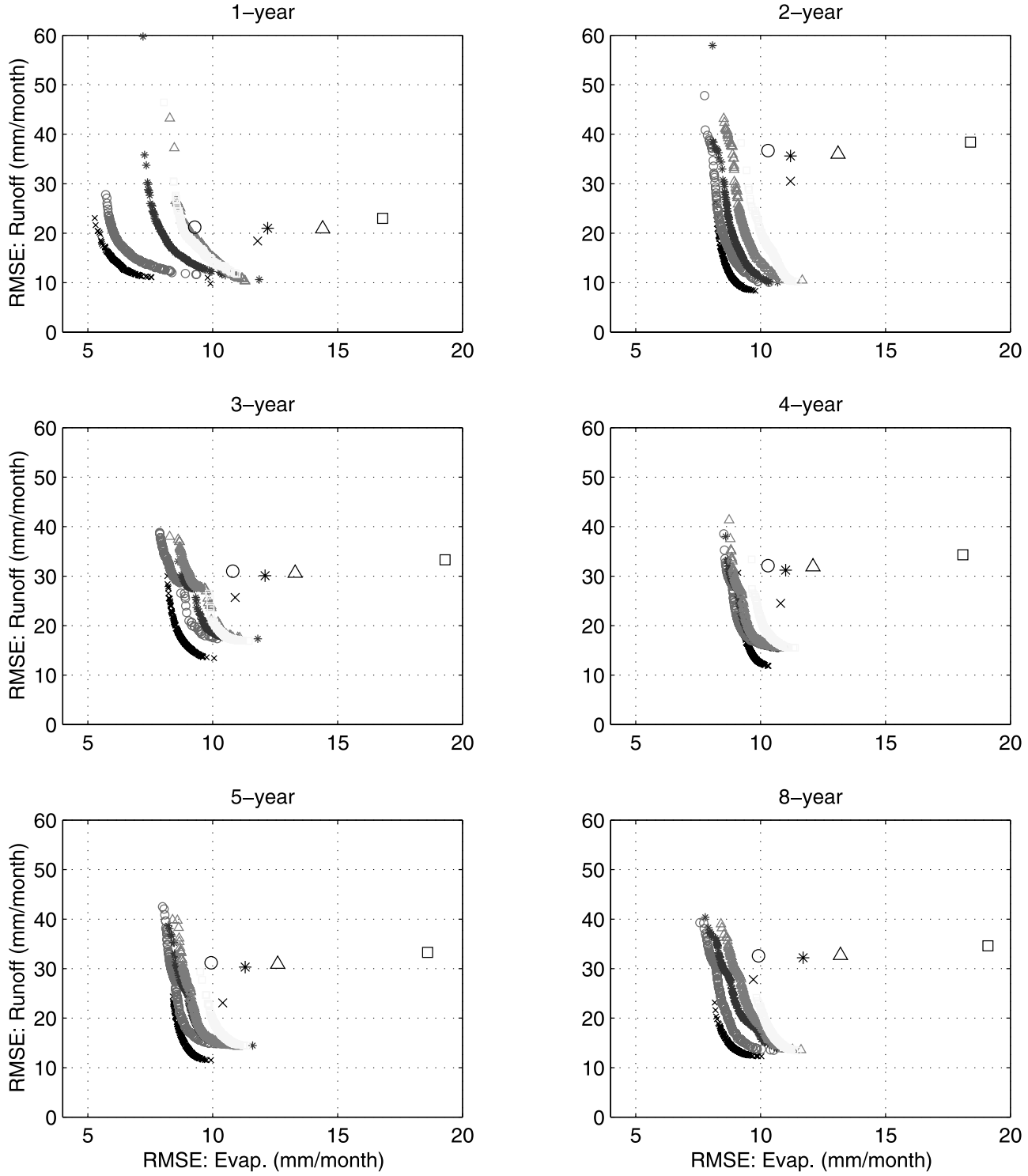


Figure 3. Same as Figure 2 except the calibration length is varied from 1 to 8 years. Only lengths 1, 2, 3, 4, 5 and 8 are shown.

the SEB configuration between modes while the subsurface hydrology parameterizations remains constant.

[21] A third result is that after calibration there are fewer differences between the modes in terms of performance compared to the default simulations. This result suggests that some of the scatter found in PILPS may be explained by the specification of parameter values rather than differences in model structure. While the exact values of parameters were specified in PILPS, our results imply that the effective values of these parameter varied significantly across models.

