Evaluation and Transferability of the Noah Land Surface Model in Semiarid Environments

TERRI S. HOGUE

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

LUIS BASTIDAS

Department of Civil and Environmental Engineering, Utah State University, Logan, Utah

HOSHIN GUPTA AND SOROOSH SOROOSHIAN*

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

KEN MITCHELL

NOAA/NWS/National Centers for Environmental Prediction, Camp Springs, Maryland

WILLIAM EMMERICH

USDA, Agricultural Research Service, Tucson, Arizona

(Manuscript received 16 February 2003, in final form 10 August 2004)

ABSTRACT

This paper investigates the performance of the National Centers for Environmental Prediction (NCEP) Noah land surface model at two semiarid sites in southern Arizona. The goal is to evaluate the transferability of calibrated parameters (i.e., direct application of a parameter set to a "similar" site) between the sites and to analyze model performance under the various climatic conditions that can occur in this region. A multicriteria, systematic evaluation scheme is developed to meet these goals. Results indicate that the Noah model is able to simulate sensible heat, ground heat, and ground temperature observations with a high degree of accuracy, using the optimized parameter sets. However, there is a large influx of moist air into Arizona during the monsoon period, and significant latent heat flux errors are observed in model simulations during these periods. The use of proxy site parameters (transferred parameter set), as well as traditional default parameters, results in diminished model performance when compared to a set of parameters calibrated specifically to the flux sites. Also, using a parameter set obtained from a longer-time-frame calibration (i.e., a 4-yr period) results in decreased model performance during nonstationary, short-term climatic events, such as a monsoon or El Niño. Although these results are specific to the sites in Arizona, it is hypothesized that these results may hold true for other case studies. In general, there is still the opportunity for improvement in the representation of physical processes in land surface models for semiarid regions. The hope is that rigorous model evaluation, such as that put forth in this analysis, and studies such as the Project for the Intercomparison of Land-Surface Processes (PILPS) San Pedro–Sevilleta, will lead to advances in model development, as well as parameter estimation and transferability, for use in long-term climate and regional environmental studies.

1. Introduction

Numerous experiments have been carried out to evaluate land surface models with a goal of facilitating advances in model development. Many of these comparison studies have been undertaken by the Project for the Intercomparison of Land-Surface Processes (PILPS; Pitman et al. 1999; Henderson-Sellers et al. 1993, 1995) under the auspices of the Global Energy and Water Cycle Experiment (GEWEX). However, few of the evaluations have been carried out in semiarid regions. Understanding the interaction of land surface processes with climate is crucial for predicting the availability of water resources (i.e., groundwater and surface water sources) in arid regions. The PILPS San Pedro–Sevilleta experiment (Bastidas et al. 2004) is being un-

* Current affiliation: Department of Civil and Environmental Engineering, University of California, Irvine, Irvine, California.

Corresponding author address: Terri S. Hogue, Department of Civil and Environmental Engineering, 5732C Boelter Hall, University of California, Los Angeles, Los Angeles, CA 90095-1593.
E-mail: thogue@seas.ucla.edu

dertaken using five semiarid vegetation sites in the southwestern United States (including two from this study). Length of data, seasonality, short-term climatic events, and data quality are all issues that can impact the estimation of model parameters. This analysis tests some of the hypotheses for the calibration and cross-validation schemes proposed for the PILPS intercomparison study.

Growing interest in the study of semiarid systems has led to the formation of several interdisciplinary groups instituted to furthering the understanding of hydrologic, ecologic, and atmospheric processes in semiarid basins, including the Semi-Arid Land-Surface-Atmosphere (SALSA) program (Goodrich et al. 2000) and, more recently, the National Science Foundation (NSF)-funded Science and Technology Center on the Sustainability of Semi-Arid Hydrology and Riparian Areas (SAHRA) (Sorooshian et al. 2002). Under these programs, several flux tower and data collection sites have been set up to investigate the coupling of surface and climate processes and to investigate the interactions of surface and groundwater systems. The availability of 4 yr of data from two distinct semiarid vegetation types allows us to study the diversity of these environments and to analyze what degree of differentiation (and, hence, parameter estimation) is needed for modeling land–atmosphere interactions in semiarid regions. The objectives of our work are threefold: 1) to rigorously evaluate the performance of the Noah model in semiarid regions, 2) to analyze transferability of the model in semiarid regions (i.e., the ability to directly transfer parameters to another site of similar climatic and vegetated conditions), and 3) to evaluate the ability of the model to capture variations in the energy and water balance due to changes in climate forcings.

A background discussion is presented in the following section. The study area and details of the region are specified in section 3. Methods are described in section 4, along with a brief overview of the Noah model and the optimization program used for parameter estimation. Results are presented in section 5, with a discussion and conclusions in section 6.

2. Background

The semiarid southwestern United States has experienced significant changes within the last century. Increasing population has had dramatic and varied impacts on the region's ecosystems. There have been significant changes in the distribution and type of vegetation found in the area, with large areas of native grasses being replaced by Chihuahuan Desert shrubs and mesquite trees. Human activities such as ranching, agriculture, urban development, fire suppression, and groundwater mining have influenced these transformations (Chehbouni et al. 2000). It is theorized that feedback influences on the local and regional climate have caused a reduction in evaporation losses from the surface to the atmosphere (Qi et al. 2000).

The scale at which surface fluxes are investigated is crucial to understanding the dynamics of land–atmosphere interactions (Rodriquez-Iturbe et al. 2001). An evaluation of the performance of a land surface model's ability to capture high-frequency variations in the water and energy budget (i.e., diurnal processes) may be better suited for study at a finer spatial resolution, or point scale, before evolving to simulations of regional-scale water balance, or global climate. Various land surface modeling studies are occurring at the point scale, but temporal aggregations are typically made to evaluate model performance at the monthly and yearly time scales (Wood et al. 1998; Schlosser et al. 2000; Boone et al. 2001). This study focuses on the point (or flux tower) scale and short time scales to allow for an unbiased evaluation of the Noah model's performance.

Parameter estimation for land surface models is still traditionally done via a global land surface classification scheme with standard values assigned for various land cover or vegetation types. In many cases, calibration is problematic because data may not exist in regions where comprehensive land surface studies are undertaken, and little parameter estimation, and even less validation, can be done. The validation of a model has been defined as the "process of demonstrating that a site-specific model is capable of making accurate predictions for periods outside a calibration period" (Refsgaard and Knudsen 1996). A model is considered validated if the accuracy and prediction during the validation period (outside of calibration) are within what are defined as acceptable errors. The importance of calibration (albeit manual or automatic) and validation within the land surface modeling community is becoming more accepted. This study will use a testing scheme adapted from Refsgaard and Knudsen (1996), along with previously developed multicriteria techniques (Bastidas 1998; Gupta et al. 1999; Bastidas et al. 2001), to obtain parameter values and conduct an objective model performance assessment (validation) at the observational time step of 20 min.

3. Study area

a. Walnut Gulch

The study sites are part of the Walnut Gulch Experimental Watershed in southeastern Arizona, a subbasin of the Upper San Pedro River basin, located in the borderland of southeastern Arizona and northeast Sonora, Mexico. The San Pedro basin is a broad, high-desert valley representing a transition between the Sonoran and Chihuahuan deserts. There is significant topographic (1100–2900 m) and vegetation variation within the basin, and the region experiences significant variability in climate. Vegetation types in the basin in-

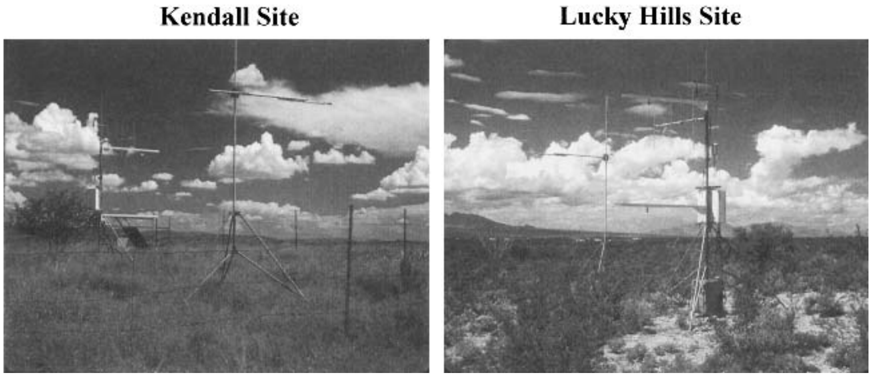


FIG. 1. Kendall grassland and Lucky Hills desert shrub sites within the Walnut Gulch Experimental Watershed [photos courtesy of B. Emmerich and U.S. Department of Agriculture (USDA) Agricultural Research Service (ARS)].

clude a desert shrub steppe, grassland, oak savannah, riparian corridor, and ponderosa pine at the higher elevations. The woody plants in the region tend to be most active in spring or autumn (using deeper soil moisture reserves), while the C_4 grass species respond quickly to upper soil moisture during the summer monsoon periods (Kemp 1983). Two specific study sites were developed within Walnut Gulch during 1996: the Lucky Hills site, with mixed desert shrub vegetation, and the Kendall site, which is a semiarid grassland (Fig. 1). The two sites are located approximately 10 km apart, with Lucky Hills at an elevation of 1372 m and Kendall at 1526 m. Further details regarding the soils and vegetation for these sites can be found in Emmerich (2003) and Scott (1999).

b. Hydroclimatology

Annual rainfall in the Upper San Pedro ranges from 300 to 750 mm yr^{-1} (Goodrich et al. 2000). The mean annual (1 July 1893–31 July 2003 for Tombstone, Arizona) precipitation in the watershed is estimated at 356 mm yr^{-1}, and mean annual temperature is 17°C (Western Region Climate Center 2003). There are several distinct climate periods in the Southwest—the winter period, which receives a significant portion of the annual precipitation, a distinct dry season in spring (March through June), and the summer monsoon, typically from July through September, which brings convective storms to the region and typically delivers over half of the annual rainfall. Potential evaporation rates in the lower parts of the basin are estimated to be 10 times the annual rainfall (Goodrich et al. 2000). Statistics were gathered on the temporal variability of the precipitation time series on the two sites. Some of these data are displayed in Table 1. For the purposes of this study, the monsoon period (MON) is defined as the months of July, August, and September, and the winter period (WIN) is defined as the months of December, January, and February. Values for the study periods are shown, along with the percent of annual precipitation falling in the monsoon for each year (in parentheses).

