Model performance, model robustness, and model fitness scores: A new method for identifying good land-surface models

Lindsey E. Gulden,[1] Enrique Rosero,[1] Zong-Liang Yang,[1] Thorsten Wagener,[2] and Guo-Yue Niu[1]

Received 22 February 2008; revised 17 April 2008; accepted 1 May 2008; published 14 June 2008.

[1] We introduce three metrics for rigorous evaluation of land-surface models (LSMs). This framework explicitly acknowledges perennial sources of uncertainty in LSM output. The model performance score (ζ) quantifies the likelihood that a representative model ensemble will bracket most observations and be highly skilled with low spread. The robustness score (ρ) quantifies the sensitivity of performance to parameter and/or data error. The fitness score (φ) combines performance and robustness, ranking models' suitability for broad application. We demonstrate the use of the metrics by comparing three versions of the Noah LSM. Using time-varying ζ for hypothesis testing and model development, we show that representing short-term phenological change improves Noah's simulation of surface energy partitioning and subsurface water dynamics at a semi-humid site. The least complex version of Noah is most fit for broad application. The framework and metrics presented here can significantly improve the confidence that can be placed in LSM predictions. **Citation:** Gulden, L. E., 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.

1. Introduction

[2] The increasing reliance of scientists, engineers, and policymakers on the predictions of land-surface models (LSMs) demands more rigorous evaluation of LSM parameterizations. Most LSMs are assessed using limited, localized, often semi-quantitative approaches [e.g., *Chen et al.*, 2007]. With few exceptions, model intercomparisons and evaluations of modified parameterizations neglect an assessment of uncertainty that extends beyond simple end-member sensitivity analyses [e.g., *Niu et al.*, 2005]. This incomplete approach is due in part to a dearth of observations and in part to evaluation procedures that are no longer state-of-the-art with respect to available computing resources. The development of robust metrics for comprehensive model evaluation is in its infancy [*Randall et al.*, 2007]. Here, we present a simple method for increasing the rigor of LSM assessment.

[3] The simplest way to assess an LSM is to evaluate performance at a single site using default parameters [e.g., *Henderson-Sellers et al.*, 1996]. LSM performance varies widely when parameters are shifted within reasonable ranges [e.g., *Gulden et al.*, 2007]. At a given site, the parameter set resulting in the best performance may significantly differ from the default. That one model equipped with default parameters does better than another is likely fortuitous. Parameters tend to be effective values, not physical quantities [*Wagener and Gupta*, 2005]. A more thorough evaluation method is to first minimize parameter error by calibrating all models and to then compare model output generated with the best parameter set [e.g., *Nijseen and Bastidas*, 2005]. Using optimal parameters does not represent the way in which LSMs are generally applied; calibration against certain criteria may worsen the simulation of other, equally important criteria [*Leplastrier et al.*, 2002]. After calibration, most equivalently complex models perform equivalently well [e.g., *Beven*, 2006]. Additional methods for LSM evaluation (e.g., the use of neural networks to benchmark LSMs [*Abramowitz*, 2005]) show promise, but to our knowledge, none has been widely adopted.

[4] Even in the rare case when we can estimate individual parameter ranges, parameter interaction and discontinuous model responses to even small shifts in parameter values undercut confidence in the realism of simulations [e.g., *Gulden et al.*, 2007; E. Rosero et al., Evaluating enhanced hydrological representations in Noah-LSM over transition zones: Part 1. Traditional model intercomparison, submitted to *Journal of Hydrometeorology,* 2008a]. The dearth of extensive validation datasets makes this limitation unlikely to soon change. To assess model performance in 'real life' settings, evaluation frameworks such as the one we present here must explicitly acknowledge these sources of uncertainty. Here we treat only parameter uncertainty, but we stress that our framework can and should be applied to incorporate uncertainty in observations. The dependence of model performance on parameter and forcing error supports the use of a probabilistic approach to evaluate LSMs. To evaluate LSMs, we propose three metrics that harness the information contained in ensemble runs of an individual model. This paper introduces the metrics themselves; E. Rosero et al. (Evaluating enhanced hydrological representations in Noah-LSM over transition zones: Part 2. Ensemble-based model evaluation, submitted to *Journal of Hydrometeorology,* 2008b) apply the metrics presented here as part of an in-depth model intercomparison.

