Evaluating Enhanced Hydrological Representations in Noah LSM over Transition Zones: Implications for Model Development

ENRIQUE ROSERO, ZONG-LIANG YANG, LINDSEY E. GULDEN, AND GUO-YUE NIU

Department of Geological Sciences, Jackson School of Geosciences, The University of Texas at Austin, Austin, Texas

DAVID J. GOCHIS

Research Applications Laboratory, National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 31 March 2008, in final form 30 December 2008)

ABSTRACT

The authors introduce and compare the performance of the unified Noah land surface model (LSM) and its augments with physically based, more conceptually realistic hydrologic parameterizations. Forty-five days of 30-min data collected over nine sites in transition zones are used to evaluate (i) their benchmark, the standard Noah LSM release 2.7 (STD); (ii) a version equipped with a short-term phenology module (DV); and (iii) one that couples a lumped, unconfined aquifer model to the model soil column (GW). Their model intercomparison, enhanced by multiobjective calibration and model sensitivity analysis, shows that, under the evaluation conditions, the current set of enhancements to Noah fails to yield significant improvement in the accuracy of simulated, high-frequency, warm-season turbulent fluxes, and near-surface states across these sites. Qualitatively, the versions of DV and GW implemented degrade model robustness, as defined by the sensitivity of model performance to uncertain parameters. Quantitatively, calibrated DV and GW show only slight improvement in the skill of the model over calibrated STD. Then, multiple model realizations are compared to explicitly account for parameter uncertainty. Model performance, robustness, and fitness are quantified for use across varied sites. The authors show that the least complex benchmark LSM (STD) remains as the most fit version of the model for broad application. Although GW typically performs best when simulating evaporative fraction (EF), 24-h change in soil wetness (ΔW_{30}), and soil wetness, it is only about half as robust as STD, which also performs relatively well for all three criteria. GW's superior performance results from bias correction, not from improved soil moisture dynamics. DV performs better than STD in simulating EF and ΔW_{30} at the wettest site, because DV tends to enhance transpiration and canopy evaporation at the expense of direct soil evaporation. This same model structure limits performance at the driest site, where STD performs best. This dichotomous performance suggests that the formulations that determine the partitioning of LE flux need to be modified for broader applicability. Thus, this work poses a caveat for simple ''plug and play'' of functional modules between LSMs and showcases the utility of rigorous testing during model development.

1. Introduction

By regulating the partitioning and horizontal distribution of water and energy fluxes, land surface processes and characteristics modulate local weather and climate (Viterbo 2002; Yang 2004). Land–atmosphere interactions are thought to be particularly strong in zones of transition between dry and wet climates, such as the U.S. southern Great Plains (Koster et al. 2004). To understand what processes are important in controlling surface-to-atmosphere fluxes and to better predict weather and climate, researchers use land surface models (LSMs) (Pitman 2003). LSMs are representations of the interactions between soil, vegetation, and the atmospheric boundary layer. LSMs also provide lower boundary fluxes of mass, energy, and momentum to weather forecasting and climate models (Nijseen and Bastidas 2005). Hence, realistic representation of key hydrological processes within LSMs is important for accurate numerical weather prediction.

Discerning which processes are essential to represent within LSMs is an ongoing effort within the research community. As our understanding of the land surface

Corresponding author address: Zong-Liang Yang, The University of Texas at Austin, 1 University Station C1100, Austin, TX 78712-0254.
E-mail: liang@mail.utexas.edu

DOI: 10.1175/2009JHM1029.1

process grows, LSMs are adapted. New parameterizations aim to improve on previous generations of models by including increasingly complex, previously neglected processes or by replacing old simplifications with newly proposed, conceptually more realistic approaches (e.g., Oleson et al. 2008).

Because they provide a source of hydrological memory, vegetation processes and anomalies in soil moisture are believed to influence precipitation and shape climate (e.g., Pielke 2001). Use of LSMs that include at least a rudimentary treatment of vegetation and soil processes tends to improve model simulations. Correct simulation of the initiation of convection depends on modeled soil temperature and moisture (Childs et al. 2006; Weckwerth and Parsons 2006). As a result, improved soil moisture representation within LSMs improves simulation of surface fluxes (Dirmeyer et al. 2000), and the use of more realistic representation of vegetation states and processes (e.g., stomatal resistance) increases the predictive power of LSMs in both offline (Niyogi and Raman 1997) and coupled simulations (Holt et al. 2006).

Further refinement of the conceptual realism of LSM soil hydrology and vegetation processes may further improve model predictive capability. When compared to more simplistic parameterizations, more complex and sophisticated LSMs have been credited with improved simulations of air temperature, runoff, snow, turbulent fluxes, and soil states (Boone et al. 2004; Bowling et al. 2003; Niu et al. 2005, 2007; Wood et al. 1998). However, other studies have demonstrated that additional complexity neither necessarily improves model performance nor reduces the uncertainty in the simulated fluxes of water and energy (Schultz and Beven 2003; Hogue et al. 2006). Additional complexity in LSM representations is perhaps unjustified when the new parameterization cannot be supported or identified with available observations (Leplastrier et al. 2002; Schultz and Beven 2003; Hogue et al. 2006).

Keeping in mind that both too parsimonious and too complex models often lead to decreased skill (e.g., Jensen 1998; Carlson and Doyle 2002), we evaluate the augmentation of the latest version of the Noah LSM (Ek et al. 2003) with two more conceptually realistic parameterizations: groundwater processes and dynamic phenology. We ask whether the new modules improve the model's capacity to simulate high-frequency turbulent fluxes and soil states and how reliable each model is when faced with parameter uncertainty. Because of the strength of the coupling, our work focuses on warm-season climates in the transition zone of the central United States.

Our primary goal is to identify whether the recent enhancements to the Noah model offer improvements in skill or robustness in simulating high-frequency fluxes and soil states, which, for this paper, we will term ''applications.'' Although LSM development enables incorporating necessary degrees of freedom to research the nature of feedbacks (e.g., the role of groundwater in long-term memory), investigate trends (e.g., phenology contrast between wet and dry years), test scenarios (e.g., carbon cycling), and so on (e.g., Dirmeyer et al. 2006; Kim and Wang 2007; Lyon et al. 2008), in our applications-focused framework, we confine our definition of a ''better'' model to one that most accurately reproduces observed high-frequency states and fluxes at the local scale.

The analysis we present here is more rigorous than the typical LSM evaluation exercise. We first evaluate the versions of Noah LSM, following the steps of a traditional model intercomparison using single-model realizations (default and calibrated runs). We then use multiple model realizations and the metrics introduced by Gulden et al. (2008) to assess model performance and reliability in conditions that more closely resemble those in which LSMs are actually applied. Our goal is to understand how and why the new parameterizations change model performance. For both segments of our evaluation, we use 45 days of high-frequency near-surface states and heat fluxes data collected as part of the International H_2O Project (IHOP_2002) (LeMone et al. 2007).

Datasets, models, and methods are described in section 2. Experimental design and methods for model performance evaluation are explained in section 3. Section 4 presents a detailed, traditional model intercomparison and sensitivity analysis. Section 5 presents an assessment of model performance under uncertainty and focuses on hypothesis testing. Section 6 discusses the implications of the results for model evaluation and development. Conclusions are summarized in section 7.

2. Models, data, and methods

a. Hydrological enhancements to Noah LSM

To alleviate known biases [e.g., dry biases in evapotranspiration and soil moisture during the warm season (e.g., Chen et al. 2007) and poor energy partitioning even after calibration (Hogue et al. 2006)], the Noah LSM (Ek et al. 2003; Mitchell et al. 2004) was augmented with modules that improve the conceptual realism of land surface processes. We compare our benchmark, the standard Noah LSM release 2.7 (Noah-STD) to (i) a version that we equipped with a short-term phenology module (Noah-DV) and (ii) one that couples a lumped, unconfined aquifer model to the model soil column (Noah-GW).

1) NOAH-DV

We added the physically based vegetation module of Dickinson et al. (1998) to Noah-STD to dynamically calculate vegetation greenness fraction. Unlike Noah-STD, which computes greenness fraction by linear interpolation between monthly climatological values, Noah-DV represents short-term phenological variation by allowing leaf biomass density to respond to environmental perturbations and to vary as a function of soil moisture, soil temperature, canopy temperature, and vegetation type. The module allocates carbon assimilated during photosynthesis to leaves, roots, and stems; the fraction of photosynthate allocated to each reservoir is a function of, among other things, the existing biomass density. The model also tracks growth and maintenance respiration and represents carbon storage. Following a modification by Yang and Niu (2003), DV explicitly makes vegetation fraction (vegfrac) an exponential function of leaf area index (LAI). STD allows LAI to only influence the computation of stomatal resistance (Rs). In addition to that, DV makes direct soil evaporation, canopy evaporation, and transpiration depend on variations in leafiness, or, more precisely, LAI.

2) NOAH-GW

Noah-GW couples a lumped unconfined aquifer model (Niu et al. 2007) to the lower boundary of the Noah-STD soil column. Water flows in both directions between the aquifer and the soil column. The modeled hydraulic potential is the sum of the soil matric and gravitational potentials. If insufficient water is available to maintain a near-surface aquifer, the water table falls below the soil column; when water is plentiful, the water table is within the soil column of the LSM. Baseflow is parameterized using an index of topography (Niu et al. 2005).

b. IHOP_2002 sites and datasets

We used data from the IHOP_2002 field campaign (Weckwerth et al. 2004) to evaluate predictions from the different versions of Noah LSM at nine sites. To enable definitive testing and development of LSMs in transition zones, IHOP_2002 collected 45 days of high-temporal-resolution, multisensor measurements of meteorological forcing, surface-to-atmosphere flux data, and near-surface measurements of soil moisture and temperature along the Kansas–Oklahoma border and in northern Texas. The interested reader is referred to LeMone et al. (2007) for details. (The authors obtained the datasets at http://www.rap.ucar.edu/research/land/observations/ihop/.)

Table 1 presents the Noah LSM soil and vegetation classes and mean meteorological values for the observation period. The nine stations were sited to obtain a representative sample of the region, which spans a strong east–west rainfall gradient.

Figure 1 shows evaporative fraction (EF) and 30-cm soil wetness (W_{30}) for sites 2 (Fig. 1a) and 8 (Fig. 1b) against the backdrop of precipitation and volumetric soil moisture (SMC) in three of the soil layers. With depth, the soil column dries at site 2 (dry) and wets at site 8 (wet). Evaporation at site 2 tends to be moisture limited; evaporation at site 8 is most often energy limited. Comparing EF at site 2 to that at site 8, we see that it peaks immediately after rainfall at site 2 but somewhat subsides immediately following precipitation at site 8; the EF does not peak until several days after the influx of rainwater to the soil.

c. Model initialization and spinup

All runs described in this paper followed the same initialization and spin-up procedures. We used downscaled North American Land Data Assimilation System (NLDAS) (Cosgrove et al. 2003) meteorological forcing, interpolated from a 60-min to a 30-min time step, to drive the simulations between 1 January 2000 and 13 May 2002. Following Rodell et al. (2005), we initialized each of the four soil layers at 50% saturation and at the multiannual mean temperature. For Noah-GW, the depth to the water table was initialized assuming equilibrium of gravitational and capillary forces in the soil profile (Niu et al. 2007). The models were subsequently driven by IHOP_2002 meteorological forcing (see Table 1) from 13 May to 25 June 2002 [day of year (DOY) 130–176].

d. Calibration datasets

To constrain and evaluate the models during the IHOP_2002 period, we used a 30-min time step and observed sensible heat flux (H), latent heat flux (LE), ground heat flux (G), ground temperature (Tg), and first layer soil moisture (SMC_{5cm}). To score the performance, we used root-mean-square error (RMSE). We scored only the last 45 days of each 2.5-yr-long model simulation, DOY 130–176.

e. Parameters calibrated

We selected 10 soil and 10 vegetation parameters that have been deemed sensitive at similar locations (Demarty et al. 2004; Bastidas et al. 2006). We included eight parameters responsible for the phenology module and four that control the aquifer model to estimate a total of 28 and 24 parameters for DV and GW, respectively. All other coefficients in the models were kept constant at the recommended values. Defaults and

TABLE 1. IHOP_2002 sites and mean meteorological forcing observed during the evaluation period (13 May–25 Jun). Noah LSM vegetation and soil types (indices in parenthesis). Rainfall is cumulative over the observation period. Mean annual precipitation (MAP), shortwave (SW) and longwave (LW) radiation, 2-m air temperature (T), surface pressure (P), specific humidity (Q2), and wind speed (U).