[22] Figure 3 shows the equivalent figures to Figure 2, but rather than calibrated against the full eight years of the observed record, are calibrated separately against 1, 2, 3 . . .

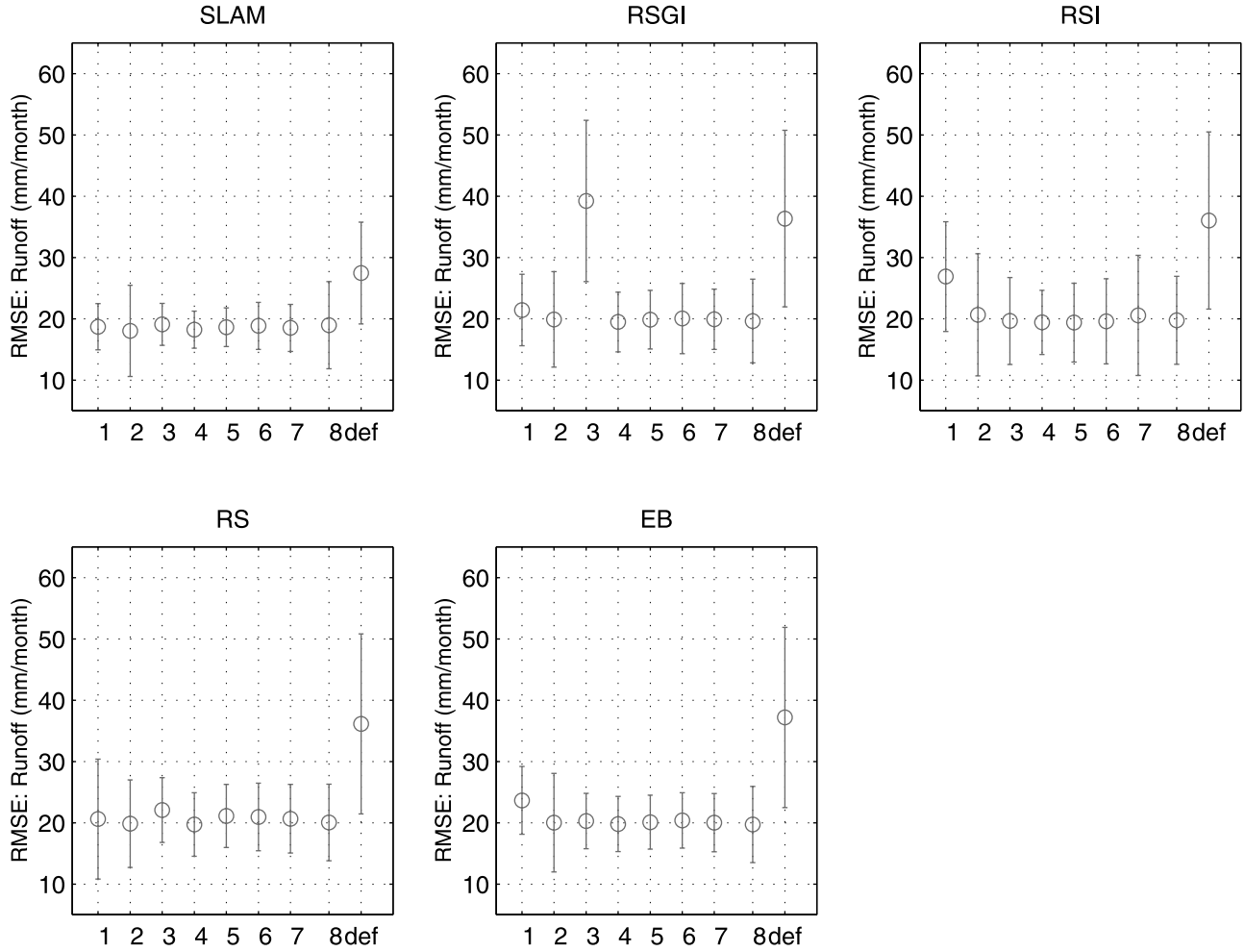


Figure 4. Ten-year annual averages for runoff RMSE (mm/month) outside the calibration period (1974–1983). Simulations for calibration lengths 1–8 years and the default are shown. The error bars represent standard deviations calculated from annual totals.