2. Example Application

[5] To demonstrate the new framework for LSM evaluation, we use an example application. We run three versions of the Noah LSM [*Ek et al.*, 2003] using meteorological

[1] Department of Geological Sciences, University of Texas at Austin, Austin, Texas, USA.

[2] Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park, Pennsylvania, USA.

0094-8276/08/2008GL033721$05.00

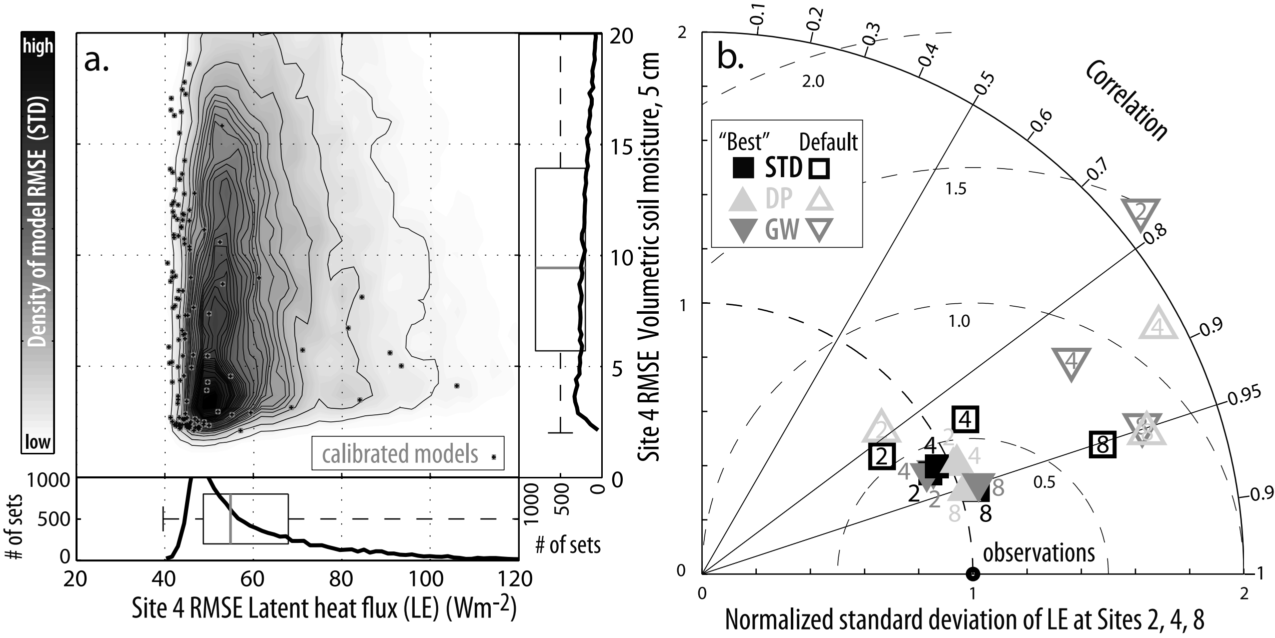

Figure 1. The range of scores obtained by varying effective parameters. (a) RMSE for 5-cm volumetric soil moisture and LE for 15,000 runs of STD at IHOP site 4. Each simulation used unique parameter sets randomly selected from uniform distributions. Also shown are calibrated-model scores. (b) Taylor diagram of LE simulated by DP, STD, and GW at 3 IHOP sites (2, 4, 8). 'Best' is the set that minimizes the L2 norm of the 5 objectives.

forcing data from the 2002 International H_2O Project (IHOP) [*LeMone et al.*, 2007] at sites covered by dry grassland (site 2), semi-arid pasture (site 4) and semi-humid grassland (site 8) (mean annual precipitation [MAP] = 540, 740, 880 mm y^{-1}, respectively). The standard version of Noah ('STD') is the benchmark against which we evaluate two newer versions: one augmented with a lumped, unconfined aquifer model ('GW') [*Niu et al.*, 2007] and a second augmented with a short-term dynamic phenology module ('DP') that allows leaf area to change in response to environmental variation on daily to seasonal time scales [*Dickinson et al.*, 1998].