Site	1	2	3	4	5	6	7	8	9
Lat (°N)	36.4728	36.6221	36.8610	37.3579	37.3781	37.3545	37.3132	37.4070	37.4103
Lon (°W)	100.6179	100.6270	100.5945	98.2447	98.1636	97.6533	96.9387	96.7656	96.5671
Vegetation type	bare ground (1)	grassland (7)	sagebrush (9)	pasture (7)	wheat (12)	wheat (12)	pasture (7)	grassland (7)	pasture (7)
Soil type	sandy clay loam (7)	sandy clay loam (7)	sandy loam (4)	loam (8)	loam (8)	clay loam (6)	silty clay loam (2)	silty clay loam (2)	silty clay loam (2)
Rain (mm)	154.5	69.1	72.4	164.5	173.6	203.6	175.4	296.6	250.8
MAP (mm)	530	540	560	740	750	800	900	880	900
SW (W m^{-2})	293.8	296.7	296.9	272.6	270.3	269.8	268.9	261.8	261.8
LW (W m^{-2})	348.3	351.8	360.6	358.1	357.9	367.5	368.5	359.3	358.3
T (°C)	21.4	21.7	22.5	20.7	20.7	21.0	20.7	20.1	19.9
P (h Pa)	914.6	915.9	924.1	955.4	955.9	966.2	970.5	965.2	963.4
Q2 (g kg^{-1})	10.3	9.9	9.8	11.2	11.9	11.7	11.9	12.1	11.9
U (m s^{-1})	7.8	7.8	6.6	6.3	5.9	5.6	5.3	5.3	5.9

feasible ranges (Table 2) for all the parameters were taken from the literature (e.g., Hogue et al. 2006).

f. Multiobjective parameter estimation technique

To calibrate the models, we used the Markov chain Monte Carlo sampling strategy of Vrugt et al. (2003). The calibration algorithm allows an initial population of parameter sets (randomly selected within preestablished, feasible ranges) to evolve until the population converges to a stable sample, which maximizes the likelihood function and fairly approximates the Pareto set (PS). The Pareto set represents the multiobjective trade-off: no member of the PS can perform better with respect to one objective without simultaneously performing worse with respect to another, competing objective (Gupta et al. 1998). The simultaneous minimization of the RMSE of multiple criteria (H, LE, G, Tg, and SMC_{5cm}) allows us to constrain the model for consistency with several types of observations. Multiobjective optimization facilitates the identification of physically meaningful parameter sets (and their underlying posterior distribution) that cause the model to mimic the processes they were designed to represent (Gupta et al. 1999; Bastidas et al. 2001; Leplastrier et al. 2002; Xia et al. 2002; Hogue et al. 2006). We used a sample of 150 parameter sets to represent the PS.

To obtain a detailed representation of the range of model performance (i.e., the objective function space), we also ran a Monte Carlo sampling of 15 000 random parameter sets, uniform within the feasible bounds (Table 2). Figure 2 shows slices of STD's objective function space at site 4. In frequentist terms, Fig. 2 suggests that, when very little is known about the parameters, the expected RMSE of STD at site 4 is most probably ~55 W m^{-2} for LE, ~3°C for Tg, and ~5% for SMC_{5cm}. Note the difference between the location of the scores most frequently (MF) obtained and the location of the low-density region where the PS resides.

3. Experimental design

We aimed to identify the model that best reproduces the physical behavior of transition zone point-scale heat fluxes and states during the warm season.

a. Traditional model intercomparison

We first compared the versions of Noah LSM using single-model realizations. To evaluate the hypothesis that increased physical realism yields an LSM that better reproduces observations, we asked, Do conceptually realistic enhancements improve the ability of LSMs to simulate fluxes and near-surface states? We compared the performance of default and multiobjectively calibrated runs using the goodness-of-fit metrics of appendix A and observations of H, LE, G, Tg, and SMC_{5cm}. In situ, high-frequency measurements are an integrated response of the land surface and therefore provide multiple data streams that we used to examine model soundness at specific locations (Bastidas et al. 2001; Stöckli et al. 2008). It is important to note that no estimates of observational uncertainty or errors in energy balance closure in the tower flux data were incorporated into the present analysis. We used the multicriteria optimization as an objective test of the underlying hypothesis that models are able to concurrently simulate all the response modes that they were designed to represent. Additionally, we compared characteristic model behaviors (obtained from extensive Monte Carlo sampling of parameter space) as a proxy for robustness. Results are presented in section 4.

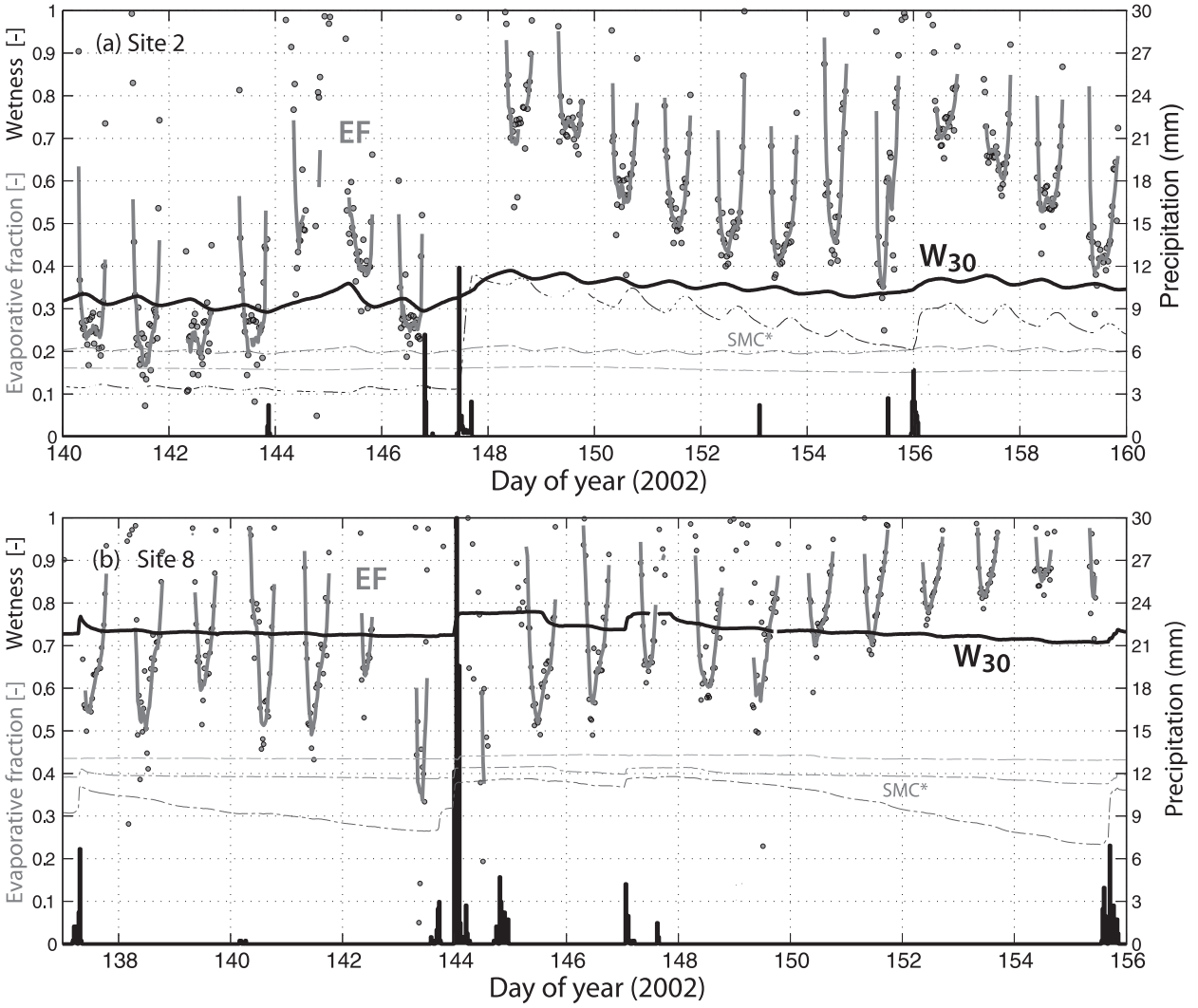

FIG. 1. Segment of the time series of EF, W_{30}, SMC, and precipitation at (a) site 2 and (b) site 8. EF is shown in two ways: 30-min data points and 3-h smoothed data (gray). EF peaks and depletes immediately after rainfall at site 2 but does not peak until several days after precipitation at site 8. Here, W_{30} shows 30%–40% at site 2 and 70%–80% at site 8. SMC measurements at 5, 15, and 60 cm below the surface are reported using gray lines: the darkest line is the SMC in the layer nearest to the surface and the lightest gray line is the soil moisture in the layer farthest from the surface.

b. Ensemble-based model intercomparison

We evaluated the hypothesis that increased physical realism in conceptual models not only improves their performance but also enhances their robustness, making them less sensitive to errant parameter values (Gulden et al. 2007a).We asked, Which version of Noah is best suited for broad application and why?

To objectively identify the model that best reproduces observations from among STD, DV, and GW, we explicitly considered uncertainty and rigorously evaluated different realizations of a model in an ensemble framework. To capture representative model behaviors (Smith 2002; Wagener and Gupta 2005), we used parameter variation to create two ensembles that we used to evaluate each model. Three metrics were used: the model performance score (ζ; quantifies skill and spread of the ensemble), the model robustness score (ρ; quantifies insensitivity to poorly known parameters), and the model fitness score (φ; enables ranking models based on suitability for broad application) (Gulden et al. 2008; equations are presented in appendix B). We used this method because it enabled us to identify shortcomings in the formulation of LSMs that hinder their capacity to simulate surface exchanges and states, even with optimized parameters. We also evaluated the hypothesis

TABLE 2. Feasible ranges of calibrated Noah LSM parameters.

Parameter	Description	Units	Min	Max
	Soil parameters			
maxsmc	Maximum volumetric soil moisture	$m^3\ m^{-3}$	0.35	0.55
psisat	Saturated soil matric potential	$m\ m^{-1}$	0.1	0.65
satdk	Saturated soil hydraulic conductivity	$m\ s^{-1}$	$\times 10^{-6}$	$\times 10^{-5}$
b	Clapp–Hornberger *b* parameter	—	4	10
quartz	Quartz content	—	0.1	0.82
refdk	Used with refkdt to compute runoff parameter kdt		0.05	3
fxexp	Bare soil evaporation exponent	—	0.2	4
refkdt	Surface runoff parameter		0.1	10
czil	Zilintikevich parameter	—	0.05	8
csoil	Soil heat capacity	$J\ m^{-3}\ K^{-1}$	1.26	3.5
	Vegetation parameters			
rcmin	Minimal stomatal resistance	$s\ m^{-1}$	40	400
rgl	Radiation stress parameter used in F1 term of canopy resistance		30	100
hs	Coefficient of vapor pressure deficit term F2 in canopy resistance		36	47
*z*0	Roughness length	m	0.01	0.1
lai	Leaf area index	—	0.1	5
cfactr	Exponent in canopy water evaporation function	—	0.4	0.95
cmcmax	Maximum canopy water capacity used in canopy evaporation	m	0.1	2.0
sbeta	Used to compute canopy effect on ground heat flux	—	−4	−1
rsmax	Maximum stomatal resistance	$s\ m^{-1}$	2 000	10 000
topt	Optimim air temperature for transpiration	K	293	303
	Dynamic phenology parameters (Noah-DV)			
fragr	Fraction of carbon into growth respiration	—	0.1	0.5
gl	Conversion between greenness fraction and LAI	—	0.1	1.0
rssoil	Soil respiration coefficient	$s^{-1} \times 10^{-6}$	0.005	0.5
tauhf	Average inverse optical depth for 1/*e* decay of light	—	0.1	0.4
bf	Parameter for present wood allocation		0.4	1.3
wstrc	Water stress parameter		10	400
xlaimin	Minimum leaf area index	—	0.05	0.5
sla	Specific leaf area	—	5	70
	Groundwater parameters (Noah-GW)			
rous	Specific yield	$m^3\ m^{-3}$	0.01	0.5
fff	*e*-folding depth of saturated hydraulic capacity	m^{-1}	0.5	10
fsatmx	Maximum saturated fraction	%	0	90
rsbmx	Maximum rate of subsurface runoff	$m\ s^{-1} \times 10^{-3}$	0.01	1

that increased physical realism in conceptual models not only improves model performance but also enhances model robustness, making them less sensitive to errant parameter values. Results are presented in section 5.

1) GENERATION OF ENSEMBLES

For each model and each of the nine IHOP_2002 sites, we generated two 150-member parameter-based ensembles: (i) a MF performing uncalibrated ensemble and (ii) a calibrated (PS) ensemble. The calibrated ensembles were drawn from the PS, which tends to provide consistent and reliable model realizations (Boyle et al. 2000). The MF ensembles were composed of 150 randomly sampled models whose RMSE was within the intersection of the spaces defined by one standard deviation around the mode of each of the five calibration objectives (H, LE, G, Tg, and SMC_{5cm}) (Fig. 2). The PS and MF ensembles characterize distinct modes of behavior and represent a signature of the LSM in the objective function space (Gupta et al. 2008). We confirmed that the parameter sets of the PS and MF samples come from distinct distributions (results not shown).

2) EVALUATION CRITERIA

For model evaluation, we use three independent verification criteria: EF, W_{30}, and change in wetness over 24 h (ΔW_{30}).