Several climatic events occurred during the period of the study data, including an El Niño period (1997/98 winter) and an extremely wet monsoon period (1999 summer). The 1997/98 El Niño significantly increased the winter rainfall totals over Arizona during this period (Buizer et al. 2000). During the 1999 monsoon, Lucky Hills received 412 mm of rain, nearly 94% of the annual rainfall during this monsoon season, while Kendall received 315 mm of rain, which is 93% of the yearly total. The precipitation time series for each year are plotted in Fig. 2. Lines are drawn at 1 July and 30 September to designate the monsoon period.

c. Flux data

Micrometeorological measurements were initiated in 1996 at both the Kendall and Lucky Hills sites using a

TABLE 1. Temperature and precipitation totals for the designated study periods at the Kendall and Lucky Hills sites. Shown in parentheses are the percent of the yearly total precipitation for the monsoon period. The monsoon period (MON) is defined as Jul, Aug, and Sep, and the winter period (WIN) is defined as Dec, Jan, and Feb.

	Kendall	Lucky Hills	Kendall	Lucky Hills
	Temperature data (°C)		Precipitation data (mm)	
4-yr avg	16.86	16.91	372	480
1997	16.63	16.79	333	495
1998	16.43	16.56	365	338
1999	17.08	16.99	337	439
2000	17.30	17.29	451	649
1997 MON	24.25	24.72	151 (45%)	241 (49%)
1998 MON	24.38	24.94	250 (68%)	197 (58%)
1999 MON	22.64	22.88	315 (93%)	412 (94%)
2000 MON	24.26	24.63	158 (35%)	282 (43%)
1997/98 WIN	6.65	6.53	132	171
1998/99 WIN	9.99	9.55	9.2	9.9
1999/00 WIN	9.95	9.4	6.1	12.4

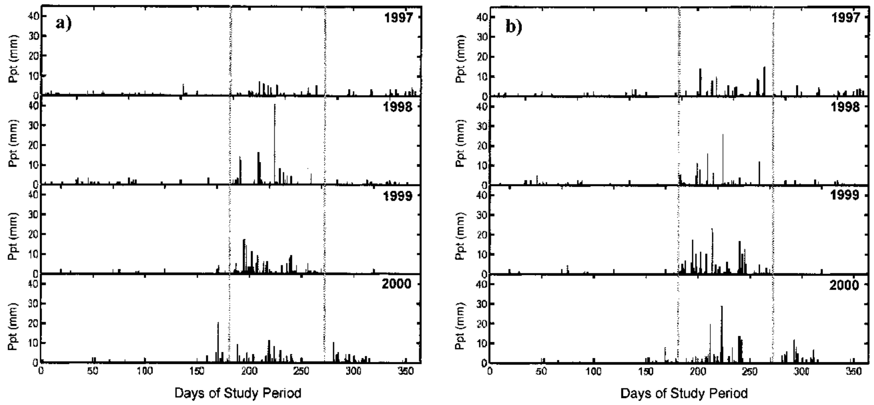
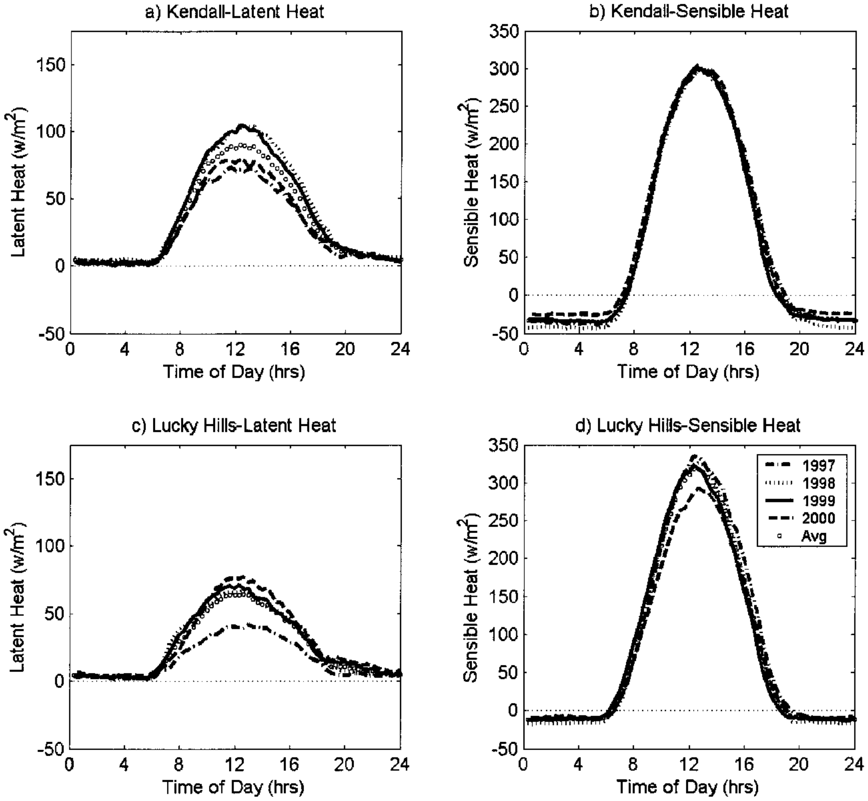


FIG. 2. Twenty-minute time series of precipitation for 4 yr at the (a) Kendall and (b) Lucky Hills sites. The monsoon time period is designated with gray vertical lines at 1 Jul and 30 Sep.

Bowen ratio energy balance system. Continuous, 20-min measurements of water and carbon vapor flux measurements were collected for the four study years (1997–2000). Along with these flux measurements, standard meteorological data (shortwave radiation, air temperature, precipitation, wind speed, precipitation, and relative humidity) were also collected. Data for both sites (4 yr) were subject to quality control with particular attention to extraneous values and Bowen ratio values approaching −1.0.

Annual mean diurnal cycles of sensible and latent heat fluxes are displayed in Fig. 3. Each line in the


a) Kendall-Latent Heat

b) Kendall-Sensible Heat

c) Lucky Hills-Latent Heat

d) Lucky Hills-Sensible Heat

FIG. 3. Annual mean diurnal cycle of sensible and latent heat at the Kendall and Lucky Hills sites: (a) Kendall: latent heat, (b) Kendall: sensible heat, (c) Lucky Hills: latent heat, and (d) Lucky Hills: sensible heat. Single-year diurnal fluxes are shown (1997, 1998, 1999, 2000) along with the daily average flux for the 4-yr period (1997–2000). Note the scale on the latent heat flux axis is half that of the sensible heat flux axis.

figures represents the aggregated daily average flux for each of the 4 yr (i.e., the yearly average at each of the 72 twenty-minute time steps throughout the day), along with the 4-yr average (note the scale on the latent heat axis is half that of the sensible heat axis). The diurnal cycle of sensible heat for each of the 4 yr reveals good consistency from year to year. However, the diurnal cycles of latent heat reveal much more variability. At the Kendall site, the average maximum varies from 70 to around 100 W m^{-2}, and has a fairly smooth diurnal cycle. At the Lucky Hills site, three of the years are fairly similar (1998, 1999, and 2000). The 1997 year at Lucky Hills was much drier overall (lowest average precipitation). Also, the monsoon period during 1997 was problematic at this site (W. Emmerich 2003, personal communication), and was removed for purposes of this study, resulting in a lower overall diurnal average for this year. Although precipitation totals for the 1997 and 1999 years are similar, the patterns in the latent heat flux are not. During 1997, the rain was more distributed throughout the year, while in 1999 there is a much larger concentration of precipitation during the monsoon, resulting in an above-average latent heat flux for this year. The increased latent heat is most likely due to more active vegetation and possible saturation of the top soil layers during the monsoons. At Kendall, 2000 is similar to 1997, with lower annual means for latent heat. This could also be due to the slightly drier monsoon period (like 1997), and more moisture later in the fall season.

4. Methods

a. Model

The Noah model (Ek et al. 2003) is one of the ever-evolving community, or multigroup, land surface models. The Noah model was chosen for evaluation for several reasons. The model has been used in previous studies by this research group and demonstrated the best performance for another semiarid site (Hogue 2003). More importantly, the model is currently parameterized for use over semiarid regions in the National Centers for Environmental Prediction (NCEP) North American Land Data Assimilation System (NLDAS) (Mitchell et al. 2004) of the National Weather Service (NWS). The model is updated periodically on the NCEP Web site (online at ftp://ftp/ncep.noaa.gov/pub/gcp/ldas/Noahlsm) and version 2.5.1 is used in this analysis (release date: 5 March 2002). The Noah model contains four soil layers: a thin 10-cm top layer, a second root zone layer of 20 cm, a deep root zone of 60 cm, and a subroot zone of 110 cm. It can be run for 13 vegetation covers (2 of which use the same parameter values) and nine different soil types (two of which also use the same parameters).