[6] We evaluate model performance against independent objectives: 3-hour-running mean evaporative fraction (EF); top 30-cm soil wetness (W_{30}); and 24-hour change in wetness (ΔW_{30}). We define EF as:

$$EF_t = LE_t/(LE_t + H_t) \tag{1}$$

where LE_t and H_t, are, respectively, the latent, and sensible heat flux, averaged over 30-minute time interval *t*. We compute W_{30} as:

$$W_{30} = \sum_{i=1}^{N_{layers}} \theta_i z_i \Bigg/ \sum_{i=1}^{N_{layers}} \omega_i z_i \tag{2}$$

where θ_i, z_i, and ω_i are, respectively, the volumetric soil moisture, thickness, and porosity of the *i*th layer of the soil column, which has N_{layers} layers (for the observations, N_{layers} = 4; for the models, N_{layers} = 2). We represent the 24-hour change in soil-moisture as:

$$\Delta W_{30,t} = W_{30,t} - W_{30,t-47}. \tag{3}$$

[7] For each site, we generate two 150-member ensembles: (1) a 'calibrated ensemble,' generated using parameters defined by the Markov Chain Monte Carlo sampler of *Vrugt et al.* [2003] while simultaneously minimizing five RMSE objectives (LE, H, ground heat flux, 5-cm soil temperature and moisture); and (2) an 'uncalibrated ensemble,' composed of runs from a subset of 15,000 generated by random sampling of uniform independent parameter distributions; the subset was defined as the group that obtained scores within one standard deviation of the mode for each RMSE (i.e., the most frequent error) (Figure 1). In this example, model performance varies widely when parameters are selected within reasonable ranges (Figure 1a); after calibration, STD, DP, and GW perform equivalently well (Figure 1b). When generating ensembles, for simplicity, we neglected data uncertainty. All realizations used a 30-minute time step to simulate 01/01/2000–06/25/2002. We treated the first 2.5 years of simulation as model spin-up; only the last 45 days of each simulation were scored.

3. Performance, Robustness, and Fitness Scores

[8] To define a metric that identifies the 'best' model or 'best' parameterization, we first define a good model. Given a representative ensemble, when the LSM is good:

[9] 1. The ensemble brackets most of the observations.

[10] 2. The ensemble is centered on the observations.

[11] 3. The ensemble has low spread when bracketing observations but high spread when not (i.e., the ensemble does not resolutely cling to an incorrect value).

[12] 4. The model is relatively insensitive to parameters that are not observable and is also not significantly affected by errors in meteorological forcing data.