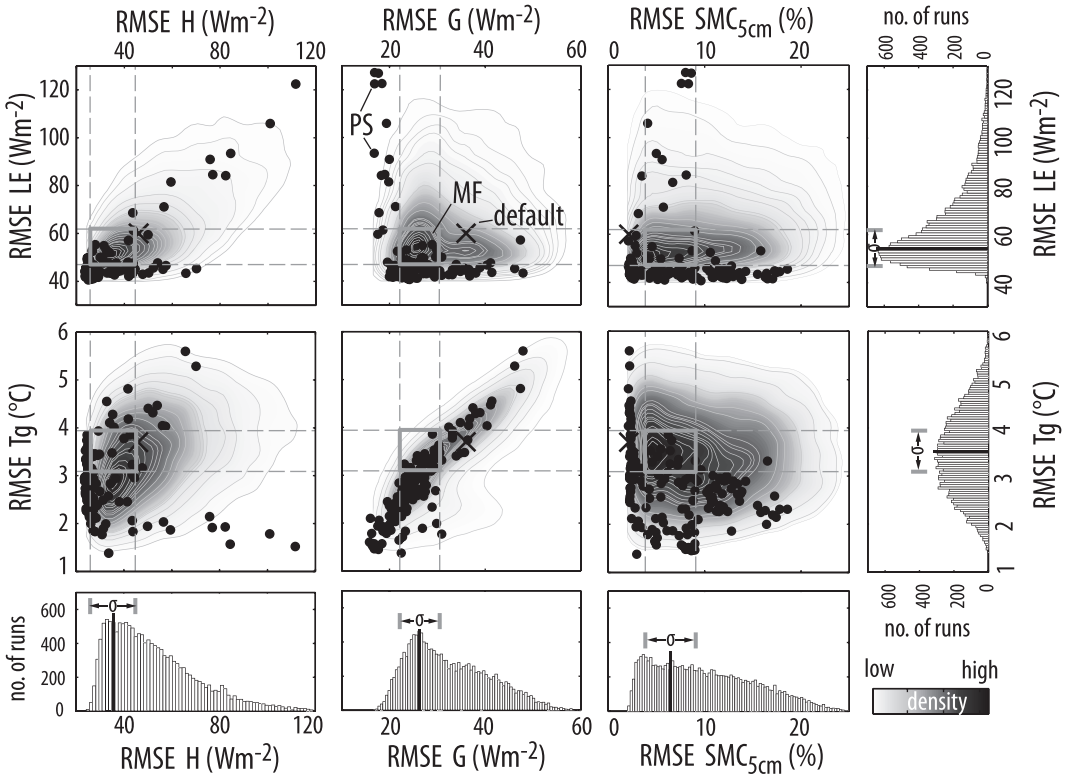

FIG. 2. Bidimensional projections of the objective function space of STD at site 4. Higher density of RMSE scores of 15 000 Monte Carlo model runs shown with darker contours. The PS, 150 calibrated parameter sets (black dots), represent the minimal uncertainty in the multiobjective trade-off (H, LE, G, Tg, and SMC_{5cm}). The MF performing models have RMSEs within the intersection of one standard deviation (σ) around the mode of each objective. Note that the relative position of default ($\times$) is no indication of the goodness of model.

4. Results of traditional model intercomparison

The traditional evaluation of model development compares the performance of a new model against a baseline model, while often neglecting parameter uncertainty. Model intercomparisons are often incomplete because they are based on ''ad hoc manual–expert model evaluation'' methods that are inadequate for highly complex models (Gupta et al. 2008). By applying customary evaluation methods to assess the potential improvement of the LSMs in simulating H, LE, G, Tg and SMC_{5cm}, we draw conclusions regarding model performance, review the strengths and limitations of typical model development procedures, and demonstrate the need for a more complete approach to thoroughly compare the models described earlier.

a. Comparison of default and calibrated runs

To illustrate the concepts of full and partial calibration, model performance (before and after augmentation with DV and GW) is presented in Figs. 3, 4. We tested the implementation of DV with the default parameter values suggested by its developers. Figure 3 shows that default STD overestimates LE flux at site 7 (wet). Because the recommended default parameters may not adequately characterize the particular conditions of the site, the new module's parameters are adjusted to better capture the desired behavior (e.g., Niu et al. 2005). The practice of adding modules and tuning only new parameters [i.e., partial tuning (xDV)] may improve model performance, yielding reduced bias (Fig. 3c), better correlation, and lower error (Fig. 3d). The improved performance may or may not be (but certainly could be) attributed to the superior nature of the new model.

The model may not achieve the desired level of improvement after partial calibration. In standard model development practice, the new model is frequently not rejected but is revised. Because of conflicting hypotheses or undesired interactions, the parameters of the host model may need to be adjusted to accommodate the new module (e.g., Gulden et al. 2007b). This is represented for SMC_{5cm} in Fig. 4. Default GW results in too-wet simulations and adjusting only its four free parameters (i.e., xGW) fails to significantly correct this bias. When the parameters of both the host model and the new module are simultaneously tuned (calibrated GW),

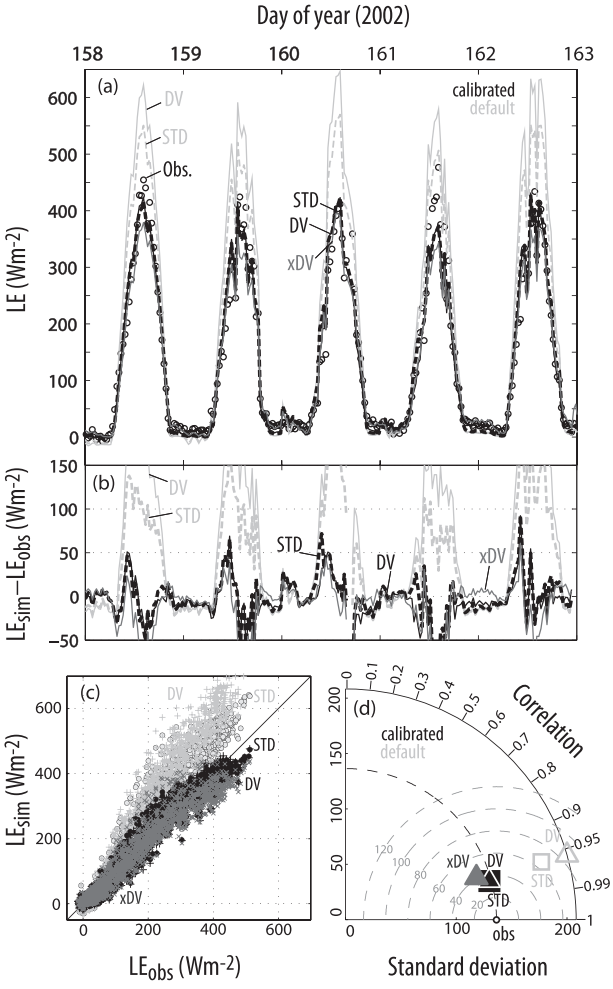

FIG. 3. Performance of Noah LSM augmented with DV in simulating LE at site 7. Figure shows (a) a segment of the time series of LE and (b) its residuals; (c) scatterplot of simulation vs observations; and (d) Taylor plot, where dark is a single-objective calibrated run and gray is the uncalibrated (default) run. Partially calibrated (xDV) stands for the tuning of the free parameters of the DV augmentation only (refer to Table 2), while the rest of the STD parameters are left fixed to its corresponding default values. (c),(d) For the entire evaluation period. Refer to Table 3 for statistics.

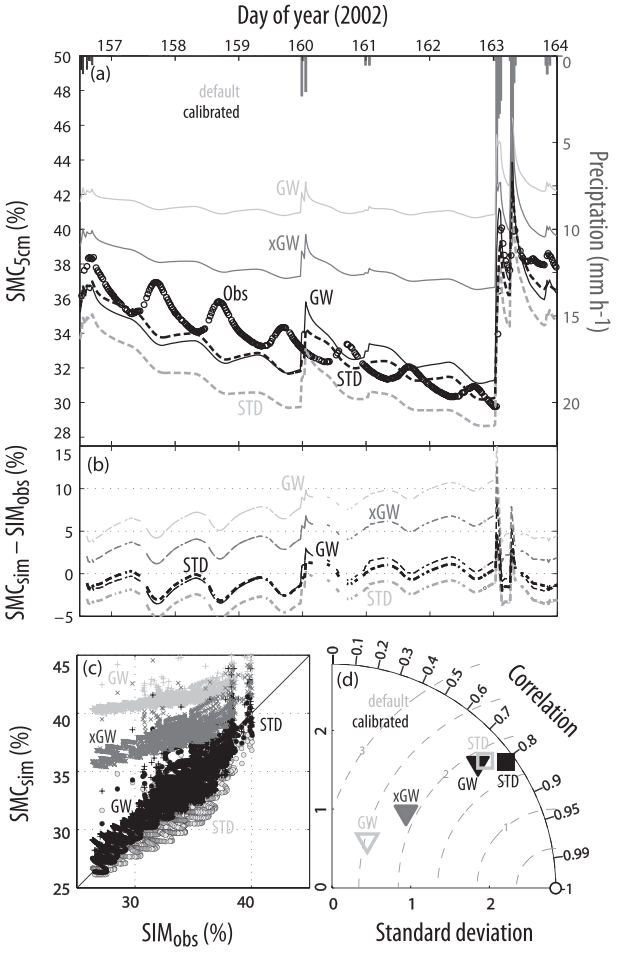

FIG. 4. As in Fig. 3, but for augmented with GW in simulating SMC_{5cm}. Refer to Table 2 for the free parameters of the GW augmentation.

the model performs at its best and surpasses the baseline established by the uncalibrated STD.

However, if we allow the calibration of the free parameters of the new models, for a fair, more consistent comparison, STD should be given the same opportunity to reach its optimal performance. For each objective, the best achievable performance of calibrated STD is also depicted in Figs. 3 and 4. Performance metrics and statistics are presented in Table 3 (see appendix A for definitions). The goodness-of-fit of calibrated STD is very similar to the best performance achieved by calibrated GW and DV. Distinguishing the models becomes nontrivial, and it is practically impossible to state which model is best based solely on these results.

To circumvent this issue, Akaike (1974) and Schwarz (1978) proposed information criteria [Akaike information criteria (AIC) and Bayesian information criteria (BIC), respectively] for model selection. They aimed to reward the model that better explains the data with the lower complexity (number of parameters). The order of preference given by the two information criteria favors STD over DV and GW (Table 3), implying that the gain in performance, if any, does not justify the additional complexity.

We do not argue that the aforementioned, generalized approach to validation within model development is fundamentally flawed, only that it is incomplete. To underscore that this indistinguishability between acceptable models (Beven and Freer 2001; Beven 2006) is not the outcome of chance nor that it is the sole consequence of demanding too little from the complex,

TABLE 3. Performance metrics and statistics for default and (fully and partially) calibrated models against LE and SMC_{5cm} at site 7 for the entire evaluation period. Models are denoted as STD, DV, and GW. Partial calibration (xDV and xGW) refers to tuning only new free parameters, while leaving all other STD parameters constant at default values. Calibrated STD is as good as calibrated DV and GW. AIC and BIC favor STD's lower complexity. Refer to appendix A for metrics definitions.

	Criterion	LE (W m^{-2})			SMC_{5cm} (%)		
		mean = 126.36; std = 136.36			mean = 33.19; std = 2.84		
Metric	Model	STD	DV	xDV	STD	GW	xGW
Mean	Default	147.14	163.24		31.52	41.29	
	Calibrated	115.38	112.82	112.01	33.18	33.07	38.27
Std dev	Default	184.39	208.34		2.53	0.76	
	Calibrated	134.35	134.53	124.57	2.72	2.39	1.33
RMSE	Default	69.01	97.18		2.22	8.46	
	Calibrated	24.27	24.66	33.46	1.26	1.48	5.48
r^2	Default	0.92	0.92		0.59	0.40	
	Calibrated	0.93	0.93	0.90	0.65	0.60	0.49
Bias	Default	31.80	49.31		−1.64	8.12	
	Calibrated	−3.55	−6.12	−6.78	0.03	−0.08	5.11
NSE	Default	0.74	0.49		0.39	−7.86	
	Calibrated	0.97	0.97	0.94	0.80	0.73	−2.72
Rank ΔAIC		1	2		1	2	
Rank ΔBIC		1	2		1	2	

multioutput models, at each site we calibrate the models simultaneously against the five objectives (H, LE, G, Tg, SMC_{5cm}). For simplicity, we selected for each calibrated model a single ''best'' set of parameters from among the PS [using minimum Euclidean norm of the vector composed by the RMSEs of the five objectives, e.g., Hogue et al. (2005)]. With this preferred compromise solution, we mimicked the common practice of using a single (''best'') parameter set during model validation.

At each location, the scores of the fully calibrated STD, GW, and DV are equivalent (Fig. 5). All calibrated models have consistently lower misfit and better correlation with observed turbulent fluxes and Tg in the wet locations. Model performance worsens as the location gets drier, and simulated SMC_{5cm} is less variable than observed. At the drier locations, scores differ slightly, particularly between DV and the rest of the models. Table 4 reports, for each site, the statistics of simulated LE by the best set for each model. Although there is some slight variation in the scores, model performance is essentially indistinguishable. Calibrated DV ranks best in terms of Nash–Sutcliffe efficiency (NSE) at four of the nine sites, calibrated GW at four sites, and calibrated STD at three sites. Note that after calibration at three of the sites (sites 4, 7, and 9), two models tie for best performance, scoring the same NSE. The maximum difference between NSE scores is 0.06 [site 1 (dry)], but most often the difference between the calibrated models' NSE scores is 0.01.

The rank of the model depends, in part, on choice of the objective (Table 4). Improvement in one evaluation metric tends to result in degradation in another (e.g., at site 3, GW has a slightly better NSE and coefficient of determination r^2 than STD and DV; however, GW has the worst bias of the three models). Good performance at one site does not guarantee reliable performance at climatologically similar sites. For instance, calibrated GW is unbiased (bias = 0.24 W m^{-2}) and has an excellent NSE (0.97) at site 7 (wet), but it is the most biased performer at site 9 (bias = –13.8 W m^{-2}) despite having the same high NSE (0.92) and r^2 (0.90) as STD.