8 years of the record. Figure 3 shows that calibrating against a variety of data lengths produces the same key result that the more complex the mode of CHASM, the better the eventual solution. There are differences between the results, but SLAM is always superior to the less complex modes and EB is usually the worse. This is not a function of more calibrated parameters in the more complex modes (the number of parameters is invariant across modes), rather it is due to the inclusion of the more explicit representation of processes.

6.2. Simulations of Runoff Outside the Calibration Periods

[23] For simulations outside the calibration period a representative parameter set from each Pareto set was selected, In selecting this parameter set, a compromise was made between the two objectives i.e. for the intermediate modes in Figures 3 and 4, a representative parameter set was not chosen halfway up the long Pareto tails in the increasing runoff RMSE direction, as these parameter sets only offer marginal decreases of evaporation RMSE at the expense of large degradations of runoff RMSE. Our choice of a parameter set was subjective but a modeler would clearly never choose extremes from the Pareto front. The same method was applied to all calibration lengths across all modes and represents each calibrated mode's behavior.

[24] The performance of an LSM also needs to be assessed using runoff data from periods to which the model has not been calibrated. Figure 4 shows the 10 year average RMSE for runoff for each mode of CHASM and for each period of calibration for calibrated and default simulations. In the case of runoff, to which CHASM was calibrated in the preceding eight years, the calibrated solutions are superior to the default solutions in all cases except the three year calibration using the RSGI mode. Calibrating each mode leads to significant reductions in both the RMSE, and the interannual error (as shown by the error bars). Even calibrating against one year of runoff data leads to significantly improved performance of CHASM when compared to observed data a decade later (note all single years 1967–1973 (not shown) produced the same result). This improvement in model performance is both in the ten year average and in the standard deviation based on each of the ten year annual averages of monthly observations. This indicates that calibration of the modes of CHASM leads to superior performance in the simulation of runoff during periods to