The Noah model has 33 parameters: 10 related to vegetation and 23 that describe soil properties (Table 2). The model also has 16 initial states (when run with four root layers). Of these 49 variables to be estimated, 32 are included throughout the parameter estimation and validation testing. Because soil moisture data were not available at this site, eight of the initial soil moisture states were included in the parameter analysis. Although optimizing initial states is not common practice in operational forecasting, it was undertaken in this study to allow for an objective "best" estimate of initial soil moisture states and to allow for a balanced comparison of model performance for short time periods (seasonal and yearly time frames). The Noah model uses a local greenness fraction from the Normalized Difference Vegetation Index (NDVI) to establish seasonality in the model for each of the 13 vegetation types. The leaf-area index (LAI) value is typically held constant (or used as a tuning parameter) instead of also being varied seasonally (Gutman and Ignatov 1998). Consequently, the LAI parameter was included as a parameter to be calibrated, while the monthly greenness fraction was obtained from NCEP and not adjusted.

b. Optimization

Calibration (or optimization) of a model involves selecting values for parameters so that the model simulates the behavior (as closely as possible) of the study site. Various automated techniques have evolved over the past few decades, with increasing success and application to hydrologic models. The multicriteria theory was more recently applied to the optimization of land surface models by Gupta et al. (1998). This work proposed that the parameter estimation problem be reformulated as a multicriteria problem that seeks a set of trade-off solutions (pareto set) instead of a single unique solution (parameter set). Yapo et al. (1997) developed the Multiple-Objective Complex Optimization Method-University of Arizona (MOCOM-UA), which has shown success in providing improved calibrations of watershed models (Boyle et al. 2000, 2001; Beldring 2002), hydrochemistry models (Meixner et al. 2000, 2002), and land surface schemes (Gupta et al. 1999; Bastidas 1998; Bastidas et al. 2001; Leplastrier et al. 2002; Xia et al. 2002). More details on the MOCOM-UA algorithm may be found in Gupta et al. (1999) and Yapo et al. (1997). We utilize root-mean-square error (rmse) as the error function for the MOCOM-UA optimization of the Noah model. For validation, both the rmse and the Nash–Sutcliffe forecasting efficiency (nse) (Nash and Sutcliffe 1970) are used. Rmse is defined as

$$\mathrm{rmse} = \sqrt{\frac{1}{n}\sum (O_t - S_t)^2}, \qquad (1)$$

where O_t is the observed value at time t, S_t is the model simulation at time t, and n is the number of observations. The nse measures the fraction of the variance of observed values explained by the model. Values can

Table 2. Noah model parameter descriptions, default values, optimized values for the 4-yr and 1999 monsoon calibrations, ranges for calibration, and flag for optimization (OPT) or fixed value (FIX).

No.	Parameter name	Flag	Default	Min value	Max value	Physical meaning of parameters
1	rcmin	OPT	400	40	1000	Minimum stomatal resistance (m)
2	rgl	OPT	100	30	150	Used in solar radiation term of canopy resistance Fx
3	hs	OPT	42	36.35	55	Used in vapor pressure deficit term of canopy resistance Fx
4	z0	OPT	0.011	0.01	0.99	Roughness length (m)
5	lai	OPT	4	0.05	6	Leaf-area index
6	cfactr	OPT	0.5	0.1	2	Canopy water parameter
7	cmcmax	OPT	5.00E−04	1.00E−04	2.00E−03	Second canopy water parameter (m)
8	sbeta	OPT	−2	−4	−1	Used in calculation of vegetation effect on soil heat flux
9	rsmax	OPT	5000	2000	10000	Maximum stomatal resistance (m)
10	topt	OPT	298	293	303	Optimum transpiration air temperature (K)
11	maxsmc	OPT	0.42	0.33	0.66	Porosity
12	drysmc	OPT	0.119	0.02	0.2	Air dry soil moisture content limits
13	psisat	OPT	0.62	0.04	0.62	Saturated soil potential
14	satdk	OPT	1.41E−05	5.00E−07	3.00E−05	Saturated soil hydraulic conductivity (m s^{-1})
15	b	OPT	4.26	3.5	10.8	The "b" parameter
16	satdw	OPT	2.33E−05	5.71E−06	2.33E−05	Saturated soil diffusivity
17	quartz	OPT	0.1	0.1	0.82	Soil quartz content
18	nroot	FIX	4	4	4	Number of root layers
19	refdk	OPT	2.00E−06	5.00E−07	3.00E−05	Reference value for saturated hydraulic conductivity
20	fxexp	OPT	2	0.2	4	Bare soil evaporation exponent used in DEVAP
21	refkdt	OPT	3	0.1	10	Reference value for surface infiltration parameter
22	czil	OPT	0.2	0.05	0.8	To calculate roughness length of heat
23	csoil	OPT	2.00E+06	1.26E+06	3.50E+06	Soil heat capacity for mineral soil component
24	zbot	FIX	−8	−3	−20	Depth of lower boundary soil temperature (m)
25	frzk	OPT	0.15	0.1	0.25	Ice threshold (above frozen soil is impermeable)
26	xnup	OPT	0.025	0.025	0.08	Threshold snow depth (100% snow cover) (m)
27	snoalb	FIX	0.75	0.3	0.75	Maximum albedo over deep snow
28	salp	FIX	2.6	2.6	2.6	Shape of distribution function of snow cover
29	slope	FIX	0.1	0.001	1	Slope
30	t1	FIX	299	265	300	Initial skin temperature (K)
31	cmc	FIX	5.00E−04	0	0.001	Intitial canopy water content (m)
32	snowh	FIX	0	0	0.1	Initial actual snow depth (m)
33	sneqv	FIX	0	0	0.1	Initial water equivalent snow depth (m)
34	sldpt1	FIX	0.1	0.1	0.1	Soil depth, layer 1 (m)
35	sldpt2	FIX	0.2	0.2	0.2	Soil depth, layer 2 (m)
36	sldpt3	FIX	0.6	0.6	0.6	Soil depth, layer 3 (m)
37	sldpt4	FIX	1.1	1.1	1.1	Soil depth, layer 4 (m)
38	stc1	FIX	297	260	300	Initial soil temperature (K)
39	stc2	FIX	293.7	260	300	Initial soil temperature (K)
40	stc3	FIX	291.5	260	300	Initial soil temperature (K)
41	stc4	FIX	290.4	260	300	Initial soil temperature (K)
42	smc1	OPT	—	0.05	0.56	Initial soil total moisture
43	smc2	OPT	—	0.05	0.56	Initial soil total moisture
44	smc3	OPT	—	0.05	0.56	Initial soil total moisture
45	smc4	OPT	—	0.05	0.56	Initial soil total moisture
46	sh2o1	OPT	—	0.05	0.56	Initial soil liquid moisture
47	sh2o2	OPT	—	0.05	0.56	Initial soil liquid moisture
48	sh2o3	OPT	—	0.05	0.56	Initial soil liquid moisture
49	sh2o4	OPT	—	0.05	0.56	Initial soil liquid moisture

range from minus infinity to 1.0, with higher values indicating better agreement. The nse is represented as

$$\mathrm{nse} = 1 - \left(\sum (O_t - S_t)^2 \Big/ \sum (O_t - O_{\mathrm{mean}})^2\right), \quad (2)$$

where O_t and S_t are defined as above, and O_{mean} is the mean of the observed values.

c. Validation

Traditional model evaluation includes a split-sample (SS) testing strategy, where parameters are estimated during one time period and are evaluated on a separate or independent time period. Few studies follow more rigorous testing schemes, such as those proposed in Klemes (1986) and Refsgaard and Knudsen (1996), including proxy basin (PB) and differential split-sample (DSS) analyses. These tests vary somewhat in approach, but, in general, parameter sets are obtained via calibration and then validated in the SS, PB, or DSS framework. The PB test involves application of a calibrated model to a similar catchment or site with no direct calibration of parameters on the validation basin.

This evaluation provides insight into the common practice of transferability of parameters within similar climatic regions. Adjustment of parameters can be undertaken to account for different conditions within the proxy basin (i.e., slightly different soil type, etc.), but the parameters are not calibrated against any observations. The DSS test involves calibration of the model based on data before a catchment change occurs, adjustment of parameters to reflect the expected changes, and validation or testing of the model after the change has occurred. These changes could involve land alterations (i.e., fire, urbanization, deforestation, etc.) or some sort of nonstationary climate event, where a climate phenomenon occurs that has not been observed in the calibration data.

