The Dissertation Committee for Enrique Xavier Rosero Ramirez certifies that this is the approved version of the following dissertation:

Evaluating Enhanced Hydrological Representations in Noah LSM over Transition Zones: An Ensemble-based Approach to Model Diagnostics

Committee:

Zong-Liang Yang, Supervisor

M. Bayani Cardenas

Rong Fu

Charles S. Jackson

Guo-Yue Niu

Mrinal Sen

Evaluating Enhanced Hydrological Representations in Noah LSM over Transition Zones: An Ensemble-based Approach to Model Diagnostics

by

Enrique Xavier Rosero Ramirez, B.S.; M.S.; M.E.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 2009

Acknowledgements

Major support for my work was provided by the NOAA National Weather Service Graduate Fellowship of the Hydrology Training Program of the Office of Hydrologic Development. The project was funded in part by a grant of the NOAA Climate Prediction Program for the Americas. I also received support for my research from the NASA Large-Scale Biosphere-Atmosphere Experiment in Amazonia and the Jackson School of Geosciences. Computational resources were provided by the Texas Advanced Computing Center.

I want to acknowledge the continued encouragement and the freedom to navigate in the gap between model development and model identification that my advisor, Dr. Zong-Liang Yang, gave me. I benefited from his experience, multitudinous memory and generosity. The insight on model conceptualization and codes of Dr. Guo-Yue Niu and Dr. Lindsey E. Gulden were critical for my research. I am especially grateful for her keen opinions and friendship. The generous and candid advice of Dr. M. Bayani Cardenas, Dr. Charles S. Jackson, Dr. Rong Fu, and Dr. Mrinal Sen, members of my committee helped me immensely.

During my career in academia my scientific thinking was challenged by a number of researchers including P.D. Dr. rer.nat. Ulf Mohrlok, Dr. Luis A. Bastidas, Dr. Mac McKee, Dr. Christa Peters-Lidard, Dr. Kenneth E. Mitchell, Dr. Fei Chen, Dr. Hoshin V. Gupta, and Dr. Pedro Restrepo. I had the opportunity to collaborate and interact with future leaders in the field of Hydrology and Earth System Sciences like Dr. Thorsten Wagener, Dr. David J. Gochis, Dr. Luis Gustavo G. de Goncalves, Dr. Enrique R. Vivoni, Dr. Soni Yatheendradas, Dr. Yasir H. Kaheil, Dr. Bethany T. Neilson, and Dr. Jasper A. Vrugt. To them I express my gratitude.

A special recognition goes to the coauthors of my manuscripts for their contributions in both substance and style. I am grateful for their time, energy and resources.

I am indebted to Philip Guerrero for keeping me in compliance with the regulations of the program.

Evaluating Enhanced Hydrological Representations in Noah LSM over Transition Zones: An Ensemble-based Approach to Model Diagnostics

Publication No._____

Enrique Xavier Rosero Ramirez, Ph.D.
The University of Texas at Austin, 2009

Supervisor: Zong-Liang Yang

This work introduces diagnostic methods for land surface model (LSM) evaluation that enable developers to identify structural shortcomings in model parameterizations by evaluating model 'signatures' (characteristic temporal and spatial patterns of behavior) in feature, cost-function, and parameter spaces. The ensemble-based methods allow researchers to draw conclusions about hypotheses and model realism that are independent of parameter choice.

I compare the performance and physical realism of three versions of Noah LSM (a benchmark standard version [STD], a dynamic-vegetation enhanced version [DV], and a groundwater-enabled one [GW]) in simulating high-frequency near-surface states and land-to-atmosphere fluxes *in-situ* and over a catchment at high-resolution in the U.S. Southern Great Plains, a transition zone between humid and arid climates. Only at more humid sites do the more conceptually realistic, hydrologically enhanced LSMs (DV and

GW) ameliorate biases in the estimation of root-zone moisture change and evaporative fraction. Although the improved simulations support the hypothesis that groundwater and vegetation processes shape fluxes in transition zones, further assessment of the timing and partitioning of the energy and water cycles indicates improvements to the movement of water within the soil column are needed. Distributed STD and GW underestimate the contribution of baseflow and simulate too-flashy streamflow.

This work challenges common practices and assumptions in LSM development and offers researchers more stringent model evaluation methods. I show that, because of equifinality, *ad-hoc* evaluation using single parameter sets provides insufficient information for choosing among competing parameterizations, for addressing hypotheses under uncertainty, or for guiding model development. Posterior distributions of physically meaningful parameters differ between models and sites, and relationships between parameters themselves change. 'Plug and play' of modules and partial calibration likely introduce error and should be re-examined. Even though LSMs are 'physically based,' model parameters are effective and scale-, site- and model-dependent. Parameters are not functions of soil or vegetation type alone: they likely depend in part on climate and cannot be assumed to be transferable between sites with similar physical characteristics.

By helping bridge the gap between the model identification and model development, this research contributes to the continued improvement of our understanding and modeling of environmental processes.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1. SCIENTIFIC SETTING AND MOTIVATION

The land surface plays a key role in the energy and water cycle and the larger climate system. By exchanging fluxes of heat, momentum and moisture with the overlying atmosphere, topographic features, seasonal vegetation cover, water stored in the ground, etc. shape weather and climate (e.g., Pielke Sr., 2001). On timescales longer than a day, anomalies in land-surface states and fluxes propagate to the atmosphere (e.g., Childs et al., 2006). This land−atmosphere coupling is thought to be particularly strong in zones of transition between wet and dry climates (Koster et al., 2004). Our ability to predict weather and climate on seasonal and interannual timescales therefore depends on our ability to quantitatively understand and accurately represent land-surface processes such as evaporation, transpiration, soil moisture dynamics, and runoff. Land-surface models (LSMs) are the numerical representation of scientific hypotheses about how different terrestrial bio-hydrological processes determine the partitioning and temporal evolution of fluxes of water and heat and their dynamical interactions with the atmosphere (Viterbo, 2002; Pitman et al., 2003; Yang, 2004; Nijssen and Bastidas, 2005; Overgaard et al., 2006). The overarching goal that motivates the research described within this dissertation is to improve the scientific community's ability to understand, model, and predict the hydrologic cycle in transition zones over short timescales.

Hydrologic models are used to synthesize past events, to predict future events, and to evaluate the effects of change on a system. The development, application, and evaluation of environmental models make up a continual and dynamic process that itself

helps researchers to identify and understand system feedbacks and interactions (Refsgaard and Henriksen, 2004). Researchers work to increase the physical realism of models as a means for increasing confidence in a model's prediction when boundary conditions change (e.g., in future prediction) (e.g., Maxwell and Miller, 2005). LSMs are a class of hydrologic models that are used to represent flows of energy, water, and momentum between the land and atmosphere and within reservoirs in the land surface. LSMs coupled with atmospheric models are used operationally for weather forecasting and climate predictions (e.g., Chen and Dudhia, 2001). Understanding and consequently representing with accuracy the hydrological processes responsible for land-memory mechanisms, such as the storage of water near the surface as soil moisture and the nature and seasonal progression of growing vegetation, still remain a challenge in land-surface modeling (Shuttleworth, 2007; Trier et al., 2008). Deficiencies in LSM parameterizations provide opportunities to improve numerical weather forecasting and climate prediction (Trenberth et al., 2003; Holt et al., 2006; Lyon et al., 2008).

Different land-surface parameterizations characterize biophysical and hydrological processes that control fluxes of moisture (interception, throughfall, infiltration, runoff and snowmelt), energy (absorption of radiation at the surface, partitioning into latent and sensible heat flux, storage of heat), and momentum (frictional drag of surface on the planetary boundary layer). More complex parameterizations are often credited with improved simulation of modeled states and fluxes (e.g., Wood et al., 1998; Bowling et al., 2003; Niu et al., 2009); however, as model complexity increases, parameter estimation becomes increasingly important. For example, Stöckli et al. (2008)

evaluated the latest enhancements made to the Community Land Model (Oleson et al., 2008a) offline at a point scale using a number of the FLUXNET (Baldocchi et al., 2001) stations around the world and found that better simulations of the hydrological cycle chiefly translate into improvements in the simulation of latent heat flux. However, they noted that the persistence of bias may result from remaining deficiencies in the parameterizations, missing processes, and/or a lack of tuning of parameters. Even so, LSM parameters are very frequently assumed to be physical quantities (not tunable coefficients) that can be measured and that have strong relationships with physical properties of the system.

The second phase of the Project for Intercomparison of Land-Surface Parameterization Schemes (PILPS) recognized that even simple manual, subjective adjustment of parameters can significantly improve model performance (Pitman et al., 1999). It is often not established by how much performance could be improved with parameter calibration relative to the improvement that could be gained by modifying model structure. For example, Niyogi et al. (2006) modified parameterizations of canopy resistance, which resulted in improved simulation of forecasted air temperature and moisture; similar improvements in controlling respiration rates can be attained by using fine-tuned parameter values (Demarty et al., 2004). Calibration has been shown to reduce errors in simulated heat fluxes by 20 to 40% at different locations around the world (Nijssen and Bastidas, 2005). Leplastrier et al. (2002) showed that, although the most complex surface-energy-balance parameterizations perform best after calibration, the relative improvement over the simpler parameterizations is minimal; the researchers

question the benefit of using more complex representations in the absence of calibration data. Hogue et al. (2006) calibrated LSMs of increasingly detailed physical parameterizations and showed that additional complexity (presumably added as means for increasing the conceptual physical realism of the model) neither necessarily improves model performance nor reduces the uncertainty in the simulated fluxes of water and energy. They suggest that only when the new parameterization can be supported and identified with available observations should the additional complexity be employed.

This dissertation comprises research endeavors that, accounting for uncertainty, evaluate the physical realism of several recently introduced representations of land-surface processes at both a point and a catchment scale. I employ the extensively used Noah LSM (Ek et al., 2003) to investigate the importance of short-term vegetation processes and aquifer dynamics in determining seasonal variation of land-surface fluxes and states. The work described here has direct relevance for weather and climate prediction, water resources assessment, and flood modeling.

1.2. MODEL EVALUATION: MODEL IDENTIFICATION VERSUS MODEL DIAGNOSTICS

LSM evaluation and development is the dynamic assessment of hypotheses about our understanding of the dominant physical mechanisms of the soil-vegetation-atmosphere system. Like other environmental models built to support scientific reasoning and testable hypotheses to improve our understanding of the Earth system, LSMs have grown in sophistication and complexity (Pitman, 2003; Niu et al., 2009). The evaluation of LSM simulations is consequently non-trivial and, especially when LSMs are to be used

in predictive mode for operational forecasting, policy assessments, or decision making, demands more powerful methods for the analysis of their behavior (Saltelli, 1999; Jakeman et al., 2006; Randall et al., 2007; Gupta et al., 2008; Abramowitz et al., 2009). For the most part, parameter estimation techniques have not been extensively used to inform LSM development even though hydrological modelers regard model calibration as a necessary step (Klemes, 1986; Kirchner et al., 1996; Gupta et al., 2005; Wagener and Gupta, 2005; Refsgaard et al., 2006). Too often, attempting to increase the realism of the parameterizations in LSMs has not been subject to rigorous model performance evaluation (Randall et al., 2007; Gupta et al., 2008). In rigorous evaluation of environmental models, an environmental model is iteratively conceptualized, identified, calibrated, and validated in methodical fashion, and meticulous assessment of model and data uncertainty is made throughout the modeling process (Beck, 1987; Jakeman et al., 2006; Refsgaard et al., 2007). In part because of a dearth of information for validation, insufficient computational power to rigorously assess large-domain models, and a pervasive belief that 'physically-based' models do not require as much parameter tuning, standard practice in LSM evaluation has proceeded in a more *ad hoc* fashion. Individual modeling groups publish model results and model development work without a rigorous assessment of true predictive uncertainty (Table 1.1).

The most concerted efforts to evaluate LSMs have been in the form of model intercomparison projects (MIPs) such as the PILPS family of research (Henderson-Sellers et al., 1993; Henderson-Sellers et al., 1995; Pitman and Henderson-Sellers, 1995; Liang et al., 1998; Pitman et al., 1999; Luo et al., 2003; Bastidas et al, 2007). MIPs depend on

the voluntary participation of members of the modeling community. Each modeling group is asked to submit one or a few model simulations for a given site/domain and time period, given prescribed meteorological forcing and ancilliary datasets. MIPs then compare output of models frequently using aggregate goodness-of-fit metrics (Legates and McCabe, 1999) or mean multiannual and seasonal bias assessments. MIP experimental designs compare model versus observation at different timescales (frequently coarse), and then they make an inference about the quality of the model. Although valuable, model-versus-observation comparisons often make it hard to infer causality. Major conclusions obtained by LSM MIPs have been that single-layer soil moisture schemes ('bucket' models) are insufficiently complex to represent hydrologic processes, that the scatter in the partitioning of energy and water fluxes among models is significant, and that, while individual land-surface schemes capture specific aspects of the cycles with reasonable accuracy, no one scheme captures the whole system satisfactorily and consistently (Pitman et al., 2003; Fox et al., 2006).

My own involvement in the PILPS 2(g) 'San Pedro' (Bastidas et al., 2006b; Rosero and Bastidas, 2007; Bastidas et al., 2007) and the LBA-MIP (Rosero et al., 2007; de Goncalves et al., 2008; Saleska et al., 2008) taught me that, although the administrative work in coordinating inputs from a diverse set of participants and a diverse group of models is significant, conclusions generated by MIPs are often subject to big uncertainties, are qualitative and general, and are unable to provide a direction for model improvement, development, or scientific advancement. MIPs are limited for several reasons. Participating groups vary in their willingness to spend time tuning their model to

the given location(s): that one model's output is 'better' than another's is often fortuitous and may not be a result of an inherently more realistic structure. Even if runs are directly comparable, MIPs are by nature formulated in a way that makes it difficult to pose questions and test hypotheses. A MIP is more of a poll of models (or a 'beauty contest') than a structured, quantitative assessment of individual model function. Additionally, in large part due to limited data, computational power, and available labor, MIPs have evaluated models by comparing time-averaged variables (e.g., annual or monthly mean temperature, monthly total evapotranspiration). Because the timescale at which the models' simulated variables are compared (monthly, annually) is often far coarser than the timescales at which the models operate (hourly or less), conclusions about model function are at best qualitative and generalized, and do not often provide much insight into how models can be improved or why model results differ.

Model identification (Step 7, Table 1.1) is seen in other hydrologic communities (e.g., the rainfall−runoff modeling community) as a process of uncertainty reduction (Wagener and Gupta, 2005; Wagener et al., 2009). It is commonly accepted that model (structure, parameters and states; also, initial and boundary conditions) and data (measurements of forcing and response) will all contain uncertainties that can affect the model predictions. These uncertainties stem from various sources and relate to our capacity to understand and measure the real-world system under study (perceptual/conceptual model uncertainty), the data (measurements errors, or the lack thereof), and the mathematical/numerical model and its components (Gupta et al., 2005). Model uncertainty is parsed into parameter-estimation uncertainty (i.e., the inability to

uniquely define a 'best' parameter set) (Gupta et al., 1998) and model structural error, which is introduced through simplifications, inadequacies, and ambiguity in the representation of real-world processes (Beck, 2002).

Current procedures for *a priori* parameter estimation are often based on semi-empirical relationships between model parameters and land (or basin) physical characteristics (i.e., soils, vegetation, topography, climate, geology, etc.). Available information about soils (e.g., texture) and vegetation (e.g., type or vegetation index) only indirectly relates to model parameters such as the parameters representing the hydraulic properties of soils (e.g., Clapp and Hornberger, 1978) and the parameterized rooting depths of vegetation (Duan et al., 2006; Wagener et al., 2006). Whether physically based model parameters are measurable physical characteristics or calibrated 'effective' quantities continues to be debated (e.g., Bastidas et al., 2006; Hogue et al., 2006). Duan et al. (2006) points out, "Estimation of hydrologic model parameters is, at present, highly problematic. Ultimately, we must deal with the fact that our models are imperfect and that one of the roles of model parameters is to 'fit' the model to the real world."

In well calibrated models: (1) the input–state–output behavior of the model is consistent with the measurements of system behavior; (2) model predictions are accurate (i.e., they have negligible bias) and precise (i.e., the prediction uncertainty is relatively small); and (3) the model structure and behavior are consistent with a current hydrological understanding of reality. (3) is often overlooked in operational settings, where the focus is generally on models that are 'useful' rather than on models that are realistic (Gupta et al. 2005; Wagener and Gupta, 2005). The ability of a parameter set to

help the model reproduce the observed system response is measured (summarized) by means of an 'objective function' (loss or cost function), which, typically, is an aggregated measure of the residuals, which are the differences between observed and simulated responses at each time step. In automatic model calibration (parameter estimation) the minimization of the objective function leads to the identification of 'optimal' parameters. In any modeling study of reasonably complex environmental systems, multiple models (and parameters) that provide predictions consistent with available observations can be found by means of calibration (Beven and Freer, 2001). Calibration results in model structure and parameter values that are either realistic or that are both unrealistic but that contain errors that compensate for one another (Kirchner et al., 1996). Furthermore, because degrees of freedom that are not constrained may worsen as a result of a rearrangement in the model structural error during calibration (Leplastrier et al., 2002), verification against datasets that are functionally equivalent to the training data can dramatically increase the number of false positives.

Decades of research into appropriate methods for hydrological model identification under uncertainty have evolved from methods to identify a 'best' model (e.g. Duan et al., 1992), toward attempting to identify all models (or model structures) that are consistent with the observed system behavior (e.g., Gupta et al., 1998; Boyle et al., 2000; Vrugt et al., 2003; Beven, 2006). Note that because the model structural space is infinite and contains no 'true' model structure, it is only possible to find a currently 'best' or 'acceptable' (i.e., 'behavioral') set of model structures by comparing each to all available observations.

Model identification approaches tend to be oriented toward finding the 'best' model (by adjusting it to best explaining the data) rather than toward understanding in which ways the model is inconsistent with the observed behavior of the natural system (Gupta et al, 2005; Wagener and Gupta, 2005). Such an evaluation framework is weak in the diagnostic sense (Gupta et al., 2008). A diagnostic approach helps determine components of a model, which when assumed to work properly, can explain the discrepancy between simulations and observations. According to Gupta et al. (2008), "At its strongest, a diagnostic evaluation will point toward the aspects of the model that need improvement, and give guidance toward the manner of improvement." This approach goes beyond identifying models that conform with the data and enables us to draw conclusions about causality not merely based on correlation (i.e., if a better fit is found then the model is superior). Assert Gupta et al., "A causal diagnostic, however, is one where the underlying theory can be used to actually predict the (observable) impact of system changes (or defects), and similarly to infer various possible causes of an observable system response (or deviation thereof)." Very recently this philosophy has begun to be applied to rainfall−runoff modeling (Yilmaz et al., 2008; Bai et al., 2009).

It is noted that a diagnostic approach is distinct from the quality assurance approach, in which a model, which is recognized as imperfect, is calibrated or bias corrected and the uncertainty bounds of its predictions are quantified (Refsgaard et al., 2005). Quality assurance approaches are best suited to applications and operations, not to model development or scientific research.

Having gleaned what is possible out of the standard MIP approach, the LSM development community must move toward a diagnostic framework of model evaluation. Such an evaluation framework should focus on testing hypotheses underlying models via the evaluation of 'signatures' (i.e., characteristic behaviors of the observed system), and accounting for sources of uncertainty (e.g., by using all the models that best conform with the observed system behavior), thereby bridging the gap between model identification and model development.

1.3. OVERVIEW OF WORK PRESENTED HERE

The work presented here was undertaken to evaluate which of three versions of the Noah LSM (the benchmark standard, a dynamic-vegetation enhanced version, and a groundwater-enabled one) better represents the near-surface land-to-atmosphere fluxes and states in transition zones, both in terms of accuracy and in terms of insensitivity to parameter and data uncertainty. The results challenge typical assumptions made as part of standard LSM evaluation practices. Looked at in full, my dissertation presents a new framework in which models are evaluated using a diagnostic approach that analyzes a model's typical behavior – or 'signature' – in cost-function space, parameter space, and feature space. This evaluation focuses on testing hypothesis behind the implementation of the models, which allow me to diagnose deficiencies in their implementation and make conclusions about the importance of short-term vegetation processes and aquifer dynamics in transition zones.

I use an ensemble approach to explicitly account for uncertainty. I identify a representative group of alternative model structures that best reproduces the observed data either by training the LSM to best reproduce primary behaviors or by selecting behavioral performing models. The evaluation investigates what structures the model takes in order to be consistent with the observations (i.e., what parameter sets are in the behavioral range), how the relationships between parameters that describe the model functioning act, how different are the model structures between constrained realizations and models with only *a priori* information, what is the typical performance and partitioning of the energy and water cycles, and how well do the models reproduce observed, defined characteristics that summarize the behavior or 'signature' of the observed system. I diagnose potential structural reasons for the shortcomings in the capacity of the model to simulate fluxes and states both in time and space that cannot be attributed to parameter uncertainty.