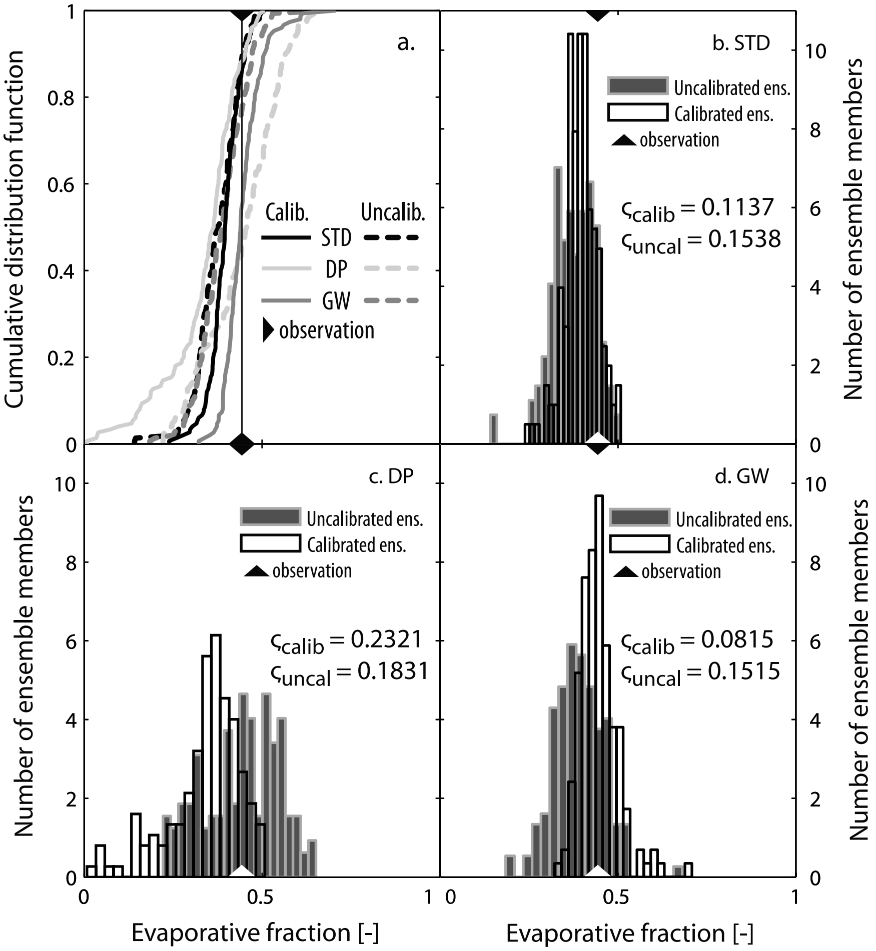

Figure 2. Distribution of EF simulated for Site 2 at 2:00 PM on Julian day 161 (2002). (a) CDFs of the calibrated and uncalibrated ensembles. (b)–(d) Performance scores. The performance score improves as the model is better described by Descriptors 1–3 (see text).

[13] 5. The model performance is consistently good (as defined by Descriptors 1–3) across sites.

[14] Descriptors 1–3 describe a model that is well suited to a given location; 4 and 5 describe a model that is robust [*Carlson and Doyle*, 2002]. The ideal model for use over regional and global domains will match all five Descriptors.

[15] At time step t, the ensemble of the best-scoring model minimizes the performance score, ζ_t:

$$\varsigma_t = \left(CDF_{ens,t} - CDF_{obs,t}\right) \Big/ \left(1 - \left(CDF_{\overline{obs}+c}\right)\right) \tag{4a}$$

where $CDF_{ens,t}$, $CDF_{obs,t}$, and $CDF_{\overline{obs}+c}$ are, respectively, the cumulative distribution functions of the variable simulated by the ensemble of models, of the observation at time t, and of all values of $\bar{o}_t$, shifted by arbitrary constant c (to prevent division by zero). $\bar{o}_t$ is the mean of all realizations of the observation at time step t. When observational uncertainty is unknown, for simplicity, ζ_t can be expressed in deterministic form as:

$$\varsigma_t = \sum_{i=1}^{N_{ens}} |x_{i,t} - o_t| \Bigg/ \sum_{i=1}^{N_{ens}} |\bar{o} - c| \tag{4b}$$

where $x_{i,t}$ is ensemble member i at time t, o_t is the observation at time t, N_{ens} is the number of ensemble members, c is an arbitrary constant that is less than all values o_t, and $\bar{o}$ is the mean of the observations. ζ_t is lowest (best) when the ensemble brackets most observations, has low spread, and is centered on the observations (Figure 2). Figure 3 shows the ensemble's time-varying performance and the corresponding ζ_t. Table 1 demonstrates that ζ_t encompasses both the commonly used ensemble spread and skill [e.g., *Talagrand et al.*, 1997].

[16] Overall insensitivity to factors that may significantly alter performance (e.g., poorly known parameters or errors in meteorological forcing data) can be expressed as:

$$\rho = \frac{|\bar{\varsigma}_{e1} - \bar{\varsigma}_{e2}|}{\bar{\varsigma}_{e1} + \bar{\varsigma}_{e2}} \tag{5}$$

where $\bar{\varsigma}_{e1}$ and $\bar{\varsigma}_{e2}$ are the time means of the performance scores for the first and second ensembles, respectively. The two ensembles should significantly differ in the way(s) in which modelers wish to assess robustness. For example, to test robustness with respect to parameter variation, the ensemble members should use parameter sets that come