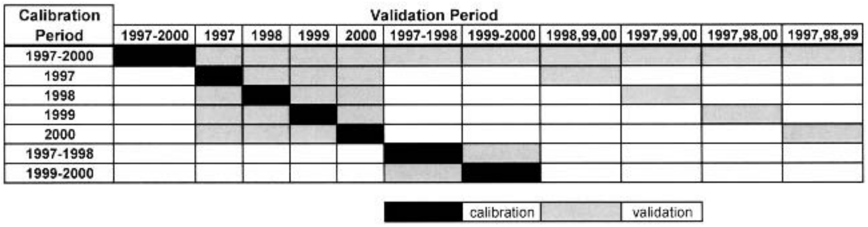
Given the length of our dataset, we can only assess short-term climate phenomena (monsoon, El Niño conditions, wet and dry years, etc.) and the impact of these periods on model simulations. Longer-term climate change and variability, and how these will affect vegetation and evaporation processes in semiarid regions, can only be undertaken through long-term model simulations. Taking into account the complexity of the above regime, the given data, and the climatic conditions of the basins in this study, a modified testing scheme is developed for this analysis as follows:

1) The SS tests will involve calibration on one time period and validation on independent time segments. However, to make this analysis more robust and to assess the influence of data length on the estimated parameters, calibration is performed over various time periods in the datasets, with corresponding validation on differing lengths of data. This is illustrated in Fig. 4.
2) The PB analysis will include the direct transfer of the estimated parameters from the calibrated models between the two similar study sites. This PB test is undertaken to test the capability of the model to simulate energy fluxes from a site for which no calibration data would be available. Direct application of the model is done, with no calibration of parameters. The two sites used in this analysis contain similar soils, but slightly different vegetation types (one grassland and one desert shrub), which respond differently to the atmospheric forcing in the region. Testing on the proxy basins will be done for the corresponding time period (i.e., same calibration and validation period).
3) A modified DSS is undertaken for this study. Because the data period of this analysis included a total of 4 yr of data, only short-term climatic and seasonal conditions were evaluated. The summer monsoon period is evaluated, as are the winter periods (the two typical precipitation periods in this region), along with the 1999 monsoon period and the 1997/98 winter (El Niño event). Parameters from the yearly calibrations are evaluated on these unusual climatic events The DSS testing scheme is illustrated in Fig. 5.

5. Results

a. General calibration

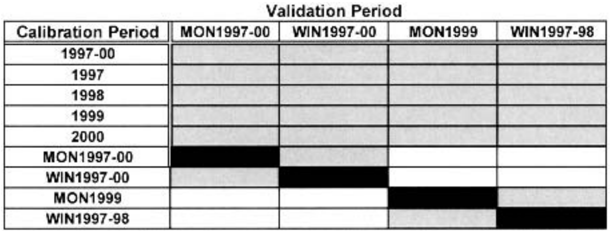
The Noah model was calibrated for each of the listed periods (Table 3) using sensible heat (SH), ground heat flux (G), and ground temperature (T_g) data (taken at a 5-cm depth), using rmse as the objective function. Due to the variability of latent heat (LE) flux and the low values found through most of the year at the Arizona sites, latent heat was not explicitly used as one of the calibration criteria, but model simulations of latent heat flux were evaluated along with the other fluxes. This selection is supported by previous work by the authors, investigating which surface fluxes are best suited to the multicriteria estimation procedures (Bastidas 1998; Bastidas et al. 2001). The multicriteria approach produces a set of solutions, or Pareto set, with the property that, moving from one solution to another, results in the improvement of one criterion while causing deterioration in another. The Pareto set represents the minimal uncertainty that can be achieved for the parameters via calibration, without subjective assignment of weights to the individual model responses (Gupta et al. 1998). For each calibration a 250 parameter set, or Pareto solution,



Calibration Period	Validation Period										
	1997-2000	1997	1998	1999	2000	1997-1998	1999-2000	1998,99,00	1997,99,00	1997,98,00	1997,98,99
1997-2000	■	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
1997		■	▨	▨	▨			▨			
1998		▨	■	▨	▨				▨		
1999		▨	▨	■	▨					▨	
2000		▨	▨	▨	■						▨
1997-1998						■	▨				
1999-2000						▨	■				

■ calibration ▨ validation

FIG. 4. The SS testing scheme. Calibration periods are listed down the left column (designated in black in the matrix). Validation periods are listed across the top with corresponding testing periods designated as gray blocks. For example, parameters from the 1997 calibration period are evaluated on the same period (1997), along with single years 1998, 1999, and 2000, and the 3-yr period of 1998, 1999, and 2000.



	Validation Period			
Calibration Period	MON1997-00	WIN1997-00	MON1999	WIN1997-98
1997-00	▨	▨	▨	▨
1997	▨	▨	▨	▨
1998	▨	▨	▨	▨
1999	▨	▨	▨	▨
2000	▨	▨	▨	▨
MON1997-00	■	▨		
WIN1997-00	▨	■		
MON1999			■	▨
WIN1997-98			▨	■

FIG. 5. The DSS testing scheme. Calibration periods are listed in the left column and the validation periods in which these parameters were tested are MON1997–2000, WIN1997–2000, MON1999, and WIN1997–98. The gray shading represents testing, white is no testing, and black is the same period as calibration.

was generated using SH, G, and T_g for criteria (based on Bastidas 1998). Along with the calibrations, default parameters were also run for each of the data periods.

Cumulative distribution plots of the objective function solutions (from the 250 values) were generated (Fig. 6). Each surface flux is labeled across the top of the respective column. Each line represents one Pareto set (represented by its objective function values). The top row in each plot (a and b) displays the distribution of results for the 1-yr calibrations, and the second row contains the seasonal calibrations (monsoons and winter periods). Each set of functions is normalized against the 4-yr calibration (1997–2000). The 4-yr set would be represented as a zero line on the plots. Therefore, if the errors during the selected calibration periods are lower than the errors from the 4-yr set, the distributions will lie to the left of zero, and errors higher than the 4-yr period will lie to the right of zero. A more vertical line represents a set of solutions that have less variation (very similar objective functions and parameter sets), while a more curved or horizontal line indicates more variations.

Results at the Kendall site are illustrated in Fig. 6a, with Lucky Hills in Fig. 6b. In general, most of the errors for the fluxes are lower during the 1-yr calibration periods than the 4-yr period (curves lie to the left of 0). There is more variability from year to year in the parameter sets for SH and T_g, while sets are more consistent for G. Errors in the LE fluxes are higher, although, as stated previously, this flux was not used in the calibration process. The seasonal calibrations at Kendall are similar to the yearly calibrations, with slight variability in the parameter sets during the seasons, but generally similar performance for SH, G, and T_g. However, model simulations did not reproduce LE for either the 4-yr monsoon calibration (MON1997–2000) or the 1999 monsoon (MON1999). Both of these time periods have much higher errors for LE than the overall 4-yr set and are off of the normalized scale (>2) for this figure. For the other fluxes, the parameter sets reveal fairly good consistency when compared to the 1997–2000 calibration set (0 line), and all seasonal calibration sets have lower errors for T_g than over the 4-yr period.

TABLE 3. Data periods used for calibration of the Noah model.

Calibration periods for study sites
1997–2000 (4-yr period)
1997–2000 (each yearly period)
1997–98 (2-yr period)
1999–2000 (2-yr period)
1997–2000 monsoon periods (MON1997–2000)
1997–2000 winter periods (WIN1997–2000)
1999 monsoon (MON1999)
1997/98 winter (WIN1997–98)

The same set of plots is shown in Fig. 6b (Lucky Hills). Again, the yearly calibration periods have much lower errors than the 4-yr set and show more consistent behavior for most of the fluxes than at Kendall. For the seasonal calibrations, SH and G fluxes show consistency and similar behavior, and errors are much lower for the winter periods than the overall 4-yr period for T_g. Again, the latent heat flux is where the model diverges. Errors are lower for the winter calibration periods (WIN1997–2000 and WIN1997–98), but are off the scale for the MON1999 and MON1997–2000. Similar to the results at Kendall, these results tell us that the model performs well during the dry periods when SH is the dominant energy component, but during the wet monsoon periods, the model does not capture the dramatic change in the energy balance and does not capture this variability in LE as well.

b. Nash–Sutcliffe efficiency

After parameter sets were generated for each of the study periods, a "best set" was chosen to use in the cross-validation testing schemes. This single-best set was generated using an L2 norm procedure (or Euclidean distance). The Euclidean distance of two points $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ in *Euclidean n space* is computed as

$$\text{L2 norm} = \sqrt{(x_1 - x_2)^2 + (x_2 - y_2)^2 + \ldots + (x_n - y_n)^2}. \tag{3}$$

The nse statistics for the selected parameter sets are shown in Fig. 7, with both the Kendall and Lucky Hills results plotted (SH, LE, and T_g). The best results are arguably for SH and T_g at both sites. There is some decrease in performance during the 4-yr winter period (WIN1997–2000) at Lucky Hills, but the trend is such that greater than 80% (>0.80 nse) of the variance of both SH and T_g is captured by the model simulations. Keeping in mind that calibration was not performed using the latent heat flux, the model performance is still less than ideal. Efficiency values range from 40%–70% at Kendall, and from 10%–70% at Lucky Hills. The high variability (and low average values) of LE at these