The following overarching questions are addressed: (1) Are the hypotheses behind the implementation of conceptually realistic enhancements to the hydrological representations of land-surface memory mechanisms adequately supported by (add value to the ability of LSMs to simulate) observed fluxes and near-surface states?; and (2) Faced with parameter uncertainty, how can deficiencies in the model structure be diagnosed and a better model identified?

I evaluate LSM performance in zones of transition between arid and humid climates in the continental U.S. Zones of transition between wet and dry climates are regions in which land-surface memory processes are particularly important to the

determination of weather and climate (Koster et al., 2004; Dirmeyer et al., 2006; Weckwerth and Parsons, 2006; LeMone et al., 2007). The research described in Chapters 2 and 3 uses hydrologic observations from the International H_2O Project (IHOP) 2002 observation campaign, which collected meteorological conditions and high-temporal-resolution surface-to-atmosphere fluxes at nine sites across the Southern Great Plains, USA, a zone of transition between the humid eastern and arid western United States. Research described in Chapter 4 employs data collected at several locations in the 611 km^2 Little Washita River watershed in south-central Oklahoma.

In Chapter 2, a traditional MIP and an ensemble-based MIP are presented. Both MIPs are used to evaluate the ability of three versions of the Noah LSM to represent land surface states and fluxes in the transitional climate of Oklahoma, USA. I demonstrate that the traditional approach to model intercomparison is insufficient to differentiate between the behavior of calibrated competing models in both cost-function space and feature space. I then build on the traditional approach with an ensemble-based method that permits the evaluation of model signatures. That is, I evaluate model behavior based on a model's typical performance (not on the performance of a single model realization) in partitioning of the energy balance and sustaining moisture during dry-down periods. I address following questions: (1) Do newly introduced, enhanced hydrologic paramterizations improve the LSM's capacity to simulate high-frequency turbulent fluxes and soil states? (2) Which versions of the models provide the right answer for the right reasons and why? (3) How reliable are the schemes when faced with parameter uncertainty? The more sophisticated, ensemble-based MIP allows me to reach

conclusions about model structure and performance across sites that are independent of parameter uncertainty, which is most often significant and unavoidable.

In Chapter 3, I focus primarily on model signatures in parameter space, an area that is often overlooked by LSM researchers. I study the effect of choice of parameters on model simulation and investigate how parameters vary by site and by model. I address the following questions: (1) What are the model parameters that contribute most to model variance in transition zones? (2) What are the dominant interactions between model parameters, and how do these change between models? (3) How do behavioral parameters change with dominant physical characteristics of the land? In the process of addressing the models' performance in parameter space, I challenge commonly held assumptions in LSM development practices and demonstrate a detailed method for variance-based quantification of model performance, linking an assessment of model performance in parameter space to that in cost-function space.

In Chapter 4, I move beyond point-scale model evaluation to catchment-scale evaluation to assess the ability of a LSM augmented with a simple groundwater model and a topography related runoff parameterization to simulate an integrated watershed characteristic, streamflow, at a daily timescale on a distributed grid at fine resolution. Building upon the work presented in Chapters 2 and 3, I apply the ensemble-based methods to address the following questions. (1) Does the hydrologically enhanced LSM improve upon the standard model's ability to represent the water cycle? (2) Are the behavioral ensembles of both models able to simulate the essential characteristics of streamflow on a daily timescale? (3) Do the behavioral ensembles accurately partition the

components of streamflow into surface and subsurface components? (4) Do the behavioral ensembles demonstrate improvement of simulations of other characteristics of the water balance (evapotranspiration and soil moisture variation)? The use of powerful, signature-based diagnostic methods to comprehensively evaluate LSMs over distributed domains presented in Chapter 4 is the first application of such techniques in the LSM field and complements the analysis presented in Chapters 2 and 3.

In Chapter 5, I recapitulate the primary conclusions and contributions of the work contained within this dissertation. The research presented here is of fundamental importance for understanding model behavior and the continued development and improvement of land-surface models.

Table 1.1 Comparison of best practices for environmental model development (Jakeman et al., 2006) and typical LSM development

Best practices	Typical LSM development
1. Define model purpose	Far-reaching purposes include: Providing lower boundary conditions to models of atmosphere that are used in numerical weather prediction and climate research; tracking land-surface states and surface-to-atmosphere fluxes under system changes; understanding system feedbacks
2. Specify modeling context (specific questions to be addressed, who will be served by the results, necessary outputs, forcing data, expected accuracy, time and space domains, etc.)	Stakeholders are frequently the broader modeling community, policymakers, or the public interest. In model development applications, questions are often 'Is the new parameterization better than the old?' (Niu et al. 2007b) or 'Is process X important to overall system behavior?' (Gulden et al. 2007). Forcing data is collected from standard, high-quality repositories.
3. Develop a conceptual understanding of the system to be represented, specify data and other prior knowledge (iteratively return to step 2, if necessary)	Researchers typically assume the correctness of the majority of conceptualizations present in existing models (e.g., Noah LSM, the Community Land Model [Oleson et al. 2008]) and then modify one or two aspects of the conceptualization thought to be deficient.
4. Select model features, family, and form of uncertainty specification	LSMs are, as a whole, semi-empirical/semi-theoretical, distributed, deterministic models. Uncertainty specification is often ignored or limited in scope; models are assumed to be correct or nearly correct because they are 'physically based.'
5. Choose method for identifying model structure and parameters. Parsimony should be the standard when selecting model structure.	With the exception of recent multi-model ensemble approaches (e.g., Niu et al. 2009), the model chosen is often assumed to be the best structure either due to modelers' preference or incentive structures that dictate the use of a given model. Parsimony is typically ignored.

6. Choose estimation/performance criteria and algorithm. The parameter estimation criteria should reflect the desired qualities of the model estimates (e.g., infectivity to outliers, etc.). When estimating parameters, the resulting model should be tested against criteria not used for calibration.	Parameter values are most often assumed to be the default or able to be transferred from similar ecosystems but are occasionally tuned using rudimentary calibration methods (e.g., Gulden et al. 2007b). Parameters are transferred between sites with similar characteristics and between models based on an assumption of 'physical realism'. Modelers typically do not allude to parameter interaction as a potential concern.
7. Identify model structure and parameter values (iteratively return to steps 4 and 5, if necessary)	See comments for 5 and 6
8. Verification, including diagnostic testing. "Once identified, the model must be 'conditionally' verified and tested to ensure it is sufficiently robust…It is also necessary to verify that the interactions and outcomes of the model are feasible and defensible, given the objectives and the prior knowledge."	LSM developers typically focus on single or a very few realizations of their models, not exploring the robustness of the model results with respect to parameter uncertainty or input data uncertainty. If the given realization provides feasible outputs for the objectives of interest, the model is deemed to be an improvement.
9. Quantification of uncertainty	LSM developers and researchers typically limit themselves to simple end-member sensitivity analyses or one-at-a-time parameter sensitivity tests. Uncertainty stemming from input data, parameter uncertainty, or structural uncertainty is rarely quantified.
10. Model evaluation or testing (iteratively return to steps 2, 3, 4, 5, and 7, as needed). Ideally this is done using data that were not used to construct the model.	LSM developers apply their model on global scales. Over time and application by multiple modeling groups to varying locations at varying timescales, the strengths and weaknesses of a given model often come to light (e.g. Mitchell et al., 2004). The community as a whole responds to the multi-site, multi-group model evaluation by using such results to target areas for future model improvement.

Chapter 2: Traditional and Ensemble-based model intercomparison[1]

2.1. ABSTRACT

We introduce and compare the performance of the unified Noah land-surface model (LSM) and its augments with physically-based, more conceptually realistic hydrologic parameterizations. We use 45 days of 30-minute data collected over 9 sites in transition zones to evaluate: (1) our benchmark, the standard Noah LSM release 2.7 ('STD'); (2) one equipped with a short-term phenology module ('DV'); and (3) one that couples a lumped, unconfined aquifer model to the model soil column ('GW'). Our model intercomparison, enhanced by multi-objective calibration and model sensitivity analysis, shows that, under the evaluation conditions, the current set of enhancements to Noah fail to yield significant improvement in the accuracy of simulated, high-frequency, warm-season turbulent fluxes and near-surface states across these sites. Qualitatively, the version 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, we compare multiple model realizations to explicitly account for parameter uncertainty. We quantify model performance, robustness, and fitness for use across varied sites. We show that the least complex, benchmark LSM (STD) remains as the most

[1]Significant portions of this chapter were first published as:
Rosero E., Z.-L. Yang, L. E. Gulden, G.-Y. Niu, and D. J. Gochis (2009), Evaluating enhanced hydrological representations in Noah-LSM over transition zones: Implications for model development. *J. Hydrometeor*., 10(3), 600-622 doi: 10.1175/2009JHM1029.1.
Works cited here are referenced in the *References* section of this dissertation.

fit version of the model for broad application. Although GW typically performs best when simulating evaporative fraction (EF), 24-hour 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 latent heat flux (LE) need to be modified for broader applicability. Thus, our work poses a caveat for simple 'plug-and-play' of functional modules between LSMs and showcases the utility of rigorous testing during model development.

2.2. 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 (Nijssen 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 land surface 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; Niu et al., 2009).

Vegetation processes and anomalies in soil moisture provide a source of hydrological memory and 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); improved soil moisture representation within LSMs improves simulation of surface fluxes (Dirmeyer et al., 2000); 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, 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; Niu et al., 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 test 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. Due to the strength of the coupling, our work focuses on warm-season climates in the transition zone of the central U.S..

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), etc. (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. (2008b) 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 2.3. Experimental design and methods for model performance evaluation are explained in section 2.4. Section 2.5 presents a detailed, traditional model intercomparison and sensitivity analysis. Section 2.6 presents an assessment of model performance under uncertainty and focuses on hypothesis testing. Section 2.7 discusses implications of the results for model evaluation and development. Conclusions are summarized in section 2.8.

2. 3. MODELS, DATA AND METHODS

2.3.1. 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], poor energy partitioning even after calibration [Hogue et al., 2006]), Noah LSM (Ek et al., 2003; Mitchell et al., 2004) has been 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 (1) a version that we equipped with a short-term phenology module ('Noah-DV') and (2) one that couples a lumped, unconfined aquifer model to the model soil column ('Noah-GW').

2.3.1.1. Augmentation of Noah with a dynamic phenology module (DV)

We added the physically-based vegetation module of Dickinson et al. (1998) to Noah-STD in order 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 an exponential function of leaf area index (LAI). STD allows LAI only to influence the computation of stomatal resistance (R_s). In

addition to that, DV makes direct soil evaporation, canopy evaporation, and transpiration depend on variations in leafiness, or, more precisely, LAI.

2.3.1.2. Augmentation of Noah with a groundwater module (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).

2.3.2. IHOP_2002 sites and datasets

We used data from the IHOP_2002 field campaign (Weckwerth et al. 2004) to evaluate predictions from the different version 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, multi-sensor 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[2]. The interested reader is referred to LeMone et al. (2007) for details[3]. Table 2.1 presents the Noah LSM soil and vegetation classes and mean meteorological values for the observation period.

[2] See Figure 3.1 for a map of the sites.
[3] The authors obtained the datasets at http://www.rap.ucar.edu/research/land/observations/ihop/.

The nine stations were sited to obtain a representative sample of the region, which spans a strong east–west rainfall gradient.

Figure 2.1 shows evaporative fraction (EF) (Eq. 2.1) and 30-cm soil wetness (W_{30}) (Eq. 2.2) for Sites 2 (Fig.2.1a) and 8 (Fig.2.1b) 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 at Site 8 somewhat subsides immediately following precipitation; the EF does not peak until several days after the influx of rainwater to the soil.

2.3.3. Model initialization and spin-up

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-minute to a 30-minute time step, to drive the simulations between January 1, 2000, and May 13, 2002. Following Rodell et al. (2005), we initialized each of the four soil layers at 50% saturation and at the multi-annual-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 2.1) between May 13, 2002, to June 25, 2002 (DOY 130 to 176).

2.3.4. Calibration datasets

To constrain and evaluate the models during the IHOP_2002 period, we used 30-minute time step, 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) (Appendix 1). We scored only the last 45 days of each 2.5-year-long model simulation, DOY 130 to 176.

2.3.5. 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., 2006a). We included 8 parameters responsible for the phenology module and 4 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 feasible ranges (Table 2.2) for all parameters were taken from the literature (e.g., Chen et al., 1996; Hogue et al., 2006).

2.3.6. Multi-objective 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 pre-established, feasible ranges) to evolve until the population converges to a stable sample, which maximizes the likelihood function and fairly approximates the Pareto set. The Pareto set (PS) represents the multi-objective

tradeoff: 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, SMC_{5cm}} allows us to constrain the model for consistency with several types of observations. Multi-objective 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 2.2). Figure 2.2 shows slices of STD's objective-function space at Site 4. In frequentist terms, Fig. 2.2 suggests that, when very little is known about the parameters, the expected RMSE of STD at Site 4 is most probably ~55 Wm^{-2} for LE, ~3°C Tg, and ~5% 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.

2.4. 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.

2.4.1. 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 the question: Do conceptually realistic enhancements improve the ability of LSMs to simulate fluxes and near-surface states? We compared the performance of default and multi-objectively calibrated runs using the goodness-of-fit metrics of Appendix 1 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 multi-criteria 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 2.5.

2.4.2. Ensemble-based model intercomparison

We evaluate the hypothesis that increased physical realism in conceptual models not only improves their performance but enhances their robustness, making them less sensitive to errant parameter values (Gulden et al., 2007a). See Appendix 2 for definitions

of the ensemble metrics. We ask the question: 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. In order 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 (quantifies skill and spread of the ensemble), the model robustness score (quantifies insensitivity to poorly known parameters), and the model fitness score (enables ranking models based on suitability for broad application) (Gulden et al., 2008b; equations are presented in Appendix 2). 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 that increased physical realism in conceptual models not only improves model performance but enhances model robustness, making them less sensitive to errant parameter values. Results are presented in Section 2.6.

2.4.2.1. Generation of ensembles

For each model and each of the nine IHOP_2002 sites, we generated two 150-member, parameter-based ensembles: (1) a most-frequent-performing (MF), uncalibrated ensemble; and (2) 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, SMC_{5cm}} (Fig. 2.2). 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).

2.4.2.2. Evaluation criteria

For model evaluation, we use three independent verification criteria: (1) evaporative fraction (EF), (2) 30-cm soil wetness (W_{30}), and (3) change in wetness over 24 hours (ΔW_{30}).

$$EF = \frac{LE}{H + LE} \tag{2.1}$$

$$W_{30} = \sum_{i=1}^{N_{layer}} \theta_i z_i \Big/ \sum_{i=1}^{N_{layer}} \omega_i z_i \tag{2.2}$$

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_{layer} layers (for the observations, $N_{layer} = 4$; for the models, $N_{layer} = 2$).

2.5. 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 above.

2.5.1. 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 on Figs. 2.3 and 2.4. First we tested the implementation of DV with the default parameter values suggested by its developers. Figure 2.3 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. 2.3c), better correlation, and lower error (Fig. 2.3d). 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 frequently is not rejected but is revised. Due to 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. 2.4. 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), the model performs at its best and surpasses the baseline established by the uncalibrated STD.

However, if we allow 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 Fig. 2.3 and 2.4. Performance metrics and statistics are presented in Table 2.3 (see Appendix 1 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 one is best based solely on these results.

To circumvent this issue, Akaike (1974) and Schwarz (1978) proposed information criteria (AIC and BIC, respectively) for model selection. They aim 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 2.3), 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 it is the sole consequence of demanding too little from the complex, multi-output models, at each site we calibrate the models simultaneously against 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 5 objectives, e.g., Hogue et al., 2005). With this preferred, compromise solution, we mimicked the common practice of using of a single 'best' parameter set during model validation.

At each location, the scores of the fully calibrated STD, GW and DV are equivalent (Fig. 2.5). 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 2.4 reports, for each site, the statistics of simulated LE by the 'best' set for each of model. Although there is some slight variation in the scores, model performance is essentially indistinguishable. Calibrated DV ranks best in terms of 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 (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 objective (Table 2.4). Improvement in one evaluation metric tends to result in degradation in another (e.g., at Site 3, GW has a slightly better NSE and 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 Wm^{-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 Wm^{-2}) despite having the same high NSE (0.92) and r^2 (0.90) as STD. Note that, given that a single solution was selected from among a population of realistic, behavioral parameters (PS), the rankings (e.g. Table 2.4) are likely to change when different parameter sets are considered.

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.

2.5.2. Comparison using multiple model realizations

2.5.2.1. 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 distributions (CDF) of 15,000 RMSE scores obtained by STD, GW and xGW are shown in Fig. 2.6. At Site 1 (dry) (Fig. 2.6a), 75% of the STD runs have a LE RMSE lower than 55 Wm^{-2} and no simulation is worse than RMSE = 90 Wm^{-2}; however, 75% of the GW runs have errors larger than 55 Wm^{-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 is RMSE = 9-30%. The behavior of GW at this dry, bare-soil site suggests significant degradation in model robustness. At Site 7 (wet) (Figs. 2.6e-f), the IQR of GW's RMSE is very similar to STD's (30-70 Wm^{-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. 2.6c): 25% of GW's runs have RMSE lower than 48 Wm^{-2}, 25% of STD runs score below RMSE=52 Wm^{-2}. However, GW performs poorly on SMC_{5cm} (Fig. 2.6d): 50% of STD runs score lower than RMSE=10%, whereas only 10% of GW runs have 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 (Fig. 2.6c-d) (LE RMSE > 110 Wm^{-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 Wm^{-2} worse than either STD or GW's top-scoring runs. In general, xGW is at least 40 Wm^{-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., 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).

2.5.2.2. Sensitivity of DV to model 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. 2.7. At Site 2 (dry) (Fig. 2.7a), the IQR of STD simulations of LE lies between RMSE=42 and RMSE=55 Wm^{-2} whereas DV's is between 50 and 67 Wm^{-2}. Fifty percent of the STD runs score below RMSE=47 Wm^{-2}. Fifty percent of the DV runs have RMSE higher than 57 Wm^{-2}. Although the best performing runs of STD and DV have RMSE =30 Wm^{-2}, only 25% of the PS of DV scores below 40 Wm^{-2}; the majority of the PS of DV scores are 15 Wm^{-2} worse than STD (results not shown). At Site 8 (wet) (Fig. 2.7b), the IQR of DV's LE (RMSE=50−70 Wm^{-2}) is very similar to that of STD

(RMSE=45−70 Wm^{-2}). Half of STD runs score below RMSE=52 Wm^{-2}, half of DV runs have RMSE lower than 57 Wm^{-2}. The best-scoring STD and DV runs at Site 8 have RMSE=30 Wm^{-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 Wm^{-2} (Fig. 2.7a). The scores of the top 10% of xDV PS are 5 Wm^{-2} worse than those of DV or STD. At Site 8 (wet), only 10% of xDV runs have RMSE<75 Wm^{-2}, while 75% of the DV and STD runs perform like the best 10% of xDV do. The top-scoring xDV has an RMSE=30 Wm^{-2} (similarly to STD, 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.

2.6. 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 at al. (2008b), summarized in Appendix 2, we show that STD is most fit for broad application.

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

Figure 2.8 shows the time variation of the performance score (ζ-see Appendix 2) of the PS ensemble for each criterion (EF, W_{30}, ΔW_{30}) and model, for Site 2 (dry) (Fig. 2.8a-c) and Site 8 (wet) (Fig. 2.8e-g). Despite calibration against {H, LE, G, Tg, SMC_{5cm}}, when simulating ΔW_{30}, all models significantly overestimate the speed at which the soil column wets and dries (Fig. 2.8c and 2.8g); 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.

2.6.2. Use of the performance score to guide model development

The ζ-score (Appendix 2) 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* (Kirchner, 2006). Here we demonstrate the use of the time-varying performance score in this way.

2.6.2.1. 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. 2.8 and 2.9.

GW achieves the best performance scores of any of the three models when simulating W_{30} at Site 8 (wet) (Fig. 2.8f). 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 – 155 (Fig. 2.8e). To reconcile this apparent contradiction, we also look at the temporal variation of ensemble bias (Fig. 2.9e) and the performance of GW when simulating ΔW_{30} (Fig. 2.8g). 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. 2.8b), 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. 2.8a), 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. 2.9b). 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 timescale (ΔW_{30} reports the difference in moisture between time *t* and 24 hours 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 timescales (months to years), the groundwater module 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.