Traditional model intercomparisons ignore the aforementioned caveats. They proceed to subjectively select models based on dependable functioning as judged by an expert (e.g., STD, GW); distinguishing solutions that fulfill predetermined criteria, such as the smallest possible RMSE with zero bias (Boyle et al. 2000); rejecting models that consistently underperform in the considered criteria (e.g., xGW, xDV); or rejecting the models whose optimal parameter values do not conform with a priori expectations, given any attributed physical meaning. Noting that a single solution was selected from among a population of realistic behavioral parameters (PS), the rankings (e.g., Table 4) are likely to change when different parameter sets are considered.

b. Comparison using multiple model realizations

1) SENSITIVITY OF GW TO MODEL PARAMETERS

GW exhibits decreased robustness at dry sites and almost the same frequency of errors as STD at wet sites. Cumulative distribution functions (CDFs) of 15 000 RMSE scores obtained by STD, GW, and xGW are shown in Fig. 6. At site 1 (dry) (Fig. 6a), 75% of the STD

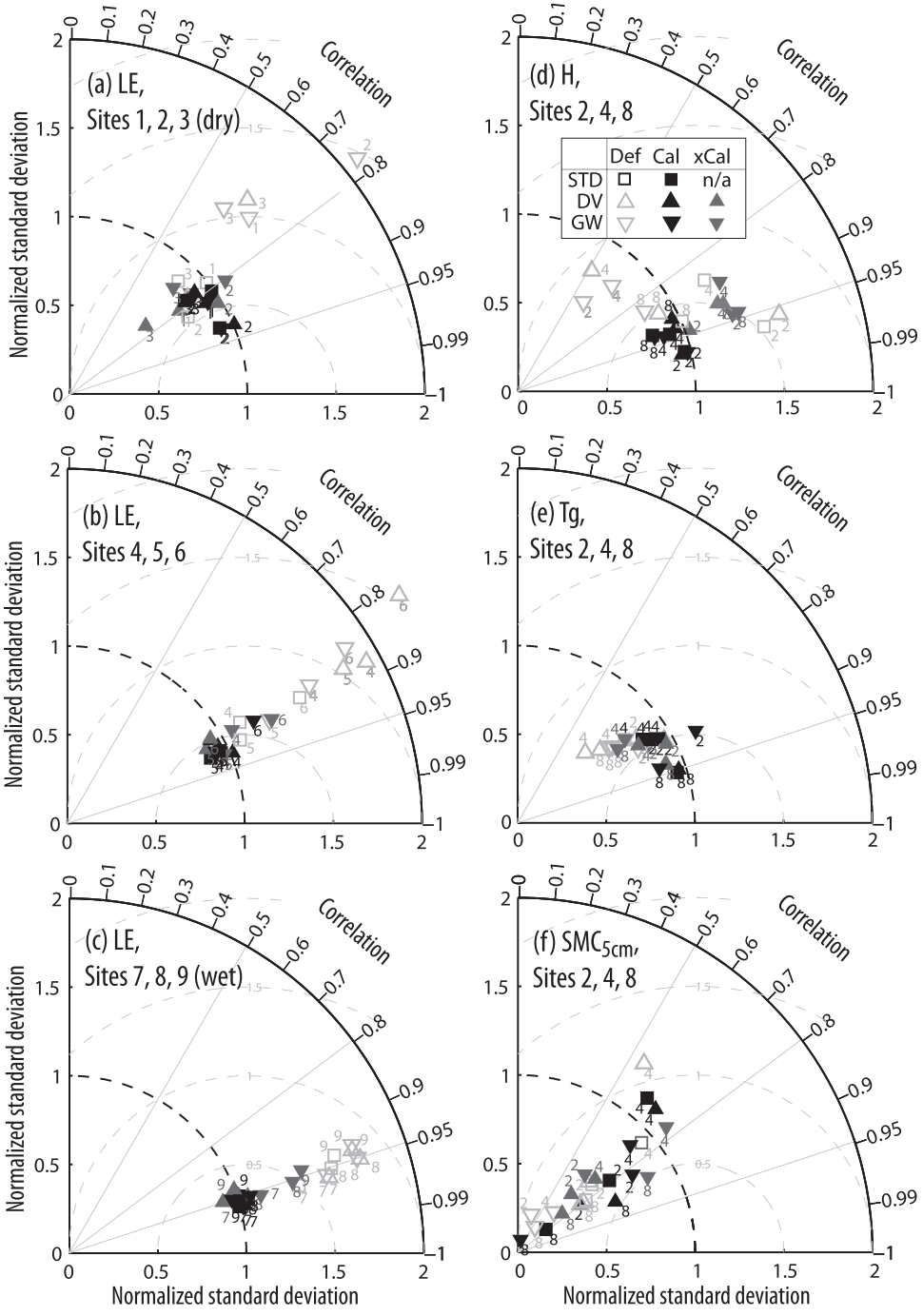

FIG. 5. Taylor diagrams of performance metrics for the entire evaluation period for (a)–(c) LE for all sites and (d) H, (e) Tg, and (h) SMC_{5cm} for sites 2, 4, and 8. Default STD, DV, and GW shown in light gray. Fully calibrated (black) and partially calibrated (dark gray) models (i.e., xDV and xGW) use a compromise "best" solution: preferred parameter set minimizes the L2 norm of the RMSE of the five objectives (H, LE, G, Tg, SMC_{5cm}). Calibrated models cluster together for any given site. Refer to Table 4 for statistics on simulated LE.

TABLE 4. Goodness-of-fit for the simulation of LE for default (def), partial, and fully (cal) calibrated models. Calibrations report only compromise solution: preferred "best" parameter set minimizes the L2 norm of the RMSE of the five objectives. Best performing model by site in bold. Number of sites a model performs the best is denoted as No. Refer to appendix A for definitions of metrics.

Metric	Model		IHOP_2002 site									No.
RMSE	STD	def	49.46	56.08	55.36	62.27	49.81	86.78	69.01	79.78	95.09	0
		cal	44.77	32.58	47.89	41.36	46.97	**42.05**	**25.50**	32.52	33.86	2
	DV	def	43.99	62.42	88.93	153.6	131.5	189.2	97.18	102.3	108.8	0
		xDV	43.99	42.68	57.16	42.99	51.48	48.98	33.46	31.08	36.30	0
		cal	**40.56**	**30.90**	48.22	**39.18**	49.14	48.39	26.15	**29.03**	34.21	4
	GW	def	87.49	157.8	90.66	113.3	69.35	138.0	98.61	102.05	112.7	0
		xGW	47.41	54.38	54.29	56.73	**43.93**	62.75	38.12	51.79	66.32	1
		cal	41.18	31.71	**47.72**	40.13	46.6	58.78	**25.09**	33.48	**33.61**	3
NSE	STD	def	0.56	0.54	0.46	0.70	0.82	0.29	0.74	0.51	0.34	0
		cal	0.64	0.84	0.59	0.87	0.84	**0.83**	**0.97**	0.92	**0.92**	3
	DV	def	0.65	0.43	−0.40	−0.80	−0.27	−2.37	0.49	0.20	0.13	0
		xDV	0.65	0.73	0.42	0.86	0.81	0.77	0.94	0.93	0.90	0
		cal	**0.70**	**0.86**	0.59	**0.88**	0.82	0.78	0.96	**0.94**	0.91	4
	GW	def	−0.39	−2.67	−0.46	0.02	0.65	−0.80	0.48	0.21	0.07	0
		xGW	0.59	0.56	0.48	0.75	**0.86**	0.63	0.92	0.80	0.68	1
		cal	0.69	0.85	**0.60**	**0.88**	0.84	0.67	**0.97**	0.91	**0.92**	4
r^2	STD	def	0.60	0.70	0.48	0.74	0.81	0.77	0.92	0.91	0.88	0
		cal	0.65	0.84	0.60	0.83	0.83	**0.81**	**0.93**	**0.92**	**0.90**	4
	DV	def	0.63	0.60	0.46	0.77	0.76	0.68	0.92	0.91	0.88	0
		xDV	0.63	0.73	0.55	0.82	0.78	0.74	0.90	0.91	0.87	0
		cal	**0.69**	**0.85**	0.60	0.85	0.79	0.75	0.93	0.91	**0.90**	3
	GW	def	0.51	0.60	0.41	0.75	0.79	0.71	0.92	0.90	0.87	0
		xGW	0.59	0.65	0.49	0.76	**0.84**	0.79	0.92	0.91	0.89	1
		cal	**0.69**	0.84	**0.62**	**0.83**	0.81	0.77	**0.93**	0.91	**0.90**	5
Bias	STD	def	9.08	−34.1	**−0.19**	7.79	**−2.82**	37.67	31.80	23.40	40.29	2
		cal	−2.38	−9.46	−10.0	−6.19	−20.6	−14.1	−7.71	−16.46	−11.4	0
	DV	def	−3.18	−34.2	34.06	79.12	49.10	96.80	49.31	37.21	48.87	0
		xDV	−3.87	2.68	−25.6	−4.59	−14.5	**−2.14**	−6.78	−10.84	**−2.05**	2
		cal	−0.61	**−1.65**	−6.63	**−1.59**	−13.5	−7.14	−4.41	−9.48	−11.6	2
	GW	def	48.23	102.4	44.87	58.45	19.99	72.08	52.64	39.38	52.47	0
		xGW	**0.10**	−10.2	−4.02	1.71	−15.6	19.13	4.76	**0.36**	19.25	2
		cal	−3.36	−7.69	−12.5	−5.57	−13.1	16.34	**0.24**	−12.6	−13.8	1

runs have a LE RMSE lower than 55 W m^{-2}, and no simulation is worse than RMSE = 90 W m^{-2}; however, 75% of the GW runs have errors larger than 55 W m^{-2}. For SMC_{5cm}, the top 10% of GW and STD runs have the same score (RMSE < 6%), but the interquartile range (IQR) of STD has RMSE = 8%–14%, whereas GW's RMSE = 9%–30%. The behavior of GW at this dry, bare soil site suggests significant degradation in model robustness. At site 7 (wet) (Figs. 6e,f), the IQR of GW's RMSE is very similar to STD's (30–70 W m^{-2}; 3%–7%). Although GW does a slightly better job when simulating LE, STD better simulates SMC_{5cm}. The good robustness of GW at the wet sites is consistent with Gulden et al. (2007a). At the intermediate site 4, STD is on average slightly worse than GW at simulating LE (Fig. 6c): 25% of GW's runs score lower than RMSE = 48 W m^{-2}, whereas 25% of STD runs score below RMSE = 52 W m^{-2}. However, GW performs poorly on SMC_{5cm} (Fig. 6d): 50% of STD runs score lower than RMSE = 10%, whereas only 10% of GW runs score lower than RMSE = 10%. The improvement gained by the addition of the particular aquifer model implemented here (comparing the CDFs of PS STD and PS GW) appears to be small (results not shown).

Partial calibration (i.e., xGW) significantly increases the probability of having large errors. At all sites, xGW shows bimodal distributions of errors. Nearly 70% of xGW runs have very poor scores. For example, at site 4 (Figs. 6c,d) (LE RMSE > 110 W m^{-2}; RMSE SMC_{5cm} > 16%), the majority of xGW runs have a larger RMSE than the worst-scoring 10% of STD runs. A very small fraction of xGW can be as good as GW. The exception is site 4, where the best 10% of xGW runs are still 10 W m^{-2} worse than either STD or GW's top-scoring runs. In general, xGW is at least 40 W m^{-2} and 5% (for LE and SMC_{5cm}, respectively) worse than the most-frequent performing models of STD and GW.

Tuning only the four new parameters (xGW) is the wrong way to calibrate GW. It leads to biased model structures. This implies that the aquifer parameters (e.g.,

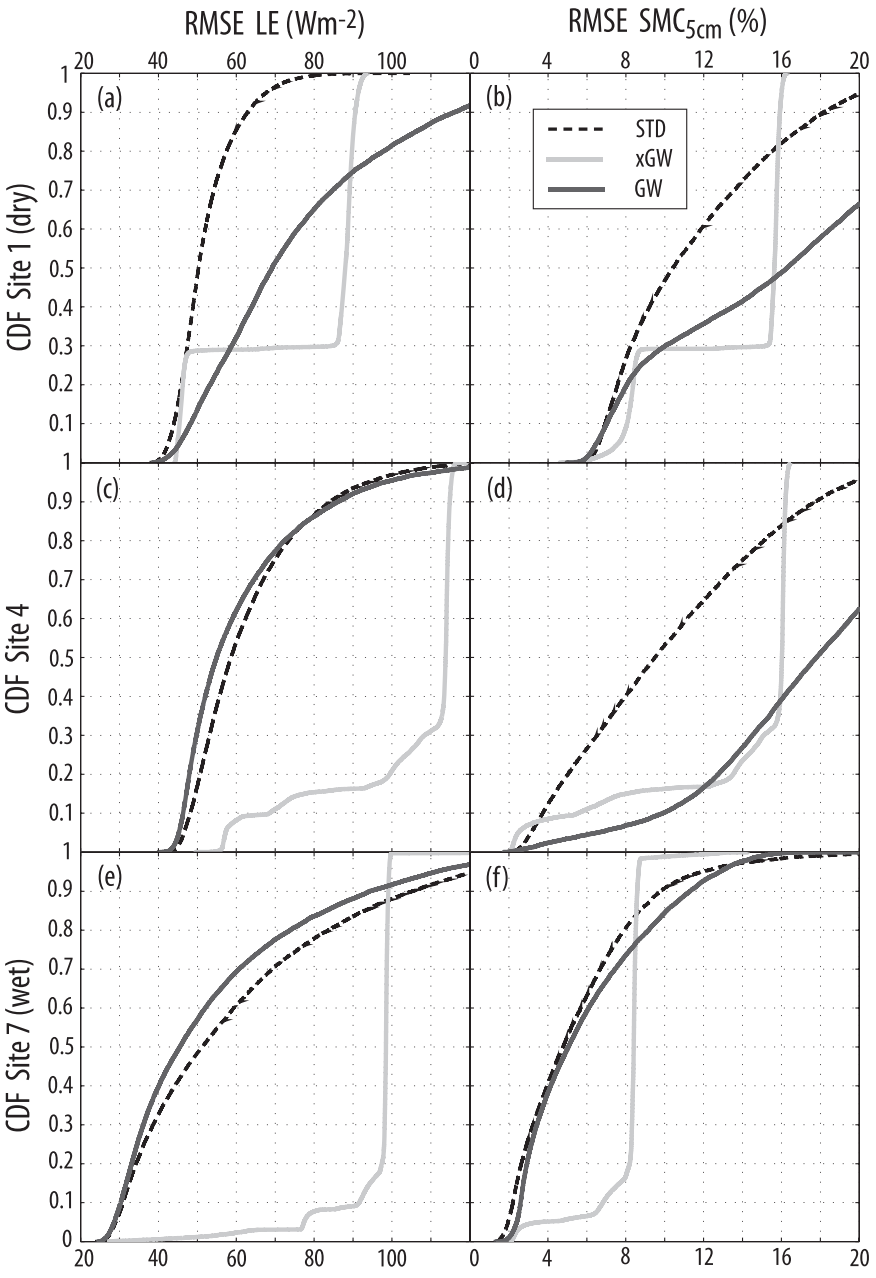

FIG. 6. CDF of 15 000 RMSE scores obtained by STD (dashed), xGW (light gray), and GW (dark gray) at (a),(b) site 1 (dry); (c),(d) site 4; and (e),(f) site 7 (wet). LE left column; SMC_{5cm} right column. Partial calibration (i.e., xGW) significantly increases the probability of having large errors. GW exhibits decreased robustness at dry sites and almost the same frequency of errors as STD at wet sites.

specific yield, exponential decay) and the STD soil parameters need to be coherent to accommodate the new structure (i.e., parameters need to be allowed to interact).