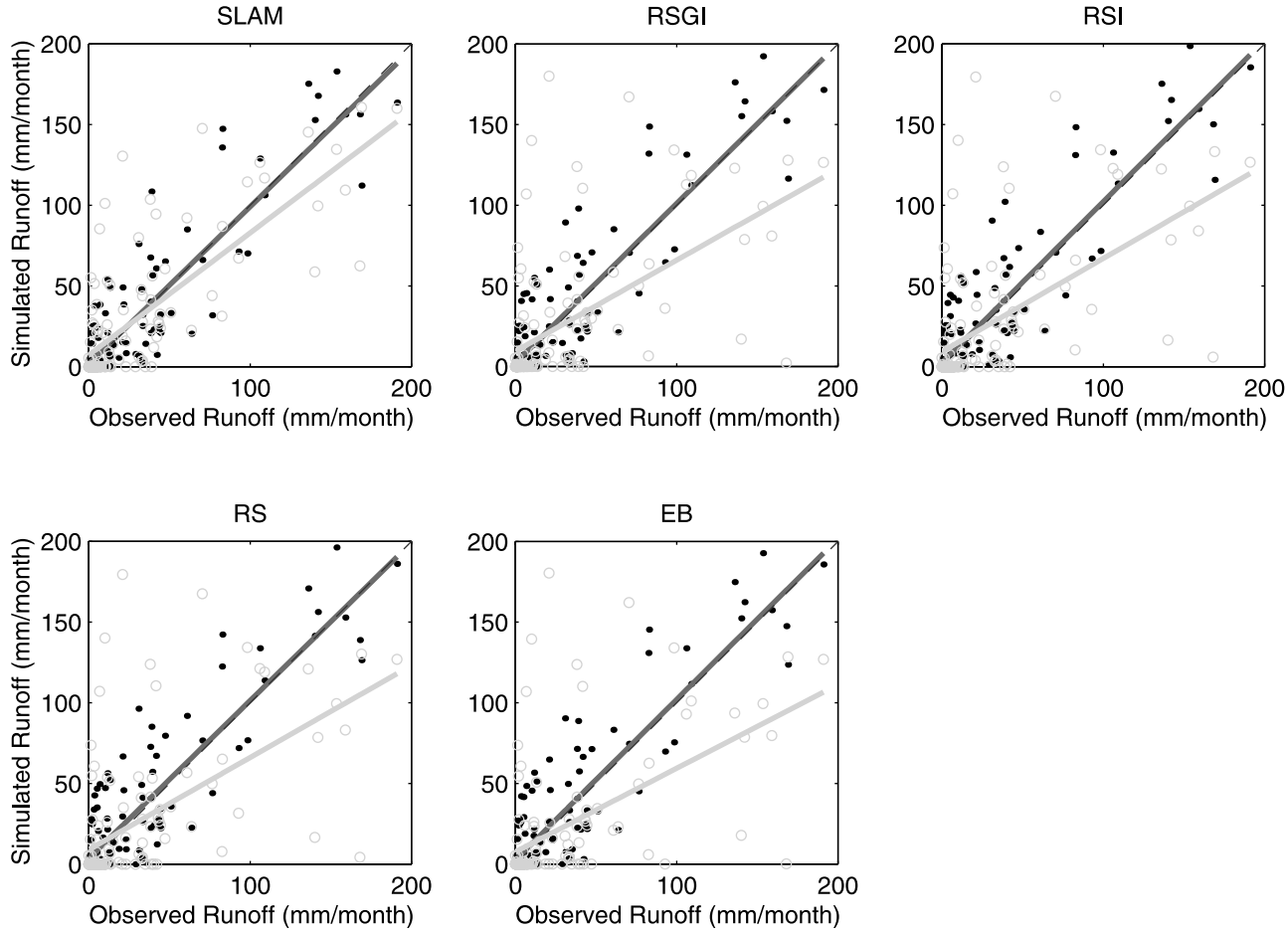


Figure 5. Monthly totals of simulated runoff (8-year calibration) against observations. Each mode simulation includes the regression line for both calibrated (black) and default (grey) simulations. Calibrated mode values are shown as black dots while default values are unfilled circles.

which the model was not calibrated, even if only one year of calibration data are available. SLAM generally performs marginally better than the other models with RMSE values of approximately 18 mm month^{-1} while the other modes perform similarly to each other with RMSEs of approximately 20 mm month^{-1}. This suggests that after calibration, performance is improved only if a very complex SEB configuration is used (e.g. SLAM). The additional complexity levels of the intermediate modes do not lead to significantly better performance than EB.

[25] Figure 5 shows the improvement in the simulation of monthly runoff following calibration of the five modes of CHASM. The open circles (filled circles) show monthly runoff over the ten year period from the default (calibrated) simulations. The scatter is clearly reduced following calibration for all modes. The regression line (Figure 5) is improved following calibration and the bias in these calibrated simulations is very small. The overall statistics for each simulation are shown for the ten-year average of simulated runoff in Figure 6. Results for the simulation of runoff outside of the calibration period following calibration against 1, 2, 5 and 8 years are shown. Figure 6 shows that the correlation coefficient (R) and slope of the line (m) are reduced over the 10-year period following a single year of calibration. This improvement is more pronounced if longer periods of calibration are used, although calibrating against 2 years of data leads to marginally better performance than 5 years of calibration. The best performance is achieved following 8 years of calibration. This result is most likely site and climate specific and is affected by the nature of the atmospheric forcing used for individual simulations. In choosing atmospheric forcing, it is beneficial to use years which contain significant interannual variability.

6.3. Simulations of Soil Moisture and Sensible Heat

[26] CHASM was not calibrated to soil moisture at any time (although the water holding capacity was calibrated). If the model is calibrated to evaporation and runoff over the first eight years of the observational record, Figure 7 shows that the performance in the simulation of soil moisture in all three intermediate models (RS, RSI, RSGI) is worse than the default results. The default version of these modes performs reasonably and is only matched by calibrating against 7 or 8 years of data using RSGI mode or 4 years of data in the RSI mode. In contrast, experiments using SLAM produce simulations of soil moisture which are comparable with the default mode almost all cases, and using EB, the calibrated results are superior to the default results provided this mode is calibrated for more than one year. The error bars which represent the