2.6.2.2. 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 (Fig. 2.1, Fig. 2.8d,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 (Fig. 2.8e-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 choice of parameters. Because the relationship expressing vegetation fraction (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: 2.3 [DV], 3.3 [STD]), slightly lower R_s values than DV (results not shown), and higher soil moisture than DV (Fig. 2.9e). 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 timescales examined here, its effect is somewhat minimal. In DV, LAI (and vegfrac) can and do vary on very short timescales (days), but this appears to not be the primary reason that 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 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).

2.6.3. Evaluation of models' suitability for broad application

2.6.3.1. Which model is most reliable for a given site and objective?

Table 2.5 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, due to the large sample sizes, differences in the 3rd 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).

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.

2.6.3.2. Which model gives the most consistent performance?

A 'robust' model is generally less impacted 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 2.6 shows the robustness (ρ) score (Appendix 2) 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.

2.6.3.3. Which model is best suited for broad application?

The model fitness (ϕ) score combines the concepts expressed by the performance and robustness scores (Appendix 2). With the exception 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 models' overall suitability for broad application by averaging their rankings for individual sites and objectives. Table 2.7 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 2.7). In the models' current configurations, and using these metrics for model fitness, the benchmark model, STD, is found to be most fit for broad application. It most consistently ranks at the top of the cohort in terms of fitness (mean rank of STD $\phi = 1.33$). GW is second-most likely to rank at the top of the cohort (mean rank of GW $\phi = 1.67$), 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 non-benchmark models' conceptual realism, the least complex version of Noah (STD) is most fit for broad application across these 9 representative sites of summer climates in the central U.S.

STD may perform better than the other models not because of a more physically realistic representation but rather 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.

2.7. DISCUSSION OF IMPLICATIONS FOR MODEL DEVELOPMENT

Although the results discussed above 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: (a) 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, (b) 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 unmeasureable, unknown, 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., CLM [Oleson et al., 2008a]) 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, 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. (2008b). 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 (that is, 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-year 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, due to 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 tradeoff between model complexity and model predictive performance. Our results have shown that when 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; Refsgard et al., 2006; Randall et al., 2007; Clarke, 2008).

2.8. 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 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 semi arid sites. Although GW typically achieves the best performance score when simulating each of the three criteria (evaporative fraction, 24-hour 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 (Site 7, 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 which use evaluation data averaged in time and uninformative misfit metrics, and which 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 to choose 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 in which 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.

2.9. ACKNOWLEDGEMENTS

The author would like to acknowledge the editor, G. Salvucci, an associate editor and an anonymous reviewer for their thoughtful comments that have contributed to the

improvement of this manuscript. I thank F. Chen at NCAR and K. Mitchell at NCEP for their insight. H. Wei, also at NCEP, provided the monthly vegetation fraction and albedo climatology values. I thank the International H_2O Project for the datasets. I appreciate the insights of Charles S. Jackson and M. Bayani Cardenas. I benefited from the computational resources at the Texas Advanced Computing Center (TACC). This project was funded by the NOAA grant no. NA07OAR4310216, the Graduate Fellowship of the Hydrology Training Program of the OHD/NWS, NSF, and the Jackson School of Geosciences.

Table 2.1. 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 (W)

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 (Wm^{-2})	293.8	296.7	296.9	272.6	270.3	269.8	268.9	261.8	261.8
LW (Wm^{-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 (hPa)	914.6	915.9	924.1	955.4	955.9	966.2	970.5	965.2	963.4
Q2 (gkg^{-1})	10.3	9.9	9.8	11.2	11.9	11.7	11.9	12.1	11.9
W (ms^{-1})	7.8	7.8	6.6	6.3	5.9	5.6	5.3	5.3	5.9

Table 2.2. Feasible ranges of calibrated Noah-LSM parameters.

Parameter	Description	units	min	max
Soil parameters				
maxsmc	Maximum volumetric soil moisture	m^3m^{-3}	0.35	0.55
psisat	Saturated soil matric potential	$m\ m^{-1}$	0.1	0.65
satdk	Saturated soil hydraulic conductivity	$m\ s^{-1}$	1E-6	1E-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	$Jm^{-3}K^{-1}$	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
z0	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} x1E-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^3m^{-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	ms^{-1} 1E-3	0.01	1

Table 2.3. Performance metrics and statistics for default and (fully and partially) calibrated models (STD, DV, and GW) against latent heat flux (LE) and first layer soil moisture (SMC_{5cm}) at site 7 for the entire evaluation period.

Partial calibration (denoted by 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 calibrated GW. AIC and BIC favor STD's lower complexity. See Appendix 1 for metrics definitions.

Metric	Criterion	LE [Wm^{-2}] mean=126.36 std=136.36			SMC_{5cm} [%] mean=33.19 std=2.84		
	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	

Table 2.4. Goodness-of-fit for the simulation of latent heat flux (LE) for default, partial and fully calibrated models.

Calibrations report only compromise solution: preferred 'best' parameter set minimizes the L2 norm of the RMSE of the 5 objectives {H, LE, G, Tg, SMC_{5cm}}. Best performing model by site in bold. No. stands for number of sites a model performs the best. See Appendix 1 for definitions of metrics.

Metric	Model		IHOP_2002 site									No.1
			1	2	3	4	5	6	7	8	9	
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

Table 2.5. Median performance (ζ) score for each ensemble, site, criterion, and model.

Lower ζ scores (Appendix 2) indicate better performance. Ensembles were constructed using 150 most-frequent performing (MF) and 150 Pareto set (PS) parameter sets.

Criterion	Site	Ensemble	STD	DV	GW
EF		MF	0.204	0.291	0.238
	2	PS	0.186	0.342	0.190
		MF	0.203	0.153	0.155
	4	PS	0.211	0.198	0.185
		MF	0.224	0.073	0.076
	8	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}		MF	0.297	0.339	0.398
	2	PS	0.291	0.575	0.299
		MF	0.330	0.299	0.247
	4	PS	0.329	0.282	0.188
		MF	0.160	0.146	0.060
	8	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}		MF	1.518	1.901	1.831
	2	PS	1.583	1.770	1.861
		MF	3.486	2.950	1.784
	4	PS	3.059	3.125	2.795
		MF	0.972	1.004	0.847
	8	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**

Table 2.6. Model robustness (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). See Appendix 2 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 2.7. Model fitness (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. See Appendix 2 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	

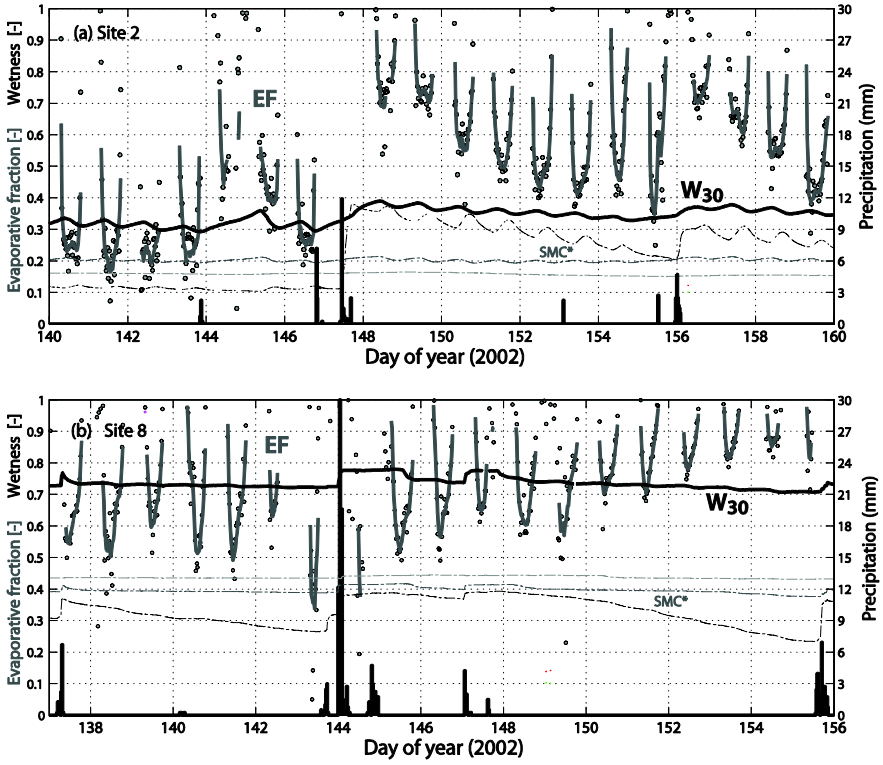

Figure 2.1. Segment of the time series of evaporative fraction (EF), 30-cm soil wetness (W_{30}), volumetric soil moisture (SMC), and precipitation.

(a) at Site 2 and (b) at Site 8. EF is shown in two ways: 30-minute data points and 3-hour 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. W_{30} is 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; the lightest gray line is the soil moisture in the layer farthest from the surface.

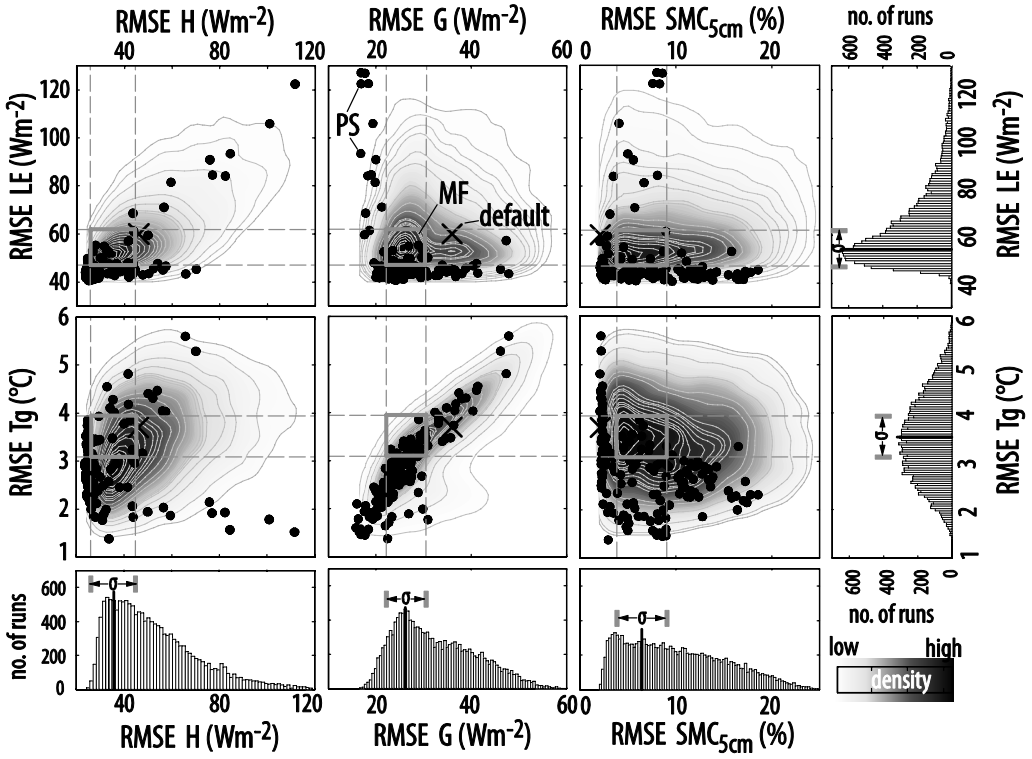

Figure 2.2. Bi-dimensional projections of the objective-function space of STD at Site 4.

Higher density of RMSE scores of 15,000 Monte Carlo model runs shown with darker contours. The Pareto Set (PS), 150 calibrated parameter sets (black dots), represent the minimal uncertainty in the multi-objective tradeoff {H, LE, G, Tg, SMC_{5cm}}. The most frequent performing (MF) models have RMSEs within the intersection of one standard deviation (σ) around the mode of each objective. Note that the relative position of 'default' (x) is no indication of the goodness of model.

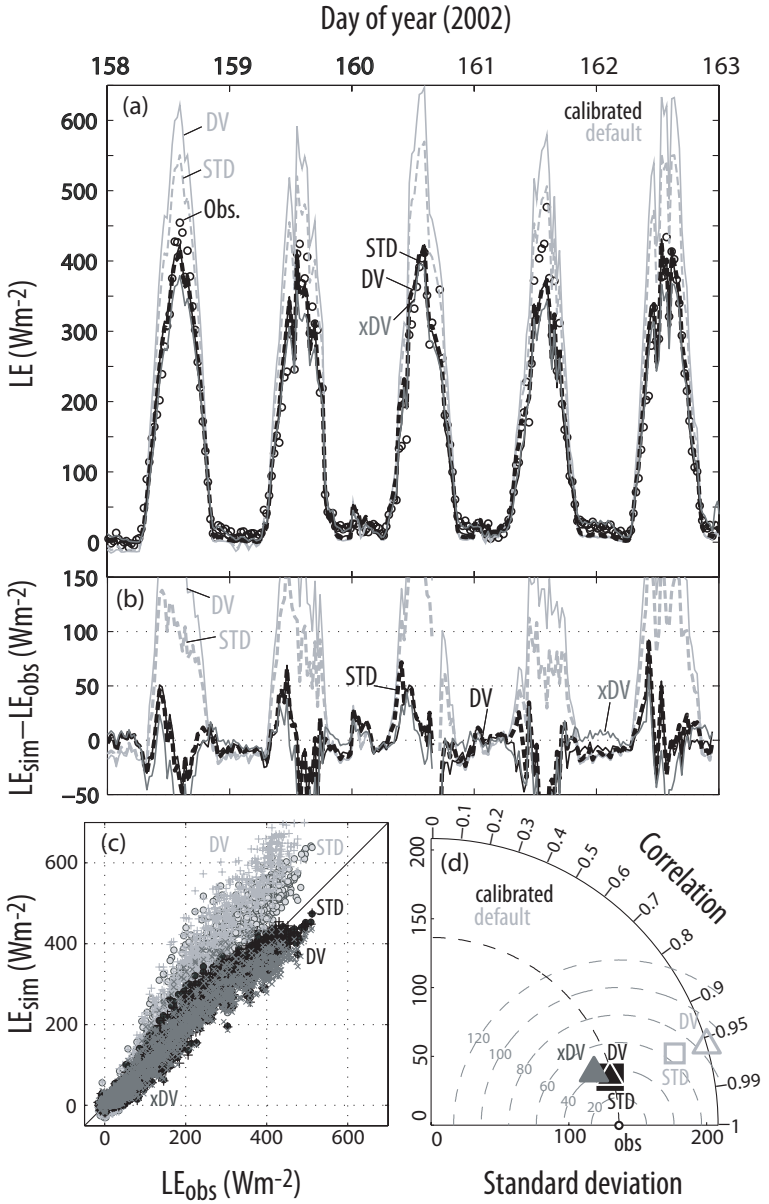

Figure 2.3. Performance of Noah LSM augmented with DV in simulating LE at Site 7.

Figure shows: (a) a segment of the time series of LE and (b) its residuals; (c) scatter plot of simulation versus 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 (see Table 2.2), while the rest of the STD parameters are left fixed to its corresponding default values. (c) and (d) are for the entire evaluation period. See Table 2.3 for statistics.

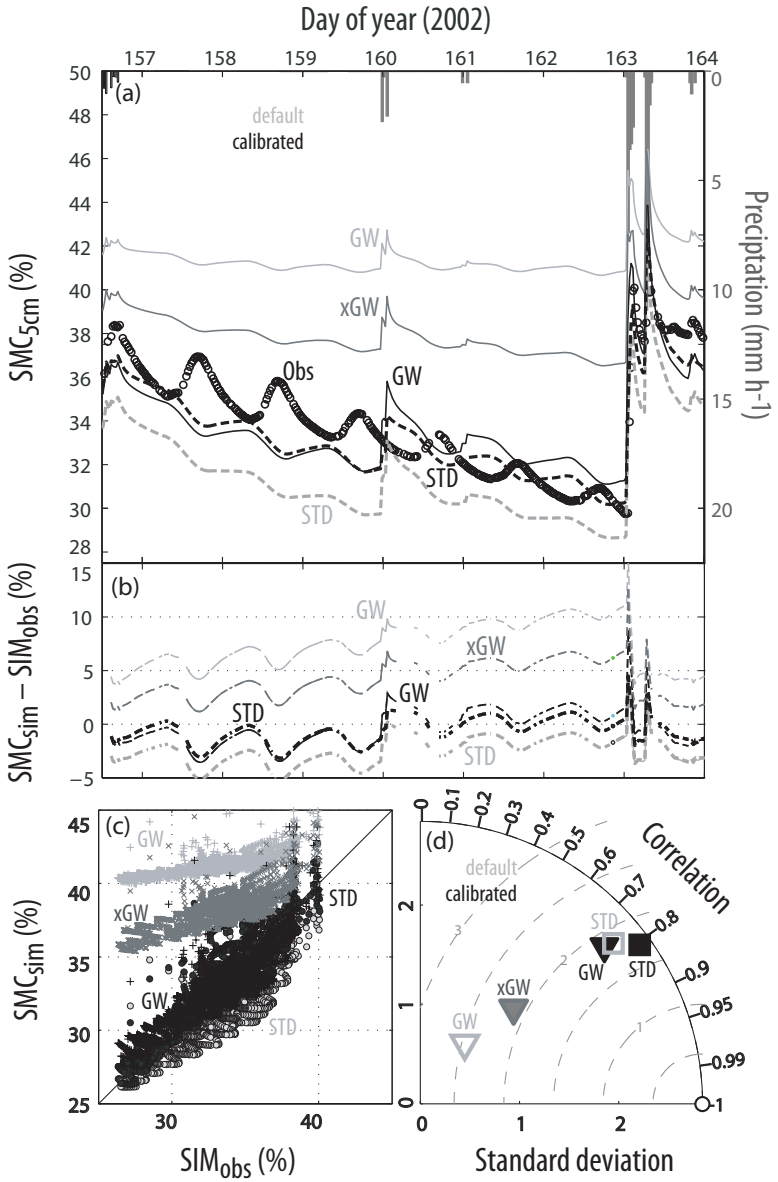

Figure 2.4. Performance of Noah LSM augmented with GW in simulating SMC_{5cm} at Site 7.

Figure shows (a) a segment of the time series of SMC_{5cm} and (b) its residuals; (c) scatter plot of simulation versus observations; and (d) Taylor plot, where dark is a single-objective calibrated run and gray is the uncalibrated (default) run. Partially calibrated (xGW) stands for the tuning of the free parameters of the GW augmentation only (see Table 2.2), while the rest of the STD parameters are left fixed to its corresponding default values. (c) and (d) are for the entire evaluation period. See Table 2.3 for statistics.

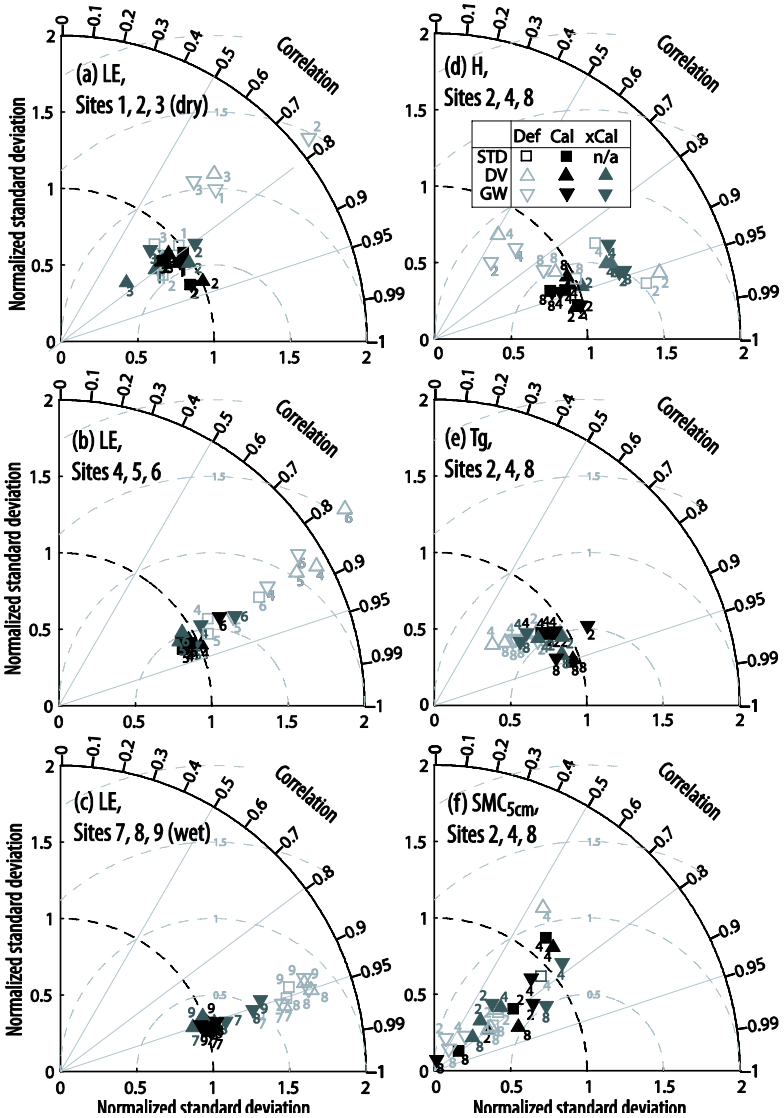

Figure 2.5. Taylor diagrams of performance metrics for the entire evaluation period.