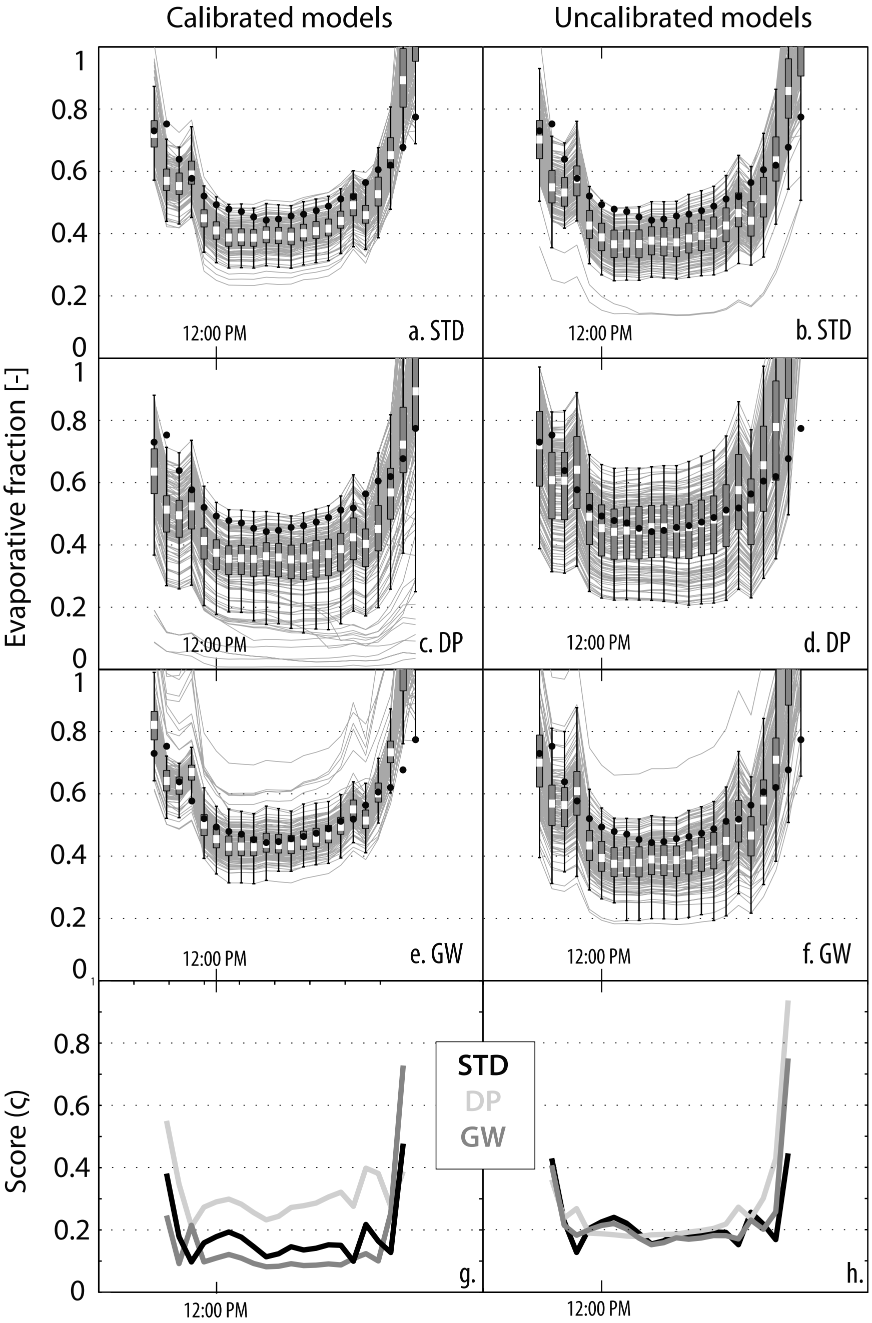

Figure 3. Relation between time-varying ensemble simulations of EF and time-varying performance score (ζ). EF simulated by each ensemble is shown for Site 2 for Julian day 161 (2002). Black circles are observations; gray lines are individual ensemble members; white bars are ensemble mean; gray bars are the ensemble interquartile range.

Table 1. Ensemble Spread, Skill, and Performance Score for EF Simulation at IHOP Site 2 at 2:00 PM Local Time on Julian Day 161 of 2002[a]

	Spread[b]		Skill[c]		Performance Score (ζ)	
	Calibrated	Uncalibrated	Calibrated	Uncalibrated	Calibrated	Uncalibrated
STD	0.00207	0.00370	0.00226	0.00470	0.114	0.154
DP	0.0105	0.0113	0.0109	6.07e−7	0.232	0.183
GW	0.00291	0.00504	9.22e−6	0.00293	0.0815	0.152

[a] See also Figure 2.