2) SENSITIVITY OF DV TO PARAMETERS

DV worsens the robustness of STD, significantly at the dry sites and slightly at wet sites. Cumulative distributions of 15 000 RMSE scores obtained by STD, DV, and xDV are shown in Fig. 7. At site 2 (dry) (Fig. 7a), the IQR of STD simulations of LE lies between RMSE = 42 and 55 W m^{-2}, whereas DV's is between 50 and 67 W m^{-2}. Fifty percent of the STD runs score below RMSE = 47 W m^{-2}. Fifty percent of the DV runs have RMSE above 57 W m^{-2}. Although the best performing runs of

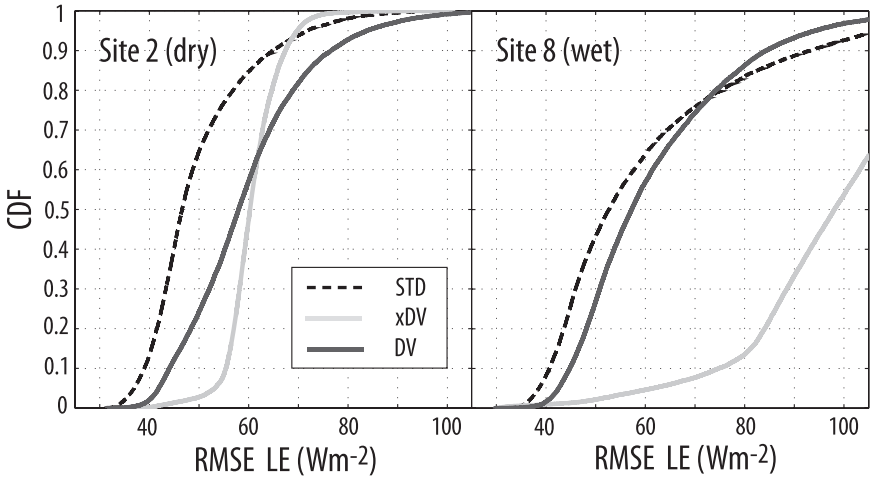

FIG. 7. CDF of 15 000 RMSE scores obtained for simulated LE by STD (dashed), xDV (light gray), and DV (dark gray) at sites 2 (dry) and 8 (wet). Partial calibration (i.e., xDV) significantly increases the probability of having larger errors.

STD and DV have RMSE = 30 W m^{-2}, only 25% of the PS of DV scores below 40 W m^{-2}; the majority of the PS of DV scores are 15 W m^{-2} worse than STD (results not shown). At site 8 (wet) (Fig. 7b), the IQR of DV's LE (RMSE = 50–70 W m^{-2}) is very similar to that of STD (RMSE = 45–70 W m^{-2}). Half of STD runs score below RMSE = 52 W m^{-2}, and half of DV runs have RMSE below 57 W m^{-2}. The best-scoring STD and DV runs at site 8 have RMSE = 30 W m^{-2} and RMSE = 1.5% (for LE and SMC_{5cm}, respectively). In general, a significant improvement in terms of better simulating LE over the reference model (STD) is not seen. The bulk of the simulations of DV are worse than the most-frequent performance of STD.

Like xGW, xDV is not an appropriate implementation of the model. At site 2 (dry), 90% of the xDV LE runs score between RMSE = 55–70 W m^{-2} (Fig. 7a). The scores of the top 10% of xDV PS are 5 W m^{-2} worse than those of DV or STD. At site 8 (wet), only 10% of xDV runs have RMSE $<$ 75 W m^{-2}, whereas 75% of the DV and STD runs perform like the best 10% of xDV. The top-scoring xDV has an RMSE = 30 W m^{-2} (similarly to STD and DV) but their SMC_{5cm} RMSE is 3% worse. We stress the need to let the parameters in the DV module interact with both vegetation and soil parameters of the host structure. This need becomes more pressing at more humid sites with more abundant vegetation.

5. Results of ensemble-based model intercomparison

We evaluate the reliability of STD, DV, and GW in simulating EF, W_{30}, ΔW_{30} when faced with parameter uncertainty. Using the framework of Gulden et al. (2008), summarized in appendix B, we show that STD is most fit for broad application.

a. Use of the performance score to evaluate time-varying model performance

Figure 8 shows the time variation of the performance score (ζ, refer to appendix B) of the PS ensemble for each criterion (EF, W_{30}, and ΔW_{30}) and model for site 2 (dry) (Figs. 8a–c) and site 8 (wet) (Figs. 8e–g). Despite calibration against H, LE, G, Tg, and SMC_{5cm}, when simulating ΔW_{30}, all models significantly overestimate the speed at which the soil column wets and dries (Figs. 8c,g); this result holds for both PS and MF ensembles. All models also overestimate the extent by which a single rainstorm increases overall soil wetness (results not directly shown). When simulating W_{30}, models typically do not identify the correct mean value. However, because individual models have their own equilibrium states, the day-to-day change in soil wetness is arguably a more important objective for models than is the modeled soil wetness (i.e., different W_{30} states in different models can yield the same ΔW_{30}). In the next paragraphs, we use the ζ score to help us understand when and why the models fail.

b. Use of the performance score to guide model development

The ζ score can be used as a tool to improve model structure and to help to assess whether a model is giving the *right answers for the right reasons*. Here we demonstrate the use of the time-varying performance score in this way.

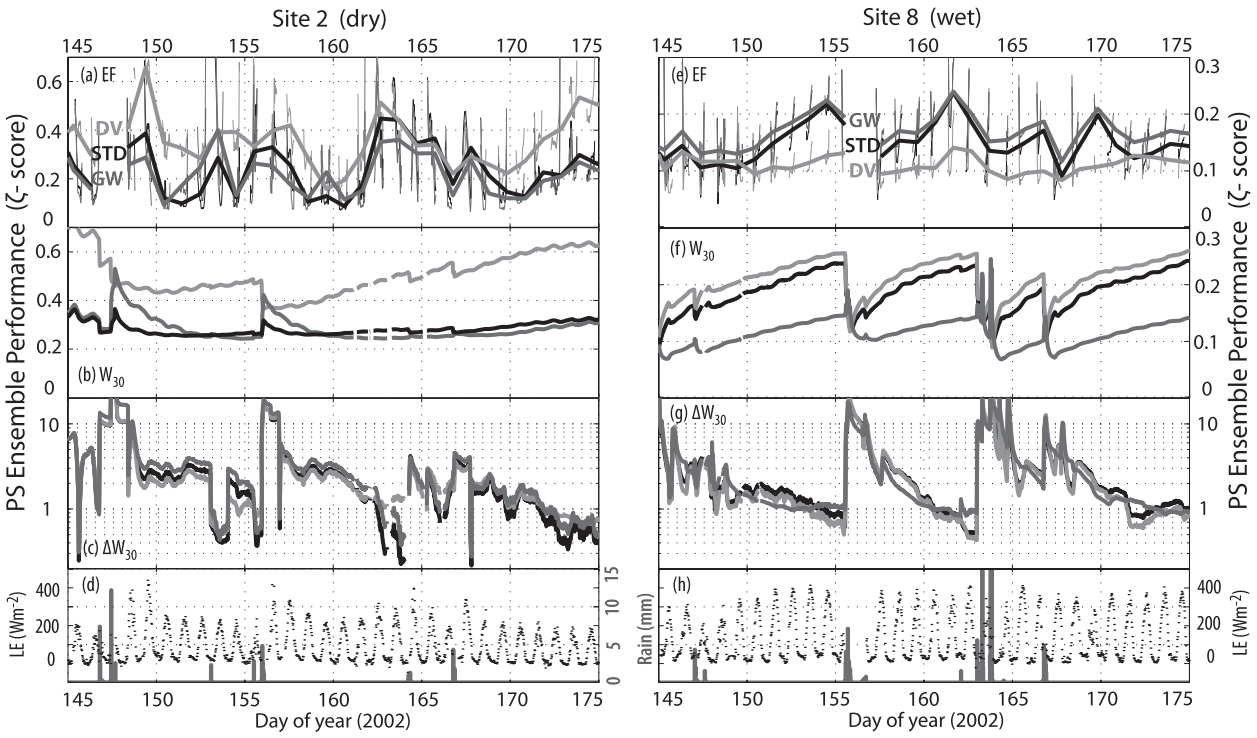

FIG. 8. Time-varying performance (ζ) scores (refer to appendix B) for STD, DV and GW PS ensembles between DOY 145 and 175 at (a)–(c) site 2 and (e)–(g) site 8 for EF, W_{30}, and ΔW_{30}, respectively; the closer the score is to zero, the better. (bottom) Precipitation and LE for (c) site 2 (dry) and (h) site 8 (wet). For ease of viewing, the EF performance score shown is also the daily mean value. Note that at site 8, periods of diverging performance (e.g., DOY 151–155) coincide with periods of increasing LE and drying soil. At site 2, unlike at site 8, DV is significantly worse than STD and GW.

1) DOES GW IMPROVE PERFORMANCE FOR THE RIGHT REASONS?

The hypothesis behind the implementation of the groundwater module is that the physical realism of the STD soil moisture profile is enhanced by improving simulated soil moisture dynamics (Niu et al. 2007). By allowing upward water flow from deep-soil stores during times of dry down or drought, the GW model presumably buffers the hydrologic cycle, alleviating the dry bias in LE in dry seasons. We examine the validity of this hypothesis with the help of Figs. 8 and 9.

GW achieves the best performance scores of any of the three models when simulating W_{30} at site 8 (wet) (Fig. 8f). However, its performance worsens as the soil dries down. This behavior is consistent with the deterioration in the performance of EF observed between DOY 150 and 155 (Fig. 8e). To reconcile this apparent contradiction, we also look at the temporal variation of ensemble bias (Fig. 9e) and the performance of GW when simulating ΔW_{30} (Fig. 8g). We assert that GW ameliorates the simulation of W_{30} by keeping the soil column wet during the overall simulation period not by improving soil moisture dynamics; hence GW is not able to improve the partitioning of surface energy at site 8 (wet). At site 2 (dry), the simulation of W_{30} by GW is comparable to that by STD (Fig. 8b), except immediately after precipitation, when STD outperforms GW. Observed EF in the dry location peaks sharply when available moisture is readily evaporated immediately after a rainstorm, but the cohort of models simulates a more muted response of EF. In terms of the partitioning of turbulent fluxes (Fig. 8a), GW's simulation degrades because the evapotranspiration can be heavily influenced by soil moisture within deep layers. We note that other structural shortcomings, such as errors in rooting depth specification or insufficient soil layer discretization, may also exist. GW shows wet bias for W_{30} after rainfall events (see DOY 148–155 in Fig. 9b). The reason GW has a good score at site 2 (dry) is likely because its mean soil moisture value is larger than that of the rest of the models in the cohort (and it therefore has a larger moisture gradient between soil and air). At the daily time scale (ΔW_{30} reports the difference in moisture between time t and 24 h prior), GW is not getting the right answers for the right reasons in the three sites reported here. It should be noted that, over longer time scales (from months to years), the groundwater module