Diagrams are shown for (a-c) latent heat flux (LE) for all sites; and, (d) sensible heat flux (H), (e) ground temperature (Tg), and (h) first layer soil moisture (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, xGW) use a compromise 'best' solution: preferred parameter set minimizes the L2 norm of the RMSE of the 5 objectives {H, LE, G, Tg, SMC_{5cm}}. Calibrated models cluster together for any given site. See Table 2.4 for statistics on simulated LE.

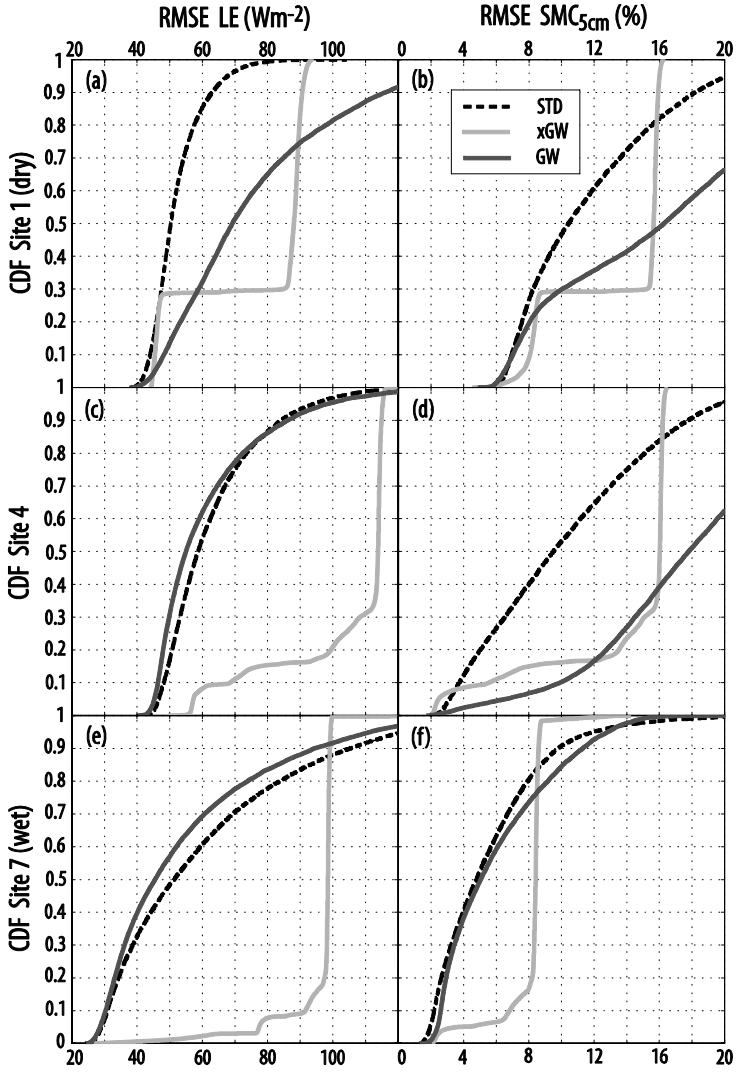

Figure 2.6. Cumulative distribution functions of 15,000 RMSE scores obtained by STD, xGW, and GW.

Cumulative distribution functions (CDF) for 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.

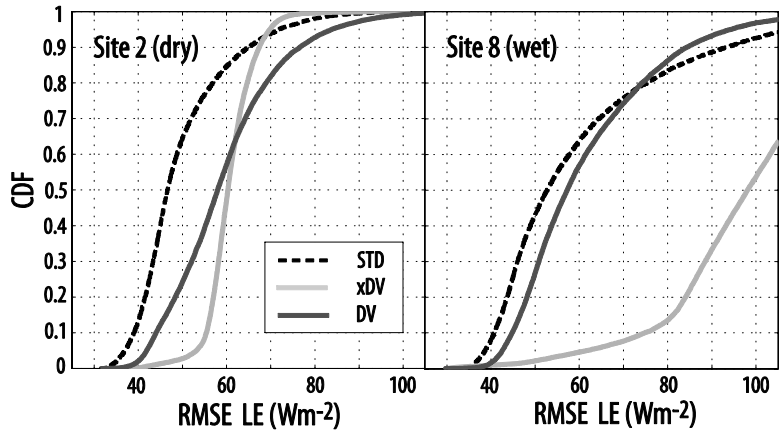

Figure 2.7. Cumulative distribution functions of 15,000 RMSE scores obtained for simulated LE by STD, xDV, and DV.

Cumulative distribution functions (CDF) for 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.

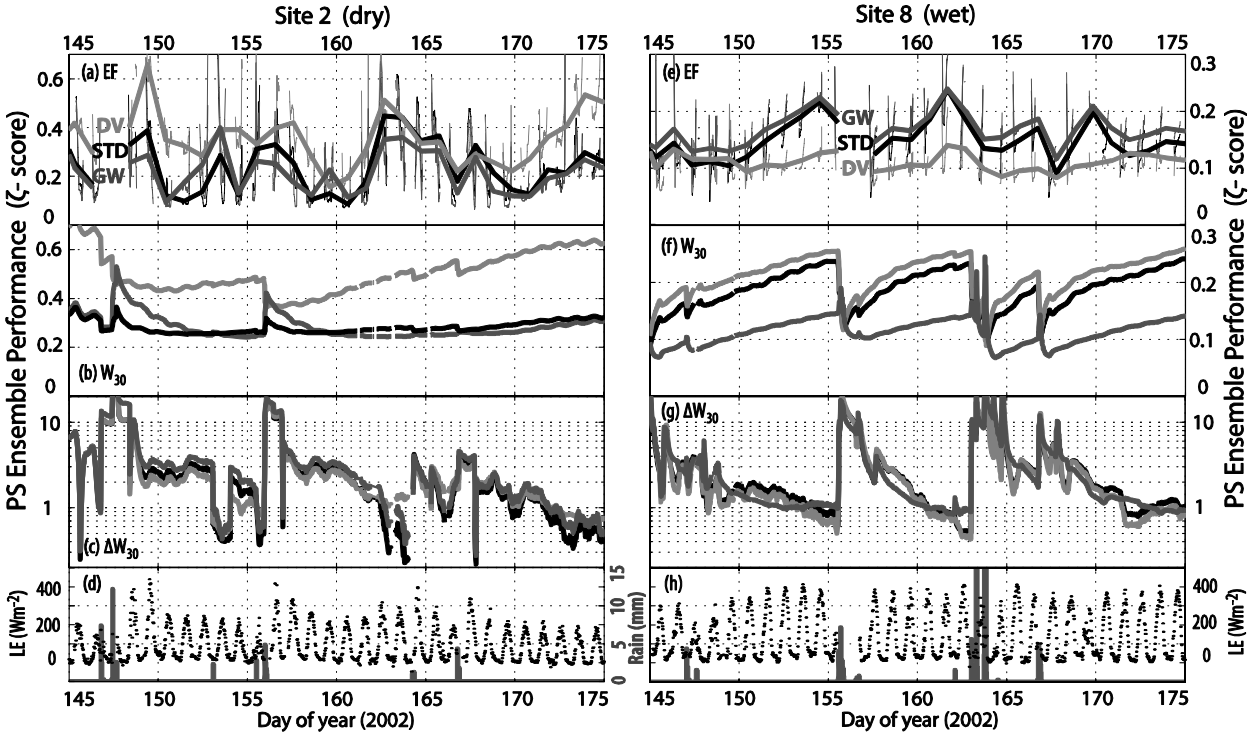

Figure 2.8. Time-varying performance (ζ scores† for STD, DV and GW Pareto set (PS) ensembles between DOY 145 and 175 at Sites 2 and 8.

Performance scores are shown 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 panels show precipitation and latent heat flux (LE) for (c) Site 2 (dry) and (h) Site 8 (wet). For ease of viewing, the EF performance score shown also as 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.

† ζ–score at time *t* is the normalized difference between the CDFs of the ensemble and of the observation. See definition in Appendix 2.

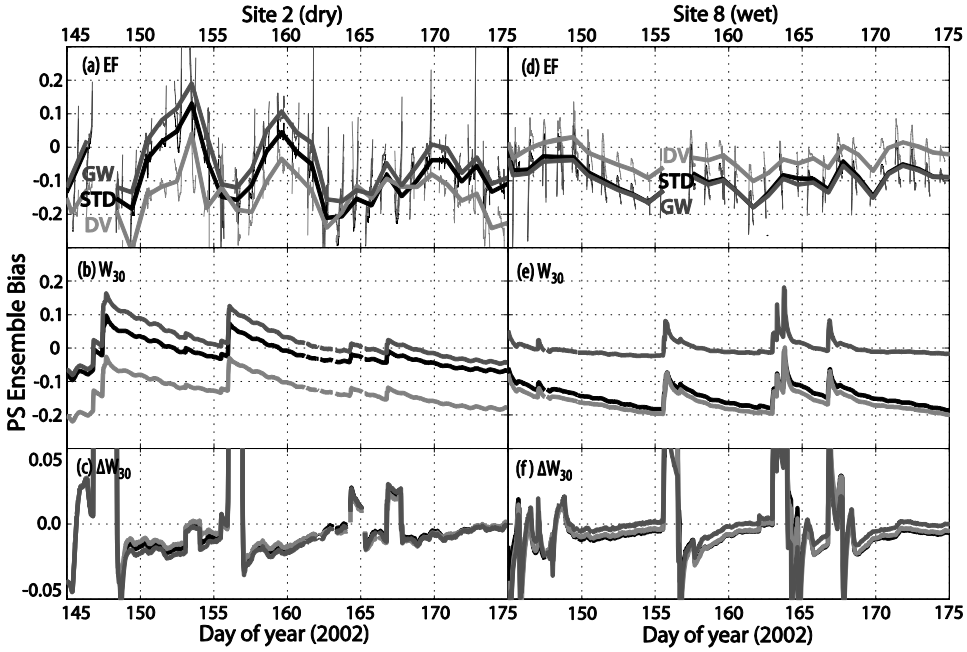

Figure 2.9. Ensemble Bias† of the STD, GW and DV Pareto set (PS) simulations of EF, W_{30} and ΔW_{30} at Sites 2 and 8.

Performance scores are shown 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.

†Bias at time t is the difference between the ensemble mean and the observation. See definition in Appendix 2.

Chapter 3: Sensitivity, Parameter Interaction and Transferability[4]

3.1. ABSTRACT

We use sensitivity analysis to identify the parameters that are most responsible for determining land surface model (LSM) simulations and to understand the complex parameter interactions in three versions of the Noah LSM: the standard version (STD), a version enhanced with a simple groundwater module (GW), and version augmented by a dynamic phenology module (DV). We use warm season, high-frequency, near-surface states and turbulent fluxes collected over nine sites in the U.S. Southern Great Plains. We quantify changes in the pattern of sensitive parameters, the amount and nature of the interaction between parameters, and the covariance structure of the distribution of behavioral parameter sets. Using Sobol's total and first-order sensitivity indexes, we show that very few parameters directly control the variance of the model response. Significant parameter interaction occurs so that not only the optimal parameter values differ between models, but the relationships between parameters change. GW decreases unwanted parameter interaction and appears to improve model realism, especially at wetter sites. DV increases undesirable parameter interaction and decreases identifiability, implying it is overparameterized and/or underconstrained. A case study at a wet site shows GW has two functional modes: one that mimics STD and a second in which GW improves model function by decoupling direct evaporation and baseflow. Unsupervised

[4]Significant portions of this chapter were accepted for publication as:
Rosero E., Z.-L. Yang, T. Wagener, L. E. Gulden, S. Yatheendradas, and G.-Y. Niu (2009), Quantifiying parameter sensitivity, interaction and transferability in hydrologically enhanced versions of Noah-LSM over transition zones during the warm season, *J. Geophys. Res*., doi:10.1029/2009JD012035 (In Press). Works cited here are referenced in the *References* section of this dissertation.

classification of the posterior distributions of behavioral parameter sets cannot group similar sites based solely on soil or vegetation type, helping to explain why transferability between sites and models is not straightforward. This evidence suggests *a priori* assignment of parameters should also consider climatic differences.

3.2. INTRODUCTION

Like other environmental models built to support scientific reasoning and testable hypotheses to improve our understanding of the Earth system, land-surface models (LSMs) have grown in sophistication and complexity (Pitman, 2003; Niu et al., 2009). The evaluation of LSM simulations is consequently non-trivial and, especially when LSMs are to be used in predictive mode for operational forecasting, policy assessments, or decision making, demands more powerful methods for the analysis of their behavior (Saltelli, 1999; Jakeman et al., 2006; Randall et al., 2007; Gupta et al., 2008; Abramowitz et al., 2009). One such method is sensitivity analysis (SA). In this article, we inform LSM development by using sophisticated SA to guide the on-going development of the commonly used Noah LSM (Ek et al., 2003).

SA is the process of investigating the role of the various assumptions, simplifications and other input (parameter) uncertainties in shaping the simulations made by a model. SA is a tool that enables the exploration of high-dimensional parameter spaces of complex environmental models to better understand what controls model performance (Saltelli et al., 2008). Monte Carlo-based SA uses multiple model realizations to evaluate the range of model outcomes and identifies the input parameters

that give rise to this uncertainty (Wagener et al., 2001; Wagener and Kollat, 2007). Used to its full potential, SA weighs model adequacy and relevance, identifies critical regions in the space of the inputs, unravels parameter interactions, establishes priorities for research, and, through an interactive process of revising the model structure, leads to simplified models and increased understanding of the natural system (Saltelli et al., 2006).

SA has been underutilized in LSM development. Approaches to quantify 'sensitivity' (the rate of change in model response with respect to a factor) very frequently are restricted to a simple exploratory analysis of the effects of factors taken one-at-a-time (OAT), without regard for their interactions. Although OAT is only justified for linear models (Saltelli, 1999; Bastidas et al., 1999; Saltelli et al., 2006), it has been used to explore the effects of parameters (e.g., Pitman, 1994; Gao et al., 1996; Chen and Dudhia, 2001; Trier et al., 2008), meteorological forcing, and ancillary data sets (e.g., Kato et al., 2007; Gulden et al., 2008a). A more powerful and sophisticated approach that implicitly accounts for parameter interactions is regionalized sensitivity analysis (RSA). RSA representatively samples the entire parameter space and provides a robust assessment of the way parameter distributions change between subjectively defined 'good' and 'bad' (i.e., behavioral and non-behavioral) model simulations (e.g., Bastidas et al., 2006a, Prihodko et al., 2008) or within the behavioral range of different models (e.g., Gulden et al., 2007a; Demaria et al., 2007). By not explicitly accounting for interactions between parameters, RSA is prone to type II errors (nonidentification of an influential parameter) (Saltelli et al., 2008). RSA does not quantify the extent to which a

parameter affects the variance of the model output, and it is typically applied with the sole purpose of identifying parameters that merit calibration (e.g., Bastidas et al., 1999). The factorial method is a global variance-based SA (VSA) that explicitly accounts for parameter interactions. It uses a set of model runs whose parameters have been perturbed from an arbitrary reference value (default) to identify parameters that affect the variance of model output. Because accounting for higher-order interactions requires a prohibitive number of model runs, factorial analyses in LSM research have been limited to two factor interactions of few selected parameters (e.g., Henderson-Sellers, 1993; Liang and Guo, 2003; Oleson et al., 2008b) and have therefore not fully characterized parameter space. When RSA and VSA are used separately, both the lack of firm conclusions regarding the effect of dominant parameters (and their interactions) on the model variance (e.g., Bastidas et al., 2006a) and the inability to draw cause-effect relationships between parameter regions and model responses (e.g., Liang and Guo, 2003) have precluded SA findings from being widely used in LSM development.

We employ SA to compare the performance and physical realism of three versions of the Noah LSM: the standard Noah (STD), a version augmented with a simple groundwater model (Niu et al., 2007) (GW), and a version augmented with an interactive canopy model (Dickinson et al., 1998) (DV) simulate the land-surface states and fluxes at nine sites in a transition zone between wet and dry climates using the datasets of IHOP_2002 (LeMone et al., 2007). Because of the strength of the land-atmosphere coupling in transition zones (Koster et al., 2004), we focus on warm-season climates of the U.S. Southern Great Plains. Neglecting uncertainty in the meteorological forcing, we

document how parameter interaction and sensitivity varies with model, site, soil, vegetation, and climate.

We use the Monte Carlo-based VSA method of Sobol' to quantify total and first-order sensitivity indexes. The method of Sobol' is more robust (it employs a representative sample of the parameter space) and efficient than factorial analysis (Saltelli et al. 2002), and it bypasses the perceived complexities (e.g., the design of the calculation matrix) often associated with factorial analysis. Note that because LSM developers have attempted to use physical principles when designing their models, the parameters of such physically-based models are assumed to correspond to unchanging physical characteristics of a system. Consequently, the level of parameter interaction can be treated as an indirect measure of the physical realism of LSMs. That is, it is assumed that physically-based models with less undesirable parameter interaction are better (i.e., more physically realistic) (Beck, 1987; Spear et al., 1994; Gupta et al., 2005). We show that only a few parameters directly control model variance and that parameter interaction is significant.

We look at the marginal distributions of behavioral parameters to investigate the ways in which 'physically meaningful' LSM parameters function within alternate model structures. We focus on selected dominant parameter interactions that dictate model response. Because LSM parameter values are assumed to be 'physically meaningful' (e.g. Dickinson et al., 1986) that can be either measured in the field (e.g., porosity) or inferred from (remotely sensed) observations (e.g. LAI), their values should not change between

models for a given site. We show that the distributions of the behavioral parameters differ between models and that the relationships between parameters change.

A priori assignment of parameters based on soil texture and vegetation type is standard practice in the application of LSMs, justified by the assumption that 'physically meaningful' parameters can be transferred between locations that share the same physical characteristics (e.g., Sellers et al., 1996). As a consequence of our SA-enabled model evaluation, we observe that LSM parameters are highly interactive and change between models and between sites, which implies that *a priori* assignment of parameters may not be justified. We use unsupervised classification to test parameter transferability. The similarity of estimated multivariate posterior distributions of behavioral parameters and their sensitivity for each site are compared to those obtained at other sites. We show that the changes between sites are not solely controlled by soil texture or vegetation types but appear to be strongly related to the climatic gradient.

This paper is organized as follows. Experimental design and driving questions are formulated in section 3.3. Datasets, models, and methods are described in section 3.4. Section 3.5 presents the patterns of sensitivity obtained by the global variance-based method of Sobol'. Section 3.6 presents a case study demonstrating the use of SA to understand the functional relationships between behavioral parameters, whose interaction serves to characterize model structure and test hypotheses that regard the formulation of model. Section 3.7 discusses implications of the results for the transferability of parameters between locations with similar physical characteristics. Conclusions are summarized in section 3.8.

3.3. DRIVING QUESTIONS AND EXPERIMENTAL DESIGN

We first ask: What are the dominant model parameters across the region? We run a suite of Monte Carlo simulations to identify parameters that exert the greatest control on the variability of simulated fluxes and states at each IHOP site for all 3 models (STD, GW and DV). We quantify sensitivity using the method of Sobol'. Our SA guides our further investigation.

We then address the question: How do the dominant parameters' interactions change between models? With our focus toward model development, we investigate the relationships between behavioral model parameters and quantify how they change between models using the estimates of the total-order sensitivity, the multivariate posterior parameter distributions, and the covariance structures.

We finally ask: How do behavioral parameters change with dominant physical characteristics of the land? We summarize the relationships between model parameters and physical characteristics by classifying the multivariate posterior parameter distributions according to sites' soil and vegetation types. Our classification provides insights into how parameters can be transferred to ungauged locations.