[b] Ensemble spread (π_t) is $\pi_t = \frac{1}{N_{ens}-1}\sum_{i=1}^{N_{ens}}(x_{i,t} - \bar{x}_t)^2$, where $\bar{x}_t$ is the ensemble mean at time t, $x_{i,t}$ is the ith ensemble member at time t, and N_{ens} is the number of ensembles.

[c] Ensemble skill (κ_t) is $\kappa_t = (\bar{x}_t - o_t)^2$, where o_t is the observed value at time t.

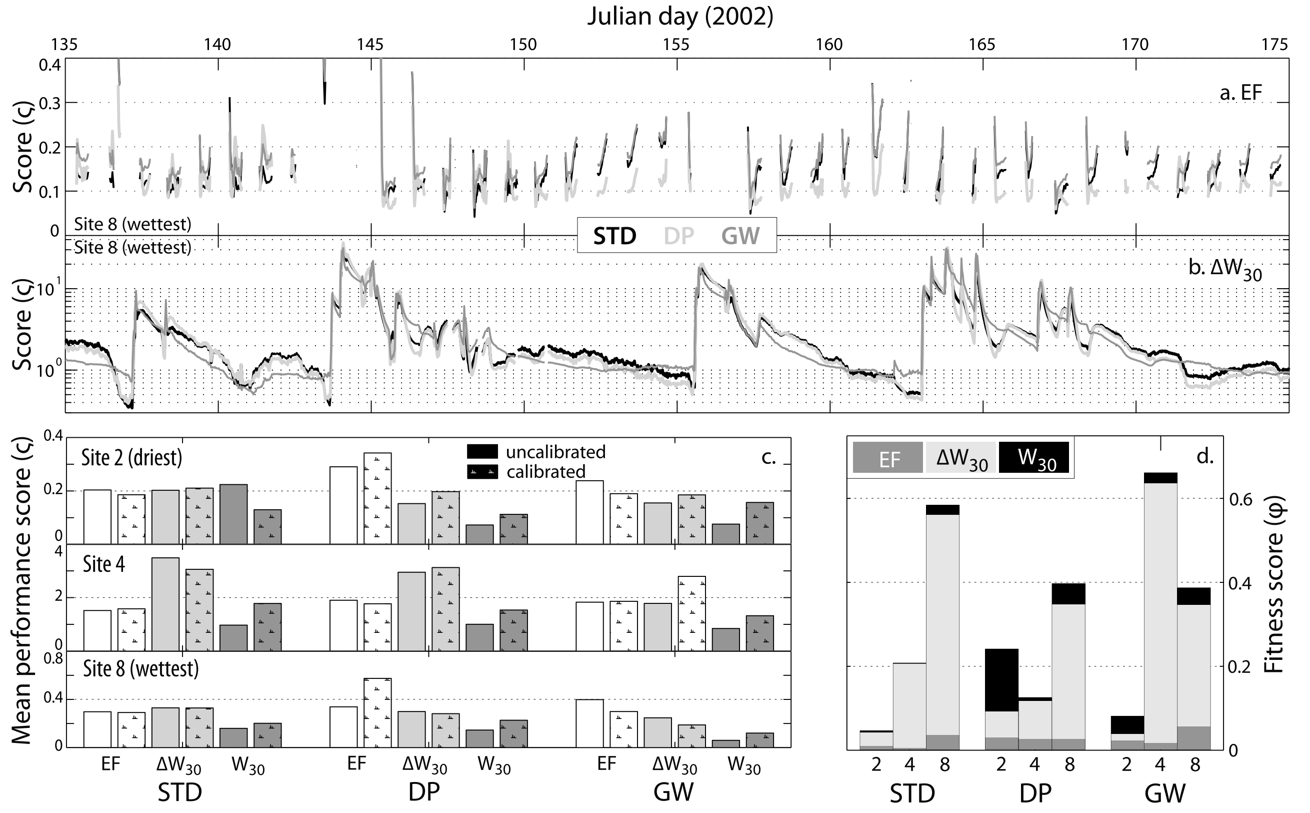

Figure 4. Utility of metrics. Time-varying performance score (ς) of STD, DP, and GW for (a) EF and (b) 24-hour change in wetness (ΔW_{30}). (c) Model performance when simulating EF, ΔW_{30}, and ΔW_{30}; change between the calibrated and uncalibrated ensemble performance scores indicates model robustness. (d) Model fitness scores for each model, criterion, and site.

from distinct distributions (as described above); to test robustness with respect to data error, the ensembles should differ in the type or level of noise by which their input data is perturbed.