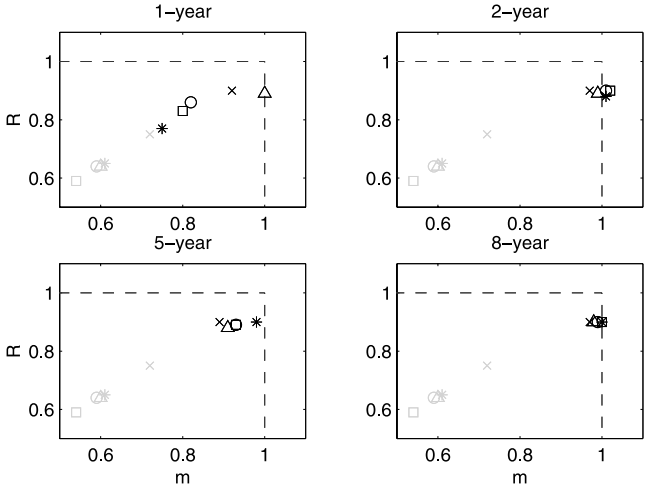


Figure 6. Correlation coefficients (R) and the corresponding slopes for the regression line (m) calculated from monthly totals of simulated runoff and monthly runoff totals from observations for the 1-year, 2-year, 5-year and 8-year calibration periods. Default simulations are grey, calibrated simulations are black. SLAM = X, RSGI = circle, RSI = star, RS = triangle and EB = square. A perfect match to observations is represented by the intersection of the two dashed lines (i.e. at the point (1,1)), where the horizontal (vertical) dashed line represents a perfect R-value (m-value).

standard deviation based on each of the ten year annual averages of monthly observations are improved in the EB mode more than in the other models. This might suggest that calibrating this simple mode adds value in both the mean and variability, however this gain may be at the expense of a poor simulation of quantities not used in calibration (see Discussion).

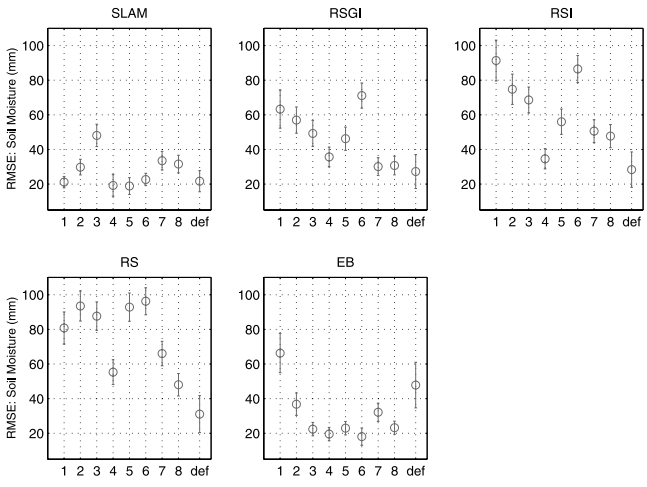


Figure 7. Ten-year annual averages for soil moisture RMSE (mm/month) outside the calibration period (1974–1983). Simulations for calibration lengths 1–8 years and the default are shown. The error bars represent standard deviations calculated from annual totals. The maximum water holding capacity is calibrated and ranged from 150–300 km m^{-2}.

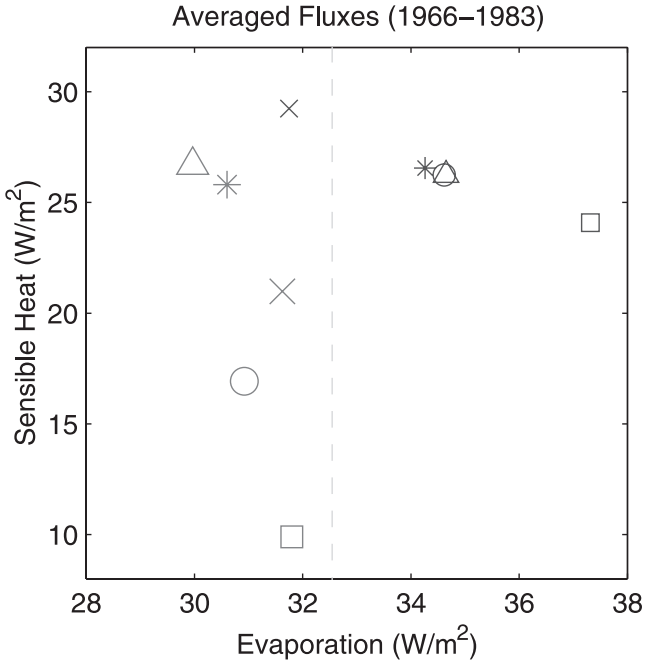


Figure 8. The 18-year averages of simulated sensible heat and latent heat fluxes. Observations for latent heat fluxes are indicated by the dashed line [from *Schlosser et al.*, 2000]. Square-EB, triangle-RS, asterix-RSI, circle-RSGI and cross-SLAM. The small symbols represent the default modes while the larger symbols represent the calibrated modes.