3.4. MODELS, DATA AND METHODS

3.4.1. Hydrologically enhanced versions of Noah LSM

We compare the standard Noah LSM release 2.7 (STD) to one that couples a lumped, unconfined aquifer model to the model soil column (GW) and a version that we equipped with a short-term phenology module (DV).

3.4.1.1. Noah standard release 2.7 (STD)

Noah (Ek et al., 2003; Mitchell et al., 2004) is a one-dimensional, medium complexity LSM used in operational weather and climate forecasting. The model is forced by incoming short- and longwave radiation, precipitation, surface pressure, relative humidity, wind speed and air temperature. The computed state variables include soil moisture and temperature, water stored on the canopy and snow on the ground. Prognostic variables include turbulent heat fluxes, and fluxes of moisture and momentum. Noah has a single canopy layer with climatologically prescribed albedo and vegetation greenness fraction. The soil profile of Noah is partitioned into 4 layers (lower boundaries at 0.1, 0.4, 1.0 and 2.0 m below the surface). The vertical movement of water is governed by mass conservation and a diffusive form of the Richard's equation. Infiltration is represented by a conceptual parameterization for the subgrid treatment of precipitation and soil moisture. Drainage at the bottom is controlled only by gravitational forces; percolation neglects hydraulic diffusivity. Direct evaporation from the top soil layer, from water intercepted by the canopy and adjusted potential Penman-Monteith transpiration are combined to represent total evapotranspiration. The surface energy balance determines the skin temperature of the combined ground-vegetation surface. Soil-layer temperature is resolved with a Crank-Nicholson numerical scheme. Diffusion equations for the soil temperature determine ground heat flux. The Noah LSM uses soil and vegetation lookup tables for static soil and vegetation parameters such as porosity, hydraulic conductivity, minimum canopy resistance, roughness length, leaf area index, etc.

3.4.1.2. Noah augmented with a simple groundwater model (GW)

GW couples a lumped unconfined aquifer model (Niu et al., 2007) to the lower boundary of the STD soil column. In GW, water flows vertically in both directions between the aquifer and the soil column. The modeled hydraulic potential is the sum of the soil matric and gravitational potentials. The relative water head between the bottom soil layer and the water table determines either gravitational drainage or upward diffusion of water driven by capillary forces. Aquifer specific yield is used to convert the water stored in the aquifer to water table depth. When water is plentiful, the water table is within the model's soil column; if water is insufficient to maintain a near-surface aquifer, the water table falls below the soil column. An exponential function of water table depth modifies the maximum rate of subsurface runoff (for computation of baseflow) and determines the fraction of the grid cell that is saturated at the land surface (for calculation of surface runoff) (Niu et al., 2005). Observed moderate recharge rates for non-irrigated agricultural ecosystems in the Southern Great Plains (Scanlon et al., 2005) warrant the simple representation of an aquifer for the simulation of surface-to-atmosphere fluxes in the region.

3.4.1.3. Noah augmented with a short-term dynamic phenology module (DV)

We coupled the canopy module of Dickinson et al. (1998) to STD in order to compute changes in vegetation greenness fraction that result from environmental perturbations. 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. Unlike STD, which computes greenness fraction by linear interpolation between monthly climatological values, DV represents short-term phenological variation by allowing leaf area to vary as a function of soil moisture, soil temperature, canopy temperature, and vegetation type. DV makes vegetation fraction an exponential function of leaf area index (LAI) (Yang and Niu, 2003). Because DV links vegetation fraction to dynamic LAI, DV makes direct soil evaporation, canopy evaporation, and transpiration more responsive to environmental conditions than STD. Unlike Dickinson et al. (1998), we parameterize the effect of water stress on stomatal conductance as a function of soil moisture deficit, not as a function of soil matric potential.

3.4.2. IHOP_2002 sites and datasets

We used datasets available at www.rap.ucar.edu/research/land/observations/ihop/ from the IHOP_2002 field campaign (LeMone et al., 2007) to evaluate the three versions of Noah LSM at nine sites along the Kansas-Oklahoma border and in northern Texas (Fig. 3.1). The nine stations were sited to obtain a representative sample of the region, which spans a strong west-east (east-west) gradient of rainfall (topography and the Bowen ratio). We used 45 days of high-frequency, multi-sensor measurements of meteorological forcing, surface-to-atmosphere fluxes, and near-surface soil moisture and temperature. Site characteristics, soil and vegetation classes, mean meteorological values,

average heat fluxes and near-surface states for the observation period are summarized in Table 3.1.

3.4.3. Model initialization and spin-up

Following Rodell et al. (2005), we initialized each of the four soil layers at 50% saturation and at the multi-annual-mean temperature. To drive the spin-up (between January 1, 2000, and May 13, 2002), we used downscaled North American Land Data Assimilation System (NLDAS) (Cosgrove et al., 2003) meteorological forcing, interpolated from a 60-minute to a 30-minute time step. The models were subsequently driven by IHOP_2002 meteorological forcing from May 13, 2002, to June 25, 2002 (DOY 130 to 176). For GW, water table depth was initialized assuming equilibrium of gravitational and capillary forces in the soil profile (Niu et al., 2007).

3.4.4. Evaluation datasets

To evaluate the models, we used sensible heat flux (H), latent heat flux (LE), ground heat flux (G), ground temperature (Tg), and first layer volumetric soil moisture (SMC_{5cm}). All data was recorded at a 30-minute time step. In situ, high-frequency flux and near-surface state measurements are an integrated response of the land surface and therefore provide useful data for examining model soundness at a specific location (Bastidas et al., 2001; Stöckli et al., 2008). To score model performance, we used root mean square error (RMSE).

3.4.5. Parameters considered in the sensitivity analysis

We study all 10 soil and 10 vegetation parameters of STD, assigned *a priori* via look-up tables. We included eight parameters responsible for the phenology module and four that control the groundwater module to analyze 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. Default values and feasible ranges (Table 3.2) for all parameters were taken from the literature (e.g., Chen and Dudhia, 2001; Hogue et al., 2006).

3.4.6. Sobol' indices for global variance-based sensitivity analysis (VSA)

We use the variance-based method of Sobol' (Sobol', 1993; 2001) to efficiently identify the factors that contribute most to the variance of a model's response. The method of Sobol' deals explicitly with parameter interaction and has recently been used to quantify model sensitivity and parameter interactions in hydrology (e.g., Tang et al., 2006, Bois et al., 2007; Ratto et al., 2007; Yatheendradas et al., 2008; van Werkhoven et al., 2008). Our review of the literature shows that it has not yet been used for LSM SA.

Sobol' indices enable researchers to distinguish the subset of independent input factors $X=\{x_1, .,x_i,., x_k\}$ that account for most of the variance of the model's response $Y=f(X)$ either by themselves (first-order) or due to interaction with other parameters (higher-order). For completeness, here we summarize the efficient Monte Carlo-based scheme presented by Saltelli (2002) to compute first-order and total Sobol' sensitivity indices.

The first-order sensitivity index (S_i) represents a measure of the sensitivity of $Y = f(x_1, x_2, ..., x_k)$ (the RMSE of a model realization evaluated against observations) to variations in parameter x_i. S_i is defined as the ratio of the variance of Y conditioned on the i^{th} factor (V_i) to the total unconditional variance (V):

$$S_i = \frac{V_i}{V(Y)} = \frac{V(E(Y \mid x_i))}{V(Y)} = \frac{\hat{U}_i - \hat{E}^2(Y)}{\hat{V}(Y)} \tag{3.1}$$

where

$$\hat{U}_i = \frac{1}{n-1}\sum_{r=1}^{n} f(x_{r1}, x_{r2}, ..., x_{rk}) f(x'_{r1}, x'_{r2}, ..., x'_{r(i-1)}, x_{ri}, x'_{r(i+1)}, ..., x'_{rk}) \tag{3.2}$$

is obtained from products of values of f computed from the sample matrix (n model realizations long) times values of f computed from another n-realizations matrix where all k parameters except x_i are re-sampled.

The estimates of the mean squared and the total variance are computed as:

$$\hat{E}^2(Y) = \frac{1}{n}\sum_{r=1}^{n} f(x_{r1}, x_{r2}, ..., x_{rk}) f(x'_{r1}, x'_{r2}, ..., x'_{rk}) \tag{3.3}$$

$$\hat{V}(Y) = \frac{1}{n}\sum_{r=1}^{n} f(x_{r1}, x_{r2}, ..., x_{rk})^2 - \hat{E}^2(Y) \tag{3.4}$$

Instead of computing all $2^k - 1$ terms of the variance decomposition:

$$V(Y) = \sum_{i}^{k} V_i + \sum_{i}\sum_{j>i} V_{ij} + ... + V_{12..k} \tag{3.5}$$

(which would require as many as $n2^k$ model runs), in addition to estimating S_i, it is customary to estimate only the total sensitivity index (S_{Ti}) associated with parameter x_i. S_{Ti} encompasses the effect that of all the terms in the variance decomposition that include

the factor x_i have on the variance of the model's response. S_{Ti} is estimated by the difference between the global unconditional variance of Y and the total contribution to the variance of Y that is caused by factors other than x_i, divided by the unconditional variance:

$$S_{Ti} = \frac{V(Y) - V\left(E\left(Y|x_{-i}\right)\right)}{V(Y)} = 1 - \frac{\hat{U}_{-i} - \hat{E}^2(Y)}{V(Y)} \tag{3.6}$$

where

$$\hat{U}_{-i} = \frac{1}{n-1}\sum_{r=1}^{n} f\left(x_{r1}, x_{r2}, ..., x_{rk}\right) f\left(x_{r1}, x_{r2}, ...x_{r(i-1)}, x'_{ri}, x_{r(i+1)}, ...x_{rk}\right) \tag{3.7}$$

is obtained from products of values of f computed from the sample matrix times the values of f computed from another matrix where only x_i is re-sampled.

A significant difference between S_{Ti} and S_i points to an important role of the interactions of the i^{th} factor (at all orders) in affecting Y (Saltelli et al., 2006). Identification of such parameter interactions can help guide model development. S_{Ti} are also useful to identify input factors that are non-influential, which can help reduce the dimensionality of the parameter estimation problem. If an S_{Ti} is negligible, then it is reasonable to fix that factor to any value within its range of uncertainty, and the dimensionality of the space of input factors or model parameters can be reduced accordingly (van Werkhoven et al., 2009).

3.4.7 Sampling strategies for sensitivity analysis

We generated representative samples of model parameters using Latin Hypercube Sampling (LH) and of the behavioral parameter sets through multi-objective calibration.

3.4.7.1 Latin Hypercube Monte Carlo sampling (LH)

We ran a total of 405,000 Monte Carlo simulations sampling random parameter sets (15,000 samples for each model and site) to obtain a representation of the range of model responses that was sufficiently detailed yet balanced computational constraints. We used LH because it combines the strengths of stratified and random sampling to ensure that all regions of the parameter space are represented in the sample (McKay et al., 1979; Helton and Davis, 2003). LH divides each parameter range into disjoint intervals of equal probability. From each hypercube, one sample value is randomly taken. We sampled uniformly within feasible bounds (Table 3.2). For each sample, we recorded the RMSE of 5 criteria: H, LE, G, Tg, and SMC_{5cm}.To create all the matrices involved in the computation of the Sobol' indices, we used a modified LH that enables replication (Tang et al., 2007).

3.4.7.2 Multi-objective Markov Chain Monte Carlo parameter estimation technique

We used the efficient Markov Chain Monte Carlo sampling strategy of Vrugt et al. (2003) to approximate the joint posterior distribution of optimal parameters. The simultaneous minimization of the RMSE of multiple criteria {H, LE, G, Tg, SMC_{5cm}} allowed us to constrain the models to be consistent with several types of observations and facilitated the identification of the underlying posterior distribution of physically meaningful behavioral parameter sets. It is hoped that sets from the posterior distribution cause the model to mimic the processes it was designed to represent (Gupta et al., 1999; Bastidas et al., 2001; Leplastrier et al., 2002; Hogue et al., 2006). The calibration

algorithm runs, in parallel, multiple chains of evolving parameter distributions to provide a robust exploration of the parameter space. These chains communicate with each other through an external population of points, which are used to continuously update the size and shape of the proposal distribution in each chain. This procedure allows an initial population of parameter sets (uniformly sampled within pre-established, feasible ranges) to converge to a stationary sample, which maximizes the likelihood function and fairly approximates the Pareto set. The Pareto set (PS) represents the multi-objective tradeoff: 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). We used a sample of 150 parameter sets to represent the posterior distribution of 'behavioral' parameter sets.

3.4.8. Hierarchical clustering for comparisons of parameter distributions

Unsupervised classification of behavioral parameter distributions allowed us to understand data similarities across locations, with specific focus on the relationships between types of parameters and sites. We used clustering methods to classify into groups the marginal posterior distributions of calibrated parameters sets. Agglomerative hierarchical clustering methods start with n groups (one object per group) and successively merge the two most similar groups until a single group is left. We used MATLAB's complete linkage algorithm, in which the maximum distance between objects, one coming from each cluster, represents the smallest sphere that can enclose all objects in the two groups within a single cluster (Hair et al., 1995). Because the distance

used to measure dissimilarity between observations (e.g., Manhattan, Euclidean, etc.) may influence the membership of samples to groups, we used the cophenetic correlation coefficient to assess the quality of the linkage (Martinez and Martinez, 2002). We used dendrograms to show the links between the objects as inverted U-shaped lines, whose height represents the distance between the objects.

3.5. WHAT PARAMETERS ARE SENSITIVE?

VSA shows that there are only a few parameters that, by themselves, exert significant influence on model predictions. In contrast, parameter interaction predominates and is hence the principal mechanism for sensitivity. Figures 3.2, 3.3, and 3.4 present, for all sites, all considered parameters, and all models, the Sobol' first-order sensitivity indexes (S_i, which is the fraction of the total variance of RMSE that can be solely attributed to the i^{th} parameter) and the residual between Sobol''s total and first-order sensitivity index ($S_{Ti} - S_i$, which is the fraction of total variance that results from the interaction of the i^{th} parameter with other parameters at all orders) for H, LE, and SMC_{5cm}, respectively. When the influence of parameters change as we would physically expect, we interpret the results as consistent with our hypothesis that, to a first order, a model adequately represents the site-to-site variation in the water and energy cycles. Site-to-site variation in the most sensitive parameters is not chiefly governed by soil or vegetation type but, similar to other studies (e.g, Liang and Guo, 2003; Demaria et al. 2007; van Werkhoven et al., 2008), appears to be of secondary importance when compared to the influence of the predominant climatic gradient. Although we cannot rule

out the potential importance of other east-west gradients (e.g., the topographic or hydrogeologic gradient), in section 3.5.1 we provide explanations for the observed patterns that are consistent with the climatological change between sites.

3.5.1. First-order sensitivity (S_i)

For several key parameters, a pattern of first-order sensitivity can be linked to the hydrology of the sites. For most sites and models, the greatest first-order control on simulated top-layer soil moisture is porosity (*maxsmc*) (Fig. 3.4a). At dry sites 1-3, where direct evaporation is presumably a major component of LE flux, for STD and GW, the bare soil evaporation exponent (*fxexp*) exerts the most first-order control on soil moisture. The LE flux simulated by GW at dry sites is controlled by *fxexp* and specific yield (*rous*), which helps control depth to the water table. *lai* directly controls transpiration and hence the surface energy budget; at the most vegetated sites (7-9), *lai* consequently shapes most of the variance of H and LE for both STD and GW (Fig. 3.2a, 3.3a). The initial value of *lai* is not important to DV's simulated H and LE because DV allows *lai* to change over time. Instead, minimum stomatal resistance (*rcmin*) exerts the most control on DV-simulated LE. Two new parameters associated with DV, *gl* and *sla*, which control the calculation of *lai*, also exert first-order control on the simulated energy fluxes. In sparsely vegetated sites (1-3), the Zilintikevich coefficient (*czil*) plays a significant role in the variance of H.

The specific parameters that control model variance change between models and between sites. In STD, as the mean annual precipitation (MAP) increases, *fxexp* becomes

less important to top-layer soil moisture (SMC_{5cm}) and *refkdt*, a parameter involved in determining maximum rates of infiltration, becomes more important (Fig. 3.4a). This pattern changes for GW, in which surface runoff is relatively de-emphasized and subsurface runoff is relatively emphasized (see discussion about GW's preferred modes of operation, Section 3.6). In GW, although *fxexp* still exerts first-order control on SMC_{5cm} at dry sites, *refkdt* has little direct influence on SMC_{5cm} at wet sites. The most sensitive parameter for SMC_{5cm} at sites 1-3 is *rous*, which controls whether aquifer water is accessible to the near-surface soil. Consistent with our expectations, soil suction (*psisat*), which in GW controls upward movement of water from the aquifer to the soil, has significant control on SMC_{5cm} within GW but not within STD, in which *psisat* plays a less dominant role in shaping soil hydraulic behavior (Fig. 3.4a).

Especially in the case of STD and DV, as sites get wetter, surface exchange coefficient *czil* exerts progressively less influence and *rcmin* progressively more influence on H (Fig. 3.2a). The shift is consistent with our expectation that at more vegetated sites, stomatal resistance will be more important to determining the surface energy balance. As a site's MAP increases, *rcmin* and *lai* increasingly shape simulated LE, and *fxexp* becomes less influential (Fig. 3.3a). Even at dry sites (1-3), DV favors larger values of vegetation fraction (*shdfac*) than are prescribed by STD and GW. As a consequence, DV stands apart from GW and STD in that *fxexp* does not directly contribute to variance of any criterion at the three driest sites (with the exception of unvegetated site 1, at which LE is controlled by *fxexp*).

Examinations of S_i that are not in line with expectations may be used to help modelers diagnose likely problems with conceptualization, forcing data, and/or model structure. For instance, in STD, *fxexp* has the highest S_i of simulated H and LE at site 6. We do not expect direct evaporation to be a relatively more significant component of the LE flux at site 6 than at climatically similar sites 4 and 5 or at the semi-arid sites 1-3. The discrepancy implies that either our conceptual understanding of the physical processes at site 6 is incorrect, that the model does not adequately represent the physical processes, and/or that our forcing and/or evaluation data are faulty at one or more of the sites.

3.5.2 Sensitivity through interactions (S_i–S_{Ti})

Interactions between parameters are responsible for most of the variance in the models' predicted H, LE, and SMC_{5cm} (Figures 3.2b, 3.3b, and 3.4b). If we assume that the parameterizations are correct, then the significant parameter interaction indicates model overparameterization (Saltelli et al., 2008; Bastidas et al., 2006a; Yatheendradas et al., 2008). Arguably, it is also possible that the observed parameter interaction results from models that are either too simplistic and/or incorrect. Although parameter interaction may not be an inherently negative trait (e.g., in porous media, we expect hydraulic conductivity and porosity to be functionally related), when there are no known functional relationships between the physical quantities that two parameters represent, interaction is likely to be indication that the model works in a way that is not consistent with the conceptual model from which the parameterizations were built.

All models exhibit the most parameter interaction at the driest sites, consistent with the findings of Liang and Guo (2003) and suggesting the need to revise the formulation of all three models for semi-arid regions (Hogue et al., 2005; Rosero and Bastidas, 2007). Especially for H and SMC_{5cm}, GW reduces parameter interaction at the middling moisture (4-6) and semi-humid sites (7-9) (e.g., Fig. 3.5b). GW's reduction of parameter interaction is evidence (although by no means conclusive) that GW is more realistic than STD at sites 4-9. This result is consistent with foregoing observations on the robustness of GW (Gulden et al., 2007a). Conversely, GW appears to increase parameter interaction at the driest sites (1-3), indicating STD better represents semi-arid processes than GW. DV parameters are much more interactive than those of STD and GW, especially at the wettest sites when simulating LE and SMC_{5cm}. The increased interaction between the DV-specific parameters and the rest of the conceptually unrelated STD parameters suggests DV is not functioning as its developers intended. The significant parameter interaction is consistent with the poor robustness of DV (Rosero et al., 2009a).

Looked at in full, the models best represent the surface water and energy balances at the intermediate moisture and wet sites, where parameter interaction tends, within a given model, to be lowest. Because it reduces parameter interaction, GW is most likely of any of the three models to be representing the key physical processes with the most realism.

3.6. HOW DO SENSITIVE PARAMETERS INTERACT AND SHAPE MODEL BEHAVIOR? CASE STUDY AT SITE 7

Toward our objective of thoroughly evaluating the physical realism of the three models presented, we perform a case study in which sensitivity analysis (SA) links model identification and model development. We follow the impact of shifted preferred values of three 'physically meaningful' parameters that made considerable contributions to variance: porosity (*smcmax*), the muting factor for vegetation's effect on thermal conductivity (*sbeta*), and minimum stomatal resistance (*rcmin*). We examine model structure at site 7 because at that site STD, DV, and GW show nearly equivalent performance when using their behavioral parameter sets (Fig. 3.5). Such 'equifinality' occurs frequently in hydrologic modeling (Beven and Freer, 2001). In this case, distinguishing a 'best' model is not trivial. It requires us not only to confront the simulations with observed behavior to test for consistency (Rosero et al., 2009a) but, mor important, to understand the underlying model structures (the relationship between parameters) that make STD, GW and DV perform equally well. We show how sensitivity analysis offers the power and the ability to discriminate between model structures that do and do not conform to our physical understanding of the systems.