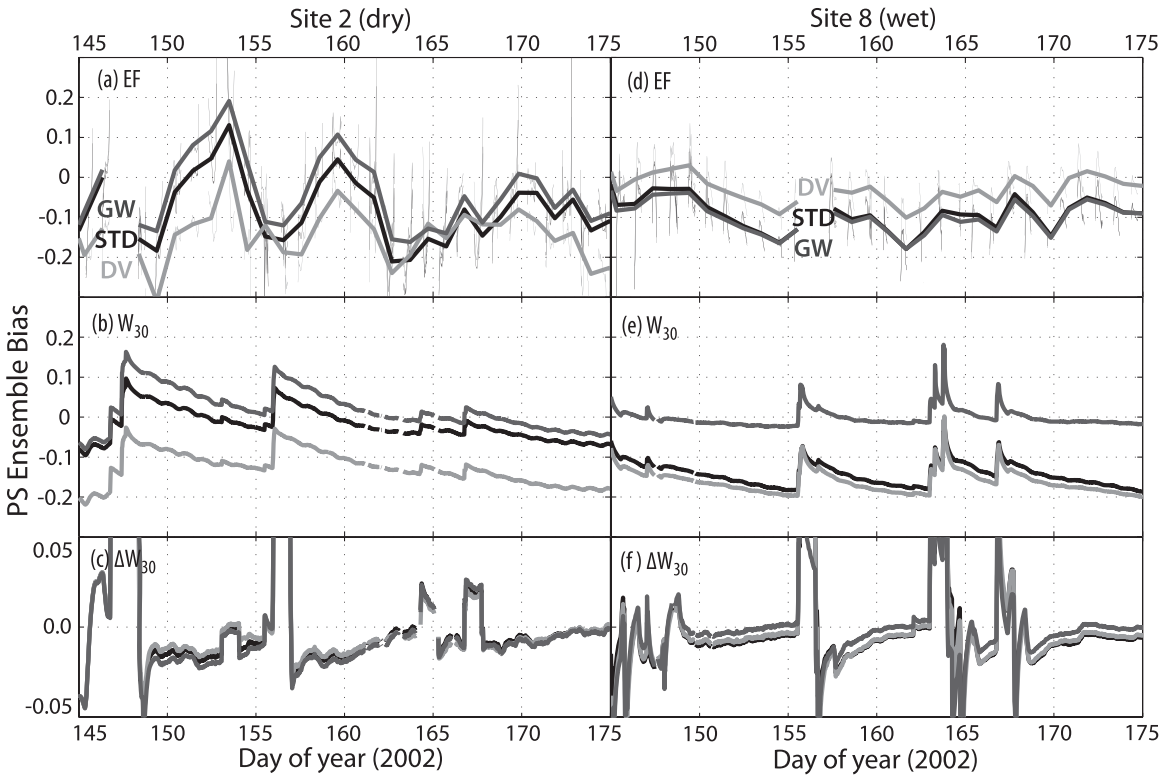

FIG. 9. Ensemble bias (refer to appendix B) of the STD, GW and DV PS simulations of EF, W_{30}, and ΔW_{30} at (a)–(c) site 2 and (d)–(f) site 8. For ease of viewing, the EF bias shown here is the daily mean value. On a diurnal scale, for all models, ensemble-mean-simulated EF typically underestimates EF at the beginning and end of the day and overestimates it during midday. Note that the 30-cm soil moisture of GW at site 8 (wet) is practically unbiased.

may yet improve the realism of vertical water transfer in the soil; however, whether the coupling of the slowly responding aquifer with high-frequency processes, such as root zone–fueled evapotranspiration, is correct has yet to be demonstrated. The dynamics of the aquifer model may be too slow and result in dampening of the variability of the soil moisture.

2) Does increased complexity of modeled vegetation improve simulation of surface energy fluxes?

DV improves model performance over STD at humid, more heavily vegetated sites (e.g., site 8) and degrades model performance at dry, sparsely vegetated sites (e.g., site 2). Sites 2 and 8 have distinct moisture and evaporation regimes (Figs. 1, 8d,h). At site 2 (dry), total LE flux peaks in the two days immediately following rainfall; at site 8 (wet), total LE flux peaks several days after the rain. We interpret this to mean that "fast" evaporation sources (canopy evaporation [E_c] and direct soil evaporation [E_{dir}]) play a larger role in shaping evaporative flux at site 2 (dry); transpiration (E_{transp}) is more significant at site 8 (wet).

At site 8 (wet), DV outperforms STD (Figs. 8e–g), especially as the soil dries after major precipitation events (e.g., DOY 153–155), when transpiration from deeper soil layers becomes the dominant source for evaporation. The relatively better performance of DV (with respect to STD) at site 8 occurs in both the MF ensemble (not shown) and the PS ensemble, underscoring the assertion that the improvement shown by DV is a structural improvement that is not related to the choice of parameters. Because the relationship expressing vegfrac as an exponential function of LAI favors vegfrac values that approach 1, DV favors a mode of behavior in which E_c and E_{transp} dominate LE flux at the expense of E_{dir}. This mode is likely more physically realistic in more densely vegetated zones (e.g., site 8). At site 8 (wet), STD's simulation of E_{dir} and E_c (the fast sources of LE flux) appears too high, and its simulated E_{transp} appears suppressed. STD tends to have higher LAI values than DV [mean LAI PS ensemble is 2.3 (DV) and 3.3 (STD)], slightly lower Rs values than DV (results not shown), and higher soil moisture than DV (Fig. 9e). Despite these transpiration-promoting conditions, because total transpiration is scaled by vegetation

fraction (0.7), STD still does not simulate as much transpiration as DV.

It should be noted that DV does explicitly link all components of the LE flux to LAI, which it allows to vary. Although this linkage may improve the conceptual physical consistency and make the seasonality and interannual variation in surface fluxes more realistic, we presume that, over the time scales examined here, its effect is somewhat minimal. In DV, LAI (and vegfrac) can and do vary on very short time scales (days), but this appears not to be the primary reason why DV improves over STD at site 8 (wet).

At site 2 (dry), DV's tendency to favor E_c and E_{transp} over E_{dir} worsens model performance. At site 2, DV supports too much evaporation too quickly from both E_c and E_{transp}. After a parameter adjustment in which the model is constrained by multiple objectives, not all of which directly improve simulation of EF, the model favors this E_c and E_{transp} mode and a second mode in which E_{dir} is strongly favored at the expense of E_c and E_{transp}. In both modes, at site 2 (dry), DV overestimates the fast sources of LE flux (E_c and E_{dir}). STD, with its forced ratio of E_c, E_{dir}, and E_{transp}, performs best at site 2. The additional degree of freedom provided by making vegfrac an exponential function of LAI makes the model very sensitive to the conversion. This sensitivity results in higher spread and less skill within the DV ensemble simulations of EF. Lastly, at site 4, STD and DV perform equivalently well in simulating EF (results not shown).

c. *Evaluation of models' suitability for broad application*

1) WHICH MODEL IS MOST RELIABLE FOR A GIVEN SITE AND OBJECTIVE?

Table 5 presents the time-median ζ score for each of the models examined, at sites 2, 4, and 8, for the PS ensemble and for the MF ensemble. The ζ score effectively combines ensemble spread and skill, hence, because of the large sample sizes, differences in the third decimal for EF and W_{30} are significant. Just as other goodness-of-fit metrics, the relative importance of a unit of difference depends on the criterion and on experience. We use the median performance score (instead of the mean) to minimize the effect of outliers, which have a relatively high chance of being the result of data outliers. As a group, the models simulate W_{30} and EF better than they simulate ΔW_{30}. Although the PS ensembles tend to perform better than the MF ensembles, this statement cannot be uniformly applied, which underscores the assertion that calibration against certain objectives may worsen the performance of the model in other, equally important, objectives (Leplastrier et al. 2002).

TABLE 5. Median model performance (ζ) score for each ensemble, site, criterion, and model. The lower the model performance score, the better the performance of the model realizations. Ensembles constructed using 150 MF and 150 PS parameter sets. Refer to appendix B for the definition of ζ score.

Criterion	Site	Ensemble	STD	DV	GW
EF	2	MF	0.204	0.291	0.238
		PS	0.186	0.342	0.190
	4	MF	0.203	0.153	0.155
		PS	0.211	0.198	0.185
	8	MF	0.224	0.073	0.076
		PS	0.130	0.113	0.157
	Average, MF		0.210	0.172	0.156
	Average, PS		0.175	0.217	0.177
	Mean, all realizations		0.193	0.195	0.167
W_{30}	2	MF	0.297	0.339	0.398
		PS	0.291	0.575	0.299
	4	MF	0.330	0.299	0.247
		PS	0.329	0.282	0.188
	8	MF	0.160	0.146	0.060
		PS	0.202	0.227	0.120
	Average, MF		0.262	0.261	0.235
	Average, PS		0.274	0.361	0.202
	Mean, all realizations		0.268	0.311	0.219
ΔW_{30}	2	MF	1.518	1.901	1.831
		PS	1.583	1.770	1.861
	4	MF	3.486	2.950	1.784
		PS	3.059	3.125	2.795
	8	MF	0.972	1.004	0.847
		PS	1.783	1.536	1.323
	Average, MF		1.992	1.952	1.487
	Average, PS		2.141	2.143	1.993
	Mean, all realizations		2.067	2.047	1.740

GW achieves the best mean performance for EF, ΔW_{30}, and W_{30}, both within its MF ensembles and within its PS ensembles. STD and DV perform equivalently well across the three criteria; however, STD tends to slightly outperform DV.

2) WHICH MODEL GIVES THE MOST CONSISTENT PERFORMANCE?

A "robust" model is generally less affected by parameter variation (Carlson and Doyle 2002; Gulden et al. 2007a) and therefore "model robustness" can provide a measure of consistent performance across ensemble members and across sites. Table 6 shows the robustness (ρ) score and rank for each model at each site and objective. The benchmark model (STD) is the most robust overall. At wet sites, DV is the most robust.

3) WHICH MODEL IS BEST SUITED FOR BROAD APPLICATION?

The φ score combines the concepts expressed by the performance and robustness scores. With the exception

TABLE 6. The ρ score and rank for each site, criteria, and model. A rank of 1 means that the model is the most robust model for that site and criterion. Mean robustness score is averaged across sites and criteria. Lower scores indicate increased robustness (lower sensitivity to errant parameters). Refer to appendix B for the definition of ρ score.

		STD		DV		GW	
Site	Criterion	Rank	ρ score	Rank	ρ score	Rank	ρ score
2	EF	1	0.046	2	0.081	3	0.113
	W_{30}	1	0.010	3	0.258	2	0.142
	ΔW_{30}	2	0.021	3	0.036	1	0.008
	Mean rank	1.33		2.67		2	
4	EF	1	0.019	3	0.127	2	0.089
	W_{30}	1	0.001	2	0.030	3	0.136
	ΔW_{30}	2	0.065	1	0.029	3	0.221
	Mean rank	1.33		2		2.67	
8	EF	2	0.266	1	0.214	3	0.347
	W_{30}	1	0.117	2	0.219	3	0.335
	ΔW_{30}	3	0.294	1	0.210	2	0.220
	Mean rank	2		1.33		2.67	
Average		1.55	0.093	2	0.134	2.44	0.179

TABLE 7. The φ score and rank for each site, criterion, and model. Lower fitness scores indicate better models. A rank of 1 means that the model is the best performing model for that site and criterion. The average rank combines performance and robustness, and it is an indication of the model's broad applicability. Refer to appendix B for the definition of φ score.

		STD		DV		GW	
Site	Criterion	Rank	φ score	Rank	φ score	Rank	φ score
2	EF	1	0.0085	3	0.0278	2	0.0214
	W_{30}	1	0.0030	3	0.1485	2	0.0424
	ΔW_{30}	2	0.0329	3	0.0635	1	0.0155
	Mean rank	1.33		3		1.67	
4	EF	1	0.0041	3	0.0252	2	0.0164
	W_{30}	1	0.0004	2	0.0085	3	0.0256
	ΔW_{30}	2	0.1997	1	0.0901	3	0.6172
	Mean rank	1.33		2		2.67	
8	EF	2	0.0346	1	0.0241	3	0.0546
	W_{30}	1	0.0235	3	0.0497	2	0.0403
	ΔW_{30}	3	0.5246	2	0.3220	1	0.2905
	Mean rank	2		2		2	
Average rank		1.55		2.33		2.11	
Variance of rank		0.53		0.75		0.61	

of site 2 (dry), the models are significantly less able to accurately represent ΔW_{30} than they are to represent EF and W_{30}. Because the models simulate some objectives more accurately than others, we evaluate the models' overall suitability for broad application by averaging their rankings for individual sites and objectives. Table 7 reports fitness scores and ranks; it also presents the individual site and criterion fitness score rankings and the mean rank of each model, averaged across sites and across criteria (see the final two lines of Table 7). In the models' current configurations, and using these metrics for model fitness, the benchmark model, STD, is found to be the most fit for broad application. It most consistently ranks at the top of the cohort in terms of fitness (mean rank of STD $\varphi = 1.33$). GW is second-most likely to rank at the top of the cohort (mean rank of GW $\varphi = 1.67$), but the variability of GW's fitness ranking is a potential caveat. DV and GW are only somewhat less fit than STD; with improvements to the realism of model physical parameterizations, guided by the time variation of the performance scores, modified versions of each of these models have the potential to outperform STD for broad application. Of the three models evaluated here (STD, DV, and GW), despite apparent increases in the nonbenchmark models' conceptual realism, the least complex version of Noah (STD) is most fit for broad application across these nine representative sites of summer climates in the central United States.

STD may perform better than the other models, not because of a more physically realistic representation but because it has fewer degrees of freedom and therefore tends to have lower ensemble spread. However, this low spread could also be an indicator of ''artificial skill'' in the context of providing an overconfident estimate. The inability of the enhanced parameterizations to outperform STD may also result from a mismatch between the level of complexity of STD and the new modules or the use of improper conceptualizations for the intended processes. For instance, the lack of a separate canopy layer in Noah may inhibit concordant functioning of Noah and the DV module. The DV module may augment the fitness of an LSM that explicitly represents canopy radiative transfer. Thus, it is possible that any of these modules may improve the fitness of other LSMs. We encourage the application of similar, thorough analyses for the same modules coupled to different LSMs as a more robust test of model performance.

6. Discussion of implications for model development

Although the results earlier may be considered model or site specific, their implications for LSM development and evaluation are significant and broad reaching. Our systematic analysis has demonstrated the limitations of traditional model evaluation techniques and has illustrated the utility of an ensemble-based framework that explicitly accounts for different sources of uncertainty in LSM predictions.