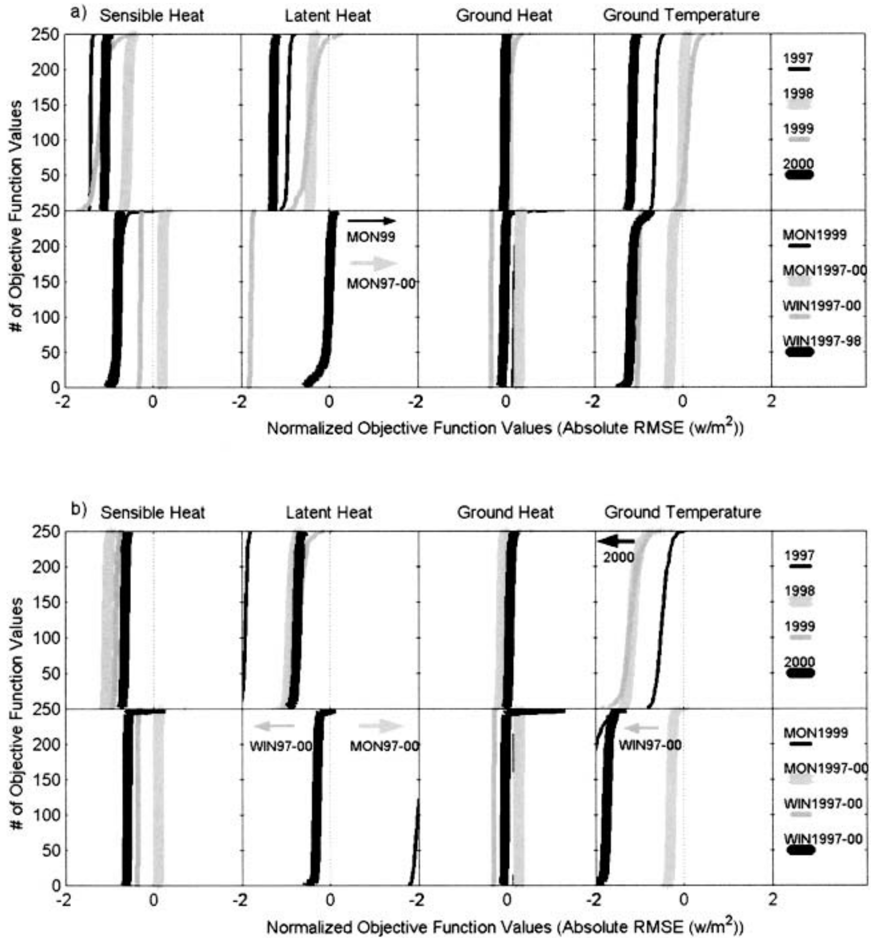


FIG. 6. Cumulative distribution functions (CDFs) of objective functions (or rmses) from the calibration periods for (a) Kendall and (b) Lucky Hills. Each box represents the errors from the calibration set for a specific flux. For example, in the upper-left corner in (a), four single-year calibrations (1997, 1998, 1999, and 2000) and the errors that resulted during the calibration for sensible heat are shown. The rmses are normalized against the 4-yr calibration and each yearly set is designated by a gray or black, thick or thin line. Objective function values that are more or less than a difference of 2 (normalized value) are offscale and are designated by an arrow.

sites results in poor model simulations, especially 1997, which overall has much lower LE values throughout the year. Low nse during the 4-yr monsoon period (MON1997–00) and the winter period (WIN1997–00) is indicative of problems with the model capturing the unusual variability of LE using parameters from a longer-time-period calibration. The model calibration from the MON1999, on the other hand, is able to capture this variability. These results suggest that the ability of the model to capture short-term, intense changes from a dry condition to a wetter, evaporative condition is limited using parameter sets from longer time periods. We theorize that this is due to the optimization procedure finding parameters that capture the dominant sensible heat flux that occurs during most of the year in this region (typically 9–10 months each year). Parameter sets calibrated specifically on the short, highly variable monsoon period do a better job of capturing the variability in these periods.

c. *Model validation*

1) SPLIT SAMPLE

The SS analysis consisted of applying the best parameter set, and default parameters, to each of the periods as shown in Fig. 4 (only annual or multiannual calibration periods). Results are presented as scatterplots in Figs. 8a and 8b. Each box in the figures represents absolute rmse (W m^{-2}) for LE (y axis) versus SH (x axis) for each of the selected 11 validation periods. For example, in the 1997–2000 plot (upper-left corner), the parameters calibrated on this time period, along with parameters from the other calibration periods (1997, 1998, 1999, 2000, 1997–98, 1999–2000), and default, are

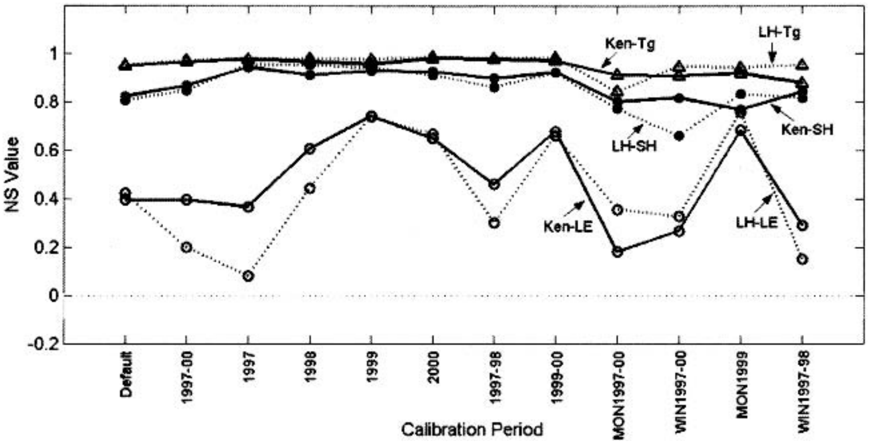


FIG. 7. Nash–Sutcliffe forecasting efficiency for the various calibration periods. The Kendall site is represented by the solid lines and Lucky Hills is represented by the dashed lines. Each calibration period is listed on the *x* axis. Sensible heat values are represented by a (●) symbol, latent heat values are represented by a (o) symbol, and ground temperature values are represented by a (Δ) symbol.

tested over the 4-yr period. Each specific parameter set is designated via symbols described in the lower-right legend, and default parameter simulation errors are designated as the dashed lines on each plot.

At the Kendall site (Fig. 8a), simulations from the default parameters generally result in higher rms errors than the calibrated parameter sets. In many cases there is a 5–10 W m^{-2} reduction in sensible heat errors and up to 20 W m^{-2} when using any calibrated parameter set rather than the default values. In the 1997–2000 period for Kendall, the default parameters resulted in rms errors of around 65 W m^{-2} for sensible heat and 50 W m^{-2} for latent heat over the 4-yr dataset (1997–2000). Although these errors may seem high, in relation to the normally large sensible heat values in this region (up to ~600 W m^{-2} during summer periods), the model errors are actually around 10% or less of this value. As has been discussed, latent heat has greater variability than sensible or ground heat fluxes in this region, with fairly low values throughout the year (near 10–15 W m^{-2}) and increasing to much higher values of around 300 W m^{-2} during the monsoon period, hence, the model errors are a significantly higher percentage of the average values. During the shorter time periods at the Kendall site (1997, 1998, 1999, and 2000), the parameter sets all perform fairly consistently on these validation periods, except for 1999, where there is more spread in the results. The parameter set calibrated over the 4-yr period, designated by the star (*), results in slightly larger errors for the latent heat fluxes when compared to the other parameter sets during three of the 1-yr periods. This result also suggests that the parameter sets from the longer time periods do not capture the short-term year-to-year variability in the data at these sites. The extremely wet monsoon period was during 1999, and the inconsistency in the parameter sets to capture this extreme is observed, with the 1999 and 2000 parameter sets performing better than other sets for this wetter year. Of the 2-yr validation periods tested, the 1999–2000 period shows more inconsistency, with only the parameter set calibrated specifically for this period performing well. The final four validation periods in the graphic consist of testing a 1-yr calibration set on the other 3 yr, along with a test of the 4-yr set and the default on these periods. Results from these runs reveal that the 1-yr set performs fairly well on the alternative 3-yr periods, and on some occasions, as well as, or slightly better than, the 4-yr set.

The Lucky Hills site (Fig. 8b) shows some of the same general trends as at the Kendall site. However, the default parameters perform better during some of the validation periods than at Kendall. This could be explained by the difference in vegetation at the two sites, with the Kendall site having a more dense vegetation cover, while the desert scrub at Lucky Hills, having bare ground exposed, may be more typical of the vegetation parameterized in these regions. Again, there is more spread in the results in the periods containing the 1999 and the 2000 periods, and the parameters calibrated from these periods perform better than the other sets tested for this time. For the 3-yr validation periods, the default parameters perform slightly better than that of the 4-yr period, and similar to the calibrated set on latent heat flux. Errors for SH are still higher in most periods using the default parameters. As stated previously, the magnitude of the errors is also related to the magnitude of the fluxes, with the sensible heat fluxes being much higher at this site for most of this year.

2) PROXY BASIN

Each parameter set discussed above, along with the four seasonal calibrations, were evaluated directly against the same time period on the other study site.