3.6.1 Focus on sensitive parameters to better understand model function

The models have distinct optimal parameter distributions for the same physical parameters (Fig. 3.6), implying not only that the parameters cannot be transferred between models but that the relationships between them are different. Even the direction of 'sensitivity' (understood as the rate of change of score with parameter value along the range of possible values of the parameter) changes between models (e.g., Fig. 3.6a). The

simulation of SMC_{5cm} by STD and DV degrades as porosity increases, while GW improves. We also note that, along the possible range, the response can be enhanced (Fig. 3.6d) or become relatively insensitive to changes in parameter value (Fig. 3.6c). The identifiability of parameters (when parameters have a clearly defined local minimum) changes between models. For example, in DV, there is a clear low point of the RMSE of LE along the range of values of the maximum water-holding capacity of the canopy (*cmcmax*), but STD and GW have less of a preference (Fig. 3.6c). The interquartile range of *rcmin* of STD is smaller than that of GW or DV (Fig. 3.6b). The fundamental implication of our observations is that although the different optimal values of parameters are important (as found during model identification), the change in the functional relationship between the parameters (the information contained in the interactions) is most relevant for purposes of model development.

*3.6.1.1 The role of porosity (***maxsmc***)*

In all three versions of Noah, higher values of *maxsmc* tend to decrease direct evaporation from the first soil layer (E_{DIR}). E_{DIR} is estimated as the product of Penman's potential evaporation (ET_{pot}), the complement of the vegetated fraction (*shdfac*), and the ratio of top-layer volumetric soil moisture (SMC_1) to *maxsmc*:

$$E_{DIR} = ET_{pot}(1 - shdfac)\left(\frac{SMC_1 - SMC_{dry}}{maxsmc - SMC_{dry}}\right)^{fx\exp} \tag{3.8}$$

where SMC_{dry} is the lowest possible volumetric water content of the top soil layer, and the bare soil evaporation exponent (*fxexp*) is a parameter ranging from 0.2 to 4.

In STD and DV, the error in simulated LE tends to be relatively small when *maxsmc* is low and relatively large when *maxsmc* is high (Fig. 3.6a). However, GW better simulates LE as *maxsmc* increases. The tendency of STD and DV to simulate LE well when *maxsmc* is low (and direct evaporation from the soil consequently tends to be high) implies that STD and DV often underestimate direct evaporation at site 7. The tendency of STD to underestimate direct evaporation was also suggested by Peters-Lidard et al. (2008), who improved results by changing the value of *fxexp* from 2 to 1. Given the same *maxsmc*, GW more easily simulates sufficient direct evaporation, perhaps because of wetter soil (Rosero et al., 2009a).

In STD and DV, *maxsmc* controls surface and subsurface runoff. Hydraulic conductivity (*wcnd*) is computed by scaling saturated hydraulic conductivity (*dksat*) by wetness (SMC/*maxsmc*), raised to an exponent containing the Clapp and Hornberger parameter (*b*):

$$wcnd = dksat\left(\frac{SMC}{maxsmc}\right)^{2b+3} \tag{3.9}$$

Lower *maxsmc* yields higher *wcnd*, which means water moves through the soil more quickly. For subsurface runoff (*Runoff2*), *wcnd* controls lateral water movement through the soil. In STD and DV, *Runoff2* is *wcnd* times the slope of the grid cell. Consequently, higher *maxsmc* decreases *Runoff2*. Higher *maxsmc* also decreases surface runoff (*Runoff1*) by increasing the maximum rate of infiltration. Both changes increase soil wetness.

GW changes the way runoff is computed; *maxsmc* does not control surface or subsurface runoff in GW, which eliminates two of the three ways that *maxsmc* controls soil moisture. *Runoff2* is represented as an exponential function of depth to water (Niu et al., 2007):

$$Runoff2 = rsbmx e^{-fff * Z_{WT}} \tag{3.10}$$

where *rsbmx* is the maximum rate of subsurface runoff, *fff* is the *e*-folding depth of saturated hydraulic conductivity, and Z_{WT} is the depth to the water table, which is computed by the model. *Runoff1* is computed using a version of the function used to compute *Runoff2* (Niu et al., 2005):

$$Runoff1 = pcpdrp * \left(fsatmx e^{-0.5 * fff * Z_{WT}} \right) \tag{3.11}$$

where *pcpdrp* is the effective incident water and the second term is the fraction of unfrozen grid cell that is saturated.

In STD (and DV), *maxsmc* couples two physically unrelated (or very weakly related) processes (direct soil evaporation and lateral surface and subsurface runoff). GW decouples these processes by eliminating the dependence of parameterized lateral runoff on *maxsmc*. This decoupling reduces the spurious parameter interaction of *maxsmc* and, within GW, nearly eliminates the tradeoff between good simulation of LE and SMC_{5cm}. GW is, in this regard, a better model for simulating fluxes at site 7.

The question remains – why does GW poorly simulate SMC_{5cm} when *maxsmc* increases? *maxsmc* is used to compute vertical hydraulic conductivity (using the same function as STD). GW uses vertical hydraulic conductivity to regulate the flow of water

between the aquifer and soil down a hydraulic gradient. Higher *maxsmc* yields lower hydraulic conductivity, which, in addition to decreasing the transfer of water between layers within the soil column, decreases the communication between the aquifer and the soil profile (that is, it decreases the flow of water between the two, increasing the potential for water to be retained near the surface). At site 7, GW best simulates SMC_{5cm} when high vertical hydraulic conductivity connects the aquifer and soil.

Consistent with the work of others (e.g., Demaria et al., 2007), parameter values and model sensitivity to *maxsmc* are not consistent between sites along a climatic gradient or even within a set of sites with similar characteristics. Conclusions about model performance are therefore difficult to generalize. This lack of continuity of behavior between sites is consistent with at least one of the following possibilities: (1) model parameterizations do not represent key aspects of the system and/or (2) our multi-objective calibration provided insufficient constraint for the estimation of behavioral parameters. We suggest the use of observed infiltration and/or runoff to increase the strength of conclusions drawn regarding the physical realism of runoff-related processes in GW.

*3.6.1.2 The role of the thermal conductivity muting factor (***sbeta***)*

All three models compute ground heat flux (G) using a flux-gradient relationship:

$$G = DF_1 \frac{STC_1 - T_1}{0.5 * ZSOIL_{(1)}} \tag{3.12}$$

In which STC_1 is the temperature at the center of the first soil layer ($0.5*ZSOIL_{(1)}$) and T_1 is the surface temperature. DF_1 is the heat conductivity of the top soil layer.

Noah assumes that, as vegetation cover increases, heat flux into the ground decreases. *sbeta* and the vegetated fraction (*shdfac*) mute DF_1:

$$DF_1 = DF_1 * e^{sbeta*shdfac} \tag{3.13}$$

At site 7, the mode of the posterior probability distribution of all three models is near the bound of the explored parameter range (-1) (Fig. 3.6d). The preference for near-bound values is more pronounced in DV, which at site 7 tends to have *shdfac* values near 1.0 (putting downward pressure on the value of *sbeta*). The skewed posterior parameter distributions suggest that an even-less-negative value of *sbeta* may have yielded better results at site 7.

The assumption that vegetation necessarily decreases the thermal conductivity of the top layer of the soil may be incorrect. If the 'vegetation effect' on thermal conductivity is real, the model underestimates the top-layer soil thermal conductivity. At site 7 (and at several other sites), there is a clear tradeoff between H and G that is mediated by the thermal conductivity. The tradeoff hints at the need for revised process understanding.

When comparing site 7 simulations to those of the other two wet sites (8 and 9), we see a roughly consistent preference for near-zero values of *sbeta*. At the drier sites (1-6), the model's strong preference for near-zero values of *sbeta* is less obvious; however, *shdfac* is closer to zero at these sites, which lowers the value of the muting factor (Eq. 3.13).

*3.6.1.3 The role of minimum stomatal resistance (*rcmin*)*

The parameter *rcmin* controls much of the variance in H and LE, especially at wetter sites. As *rcmin* increases, the ratio of actual to potential evapotranspiration decreases. *rcmin* has a more consistent influence on the variance of H than on that of LE.

At site 7, all three models perform best with a low *rcmin* (Fig. 3.6b), which increases LE for a given potential evapotranspiration; however, *rcmin* is less identifiable in GW and DV. The mode of the *rcmin* distribution is higher for GW than for STD, perhaps because GW tends to have a wetter soil and a more robust simulation of LE. The spread of the posterior parameter distribution of *rcmin* for DV is significantly larger than that for STD, although both models share the same mode. This decrease in identifiability of parameters functionally related to *lai* (as is *rcmin*) is consistent with the added degrees of freedom allowed by DV (DV parameters *gl* and *sla* are most important in predicting *lai* [Fig. 3.2]). Because DV simulations include a wider spread of *lai* states, they also have a wider spread of 'good' *rcmin* values.

3.6.2 What changes in GW to make it work better than or as well as STD at Site 7?

The response surface of RMSE SMC_{5cm} changes between STD and GW (Fig. 3.7; e.g., see *maxsmc* vs. *psisat*). For GW, the shape of the bivariate posterior distributions of soil parameters that are shared with STD is significantly different, presumably because of interaction of the GW parameters and module with those of STD. Such shifts in model function affect the model covariance structure (Table 3.3).

After multi-objective parameter estimation, at site 7, GW functions in one of two preferred modes (Fig. 3.7b). In the slightly preferred first mode (*m1*), the parameters work together to help GW function as the developers likely intended. Strong communication between the aquifer and the soil column is supported by relatively high values of saturated hydraulic conductivity (*satdk*), low values of the reciprocal of the *e*-folding depth of hydraulic conductivity (*fff*), and low porosity (*maxsmc*). A relatively low surface runoff scaling factor (*fsatmx*) and relatively high subsurface runoff scaling factor (*rsbmx*) ensure that subsurface runoff dominates surface runoff. Mimicking nature, high soil suction (*psisat*) pulls water upward. A high aquifer specific yield (*rous*) deepens the water table (weakening the direct influence of the saturated zone on the model soil column) and transforms more water to runoff rather than to recharge.

In the second mode (*m2*), GW adopts parameter values that make the model work as one would expect STD to function (i.e., the model operates with parameters that render GW nonfunctional) (Fig. 3.7b). Relatively high values of *fff* effectively seal the bottom of the soil column, limiting communication between the aquifer and the soil column; high *maxsmc* decreases the vertical conductivity, further inhibiting the already poor communication between the soil and aquifer. High *maxsmc* favors decreased direct evaporation. Surface runoff is augmented by a relatively high *fsatmx*; subsurface runoff is lessened by the relatively low *rsbmx*.

These alternative behaviors are a possible explanation for the issue identified by Rosero et al. (2009a), who showed that despite very good performance of calibrated GW,

the model suffered from low robustness (i.e., a high sensitivity to unmeasurable parameters).

3.6.3 What changes in DV to make it work better than or as well as STD at site 7?

STD and DV functionally differ in two ways: 1) STD prescribes *shdfac* using monthly climatological values (~0.7 at site 7), while DV predicts it as a function of environmental variation in moisture and radiation availability; and 2) STD treats *lai* as a parameter (a constant throughout the simulation), while DV uses *shdfac* to predict variable *lai* variation using a functional relationship:

$$lai = \max\left(xlai_{\min}, \frac{-1}{gl}\log^{-1}(1 - shdfac) \right) \tag{3.14}$$

Vegetation affects all components of LE flux (via *shdfac*): vegetation shades the soil, modulating direct evaporation (E_{dir}); vegetation retains water above the soil, contributing to evaporation from the canopy (E_c); vegetation fuels transpiration (E_{transp}). In DV, a high value of conversion parameter *gl* fixes *shdfac* near 1 and yields a regime in which E_c and E_{transp} are strongly favored over E_{dir}. Low values of *gl* fix *shdfac* near zero and promote a regime in which E_{dir} is the dominant component of LE. When *shdfac* is near zero, both E_c and E_{transp} are minimized. At sites with sufficient vegetation, DV enables the model to correctly give more weight to E_{transp}. STD, unable to change the value of *shdfac* to shift the balance of components of LE, favors higher *lai* (which decreases stomatal resistance and increases E_{transp}) as means for increasing total LE.

When compared to STD, DV can achieve 'good' model performance using a wider range of values for *shdfac* and *lai*. We see this decreased identifiability of DV parameters when comparing the bivariate posterior parameter distributions of STD to those of DV at site 7 (Fig. 3.8). The identifiability in the response surface of RMSE LE has changed (e.g. *lai* vs. *rcmin*) (Fig. 3.8). The decrease in identifiability of parameters that are functionally related to *shdfac* and/or *lai* can be seen across the IHOP sites (results not shown). The interplay of the parameters of the DV module also leads to changes in parameter densities of STD and DV (Fig. 3.8). We see additional evidence for increased interaction between parameters in DV when we note that the models' covariance structure has been altered (Table 3.4). For example, *rcmin* and *maxsmc* are positively correlated in STD, but in DV they have a very slight negative correlation.

Although the increased flexibility of *lai* and *shdfac* values may improve the model's simulation of seasonal and interannual variation in surface fluxes, over timescales examined here, DV does not appear to improve the model. The constraints imposed by the turbulent and near-surface states may be insufficient for the complexity of the model and/or DV's degrees of freedom may need to be constrained with observations of carbon fluxes and plant growth. When there is little vegetation (e.g., at sites 1-3), DV may be failing to consider special water use features associated with the semi-arid vegetation (Unland et al., 1996). The function of the DV module may be hindered by Noah's lack of a separate canopy layer (Rosero et al., 2009a) or the absence of a more complex Ball-Berry type of stomatal conductance formulation (Niu et al., 2009).

3.7. WHAT ARE THE IMPLICATIONS OF OUR SENSITIVITY ANALYSIS FOR PARAMETER TRANSFERABILITY?

Our foregoing assessments have shown that parameter interaction is a significant contributor to model variance (section 3.5) and that the behavioral posterior parameter distributions for a given site change between models (section 3.6) and for a given model between sites (not shown; see Fig. 3.9). These observations challenge a long-standing assumption of land-surface modeling: i.e., LSM parameters are physically meaningful quantities. Because developers have attempted to use physical principles when designing LSMs, physically based model parameters have been assumed to correspond to unchanging physical characteristics of a system (e.g. Dickinson et al., 1986), which can be either measured in the field (e.g., porosity) or inferred from (remotely sensed) observations (e.g. LAI). Identical LSM parameters are used in locations that share the same physical characteristics (e.g., Sellers et al., 1996). 'Parameter transferability,' *a priori* assignment of parameter values based on a site's physical characteristics (e.g., soil and vegetation type), depends on the above assumption. By making sets of vegetation-related (soil-related) parameters functions of vegetation (soil) type, LSMs contain the implicit assumption that vegetation (soil) type solely determines the ideal values of vegetation (soil) parameters.

The joint multivariate posterior distribution summarizes much of the information regarding the relationships between model parameters (i.e., the model structure) at a particular location given observed datasets. We compare the similarity of the marginal posterior distributions of the behavioral parameter sets across sites to test the assumption

that parameters and parameter relationships directly relate to physical characteristics of the sites. We also evaluate the extent to which climate determines the similarity of parameters between locations.

3.7.1 Testing parameter transferability between sites using soil textures and vegetation types

If parameters were readily 'transferable' between sites solely based on the sites' vegetation type, we would expect the distributions of the vegetation parameters at two sites with the same vegetation type but different climatic regime (e.g., sites 2 and 8) to be more similar than the distributions of the same parameters at two sites with different vegetation but similar climate (e.g., sites 2 and 1). This expectation is in general not supported by evidence. The distributions of *rcmin* and *lai* (Fig. 3.9a, 3.9b) and *rsmax* and *z0* are more similar between sites with similar climate (dry) than they are between sites with the same vegetation (grass). *hs* and *cmcmax* show a similar lack of transferability. Only *sbeta* shows 'transferability' (i.e., there are smaller differences between the distributions from sites with the same vegetation cover) for all models (Fig. 3.10c). Parameter *cfactr* is transferable, but only for STD. *rgl* could be considered 'transferable,' but only for DV and GW. The IHOP dataset does not enable us to test parameter transferability using two sites with the same soil texture but different climatology.

The case studies above are by no means conclusive, but they do not support the hypothesis that parameters are transferable solely based on vegetation type. The results instead suggest that LSM parameters are more sensitive to climatic forcing than to a specific land-cover classification. Our results support similar observations for other

hydrologic models (Demaria et al., 2007; van Werkhoven et al., 2008), for the Noah LSM, and using single optimal parameter sets (Hogue et al., 2005; Rosero and Bastidas, 2007; Gutmann and Small, 2007).

3.7.2 Synthesizing sensitivity to site, soil and vegetation classes by means of clustering

In order to more quantitatively synthesize knowledge gained through sensitivity analysis for use at ungauged locations, we build upon the aforementioned idea of comparing the similarity of parameter distributions across sites (Rosero and Bastidas, 2007) by complementing the approach with unsupervised, agglomerative hierarchical clustering methods.

For each IHOP site, we obtained a stable, multivariate probability distribution χ of behavioral parameter sets $X=\{x_1, x_2, \ldots, x_i, \ldots x_k\}$ using multi-objective MCMC sampling. The marginal probability distribution for the i^{th} parameter is χ_i. To circumvent comparing sites two at a time, as done in section 3.7.1, we define a triangular probability distribution D_i as a reference distribution for each parameter. $D_i=1$ when the value of parameter x_i is the "default" for the site. $D_i=0$ when x_i is at either edge of the feasible range. This step allows us to introduce the assumption that the parameters relate to soil and vegetation types.

For each parameter, and at each site, we quantify the closeness (similarity) between the cumulative distribution of the 'optimal' values of x_i (i.e, the marginal χ_i) and the reference D_i. We use the Hausdorff norm to quantify the difference $\chi_i - D_i$. For each

model, the matrix of 'signatures' of the marginal distributions of k parameters at all the n evaluation sites is:

$$S = \begin{bmatrix} \chi_{11} - D_{11} & \ldots & \chi_{k1} - D_{k1} \\ \ldots. & \ldots & \ldots \\ \chi_{k1} - D_{k1} & \ldots & \chi_{kn} - D_{kn} \end{bmatrix} \tag{3.15}$$

S can be used to identify groups of parameters that are similar between locations or to identify locations where groups of parameters behave alike. We then use the unsupervised, agglomerative hierarchical clustering algorithm (described in Section 3.3.8) to find these groups without making any further assumptions about the number of groups.

If the previously described assumption of parameter transferability based on site characteristics holds (and if IHOP vegetation classifications are correct), then, given the set of signature vectors created using the set of vegetation parameter distributions $S(x_{\text{veg},1..n})$, a clustering procedure should be able to classify similar sites in groups that resemble the IHOP vegetation type groupings (Table 3.1). Similarly, clustering of $S(x_{\text{soil},1..n})$ would result in sites grouped according to the IHOP soil texture classification (Table 3.1).

Applying a suite of distance metrics (e.g., Manhattan, Euclidean, Cosine, etc), neither soil nor vegetation parameters render groups of sites that partition solely based on the expected soil or vegetation classifications. Figure 3.10a shows the classification tree (dendrogram) for STD using the Euclidean distance, which maximizes the cophenetic correlation coefficient of the linkage (also shown). None of the distance metrics allowed

us to classify $S(x_{veg,1..n})$ by location in a way that matched the IHOP vegetation classifications. Given the subset $S(x_{soil,1..n})$, composed of the signature vectors of the 10 soil parameters at all sites, classification of the IHOP sites according to soil characteristics was also not feasible (Fig. 3.10b). Using signature vectors for STD, GW, and DV, some (but not all) of the distances identified sites 7, 8, and 9 as having the same soil and same vegetation type (although, because they also share the same climate type, we are unable to definitively attribute such classification to shared vegetation type). The rest of the sites do not strongly coalesce according to physical properties. For example, the pasture sites are not distinctively grouped; sites 5 and 6 (wheat crops) are never grouped according to vegetation (Fig. 3.10a). Sites 1 and 2 (sandy clay loam) and sites 4 and 5 (loam) do not cluster together using soil parameters (Fig. 3.10b). These results are consistent with earlier findings presented here, which suggest that interaction between soil and vegetation parameters is significant (section 3.5), to the point that it shapes the posterior parameter distributions (section 3.6). These results also suggest that soil or vegetation type are not, by themselves, good physical characteristics by which to transfer parameters.