Standard evaluation methods are inadequate for highly complex models, such as LSMs. All models require

parameter estimation (Jakeman et al. 2006). Regarding models that require calibration as inferior is not practical (Beck 2002). We have shown that the improvement gained by calibration from an initial ''default'' state should not be used as a measure of the quality of the model for two reasons: (i) default parameters are educated guesses made by developers, (Dickinson et al. 1998; Shuttleworth 2007) or are model-dependent values adopted by modelers after extensive testing (which makes the score of the model applied to analogous settings fortuitous); and (ii) using ''improvement'' gained by calibration as a ''measure'' of overall model goodness is not correct. Models often adapt their structural error when undergoing calibration (Kirchner et al. 1996; Leplastrier et al. 2002). For that reason, even elevating models to their ''optimal'' performance before comparison is an incomplete and information-limited approach for model intercomparison. We have shown that conclusions regarding model quality should not be drawn using a single set of parameters (whether with default or ''best'' parameters). Single-realization model intercomparisons provide insufficient information to choose among competing models. Furthermore, such exercises offer limited help in diagnosing model structural deficiencies and do not fully explain why models differ and are therefore insufficient to guide model development.

We used sensitivity analysis to show that significant uncertainty comes from immeasurable, unknown, and effective parameters (e.g., the *e*-folding depth of saturated hydraulic conductivity, or the transformation factor for LAI to vegetation greenness). Our results are consistent with the notion that parameter values are model dependent (Wagener and Gupta 2005; Hogue et al. 2006) and that there is no straightforward transferability of the values between models and/or, potentially, sites (Hogue et al. 2005). The resulting implication is that default parameter values tested for a model component (e.g., GW) within one LSM [e.g., community land model (CLM); Oleson et al. 2008] will likely not be the same as those that yield the best—or even good—performance when the same module is used within a different LSM (e.g., Noah). This poses a caveat for simple ''plug and play'' use of functional modules between LSMs.

Additionally, we showed that tuning only the parameters associated with new modules leads to biased model structures and significantly increases the chance of poor performance. We assert that parameters in the host model need to be modified coherently and in unison with the new parameters to allow for interactions in the soil–vegetation system that control responses to meteorological forcing.

Because of these limitations and because of the dearth of spatially and temporally extensive evaluation and validation data, modeling for the foreseeable future will have to contend with significant parameter uncertainty. We assert that, especially when LSMs are to be used operationally (for short-term weather forecasting), the community needs to employ an evaluation technique that explicitly accounts for sources of uncertainty that are inherent to modeling (e.g., parameters and data). For the purposes of model development, evaluation techniques should identify, in time, the model shortcomings that hinder its capacity to simulate surface exchanges and states, even with optimized parameters.

To effectively capture a more complete spectrum of model behaviors, we employed the ensemble-based evaluation framework of Gulden et al. (2008). Comparison of the performance of the MF and PS ensembles enabled us to draw conclusions regarding model structure that were independent of parameter uncertainty. The framework also allowed us to evaluate models rigorously and to consider model robustness as a criterion when selecting models best suited to operational use (i.e., when possible, we wanted to choose the best-performing LSMs that were also less sensitive to parameter variation). Finally, because model rank depends on criteria and reliability cannot be guaranteed for similar sites, the use of fitness scores gave us an objective way to compare models.

One major caveat to this study is that we have neglected the uncertainty in the data, but we assert that the framework used here can and should accommodate both data and parameter uncertainty. Uncertainty in model output that stems from uncertain initial conditions is relatively unimportant when compared to uncertainty in parameter values, so long as reasonable initial conditions are used or the model is properly spun up (Bastidas et al. 2001; Abramowitz et al. 2006; De Lannoy et al. 2006). We assume that this relative unimportance of initial data, combined with our 2.5-yr spin-up period before the calibration/evaluation period, allows us to neglect uncertainty in initial conditions in this analysis. A less trivial source of uncertainty is uncertainty in meteorological forcing data. Model sensitivity to errors in boundary forcing data should be a criterion for model evaluation; however, because of computational constraints, we also neglect forcing data uncertainty. Next-step work should encompass ensembles of simulations in which both parameters and input data are perturbed for each model run.

This study illustrates how increased physical realism does not necessarily yield an LSM that better reproduces observations. Thus, our results are consistent with the notion that increasing complexity (and therefore degrees of freedom) can significantly increase the modeler's risk that his model will not perform as expected (e.g., Gulden

et al. 2007a). We recognize that nature is inherently complex, and that models must be sufficiently complex to represent key processes and feedbacks; however, especially when models are being used for prediction, because of parameter and structural uncertainty, researchers should be aware that there often exists a trade-off between model complexity and model predictive performance. Our results have shown that while adding more conceptually realistic components reduces error in model simulations, additional information-based criteria often do not deem the improvement to be worth the additional complexity. Hence, modelers must increase the precision of their definition of ''improvement'' (Smith 2002) to include a broad, multivariate suite of metrics. Results presented here illustrate that lack of rigorous testing can preclude significant model development efforts. Raising the standards for objective comparison against benchmarks using strict, relevant tests will reward developers and foster confidence of the public and policymakers (Kirchner et al. 1996; Jakeman et al. 2006; Refsgaard et al. 2006; Randall et al. 2007; Clarke 2008).

7. Summary and conclusions

We compare three versions of the Noah LSM (benchmark STD, dynamic vegetation-enhanced DV, and groundwater-enabled GW) using an analysis that employs high-frequency, local-scale turbulent fluxes and near-surface states while taking into account both model structure and uncertainty in model parameters. When using either default model parameters or a single calibrated set of parameters, the performance of STD, DV, and GW is not distinguishable. After a detailed analysis that takes into account parameter uncertainty, our primary conclusion is that, of the three models examined, the benchmark model (STD) is the best suited for reproducing observed high-frequency heat fluxes and soil states. It is significantly more fit than other models at arid and semiarid sites. Although GW typically achieves the best performance score when simulating each of the three criteria (evaporative fraction, 24-h change in soil wetness, and soil wetness), GW is only about half as robust as the benchmark model (STD). DV is reasonably well suited for broad application in wet regions. It significantly improves the model's ability to correctly partition net radiation at the site 8 (wet), even when good model parameters cannot be identified.

We further conclude that although GW has the best average performance of any models in simulating all three criteria, its superior performance results from correcting the mean model state and is not due to improved short-term soil moisture dynamics. All three models are too quick to wet and too quick to dry; GW does not appear to significantly correct this problem. When compared to STD (and GW), DV improves simulation of EF at site 8 (wet) because its partitioning of LE flux favors transpiration and canopy evaporation over direct soil evaporation. At site 2 (dry), DV's increased emphasis on canopy evaporation and transpiration leads to model degradation.

Our results do not provide definitive evidence regarding the role of conceptual realism in shaping model robustness. At wetter sites (sites 7 and 8), DV and GW often perform better and are slightly more robust than STD; at drier sites, GW and DV do not perform as well as STD and are less robust than STD. Therefore, the present formulations of DV and GW may be considered less conceptually realistic for use when simulating arid sites.

Although the results discussed above may be model and site specific, the implications of our work are not. We have shown that traditional LSM evaluation methods that use evaluation data averaged in time and uninformative misfit metrics and that do not account for parameter uncertainty are, in many cases, insufficient for confident assessment of model performance. Ad hoc evaluation using single parameter sets provides insufficient information for choosing among competing models. It neither helps in diagnosing deficiencies nor explains why models differ, and it is insufficient to guide model development. We have demonstrated a need for increased rigor in LSM evaluation using techniques that explicitly account for multiple sources of uncertainty and that can identify in time the shortcomings in the formulations of LSMs. Because default parameters are at best an educated guess and because models are frequently not distinguishable when all are given ''ideal'' parameters, it may be necessary to revisit conclusions drawn from model evaluation studies that have not fully accounted for parameter uncertainty. Plug-and-play use of new modules, in which the new module's parameters are either not calibrated or only parameters within the new module are calibrated, does not reliably yield optimal model performance. Adding complexity to models (although crucial for research endeavors) entails a significant risk in decreasing model robustness, which can lessen the model's overall fitness for broad application in operational settings.

We recommend that the approach used here be widely adopted by model intercomparison projects, which, in part because of a lack of stringent evaluation metrics, have often been plagued by a lack of firm conclusions. We encourage other modeling groups to perform similar analyses with their models. Finally, we advocate for a cooperative approach between the parameter estimation

and model development communities as a way to ensure rapid, continued improvement of our understanding and modeling of environmental processes.

Acknowledgments. The authors would like to acknowledge editor G. Salvucci, an associate editor, and an anonymous reviewer for their thoughtful comments that have contributed to the improvement of this manuscript. We thank F. Chen at NCAR and K. Mitchell at NCEP for their insight. H. Wei, also at NCEP, provided us with monthly vegetation fraction and albedo climatology values. We thank the International H_2O Project for the datasets. We appreciate the insights of Charles S. Jackson and M. Bayani Cardenas. We benefited from the computational resources at the Texas Advanced Computing Center (TACC). This project was funded by NOAA (Grant NA07OAR4310216), the Graduate Fellowship of the Hydrology Training Program of the OHD/NWS, NSF, and the Jackson School of Geosciences.

APPENDIX A

Statistics and Goodness-of-Fit Metrics

For the following definitions, P_t is the prediction at time t, o_t is the observation at time t, and T is the number of time steps (t) in the series. Here, k is the number of free parameters in the model (Legates and McCabe 1999; Akaike 1974; Schwarz 1978).

$$\text{Observation mean: } \overline{O} = \frac{1}{T}\sum_{t=1}^{T}(O_t). \tag{A1}$$

$$\text{Model mean: } \overline{P} = \frac{1}{T}\sum_{t=1}^{T}(P_t). \tag{A2}$$

Observation standard deviation:

$$\text{std dev} = \left[\frac{1}{T}\sum_{t=1}^{T}(O_t - \overline{O})^2\right]^{0.5}. \tag{A3}$$

Model standard deviation:

$$\text{std dev} = \left[\frac{1}{T}\sum_{t=1}^{T}(P_t - \overline{P})^2\right]^{0.5}. \tag{A4}$$

Root-mean-square error:

$$\text{RMSE} = \left[\frac{1}{T}\sum_{t=1}^{T}(O_t - P_t)^2\right]^{0.5}. \tag{A5}$$

Coefficient of determination:

$$r^2 = \left\{\frac{\sum_{t=1}^{T}(O_t - \overline{O})(P_t - \overline{P})}{\left[\sum_{t=1}^{T}(O_t - \overline{O})^2\right]^{0.5}\left[\sum_{t=1}^{T}(P_t - \overline{P})^2\right]^{0.5}}\right\}^2. \tag{A6}$$

$$\text{Bias: bias} = \frac{1}{T}\sum_{i=1}^{T}(P_i - O_i). \tag{A7}$$

$$\text{Nash–Sutcliffe efficiency: NSE} = 1 - \frac{\sum_{t=1}^{T}(O_t - P_t)^2}{\sum_{t=1}^{T}(O_t - \overline{O})^2}. \tag{A8}$$

Akaike information criteria:

$$\text{AIC} = 2k + T\ln\left(\frac{\text{RMSE}}{T} + 1\right) + \frac{2k(k+1)}{T-k-1}. \tag{A9}$$

Bayesian information criteria:

$$\text{BIC} = T\ln\left(\frac{\text{RMSE}}{T}\right) + k\ln(T). \tag{A10}$$

APPENDIX B

Ensemble Metrics

For the following definitions, $x_{i,t}$ is the ensemble member i at time t; o_t is the observation at time t; N_{ens} is the number of ensembles at time t; and T is the number of time steps (t) in the series (Talagrand et al. 1997).

$$\text{Ensemble mean: } \overline{x}_t = \frac{\sum_{i=1}^{N_{\text{ens}}} x_{i,t}}{N_{\text{ens}}}. \tag{B1}$$

$$\text{Ensemble bias: } \beta_t = \overline{x}_t - o_t. \tag{B2}$$

$$\text{Ensemble skill score: } \kappa_t = (\overline{x}_t - o_t)^2. \tag{B3}$$

$$\text{Ensemble spread: } \pi_t = \frac{\sum_{i=1}^{N_{\text{ens}}}(x_{i,t} - \overline{x}_t)^2}{N_{\text{ens}} - 1}. \tag{B4}$$

Metrics for model evaluation

MODEL PERFORMANCE (ζ)

For time step t, the best performing model will have the lowest performance score (Gulden et al. 2008):

$$\zeta_t = \frac{\mathrm{CDF}_{\mathrm{ens},t} - \mathrm{CDF}_{\mathrm{obs},t}}{1 - \mathrm{CDF}_{\overline{\mathrm{obs}}}}, \tag{B5}$$

where $\mathrm{CDF}_{\mathrm{ens},t}$ is the CDF of the ensemble at time t, $\mathrm{CDF}_{\mathrm{obs},t}$ is the CDF of the observations at time t, and $\mathrm{CDF}_{\overline{\mathrm{obs}}}$ is the CDF of the time mean of observation time series. As ζ_t decreases, model performance at time t increases. Inspired by ensemble verification metrics, model performance score ζ_t is lowest (i.e., best) when the parameter set ensemble brackets observations and when the ensemble is highly skilled (ensemble mean closer to the observation) and has low spread. It rewards near misses and penalizes overly uncertain prediction bounds. Note that when no uncertainty information is available for the observations, $\mathrm{CDF}_{\mathrm{obs},t}$ is a step function. Denominator $1 - \mathrm{CDF}_{\overline{\mathrm{obs}}}$ scales the score to enable cross-criterion and cross-site comparison along a time series. Note that if the modeler would like to penalize one criterion more heavily than another, the denominator can be modified: for example, using a denominator of $1 - \mathrm{CDF}_{\mathrm{obs},t}$ would increase the stringency of the score more when observations are low than when observations are high.