[27] Figure 8 shows the 18-year averages for sensible heat and evaporation (converted to latent heat) fluxes for all modes (calibrated and default). Unfortunately there are no observational data available for sensible heat but the range simulated by the calibrated modes (9–29 W m^{-2}) is much larger compared to the range simulated by the default modes (24–29 W m^{-2}). Sensible heat simulated by EB were approximately 2.5 times less than its default version and half the averages simulated by the calibrated SLAM mode. Thus, although the calibrated versions of EB and SLAM have performed similarly for runoff, evaporation and soil moisture, their simulations vary significantly for the 18-year average of sensible heat fluxes. This result casts some doubts to the effectiveness of calibrating CHASM to just one aspect of the land surface such as the hydrological quantities, runoff and evaporation. *Gupta et al.* [1999] found a similar result using BATS for grassland and semi-arid sites. They found that overall performance was only increased after calibrating to the dominant heat flux and an appropriate state variable.

7. Discussion

[28] This paper has used five modes of CHASM to assess the model's calibrated and noncalibrated performance using observed data from Valdai. The most significant result is that additional complexity in the SEB representation appears to improve model performance and lead to better validation within the calibration period. SLAM, when calibrated, always performs best of the five modes within the calibration period and the simplest mode (EB) performs

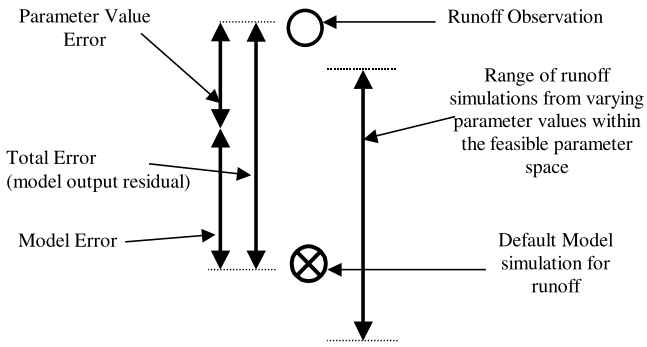


Figure 9. Hypothetical case illustrating the contributions from parameter value error and model error that create the model output residual or total error for runoff, which is the distance between the model simulation (circle/cross) and the runoff observation (circle). The relative range of the runoff simulation due to changes to parameter values from the feasible parameter space is also shown.

worst. This is not related to the number of parameters being calibrated which do not vary between modes. In all cases, calibration improves model performance over estimated parameter values when compared to variables used in the calibration.

[29] When assessing the predictive skill of each mode against runoff data outside the calibration period, all modes generally out-performed their default modes regardless of the length of the calibration period. However, there was no advantage in performance by increasing the SEB complexity from the simplest structure (EB) unless the most complex structure (SLAM) was used. In addition, SLAM was only marginally superior compared with the other modes suggesting that there is little advantage to increasing complexity in the representation of the SEB in LSMs (at least for the simulations of water balance terms at this midlatitude grassland site), if the model is to be used in a predictive mode.

[30] We have shown that increasing SEB complexity leads to some improvement in calibration, however, there may be other factors that enable a model to calibrate close to observations. According to *Gupta et al.* [1998], the total error in the system, as represented by the distance between the model output and observations, is equal to the aggregation of model error, measurement error and parameter uncertainty. If measurement error is ignored for simplification, the total error can be represented by Figure 9. Using the multicriteria method, parameters are derived to reduce the total error. The degree to which total error is reduced via the choice of parameter values will depend on how sensitive the modes are to small changes in parameter values and to the relative size of the model error compared to the model output residual. If the relative size of model error is large compared to parameter uncertainty, then parameter values are derived based on inefficiencies in model structure rather than only parameter uncertainty and may lead to the derivation of inaccurate parameter values. A combination of these factors may lead to a good performance to the quantities calibrated to but may decrease performance when simulating quantities not used in calibration (e.g. soil moisture and sensible heat).

[31] The poor performance by the intermediate modes of CHASM for soil moisture simulations and the erratic simulations of sensible heat across all the modes after calibration may be explained by the combination of these two factors. It may highlight a problem with calibrating to one component of the land surface, and indicates that a range of observed quantities may be needed to improve overall performance. Figures 10 and 11 show a see-saw analogy where the size of the fulcrum represents the model output residual. To improve overall model performance via calibration, the parameter error must be isolated in the calibration process so that the beam lowers evenly as the part of the fulcrum representing parameter uncertainty is reduced. In reality it is impossible to isolate parameter uncertainty completely so it is necessary to calibrate to quantities that effectively lowers the beam evenly. This may have been achieved with the calibration to runoff and evaporation for SLAM as shown in Figure 10. However for the other modes, calibrating to these quantities may have unevenly lowered the beam (through derivation of inaccurate parameter values) (Figure 11).