To account for interdependence between soil and vegetation parameters, we classified the entire matrix $S(x_{soil},x_{veg})$. If parameters can be transferred based on shared vegetation and soil type, then the clustering of the entire matrix should identify groups of sites with the same vegetation and soil type (e.g, sites 7-9). Figure 3.10c shows a pattern (found with several distances) that is consistent across models: sites 7-9 cluster together. Sites 7, 8, and 9 have also similar climates, and the classification of the sites shows

strong resemblance to the climatic gradient. Given this dataset, we cannot disprove the contention that parameters can be transferred between sites that have both the same vegetation and soil type.

If we instead cluster *S* looking for groups of parameters, we expect that x_{soil} will as a whole behave in a similar way across sites. In other words, one can produce a map of sensitivity to characterize which parameters are most similar to their default (prior distribution) and which are not. Figure 3.11 shows representative groupings of the behavioral, marginal posterior distributions of STD and GW parameters at all sites. Using a suite of distances, we were unable to identify definitive clusters of soil and vegetation parameters within the set of signature vectors *S*, meaning that individual parameters are not sensitive in groups that primarily relate to soil alone or to vegetation alone. The new GW parameters do behave in a way that is similar to other soil parameters (Fig. 3.11b), which informs the estimation of GW parameters for distributed applications.

We conclude that the primary site-to-site control on parameters values is not a site's soil or vegetation type alone. This result is consistent with the notion that LSM parameters, which must represent physical processes across multiple scales, are "effective" values rather than physically derived quantities (Wagener and Gupta, 2005). It is also consistent with the assertion that interaction between classes of parameters (e.g., 'soil' parameters and 'vegetation' parameters) is very important. Our clustering analysis suggests that climate is a major control of site-to-site variation in parameter values and supports recommendations that climate be considered when transferring parameter values between sites (Liang and Guo, 2003; Demaria et al., 2007; van Werkhoven et al., 2008).

3.8. SUMMARY AND CONCLUSIONS

Sensitivity analysis allows us to draw conclusions regarding land-surface model (LSM) development and model assessment practices, the functioning of three versions of the widely used Noah LSM, and the *a priori* assignment of parameter values. Our work yields several conclusions that can be generalized to all LSM and to other environmental models and several others that are specific to the Noah LSM.

We show that the clear patterns of parameter importance identified by variance-based sensitivity analysis (VSA) are consistent with site-to-site variation in climate and with model-to-model changes in physical parameterization. VSA shows that parameter interactions within models exert significant control on model variance. Shifts in parametric control on variance and covariance hint at whether a model represents the water and energy cycles in a way that is consistent with expectations. Although the optimal value of a parameter is useful information, the change in the functional relationship between parameters is more relevant for model development and hypothesis testing.

Transfer of parameters based solely on shared vegetation type or on shared soil texture is not a viable method for *a priori* parameter assignment. The work presented here shows that vegetation type and soil texture are not the most significant contributors to site-to-site variance in optimal parameter values. Interaction between soil and vegetation parameters is significant and varies between sites; parameter interaction at least partially explains why transfer of parameters based solely on shared vegetation type or soil texture does not work. The primary factor controlling site-to-site variation in

parameters is likely climate, although, given the dataset used here, the combination of a site's vegetation type and soil texture or some unidentified factor cannot be ruled out as the dominant controlling factor. The lack of viability of parameter transfer based solely on soil and vegetation type is a conclusion that has significant implications for the field of regional and global land surface modeling, which depends on parameter transfer based on stand-alone vegetation type and soil texture as a means for *a priori* parameter assignment.

Looking specifically at the performance of the three versions of the Noah LSM used here (STD, GW, and DV), we make several non-site-specific conclusions regarding model behavior. All three models exhibit significant parameter interaction, indicating that the models are overparameterized and/or underconstrained. All three show the least parameter interaction at the middling-moisture and wet sites and the most parameter interaction at the three driest sites. This difference suggests a need for reformulation of Noah LSM such that semi-arid regions are more realistically represented. On the whole, GW has less parameter interaction than STD (except at dry sites), indicating that it represents land-surface system with the most realism of any of the three models. GW is also least sensitive to errant parameters at the wettest sites (where groundwater is likely the most influential). DV has much more parameter interaction than STD, which provides evidence that the model is not performing as its developers intended, does not add value to STD, and/or requires additional constraint. Specific to site 7, we make the following observations: STD and DV tend to underestimate direct evaporation from the soil; GW does not (maybe because of wet soil). The assumption that vegetation decreases the

thermal conductivity of the top layer of the soil is not well supported by data (this conclusion can be roughly generalized to other sites, especially the wet sites). At site 7, GW functions in one of two modes – the slightly preferred mode works in a way that mirrors what the developers likely intended; the second mode makes GW function as one might expect STD to work. Constraining runoff may isolate the more realistic mode. GW has less spurious parameter interaction in part because it decouples direct evaporation and subsurface runoff (which are coupled via porosity in STD and DV). This decoupling appears to make the model function more realistically, with less tradeoff between the simulation of soil moisture and LE. Adding modules (DV, GW) decreases the identifiability of minimum stomatal resistance, although all three models prefer low minimum stomatal resistance (thus increasing LE for a given set of conditions). Across several sites, DV functions in one of two modes: the first emphasizes direct soil and canopy evaporation over transpiration; the second emphasizes transpiration over direct evaporation from the soil and canopy.

Our approach to sensitivity analysis complements new methods for characterizing typical modes of LSM behavior (Gulden et al., 2008b; Rosero et al., 2009a) within a model diagnostic framework (Gupta et al., 2008) that helps bridge the gap between model identification and development. We encourage other modeling groups to perform similar analyses with their models as a way to ensure rapid, continued improvement of our understanding and modeling of environmental processes.

3.9. ACKNOWLEDGEMENTS

I thank Pedro Restrepo at OHD/NWS, Dave Gochis at NCAR and Ken Mitchell at NCEP for their insight. I appreciated suggestions by M. Bayani Cardenas, Charles S. Jackson and Yasir H. Kaheil. I acknowledge the International H_2O Project for the datasets. I benefited from the computational resources at the Texas Advanced Computing Center (TACC). The author was supported by the Graduate Fellowship of the Hydrology Training Program of the OHD/NWS. This project was also funded by the NOAA grant no. NA07OAR4310216, NSF, and the Jackson School of Geosciences.

Table 3.1. Average meteorology, near-surface states and turbulent fluxes observed during the calibration period (13 May−25 Jun) at the nine IHOP_2002 sites.

See Figure 3.1. Indices of vegetation types and soil classes are in parenthesis. Rainfall is cumulative over the observation period. Dry, sparsely vegetated sites (1-3) receive almost half of the amount of mean annual precipitation (MAP) than wet sites (7-9), with lush vegetation. Mean 2-m air temperature (Ta), Bowen ratio (β), sensible (H), latent (LE) and ground (G) heat flux, ground temperature (Tg) and soil moisture content at 5-cm (SMC_{5cm}).

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 texture	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)
Elev. (m)	872	859	780	509	506	417	382	430	447
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
Ta (°C)	21.4	21.7	22.5	20.7	20.7	21.0	20.7	20.1	19.9
β (-)	1.08	0.92	1.11	0.41	0.46	0.63	0.20	0.14	0.24
H (Wm^{-2})	70.5	70.7	75.7	43.9	51.9	61.4	25.9	17.1	27.9
LE (Wm^{-2})	65.1	76.1	68.2	106.2	111.2	97.1	126.4	122.8	115.3
G (Wm^{-2})	-10.4	-6.4	-9.3	-2.7	-5.1	-7.5	-5.6	-12.1	-10.5
Tg (°C)	24.1	24.1	25.8	23.2	21.9	22.9	22.3	22.4	22.7
SMC_{5cm} (%)	15.4	18.0	7.0	18.0	18.1	19.0	33.2	32.8	34.0

Table 3.2. Feasible ranges of Noah parameters considered in the sensitivity analysis.

Parameter	Description	units	min	max
Soil parameters				
maxsmc	Maximum volumetric soil moisture	m^3m^{-3}	0.35	0.55
psisat	Saturated soil matric potential	$m\ m^{-1}$	0.1	0.65
satdk	Saturated soil hydraulic conductivity	$m\ s^{-1}$	1E-6	1E-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	$Jm^{-3}K^{-1}$	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
z0	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	Optimum air temperature for transpiration	K	293	303
Dynamic Phenology parameters (Noah-DV only)				
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} x1E-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 only)				
rous	Specific yield	m^3m^{-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	ms^{-1} 1E-3	0.01	1

Table 3.3. Spearman rank correlation coefficients between parameter sets belonging to the behavioral set for STD and GW.

STD is above the diagonal; GW is below the diagonal. Note the change in the covariance structure in Fig. 3.7. See Table 3.1 for abbreviations of parameter names.

	STD						
GW	*maxsmc*	*psisat*	*satdk*	*fxexp*	*rous*	*fff*	*fsatmx*
maxsmc		-0.10	-0.40	0.29			
psisat	-0.33		-0.14	-0.32			
satdk	-0.09	0.49		0.22			
fxexp	-0.26	0.41	0.23				
rous	-0.01	0.26	0.24	0.14			
fff	0.11	-0.46	-0.45	-0.49	-0.37		
fsatmx	-0.22	-0.04	-0.17	0.09	-0.37	0.17	
rsbmx	0.11	-0.25	-0.13	-0.21	0.32	0.08	-0.24

Table 3.4. Spearman rank correlation coefficients between parameter sets belonging to the behavioral set for STD and DV.

STD is above the diagonal; GW is below the diagonal. Note the change in the covariance structure in Fig. 3.8. See Table 3.1 for abbreviations of parameter names.

	STD						
DV	*rcmin*	*hs*	*maxsmc*	*psisat*	*fragr*	*bf*	*xlaimin*
rcmin		-0.35	0.44	0.02			
hs	0.30		-0.15	-0.36			
maxsmc	-0.16	-0.29		-0.10			
psisat	0.50	0.36	-0.21				
fragr	0.58	0.24	-0.02	0.10			
bf	0.61	0.30	-0.19	0.59	0.40		
xlaimin	-0.72	-0.31	0.10	-0.31	-0.62	-0.54	
sla	0.80	0.21	-0.15	0.35	0.66	0.45	-0.67

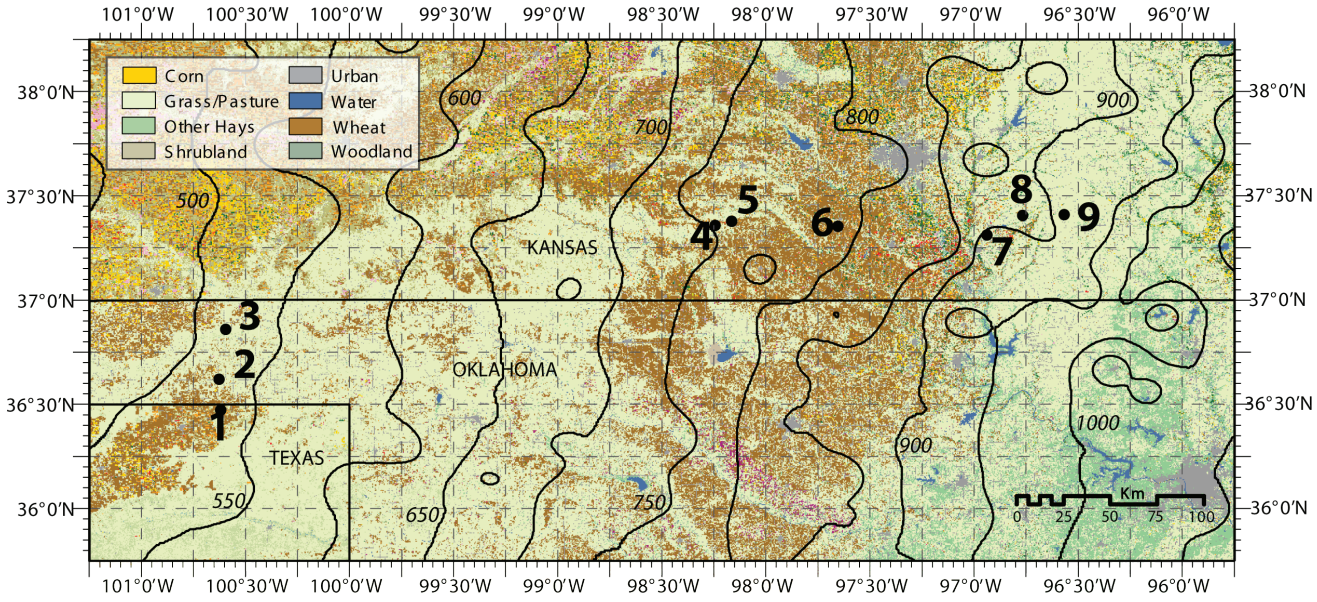

Figure 3.1. IHOP_2002 near-surface state and flux stations.

The contours show the strong east−west mean annual precipitation (MAP) gradient. The nine sites were located in representative land covers (see Table 3.1): six on grassland of varying thickness, two on winter wheat, one on bare ground, and one on shrubland. The surface temperature of the dry (MAP=550 mm), sparsely vegetated sites (1-3) is mainly linked to the soil moisture. In contrast, the green, lush vegetation of the wet sites (7-9) (MAP=900 mm) controls the surface temperature. In sites 4-6 (MAP=750 mm), a mix of winter wheat and grassland, the surface temperature is influenced by both soil moisture and vegetation.

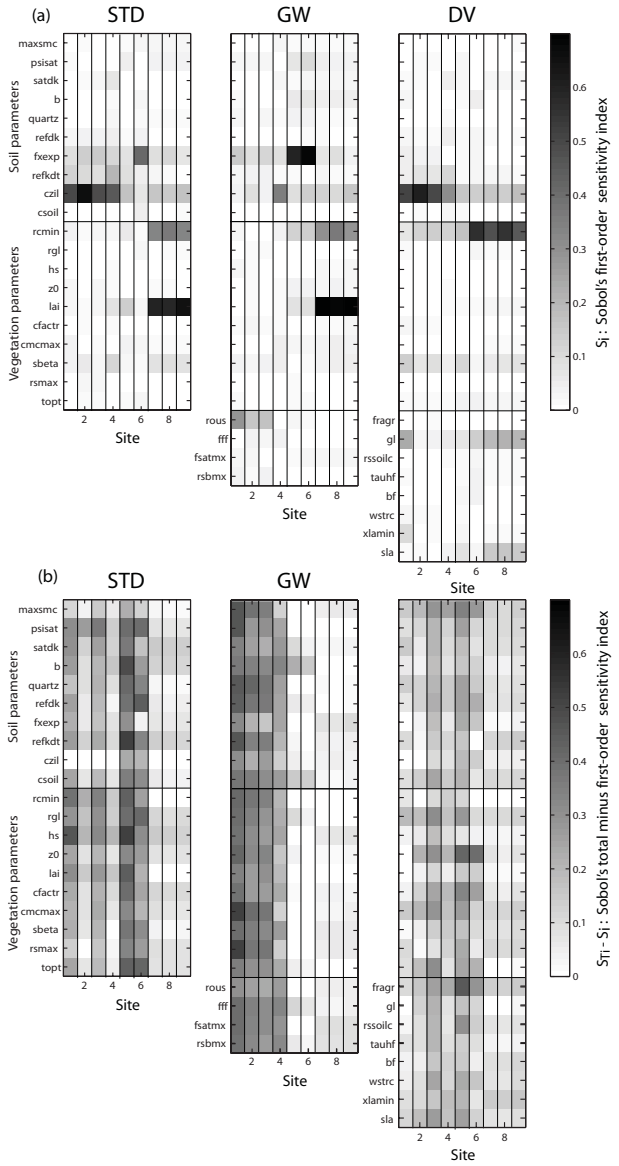

Figure 3.2. First-order sensitivity indices (S_i) and difference between total sensitivity index and Si for H for the parameters of STD, GW and DV at all sites.

(a) First-order Sobol' sensitivity indices. S_i is the individual contribution of a parameter to the variance of the RMSE of H. (b) Difference between Sobol''s total sensitivity index and S_i. S_{Ti}-S_i is the contribution to the variance through interactions with other parameters. Parameters grouped by soil and vegetation. Table 3.2 lists abbreviations of parameter names. Regional sensitivity patterns from semi-arid (MAP=550 mm), sparsely vegetated sites (1−3) to semi-humid (MAP=900 mm) sites (7−9) with green, lush vegetation, are easily distinguishable.

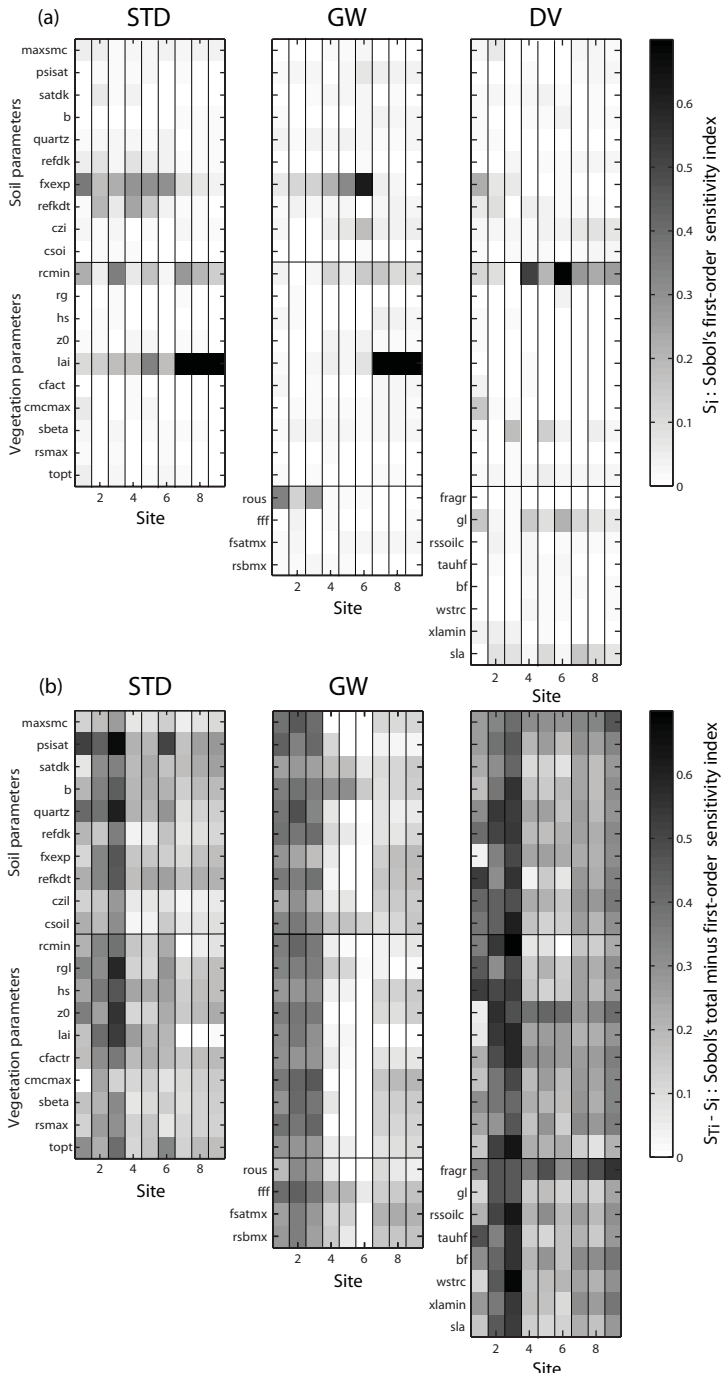

Figure 3.3. First-order sensitivity indices (S_i) and difference between total sensitivity index and S_i for LE for the parameters of STD, GW and DV at all sites.

Same as Figure 3.2 but for LE.