[17] Given the assumption that spatially varying characteristics of the land surface shape surface-to-atmosphere fluxes and near-surface states [e.g., *Dickinson*, 1995], we note that there is an inherent contradiction between Descriptors 4 and 5 above. A tradeoff exists between a model that is completely insensitive to parameter variation and one that consistently does well across sites. A 'compromise ideal' model is insensitive to the parameters that cannot be easily identified; it is at least somewhat sensitive to parameters that are physically realistic (e.g., vegetation type) and when model and measurement scales are similar.

[18] For a given site and criterion, a model's overall fitness score, ϕ, quantifies its suitability for broad application. We define ϕ as:

$$\phi = \bar{\varsigma}\rho \tag{6}$$

where $\bar{\varsigma}$ is the mean performance score (Equation 4) of the most representative ensemble and ρ is robustness (Equation 5). The more sites and criteria used, the more confident we can be of model performance (Descriptor 5). The best model (m) from a set of M models tested across N sites minimizes $\frac{1}{N}\sum_{i=1}^{N} \phi_{i,m}$. For fair cross-criterion comparison, modelers should first rank ϕ for a given criterion and should then compare average fitness rankings of the models.

[19] Figures 4a and 4b show the utility of this framework for improving LSM physical structure. DP, which is different from benchmark model STD only in that it allows leaf area index to vary over short time scales in response to environmental variation, consistently better simulates EF at semi-humid site 8 (Figure 4a), a result that is consistent with the hypothesis that short-term phenology can alter surface energy partitioning (see Rosero et al. (submitted manuscript, 2008b) for detailed analysis). DP tends to have the best ς_t when simulating 24-hour change in wetness (Figure 4b and Table 2). When used with other model output characteristics (e.g., anomalies, bias), ς_t helps determine when and where the model is most likely to succeed or fail. It can also be used across criteria (Figure 4c); the combination of ς and ρ yields an assessment of overall model fitness (Figure 4d and Table 2). Overall, GW performs best but is less robust than STD. STD is the fittest model, but tradeoffs in fitness between criteria make all models equivalently fit at the most-humid site 8 (Table 2).

4. Summary and Implications

[20] We introduce three metrics for rigorous and realistic evaluation of LSM performance within a framework that explicitly acknowledges perennial sources of LSM output uncertainty (e.g., sensitivity of output to parameters that are

Table 2. Model Fitness and Average Ranking[a]

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

[a] The rank of model fitness score among the cohort for each given site and criterion is shown in parentheses.

impossible to specify and to errors in meteorological forcing data). The model performance score (ζ) quantifies the likelihood that a representative model ensemble will be highly skilled with low spread; the robustness score (ρ) quantifies the sensitivity of model performance to changes in parameters (as shown in the example here) and/or perturbations to meteorological forcing. Our framework treats the relative insensitivity of an LSM to both parameter variability and to forcing error as beneficial characteristics in the face of the less-than-perfect settings in which LSMs are applied. The fitness score (ϕ) combines the concepts of good performance and robustness and is used to rank models' suitability for broad application.

[21] We demonstrate the use of the metrics using three versions of the Noah LSM to simulate summer in Oklahoma. Our example shows that the least complex version of Noah is most fit for broad application. We use the time-varying ζ (a tool for model evaluation and development) to show that allowing leaf area index to vary on short time scales improves Noah's simulation of surface energy partitioning and subsurface water dynamics at the semi-humid site. Standard computational resources are now such that the presented framework can be applied to several models (or a single model with candidate parameterizations) and numerous flux tower sites as a means for more thorough and informative model evaluation: on a single 2.66 GHz processor, to run one model for 2.5 years 15,000 times (as we did for each Monte Carlo sampling) required less than 2 hours of computing time.