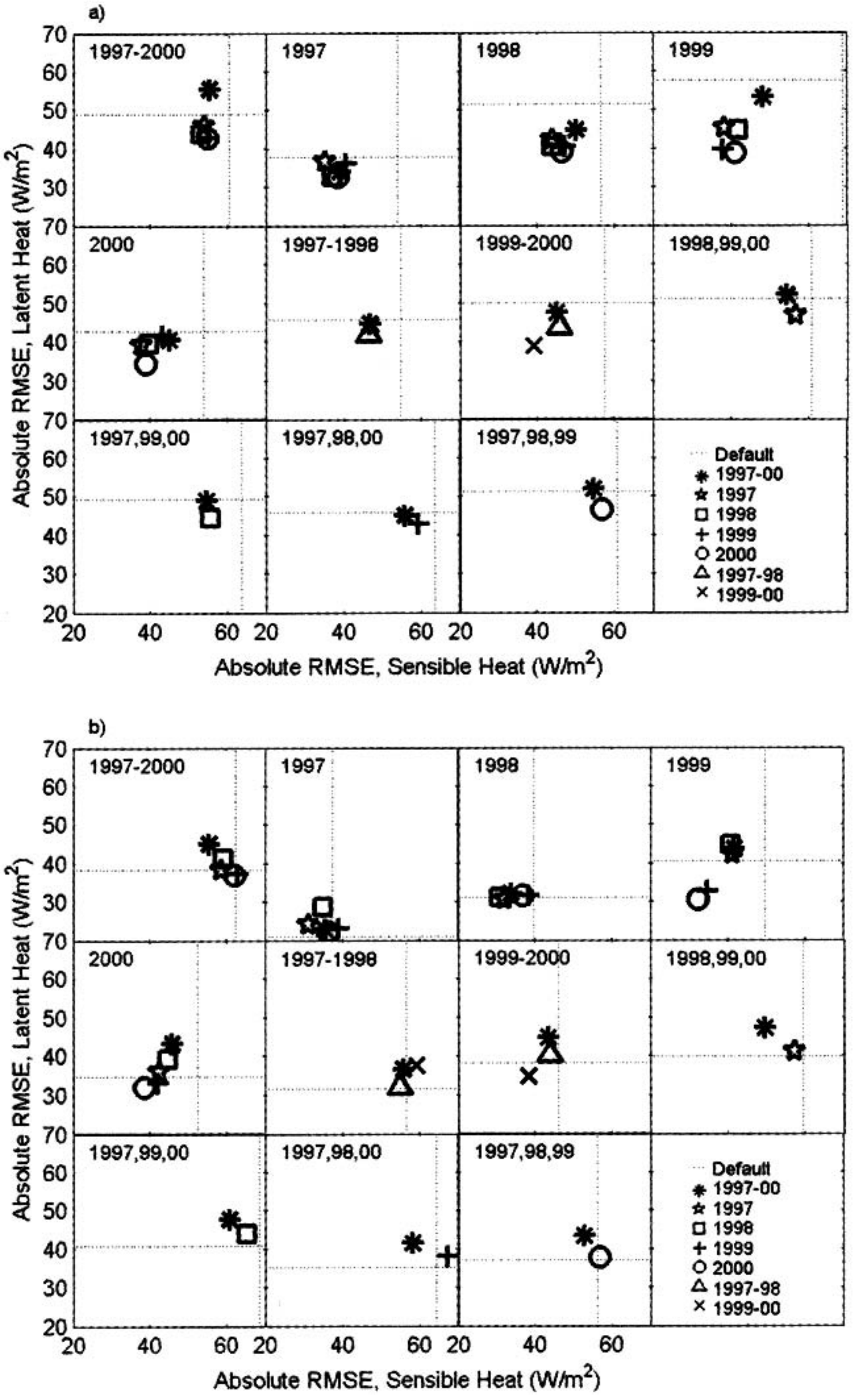


FIG. 8. The SS results for the (a) Kendall and (b) Lucky Hills sites. Each box represents one of the validation periods and is a scatterplot of the rmse for latent heat and sensible heat flux for each parameter set. Default parameters: dotted gray lines; various calibration sets: symbols listed in legend.

Direct application of the parameter sets is used to simulate surface fluxes for the same time period at a proxy site (Kendall to Lucky Hills, and Lucky Hills to Kendall), with no adjustment of parameters. Results from this site-to-site cross-validation study are presented in Fig. 9. Results for the calibrated parameter sets (stars), default (dashed lines), and the proxy basin parameter set (squares) are shown in each plot.

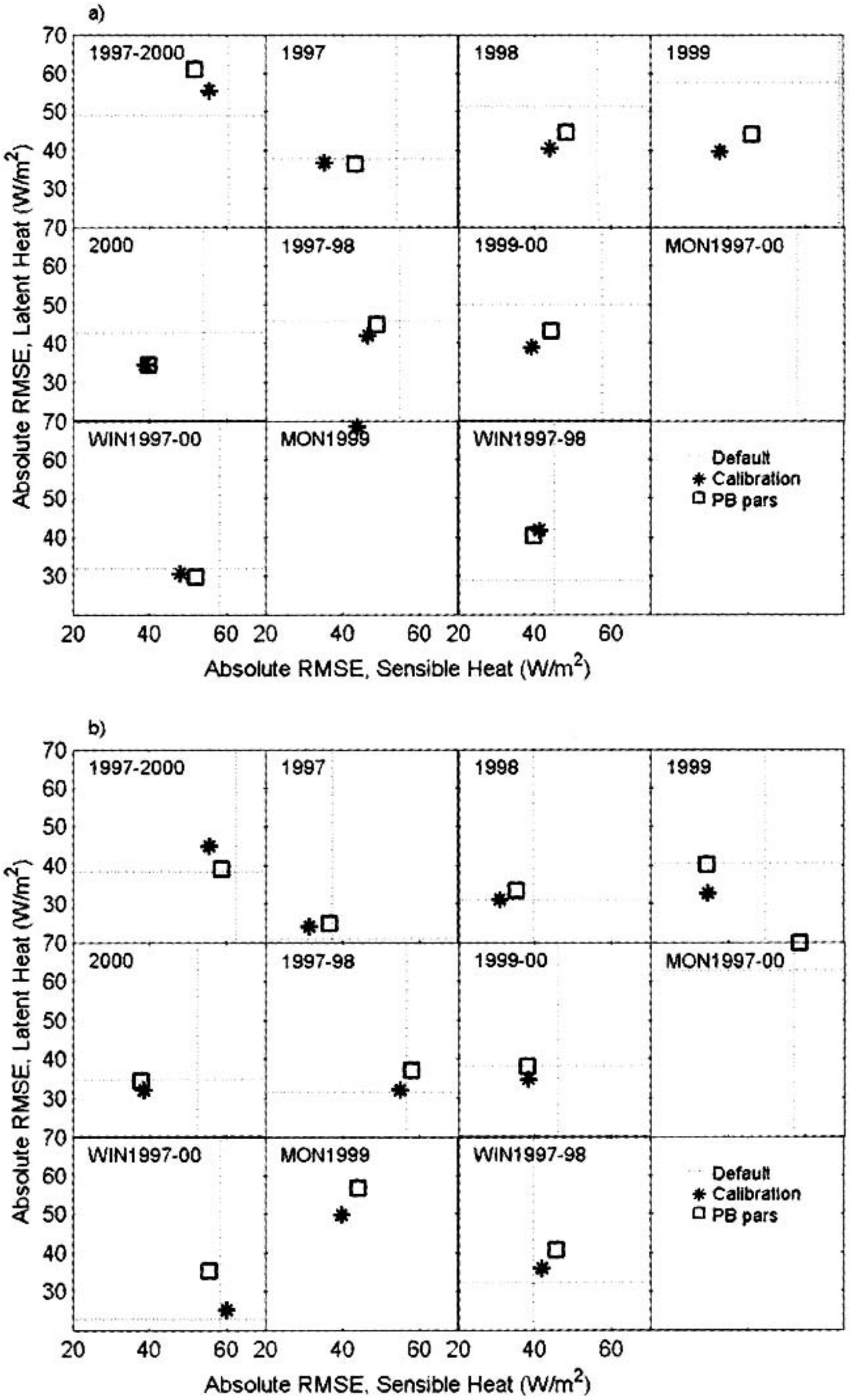


Fig. 9. Same as in Fig. 8, but for the PB results. Again, the (a) Kendall and (b) Lucky Hills sites are shown.

When applying the Lucky Hills parameters to the Kendall site (Fig. 9a), there are a few validation periods when the default parameters perform as well as the calibrated sets. However, the default set also results in extremely poor performance for several periods (1997, 1998, 1999, 2000, 1999–2000) when compared to the calibrated or proxy parameter sets. More significantly, in most cases at these sites, proxy basin parameters result in a slight decrease in model performance when compared to the site-specific calibrated set (increase of

5–20 W m^{-2} in flux errors). Two of the seasonal calibrations—the 1997–2000 monsoon period (MON1997–2000), and the 1999 monsoon (MON1999)—have proxy basin and calibration sets that result in errors greater than the 70 W m^{-2} on the given figure.

At the Lucky Hills site (Fig. 9b) the proxy basin parameters (Kendall sets) generally also result in poorer performance than the site-specific set for several time periods (1999, 1997–1998, MON1999, WIN1997–2000, and WIN1997–1998). Only during 2000 do the proxy basin parameters yield similar results to the site-specific calibration. Default parameters at the Lucky Hills typically result in larger flux errors than either of the calibrated sets, but do perform well for two winter periods at Lucky Hills (WIN1997–2000 and WIN1997–98). The monsoon periods at Lucky Hills show similar problems as at the Kendall site, with generally higher errors for these periods.

3) Differential split sample

Four distinct climatic periods were selected (MON1997–2000, WIN1997–2000, MON1999, and WIN1997–98) to evaluate how parameter sets from various calibration periods would perform on these atypical climatic events found in the southwest. Default parameters were also run. Results are presented in Fig. 10. Findings from the Kendall site are displayed in the top row of plots (Fig. 10a), with Lucky Hills in the bottom series (Fig. 10b). Note the difference in scale for the rmse (range of 20–100 W m^{-2}) from Figs. 8 and 9 (range of 20–70 W m^{-2}). The parameter sets show consistent performance during the winter periods at both sites with a clustering of errors in the same region. These sets (and default parameters) result in fairly low LE errors (20–40 W m^{-2}). The parameter sets have much greater variability during the monsoon periods. For the monsoon 1997–2000 period (Kendall site), the parameter sets yield similar errors for SH, but much more variability in errors for LE. For the MON1999, all parameter sets show a large spread in both SH and LE. Interestingly, the 4-yr parameter set (1997–2000) does capture this period fairly well, along with the 1999 and MON1999 set (which would be more expected). At Lucky Hills, similar results are observed, with parameter sets performing well on the drier winter periods, and showing poorer performance on the wetter monsoon periods. Simulations during these wet periods (with high LE fluxes) seem to be very dependent on the parameter values, and calibration of parameters specific to these periods appears critical for the model to yield adequate simulations. Again, using parameter sets from longer time frames or periods where the sensible heat flux is dominant (which is most of the year in this region), contributes to poorer model performance during short, wet periods.