MODEL ROBUSTNESS (ρ)

A robust model is insensitive to errant parameters: its performance is not significantly degraded when performing with suboptimal parameters (Carlson and Doyle 2002). We describe the sensitivity of model output to parameter choices as:

$$\rho = \frac{\left|\overline{\zeta}_{\mathrm{ps}} - \overline{\zeta}_{\mathrm{mf}}\right|}{\overline{\zeta}_{\mathrm{ps}} + \overline{\zeta}_{\mathrm{mf}}} \tag{B6}$$

where $\overline{\zeta}_{\mathrm{ps}}$ is the time median performance score of the Pareto set (PS) ensemble; $\overline{\zeta}_{\mathrm{mf}}$ is the time median performance score of the most-frequent performing (MF) ensemble.

MODEL FITNESS (φ)

The ζ score can be combined with a measure of model robustness to evaluate overall model fitness. We quantify each model's overall suitability for broad application using

$$\varphi = \rho\overline{\zeta}_{\mathrm{ps}}, \tag{B7}$$

where ρ is the robustness score for a given model where $\overline{\zeta}_p$ is the time median of the performance score for the PS ensemble of that model. For a given site and objective, the model with the lowest value of φ is considered most suitable for broad application.

REFERENCES

Abramowitz, G., H. Gupta, A. J. Pitman, Y. Wang, R. Leuning, and H. Cleugh, 2006: Neural Error Regression Diagnosis (NERD): A tool for model bias identification and prognostic data assimilation. *J. Hydrometeor.,* **7,** 160–177.

Akaike, H., 1974: A new look at the statistical model identification. *IEEE Trans. Autom. Control,* **19,** 716–723.

Bastidas, L. A., 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 Series, Vol. 3, Amer. Geophys. Union, 65–76.

——, T. S. Hogue, S. Sorooshian, H. V. Gupta, and W. J. Shuttleworth, 2006: Parameter sensitivity analysis for different complexity land surface models using multicriteria methods. *J. Geophys. Res.,* **111,** D20101, doi:10.1029/2005JD006377.

Beck, M. B., Ed., 2002: *Environmental Foresight and Models: A Manifesto*. Developments in Environmental Modelling Series, Vol. 22, Elsevier Science, 473 pp.

Beven, K. J., 2006: A manifesto for the equifinality thesis. *J. Hydrol.,* **320,** 18–36.

——, and J. Freer, 2001: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems. *J. Hydrol.,* **249,** 11–29.

Boone, A., and Coauthors, 2004: The Rhône-Aggregation Land Surface Scheme intercomparison project: An overview. *J. Climate,* **17,** 187–208.

Bowling, L. C., and Coauthors, 2003: Simulation of high-latitude hydrological processes in the Torne–Kalix basin: PILPS Phase 2(e). 1: Experiment description and summary intercomparisons. *Global Planet. Change,* **38,** 1–30.

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.

Carlson, J. M., and J. Doyle, 2002: Complexity and robustness. *Proc. Natl. Acad. Sci. USA,* **99,** 2538–2545.

Chen, F., and Coauthors, 2007: Description and evaluation of the characteristics of the NCAR high-resolution land data assimilation system. *J. Appl. Meteor. Climatol.,* **46,** 694–713.

Childs, P., A. Qureshi, S. Raman, K. Alapaty, R. Ellis, R. Boyles, and D. Niyogi, 2006: Simulation of convective initiation during IHOP_2002 using the Flux-Adjusting Surface Data Assimilation System (FASDAS). *Mon. Wea. Rev.,* **134,** 134–148.

Clarke, R. T., 2008: Issues of experimental design for comparing the performance of hydrologic models. *Water Resour. Res.,* **44,** W01409, doi:10.1029/2007WR005927.

Cosgrove, B. A., and Coauthors, 2003: Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) project. *J. Geophys. Res.,* **108,** 8842, doi:10.1029/2002JD003118.

De Lannoy, G. J. M., P. R. Houser, V. R. N. Pauwels, and N. E. C. Verhoest, 2006: Assessment of model uncertainty for soil moisture through ensemble verification. *J. Geophys. Res.,* **111,** D10101, doi:10.1029/2005JD006367.

Demarty, J., C. Ottlé, I. Braud, A. Olioso, J. P. Frangi, L. Bastidas, and H. V. Gupta, 2004: Using a multiobjective approach to

retrieve information on surface properties used in a SVAT model. *J. Hydrol.,* **287,** 214–236.

Dickinson, R. E., M. Shaikh, R. Bryant, and L. Graumlich, 1998: Interactive canopies for a climate model. *J. Climate,* **11,** 2823–2836.

Dirmeyer, P. A., F. J. Zeng, A. Ducharne, J. C. Morrill, and R. D. Koster, 2000: The sensitivity of surface fluxes to soil water content in three land surface schemes. *J. Hydrometeor.,* **1,** 121–134.

——, R. D. Koster, and Z. C. Guo, 2006: Do global models properly represent the feedback between land and atmosphere? *J. Hydrometeor.,* **7,** 1177–1198.

Ek, M. B., and Coauthors, 2003: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *J. Geophys. Res.,* **108,** 8851, doi:10.1029/2002JD003296.

Gulden, L. E., E. Rosero, Z.-L. Yang, M. Rodell, C. S. Jackson, G.-Y. Niu, P. J.-F. Yeh, and J. Famiglietti, 2007a: Improving land-surface model hydrology: Is an explicit aquifer model better than a deeper soil profile? *Geophys. Res. Lett.,* **34,** L09402, doi:10.1029/2007GL029804.

——, Z.-L. Yang, and G.-Y. Niu, 2007b: Interannual variation in biogenic emissions on a regional scale. *J. Geophys. Res.,* **112,** D14103, doi:10.1029/2006JD008231.

——, E. Rosero, Z.-L. Yang, T. Wagener, and G.-Y. Niu, 2008: Model performance, model robustness, and model fitness scores: A new method for identifying good land-surface models. *Geophys. Res. Lett.,* **35,** L11404, doi:10.1029/2008GL033721.

Gupta, H. V., S. Sorooshian, and P. O. Yapo, 1998: Toward 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 multicriteria methods. *J. Geophys. Res.,* **104** (D16), 19 491–19 504.

——, T. Wagener, and Y. Liu, 2008: Reconciling theory with observations: Elements of a diagnostic approach to model evaluation. *Hydrol. Processes,* **22,** 3802–3813, doi:10.1002/hyp.6989.

Hogue, T. S., L. A. Bastidas, H. V. Gupta, S. Sorooshian, K. Mitchell, and W. Emmerich, 2005: Evaluation and transferability of the Noah land surface model in semiarid environments. *J. Hydrometeor.,* **6,** 68–84.

——, ——, ——, and ——, 2006: Evaluating model performance and parameter behavior for varying levels of land surface model complexity. *Water Resour. Res.,* **42,** W08430, doi:10.1029/2005WR004440.

Holt, T., D. Niyogi, F. Chen, K. Manning, M. A. LeMone, and A. Qureshi, 2006: Effect of land–atmosphere interactions on the IHOP 24–25 May 2002 convection case. *Mon. Wea. Rev.,* **134,** 113–133.

Jakeman, A. J., R. A. Letcher, and J. P. Norton, 2006: Ten iterative steps in development and evaluation of environmental models. *Environ. Modell. Software,* **21,** 602–614, doi:10.1016/j.envsoft.2006.01.004.

Jensen, M. J. W., 1998: Prediction error through modelling concepts and uncertainty from basic data. *Nutr. Cycling Agroecosyst.,* **50,** 247–253.

Kim, Y., and G. Wang, 2007: Impact of vegetation feedback on the response of precipitation to antecedent soil moisture anomalies over North America. *J. Hydrometeor.,* **8,** 534–550.

Kirchner, J. W., R. P. Hooper, C. Kendall, C. Neal, and G. Leavesley, 1996: Testing and validating environmental models. *Sci. Total Environ.,* **183,** 33–47.

Koster, R. D., and Coauthors, 2004: Regions of strong coupling between soil moisture and precipitation. *Science,* **305,** 1138–1140.

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

LeMone, M. A., and Coauthors, 2007: NCAR/CU surface, soil, and vegetation observations during the International H_2O Project 2002 field campaign. *Bull. Amer. Meteor. Soc.,* **88,** 65–81.

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.

Lyon, S. W., and Coauthors, 2008: Coupling terrestrial and atmospheric water dynamics to improve prediction in a changing environment. *Bull. Amer. Meteor. Soc.,* **89,** 1275–1279.

Mitchell, K. E., and Coauthors, 2004: The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *J. Geophys. Res.,* **109,** D07S90, doi:10.1029/2003JD003823.

Nijseen, B., and L. A. Bastidas, 2005: Land–atmosphere models for water and energy cycle studies. *Encyclopedia of Hydrological Sciences, Part 17: Climate Change,* M. G. Anderson and J. J. McDonnell, Ed., Vol. 5, John Wiley, 3089–3102.

Niu, G.-Y., Z.-L. Yang, R. E. Dickinson, and L. E. Gulden, 2005: A simple TOPMODEL-based runoff parameterization (SIMTOP) for use in global climate models. *J. Geophys. Res.,* **110,** D21106, doi:10.1029/2005JD006111.

——, ——, ——, ——, and H. Su, 2007: Development of a simple groundwater model for use in climate models and evaluation with Gravity Recovery and Climate Experiment data. *J. Geophys. Res.,* **112,** D07103, doi:10.1029/2006JD007522.

Niyogi, D. S., and S. Raman, 1997: Comparison of stomatal resistance simulated by four different schemes using FIFE observations. *J. Appl. Meteor.,* **36,** 903–917.

Oleson, K. W., and Coauthors, 2008: Improvements to the Community Land Model and their impact on the hydrological cycle. *J. Geophys. Res.,* **113,** G01021, doi:10.1029/2007JG000563.

Pielke, R. A., Sr., 2001: Influence of the spatial distribution of vegetation and soils on the prediction of cumulus convective rainfall. *Rev. Geophys.,* **39,** 151–177.

Pitman, A. J., 2003: The evolution of, and revolution in, land surface schemes designed for climate models. *Int. J. Climatol.,* **23,** 479–510, doi:10.1002/joc.893.

Randall, D. A., and Coauthors, 2007: Climate models and their evaluation. *Climate Change 2007: The Physical Science Basis*, S. D. Solomon et al., Eds., Cambridge University Press, 589–662.

Refsgaard, J. C., J. P. van der Sluijs, J. Brown, and P. van der Keur, 2006: A framework for dealing with uncertainty due to model structure error. *Adv. Water Resour.,* **29,** 1586–1597.

Rodell, M., P. R. Houser, A. A. Berg, and J. S. Famiglietti, 2005: Evaluation of 10 methods for initializing a land surface model. *J. Hydrometeor.,* **6,** 146–155.

Schultz, K., and K. Beven, 2003: Data-supported robust parameterizations in land surface–atmosphere flux predictions: Towards a top-down approach. *Hydrol. Processes,* **17,** 2259–2277.

Schwarz, G., 1978: Estimating the dimension of a model. *Ann. Stat.,* **6,** 461–464.

Shuttleworth, W. J., 2007: Putting the ''vap'' into evaporation. *Hydrol. Earth Syst. Sci.,* **11,** 210–244.

Smith, L. A., 2002: What might we learn from climate forecasts? *Proc. Natl. Acad. Sci. USA,* **99,** 2487–2492, doi:10.1073/pnas.012580599.

Stöckli, R., and Coauthors, 2008: Use of FLUXNET in the community land model development. *J. Geophys. Res.,* **113,** G01025, doi:10.1029/2007JG000562.

Talagrand, O., R. Vautar, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proc. ECMWF Workshop on Predictability*, Reading, United Kingdom, ECMWF, 1–25.

Viterbo, P., 2002: A review of parameterization schemes for land surface processes. Meteorological Training Course Lecture Series, ECMWF. [Available online at http://www.ecmwf.int/newsevents/training/lecture_notes/pdf_files/PARAM/Land_surf.pdf.]

Vrugt, J. A., H. V. Gupta, L. A. Bastidas, W. Bouten, and S. Sorooshian, 2003: Effective and efficient algorithm for multiobjective optimization of hydrologic models. *Water Resour. Res.,* **39,** 1214, doi:10.1029/2002WR001746.

Wagener, T., and H. V. Gupta, 2005: Model identification for hydrological forecasting under uncertainty. *Stochastic Environ. Res. Risk Assess.,* **19,** 378–387.

Weckwerth, T. M., and D. B. Parsons, 2006: A review of convection initiation and motivation for IHOP 2002. *Mon. Wea. Rev.,* **134,** 5–22.

——, and Coauthors, 2004: An overview of the International H_2O Project (IHOP_2002) and some preliminary highlights. *Bull. Amer. Meteor. Soc.,* **85,** 253–277.

Wood, E. F., and Coauthors, 1998: The Project for Intercomparison of Land-Surface Parameterization Schemes (PILPS) Phase 2(c) Red–Arkansas River basin experiment: 1. Experimental description and summary intercomparisons. *Global Planet. Change,* **19,** 115–135.

Xia, Y., A. J. Pitman, H. V. Gupta, M. Lepastrier, 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.

Yang, Z.-L., 2004: Modeling land surface processes in shortterm weather and climate studies. *Observation, Theory and Modeling of Atmospheric Variabilit*y, X. Zhu et al., Eds., World Scientific Series on Meteorology of East Asia, Vol. 3, World Scientific, 288–313.

——, and G.-Y. Niu, 2003: The versatile integrator of surface and atmosphere processes. Part 1: Model description. *Global Planet. Change,* **38,** 175–189.