[22] Researchers are often quick to assume that the model is performing well because it is more complex (i.e., has more parameters), a characteristic often equated with increased physical realism. This method for evaluation should not be used to 'prove' that one representation is more physically realistic than another. Although it is likely that improved conceptual realism will improve model performance, the converse is not necessarily valid. Regardless, the models examined here have so many degrees of freedom that it is difficult to parse whether strong model performance is the result of cancelling errors or the result of physical correctness; however, the performance and fitness scores presented here allow for hypothesis testing and model development that gives a realistic treatment to uncertainty. Because the results are obtained using ensemble simulations, we can be more confident that this improvement is indeed the result of the altered model structure and is not the simple result of a lucky guess of parameters.

[23] Land-surface modelers are unlikely to ever know the 'right' parameters for any site on which their models are applied; they will not be able to eliminate error in meteorological forcing data. Evaluation of LSM performance must take a realistic view of these limitations. The metrics proposed here enable more objective comparison of LSM performance in a framework that more accurately represents the ways in which LSMs are applied. Use of this framework and metrics will strengthen modelers' conclusions and will improve confidence in LSM predictions.

[24] **Acknowledgments.** We thank the Texas Advanced Computing Center, NOAA (grant NA07OAR4310216), the NSF, the NWS/OHD, and the Jackson School of Geosciences; we are also grateful for the insightful comments of two anonymous reviewers.

References

Abramowitz, G. (2005), Towards a benchmark for land surface models, *Geophys. Res. Lett.*, *32*, L22702, doi:10.1029/2005GL024419.

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

Carlson, J. M., and J. Doyle (2002), Complexity and robustness, *Proc. Natl. Acad. Sci. U. S. A.*, *99*, Suppl. 1, 2538–2545.

Chen, F., et al. (2007), Description and evaluation of the characteristics of the NCAR high-resolution land data assimilation system, *J. Appl. Meteorol. Climatol.*, *46*, 694–713, doi:10.1175/JAM2463.1.

Dickinson, R. E. (1995), Land-atmosphere interaction, *Rev. Geophys.*, *33*, 917–922.

Dickinson, R. E., et al. (1998), Interactive canopies for a climate model, *J. Clim.*, *11*, 2823–2836.

Ek, M. B., K. E. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno, and J. D. Tarpley (2003), Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model, *J. Geophys. Res.*, *108*(D22), 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 (2007), 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.

Henderson-Sellers, A., K. McGuffie, and A. J. Pitman (1996), The Project for Intercomparison of Land-Surface Parameterization Schemes (PILPS): 1992 to 1995, *Clim. Dyn.*, *12*, 849–859.

LeMone, M. A., et al. (2007), NCAR/CU surface, soil, and vegetation observations during the International H_2O Project 2002 Field Campaign, *Bull. Am. Meteorol. 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*(D20), 4443, doi:10.1029/2001JD000931.

Nijseen, B., and L. A. Bastidas (2005), Land-atmosphere models for water and energy cycle studies, in *Encyclopedia of Hydrological Sciences*, vol. 5, part 17, edited by M. G. Anderson, chap. 201, pp. 3089–3102, John Wiley, Hoboken, N. J.

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.

Niu, G.-Y., Z.-L. Yang, R. E. Dickinson, L. E. Gulden, 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.

Randall, D. A., et al. (2007), Climate models and their evaluation, in *Climate Change 2007: The Physical Science Basis. Contribution of Working Group 1 to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by S. Solomon et al., Cambridge Univ. Press, Cambridge, U.

Talagrand, O., R. Vautard, and B. Strauss (1997), Evaluation of probabilistic prediction systems, paper presented at ECMWF Workshop on Predictability, Eur. Cent. for Med. Range Weather Forecasts, Reading, U. K.

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*(8), 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.

L. E. Gulden, E. Rosero, Z.-L. Yang, and G.-Y. Niu, Department of Geological Sciences, 1 University Station C1100, University of Texas at Austin, Austin, TX 78712-0254, USA. (liang@mail.utexas.edu)

T. Wagener, Department of Civil and Environmental Engineering, The Pennsylvania State University, 212 Sackett Building, University Park, PA 16802, USA.