[32] To improve the overall performance of a mode after calibration with the multicriteria method, the land surface quantities either side of the fulcrum in Figures 10 and 11 need to be identified and used in the calibration process. This will lead to the derivation of realistic parameter values and therefore minimize poorer simulations of quantities not calibrated to. However there are complications in exploring this issue. From these results, the observational quantities

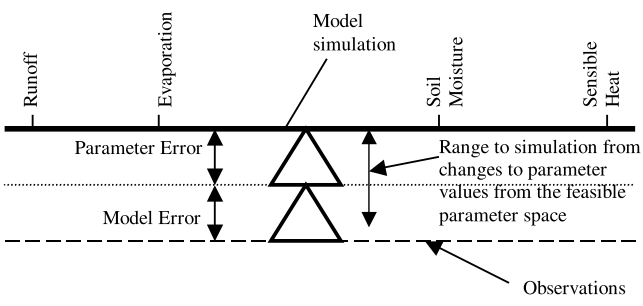


Figure 10. Analogy of multicriteria calibration problem before calibration.

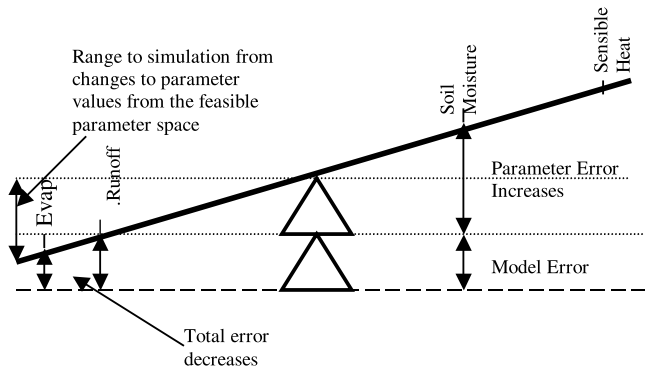


Figure 11. Analogy of multicriteria calibration problem after calibration where the total error is reduced for evaporation and runoff while the parameter error increases for soil moisture and sensible heat.

needed may be model and location specific and many observational quantities may be needed to identify the best ones for calibration. This may place an unrealistic requirement of observational data sets. However, *Sen et al.* [2001] indicate that local scale calibration appears to improve the simulation of the global scale climate hence there is some evidence that gains are to be found by using calibration.

[33] Since large model error is more likely to cause the derivation of unrealistic parameter values (Figures 9, 10 and 11) and subsequently cause inferior simulations for quantities not used in the calibration process, the choice of a model with small model error is preferable. The degree to which the simulations for quantities not used in the calibration process, such as sensible heat or soil moisture, are worsened by calibration may indicate the size of model error. Based on this hypothesis, SLAM performed most consistently after calibration to runoff and evaporation as the simulations of soil moisture and sensible heat remained similar to the default simulations. This is an indication that SLAM has the least model error and therefore implies that more complex SEB parameterizations perform better overall compared to simpler methods. To make any conclusions to the relationship between complexity and performance for the other modes would need rigorous assessment against sensible heat observations and other land surface quantities not available at Valdai.

8. Conclusions

[34] The multicriteria method has provided an efficient means to minimize the model output residuals of evaporation and runoff through an objective selection of parameter values. It has worked well with each of the modes of CHASM and has worked well against monthly data for various calibration lengths from one to eight years. The multicriteria method can be applied to other LSMs and to different data sets that vary in length and resolution, and could be used in future intercomparison work (e.g. PILPS) for isolating differences resulting from parameterizations by removing differences resulting from parameter values. Such an approach would need to be followed in parallel with noncalibrated intercomparisons.

[35] The results presented are based on the application of the multicriteria method using quantities which are not ideal (according to *Gupta et al.* [1999]) to properly constrain the models. However, calibrating to these quantities provided a useful insight to the relationship between model error and complexity. Our key conclusions are as follows:

1. Calibration significantly improves model performance over default simulations for quantities used in the calibration process, regardless of the mode or calibration length (for calibration lengths one to eight years).

2. Over the calibration period and for quantities used in the calibration process, performance is better with increasing SEB complexity.