4) Time series

Time series of selected optimizations from the Kendall site are presented in Figs. 11 and 12, with 10-day periods from the dry season (1 June 1998) and the monsoon period (1–10 August 1999), respectively. Both figures show the default (dark solid line) simulation, the 250 trade-off or Pareto set solution (gray area), and the observed data for the period (circles). The scatterplots to the right of the time series correspond to the entire period of calibration (i.e., observed versus 4-yr calibration results). For the 10-day period during June (Fig. 11), it is observed that the model tracks well for SH, G, and T_g. The trade-off solution encompasses the observed data throughout the diurnal cycle for these three fluxes. The width of the solutions is also fairly narrow, indicating less parameter uncertainty for these simulations. The default parameters tend to overestimate on the sensible heat and undersimulate for T_g and G. For

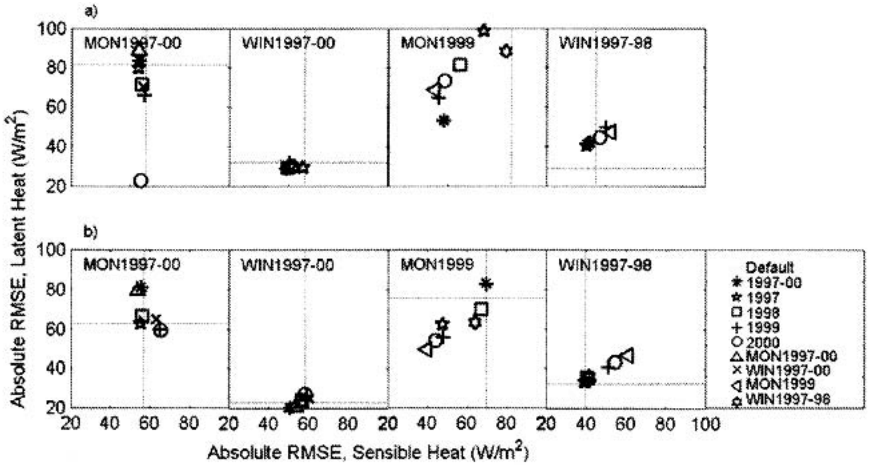


FIG. 10. Same as in Fig. 8, but for the DSS results. Both the (a) Kendall and (b) Lucky Hills sites are shown.

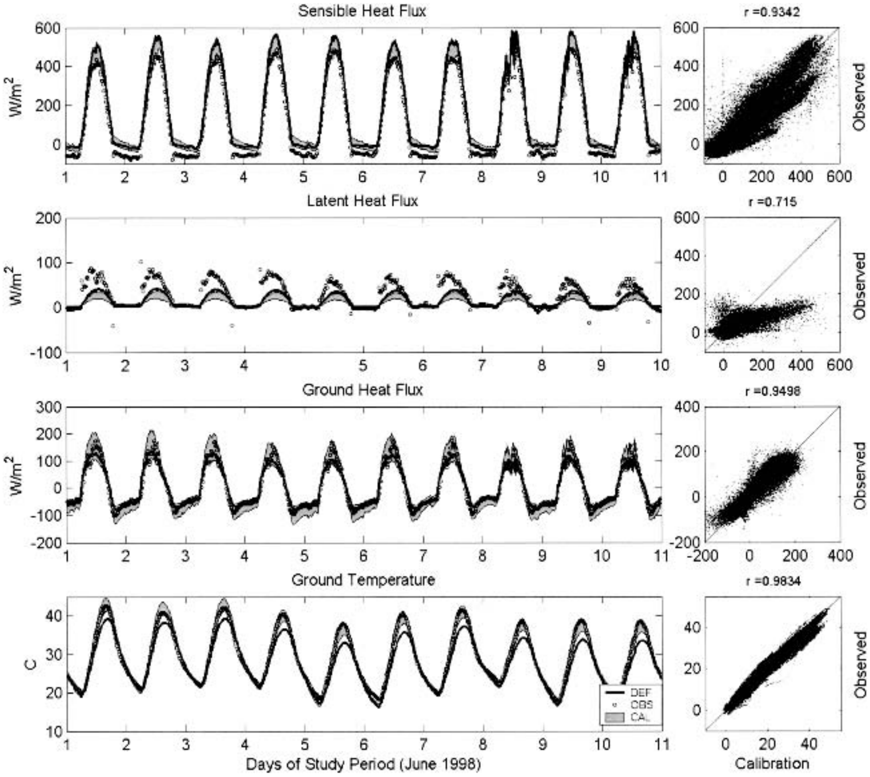


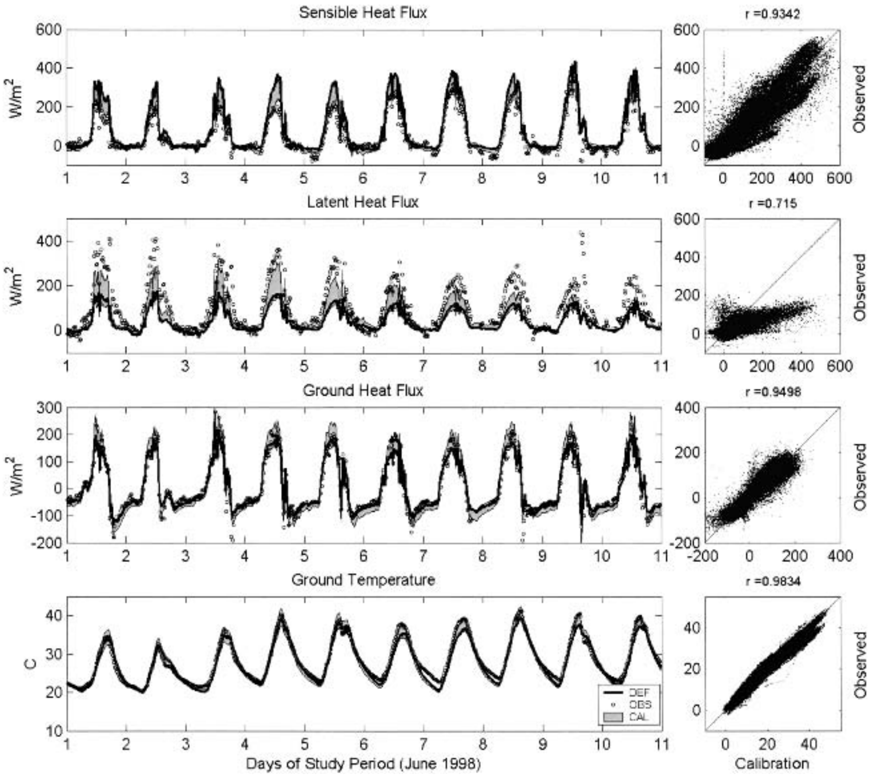
FIG. 11. Kendall site model simulations for an 11-day period during the dry season (Jun 1998). Gray shaded area: trade-off bounds (Pareto set) for 1997–2000 parameters (4-yr calibration); circles: observed data; and solid line: default parameters.

latent heat, although the daily flux values are very low, neither the narrow trade-off solution, nor the default, are capturing the peak flux during the height of the day. The situation is much different during the monsoon period displayed in Fig. 12. Sensible heat fluxes trend much lower during the day, while latent heat fluxes are higher and more variable. The trade-off solution still tracks SH with a fairly narrow range of uncertainty, and captures G and T_g well. Default parameters oversimulate SH during this period, but match G and T_g as well as the pareto set. However, there are obvious differences when evaluating LE. The uncertainty associated with the trade-off solution is much greater during this period. The range of solutions tracks better than the default parameters, but still slightly underestimates the peak flux. This is also evident in the scatterplots of the observed and simulated LE. Although the correlation is not extremely poor (r = 0.715), the trend of the model to undersimulate higher values is apparent. These results reinforce earlier statements, which is the inability of the model to capture latent heat fluxes during the monsoon period, and especially with parameter sets from longer-time-period optimizations.

6. Discussion and conclusions

The goal of this investigation was to rigorously evaluate the performance of the Noah land surface model in a semiarid region using an objective, systematic, calibration and evaluation procedure. We sought to answer several questions, including the following: 1) How well does a common land surface model, such as the NCEP Noah model, perform in semiarid regions? 2) Do parameters calibrated at a proxy site lead to reasonable simulations at other sites with similar climate and vegetation? and 3) How do parameters perform when tested under various climatic conditions that were not included as part of the calibration period? Few studies have addressed these issues with long-term datasets or in a semiarid region.

In general, the Noah model accurately reproduces the sensible heat, ground heat, and ground temperature observations. Soil moisture observations were not available for this study, and, hence, no conclusions can be directly stated regarding the tracking of this variable. Follow-up studies may be needed to address the issue of initialization of prognostic land states without observation data, and the potential impact on optimization of parameter values. However, latent heat flux values were available and it is observed that the model does not reproduce this variable as well, especially during the very significant monsoon period. The dramatic increases in LE during this period (and decrease in sensible heat) are problematic for the model, and hence, higher errors are observed. The latent heat flux in the



Sensible Heat Flux

Latent Heat Flux

Ground Heat Flux

Ground Temperature

FIG. 12. Same as in Fig. 11, but for Kendall site model simulations for an 11-day period during the monsoon season (Aug 1998).

Noah model is strongly dependent on the greenness fraction value obtained from long-term remote sensing data (Kurkowski et al. 2003). This fraction may be a major source of poor performance during the wet monsoon summers, because the model has no mechanism to adjust to abnormal or abrupt changes in vegetation response (Kurkowski et al. 2003). This season is critical in the southwest, because vegetation processes are fully active for most of the grasses and other plant species in the region, and the evaluation of the evaporative component in these basins (and land surface models) is crucial to the ongoing consumptive use studies in the region.