3. Calibration significantly reduces the scatter between the simulations from the modes of CHASM after calibration. We infer from this that much of the scatter between the PILPS Phase 2d models could be explained by the specification of parameter values.

4. For quantities used in the calibration and based on 10-year averages outside the calibration period, the most complex mode performs best, but is only slightly better than the simpler modes. This suggests that there is little improvement to simulations from additional complexity above the simplest mode.

5. Improvements in model performance via calibration depend on small model error and calibrating to quantities that properly constrain a model. Identifying these quantities may be model specific. If so, a more complex model such as SLAM is more likely to perform well overall regardless of the quantities used in the calibration process.

[36] Thus, in summary the multicriteria method proved useful in investigating the relationship between model performance and model complexity. Any calibration leads to superior model performance in comparison to no calibration. We do not advocate the replacement of noncalibrated intercomparison exercises with the approach followed here. Rather we suggest that the use of calibration to minimize the impact of variations in the effective parameter values may, if used in conjunction with noncalibrated intercomparisons, aid the identification of relationships between choices of physical parameterizations and model performance.

[37] Overall, we find that complexity improves model performance during the calibration period. However, in the absence of more observed data, which properly constrain the model during calibration, our results suggest that additional SEB complexity adds little value in predicting the future values of quantities not used in the calibration process.

References

Bastidas, L. A., H. V. Gupta, S. Sorooshian, W. J. Shuttleworth, and Z. L. Yang, Sensitivity analysis of a land surface scheme using multicriteria methods, *J. Geophys. Res.*, *104*, 19,481–19,490, 1999.

Budyko, M. I., Heat balance of the Earth's surface (in Russian), *Gidrometeoizdat*, 255 pp., 1956.

Chen, T. H., et al., Cabauw experimental results from the project for intercomparison of land surface parameterization schemes, *J. Clim.*, *10*, 1144–1215, 1997.

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

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

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

Federov, S. F., A study of the components of the water balance in forest zone of the European part of the USSR (in Russian), *Gidrometeoizdat*, *102*, 264 pp., 1977.

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

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

Gupta, H. V., L. A. Bastidas, S. Sorooshian, W. J. Shuttleworth, and Z. L. Yang, Parameter estimation of a land surface scheme using multicriteria methods, *J. Geophys. Res.*, *104*, 19,491–19,504, 1999.

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

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

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

Schlosser, C. A., A. Robock, K. Y. Vinnikov, N. A. Speranskaya, and Y. Xue, 18-year land surface hydrology model simulations for a midla-

titude grassland catchment in Valdai, Russia, *Mon. Weather Rev.*, *125*, 3279–3296, 1997.

Schlosser, C. A., et al., Simulations of a boreal grassland hydrology at Valdai, Russia: PILPS Phase 2(d), *Mon. Weather Rev.*, *128*, 301–321, 2000.

Sellers, P. J., W. J. Shuttleworth, J. L. Dorman, A. Dalcher, and J. M. Roberts, Calibrating the simple biosphere model (SiB) for amazonian tropical forest using field and remote sensing data, 1, Average calibration with field data, *J. Appl. Meteorol.*, *28*, 728–756, 1989.

Sellers, P. J., D. A. Randall, C. J. Collatz, J. A. Berry, C. B. Field, D. A. Dazlich, C. Zhang, G. Collelo, and L. Bounoua, A revised land-surface parameterization (SiB2) for atmospheric GCMs, part 1, Model formulation, *J. Clim.*, *9*, 676–705, 1996.

Sen, O. L., L. A. Bastidas, W. J. Shuttleworth, Z.-L. Yang, H. V. Gupta, and S. Sorooshian, Impact of field calibrated vegetation parameters on GCM climate simulations, *Q. J. R. Meteorol. Soc.*, *127*, 1199–1223, 2001.

Vinnikov, K. Y., A. Robock, N. A. Speranskaya, and C. A. Schlosser, Scales of temporal and spatial variability of midlatitude soil moisture, *J. Geophys. Res.*, *101*, 7163–7174, 1996.

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

H. Gupta, Sustainability of Semi-Arid Hydrology and Riparian Areas (SAHRA), Department of Hydrology and Water Resources, 1133 E. North Campus Drive, Harshbarger Building, University of Arizona, Tucson, AZ 85721, USA.

M. Leplastrier, A. J. Pitman, and Y. Xia, Department of Physical Geography, Macquarie University, North Ryde, New South Wales 2109, Australia. (apitman@penman.es.mq.edu.au)