Parameter sets from the optimization procedure show improved performance over nearly all of the default simulations with model error reduced by as much as 20–40 W m^{-2} using the calibrated parameters. Climatic events such as El Niño and the seasonal monsoon period require parameters that are specifically calibrated during these events (or include similar events) for adequate representation of the latent heat fluxes. This finding has significance for the range of studies comparing the performance of land surface models. The elimination of parameter identification errors and a systematic evaluation procedure allows for a more realistic and fair intercomparison of model performance.

In this study, single-year calibrations have lower errors during that period than using parameters from a longer calibration period. Also, when applying shorter time period calibrations (1 yr) to longer time series (3 or 4 yr), performance does not decline significantly. This analysis supports that for long-term simulations, it may be acceptable to select a representative data period for application to a longer period. There is also a demonstrated need to include both a "wet" and a "dry" period in this calibration period. Using parameter sets from longer time-frame periods (2 or 4 yr) results in decreased performance during the nonstationary climatic events, especially for the more extreme monsoon periods (such as the 1999 monsoon period). To capture the variability of these climatic conditions, there is a need to include wet periods where latent heat is markedly pronounced to initiate all physical processes in the model and to improve the estimation of model parameters.

In general, it can be stated that the application of proxy site (transferred) parameters results in slightly larger errors, from 5 to 20 W m^{-2}, over a parameter set specific to the flux site. Results also indicate that this conclusion is specific to the validation time period, with model performance declining by as much as 40–60 W m^{-2} (sensible heat) when using proxy parameters over some time periods. In most cases, however, the proxy site parameters are still an improvement over default values at these sites. In the absence of calibration data, a proxy basin set of parameters can be applied with a moderate decline in performance. Although these results are specific to the Arizona sites and, in theory, can not be generalized to other pairs of proxy sites, we suspect that these results may hold for other case studies. As longer-term datasets become available, further proxy site studies will allow greater insight into this hypothesis. At the minimum, this analysis provides a first attempt to quantify the range of errors that can be expected when applying proxy site parameters to other similar climatic regions.

This is one of the first studies to address data length and quality on parameter estimates in a land surface model, and the transferability of those estimates in semiarid regions. The Noah model was selected as a surrogate in this study for other common land surface models and did have some simulation problems during the wet season. However, based on our analysis of model performance in semiarid and other regions (Hogue 2003), we expect other models to behave similarly under these conditions. Land surface models need to be able to simulate long-term change, but one would also hope they would also capture seasonal variability. An assessment of this ability was tested in this analysis. Our goal is that rigorous calibration and validation studies performed on models, such as the NCEP Noah model, will lead to improved parameter estimates for use in long-term climate and regional environmental studies. A companion paper is in progress, further analyzing the parameter behavior and sensitivity at these two sites, as well as at five other climatic regions.

Acknowledgments. The authors would like to thank the reviewers for their thoughtful comments that have contributed greatly to the improvement of this manuscript. Partial financial support for this research was provided by grants from the NASA EOS Graduate College Fellowship and SAHRA (Sustainability of Semi-Arid Hydrology and Riparian Areas) under the STC program of the National Science Foundation, Agreement EAR-9876800.

REFERENCES

Bastidas, L. A., 1998: Parameter estimation for hydrometeorological models using multi-criteria methods. Ph.D. dissertation, The University of Arizona, 204 pp.

——, H. V. Gupta, and S. Sorooshian, 2001: Bounding the parameters of land-surface schemes using observational data. *Land Surface Hydrology, Meteorology and Climate: Observations and Modeling,* V. Lakshmi, J. Albertson, and J. Schaake, Eds., *Water Science and Application,* Vol. 3, Amer. Geophys. Union, 65–76.

——, B. Nijssen, W. Emmerich, H. V. Gupta, and E. Small, 2004: Description of the PILPS 2g experiment: Model comparison over semi-arid areas. GLASS Science Panel of GEWEX Proposal, 19 pp.

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

Boone, A., F. Habets, and J. Noilhan, 2001: The Rhone-AGGregation Experiment. *GEWEX News,* Vol. 11, No. 3, International GEWEX Project Office, Silver Spring, MD, 3–5.

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

——, ——, ——, V. Koren, Z. Y. Zhang, and M. Smith, 2001: Toward improved streamflow forecasts: Value of semidistributed modeling. *Water Resour. Res.,* **37,** 2749–2759.

Buizer, J. L., J. Foster, and D. Lund, 2000: Global impacts and regional actions: Preparing for the 1997–98 El Niño climate and societal interactions. *Bull. Amer. Meteor. Soc.,* **81,** 2131–2139.

Chehbouni, A., and Coauthors, 2000: A preliminary synthesis of major scientific results during the SALSA program. *Agric. For. Meteor.,* **105,** 311–323.

Ek, M. B., K. E. Mitchell, Y. Lin, P. Grunmann, E. Rogers, G. Gayno, and V. Koren, 2003: Implementation of the upgraded Noah land-surface model in the NCEP operational mesoscale Eta model. *J. Geophys. Res.,* **108,** 8851, doi:10.1029/2002JD003296.

Emmerich, W. E., 2003: Carbon dioxide fluxes in a semiarid environment with high carbonate soils. *Agric. For. Meteor.,* **116,** 91–102.

Goodrich, D. C., and Coauthors, 2000: Preface paper to the Semi-Arid Land-Surface-Atmosphere (SALSA) Program special issue. *Agric. For. Meteor.,* **105,** 3–20.

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

——, L. A. Bastidas, S. Sorooshian, W. J. Shuttleworth, and Z. L. Yang, 1999: Parameter estimation of a land surface scheme using multi-criteria methods. *J. Geophys. Res.,* **104,** 19 491–19 504.

Gutman, G., and A. Ignatov, 1998: The derivation of the green vegetation fraction from NOAA/AVHRR data for use in numerical weather prediction models. *Int. J. Remote Sens.,* **19,** 1533–1543.

Henderson-Sellers, A., Z.-L. Yang, and R. E. Dickinson, 1993: The Project for Intercomparison of Land-surface Parameterization Schemes. *Bull. Amer. Meteor. Soc.,* **74,** 1335–1349.

——, 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.

Hogue, T. S., 2003: A multi-criteria evaluation of land-surface models and application to semi-arid regions. Ph.D. dissertation, The University of Arizona, 247 pp.

Kemp, P. R., 1983: Phenological patterns of Chihuahuan desert plants in relation to the timing of water availability. *J. Ecol.,* **71,** 427–436.

Klemes, V., 1986: Operational testing of hydrological simulation models. *Hydrol. Sci. J.,* **31,** 13–24.

Kurkowski, N. P., D. J. Stensrud, and M. E. Baldwin, 2003: Assessment of implementing satellite-derived land cover data in the Eta Model. *Wea. Forecasting,* **18,** 404–416.

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

Meixner, T., R. C. Bales, M. W. Williams, D. H. Campbell, and J. S. Baron, 2000: Stream chemistry modeling of two watersheds in the front range, Colorado. *Water Resour. Res.,* **36,** 77–87.

——, L. A. Bastidas, H. V. Gupta, and R. C. Bales, 2002: Multicriteria parameter estimation for models of stream chemical composition. *Water Resour. Res.,* **38,** 1027, doi:10.1029/2000WR000112.

Mitchell, K. E., and Coauthors, 2004: The multi-institution North American Land Data Assimilation System (NLDAS) project: Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *J. Geophys. Res.,* **109,** D07S90, doi:0.1029/2003JD003823.

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

Pitman, A. J., and Coauthors, 1999: Key results and implications from phase 1(c) of the Project for Intercomparison of Land-surface Parameterization Schemes. *Climate Dyn.,* **15,** 673–684.

Qi, J., and Coauthors, 2000: Spatial and temporal dynamics of vegetation in the San Pedro River basin area. *Agric. For. Meteor.,* **105,** 55–68.

Refsgaard, J. C., and J. Knudsen, 1996: Operational validation and intercomparison of different types of hydrological models. *Water Resour. Res.,* **32,** 2189–2202.

Rodriquez-Iturbe, I., A. Porporato, F. Laio, and L. Ridolfi, 2001: Plants in water-controlled ecosystems: Active role in hydrologic processes and response to water stress. *Adv. Water Resour.,* **24,** 695–705.

Schlosser, C. A., and Coauthors, 2000: Simulations of a boreal grassland hydrology at Valdai, Russia: PILPS phase 2(d). *Mon. Wea. Rev.,* **128,** 301–321.

Scott, R. L., 1999: Riparian and rangeland sol-vegetation-atmosphere interactions in southeastern Arizona. Ph.D. dissertation, The University of Arizona.

Sorooshian, S., R. Bales, H. Gupta, G. Woodard, and J. Washburne, 2002: A brief history and mission of SAHRA: A National Science Foundation Science and Technology Center on "Sustainability of semi-arid hydrology and riparian areas" (invited commentary). *Hydrol. Processes,* **16,** 3293–3295.

Western Region Climate Center, cited 2003: Historical climate information. [Available online at http://www.wrcc.dri.edu/CLIMATEDATA.html.]

Wood, E. F., X. Liang, D. Lohmann, and D. P. Lettenmaier, 1998: The Project for Intercomparison of Land-surface Parameterization Schemes (PILPS) phase-2(c) Red-Arkansas River Experiment: 1. Experiment description and summary intercomparisons. *J. Global Planet. Change,* **19,** 115–135.

Xia, Y., A. J. Pittman, H. V. Gupta, M. Leplastrier, A. Henderson-Sellers, and L. A. Bastidas, 2002: Calibrating a land surface model of varying complexity using multicriteria methods and the Cabauw dataset. *J. Hydrometeor.,* **3,** 181–194.

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