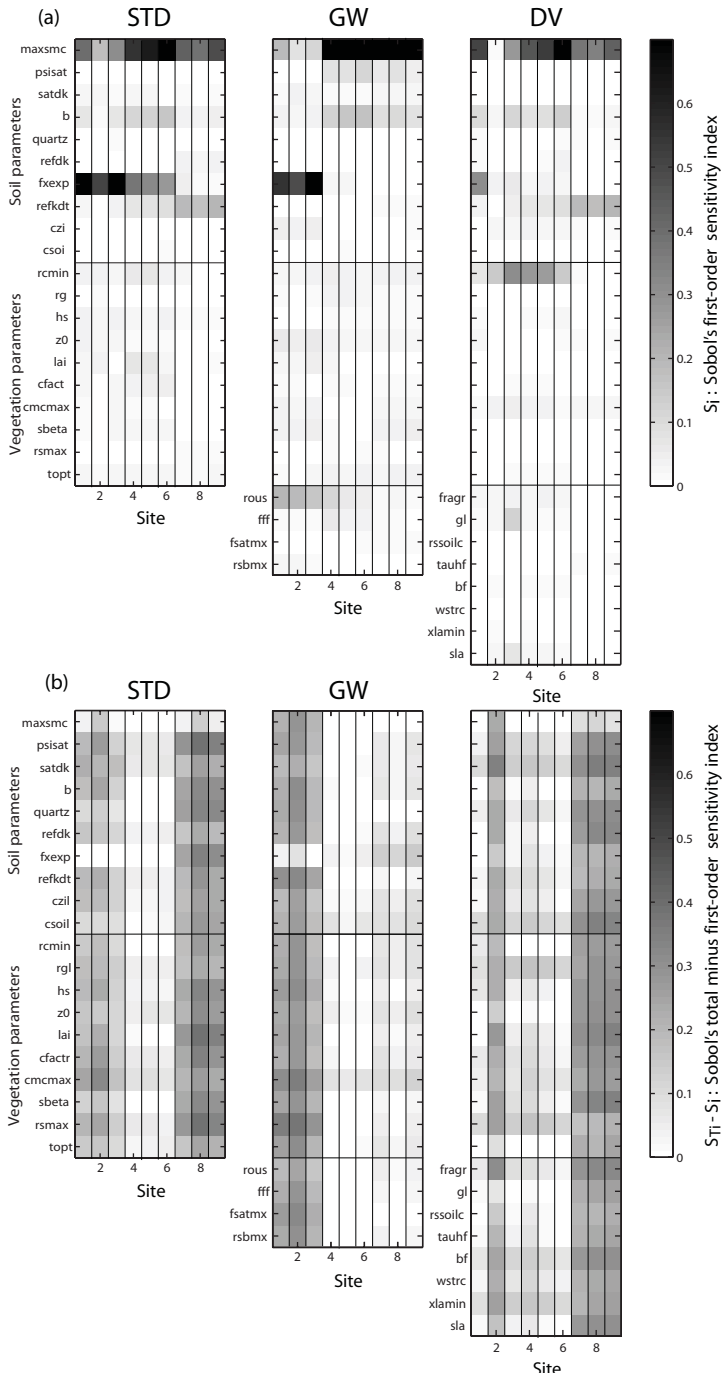

Figure 3.4. First-order sensitivity indices (S_i) and difference between total sensitivity index and S_i for SMC_{5cm} for the parameters of STD, GW and DV at all sites.

Same as Figure 3.2 but for SMC_{5cm}.

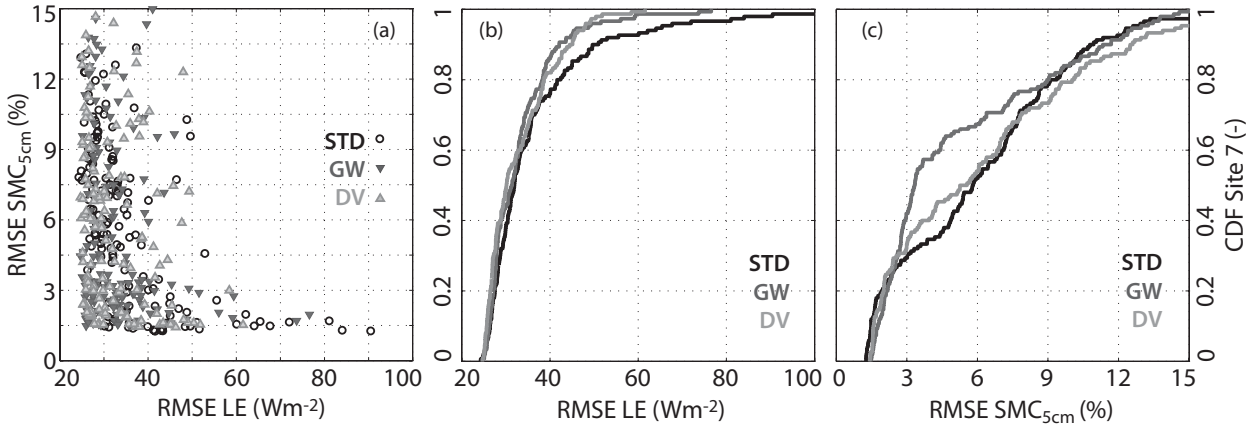

Figure 3.5. Tradeoff LE-SMC_{5cm} and cumulative distribution functions (CDF) of scores of behavioral STD, GW and DV at Site 7.

(a) Scatterplot in objective function space of parameter sets that maximize the likelihood function after multi-objective calibration against {H, LE, G, Tg, SMC_{5cm}}. CDF of root mean squared errors (RMSE) of behavioral runs evaluated against observed (b) LE, and (c) SMC_{5cm}. GW (dark grey), DV (light gray) perform as good as or better than STD (black).

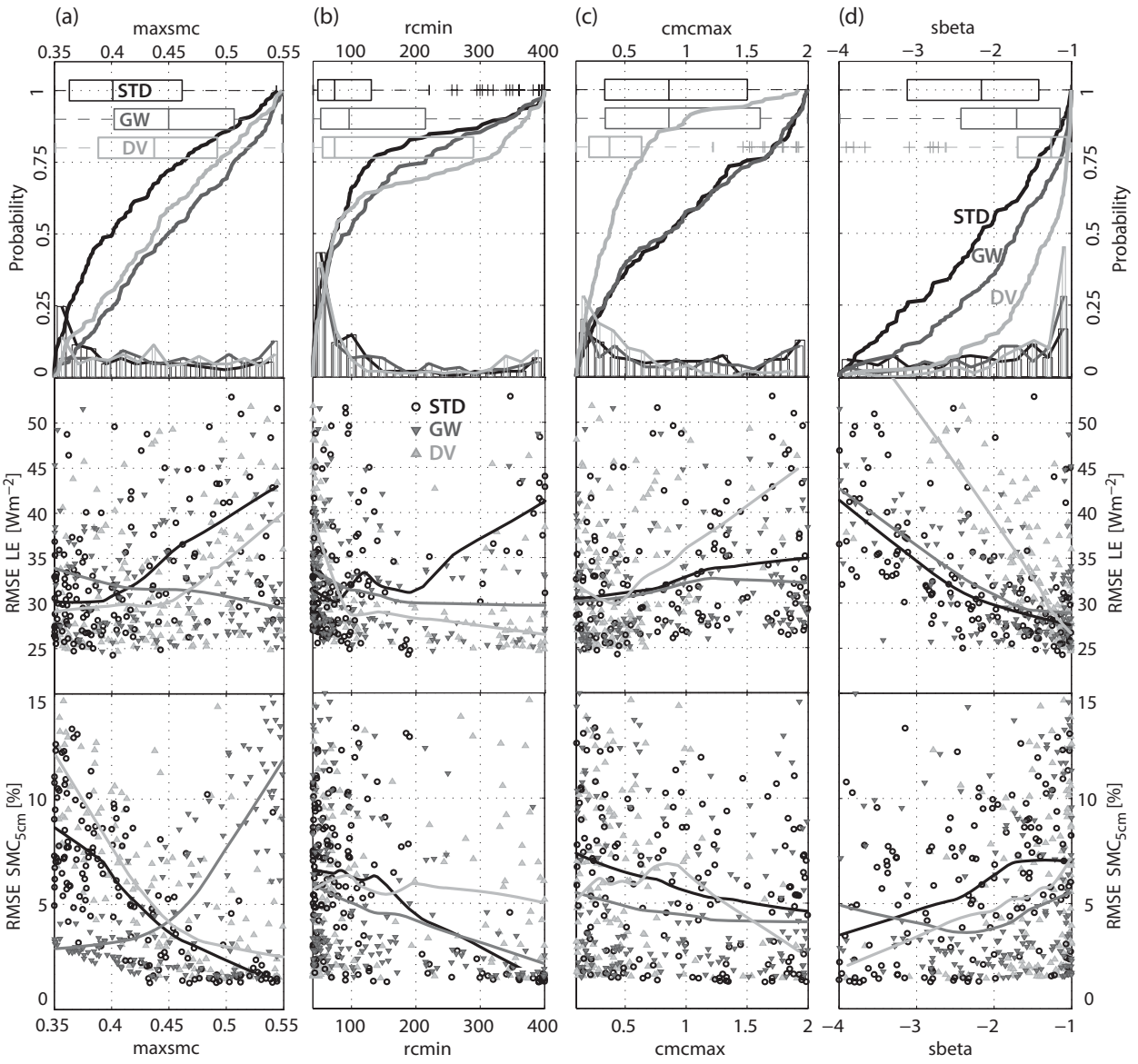

Figure 3.6. Marginal cumulative distribution functions (CDF) of the posterior distribution of selected behavioral parameter sets at Site 7.

(a) Porosity [*maxsmc*], (b) minimum stomatal resistance [*rcmin*], (c) maximum water holding capacity of the canopy [*cmcmax*], and (d) effect of the vegetation on ground heat flux [*sbeta*]. Along with the CDFs, the histograms and interquartile ranges are also shown. The trend in the scatterplots of RMSE of LE and SMC_{5cm} is shown by fitting a minimum complexity polynomial. Note that in all subpanels GW (dark grey), DV (light gray) and STD (black) are shown.

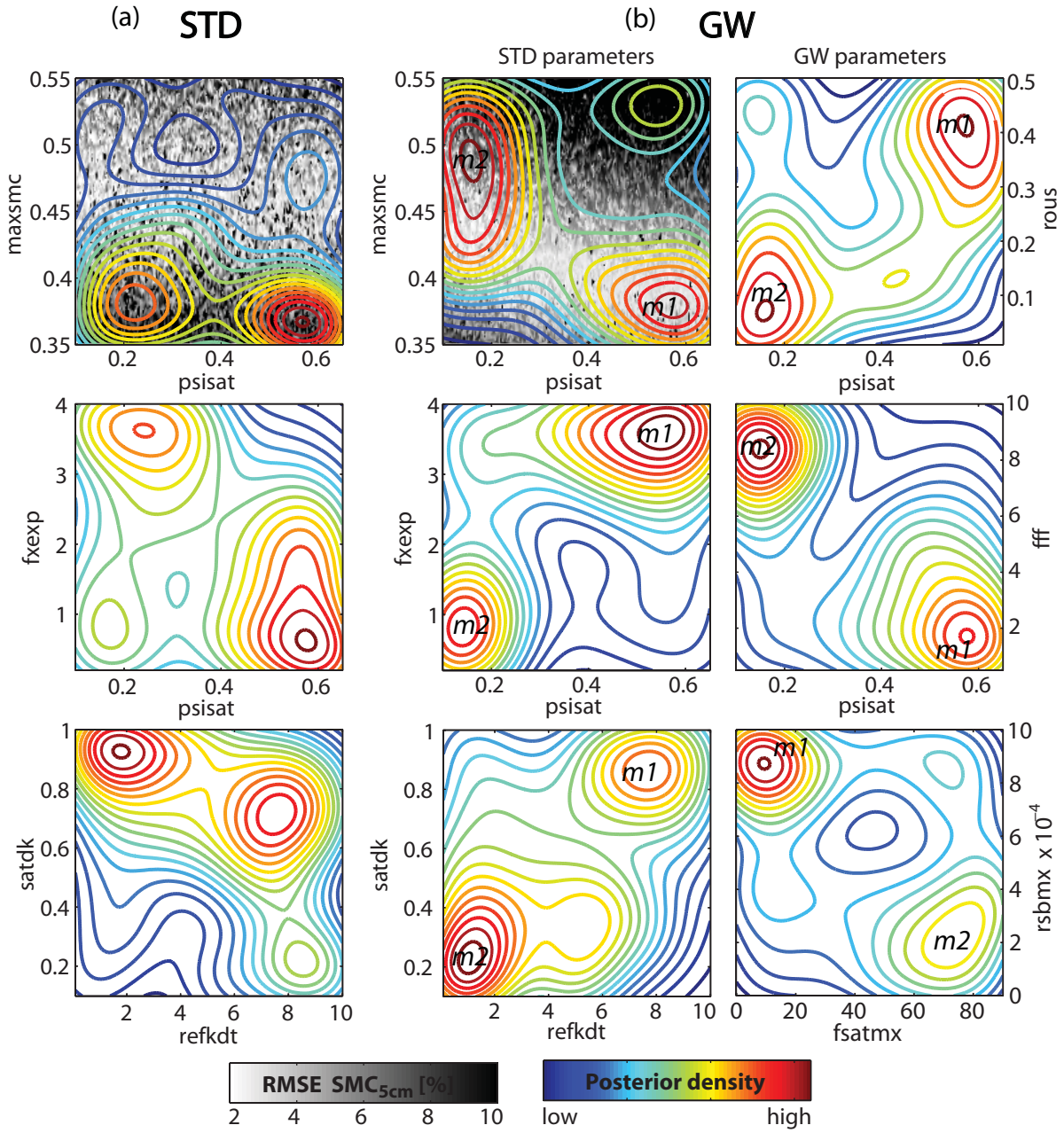

Figure 3.7. Multivariate posterior distribution of the behavioral parameters of STD and GW at site 7 shown for selected parameter combinations in bivariate plots.

Higher density of parameter values are indicated with increasingly redder contours. The response surface of SMC_{5cm} is shown in the back; darker regions have higher errors. The bi-modal behavior of GW is signaled by *m1* and *m2*. See text for explanation.

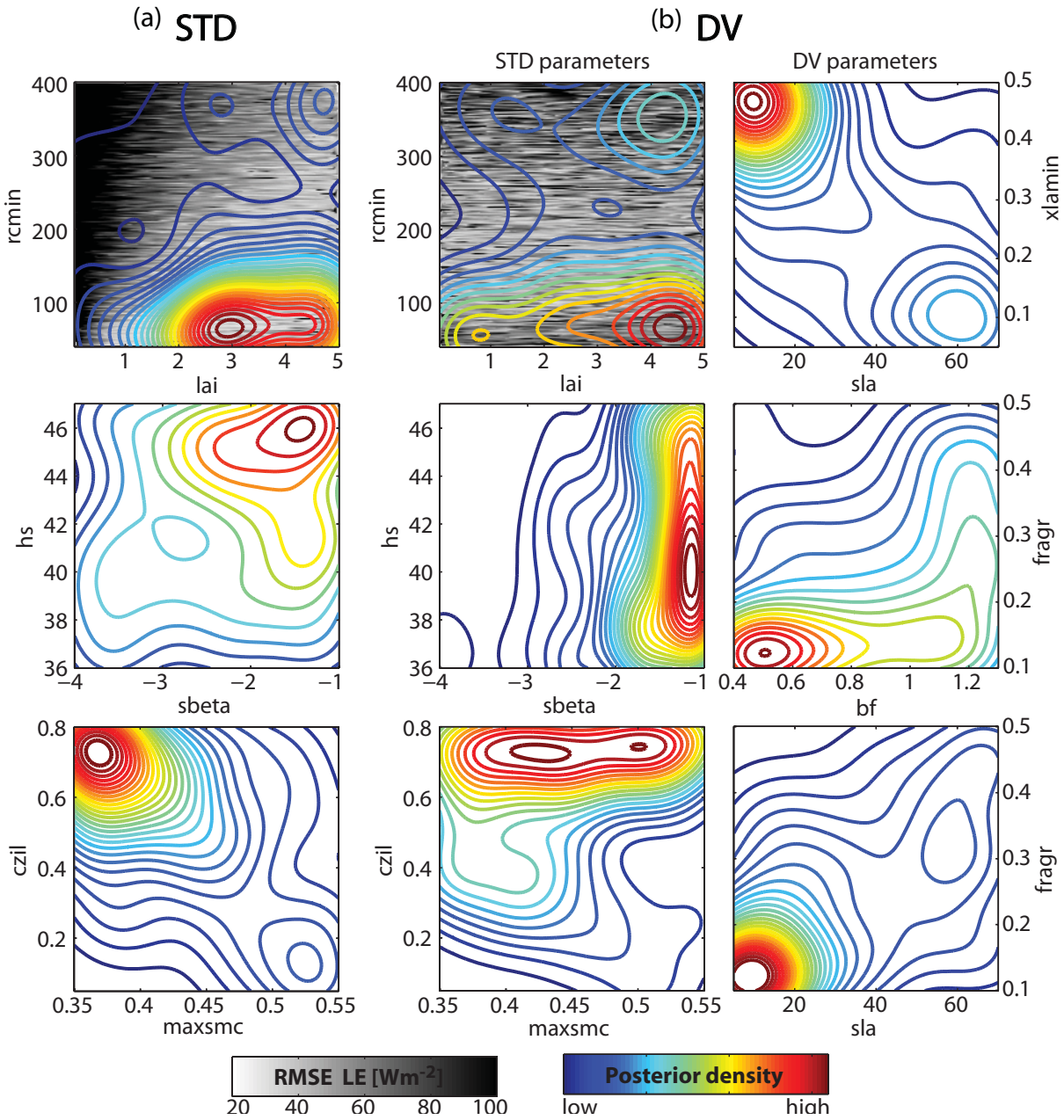

Figure 3.8. Bivariate depiction of the posterior distribution of behavioral parameters of STD and DV at Site 7.

Higher density of parameter values are indicated with red contours. The response surface of LE is shown in the back; darker regions have higher errors. Note the significant change in the identifiability of *hs* and *maxsmc*.

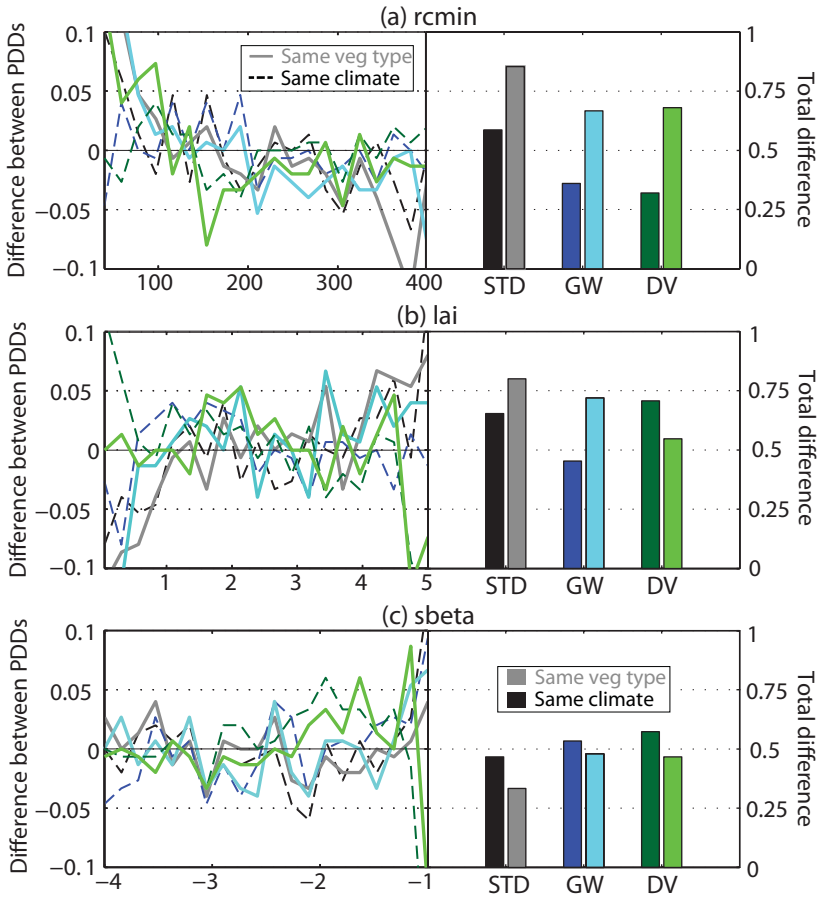

Figure 3.9. Difference between the marginal posterior parameter distributions.

For selected, sensitive vegetation parameters ((a) *rcmin*, (b) *lai*, and (c) *sbeta*), the left panels show the difference between the marginal posterior parameter distributions (PPD) obtained at sites with the same vegetation but different climate (sites 2 and 8) (continuous, bright lines) and the difference between the marginal posterior parameter distributions obtained at sites with similar climate but different vegetation (sites 1 and 2) (dashed, dark lines). As shown in the bar graphs at right, the total difference between parameter distributions at sites with the same vegetation but different climate (brightly colored bars) is generally not smaller than the difference of distributions of the same parameters between contiguous sites with similar climate but different vegetation (dark colored bars).

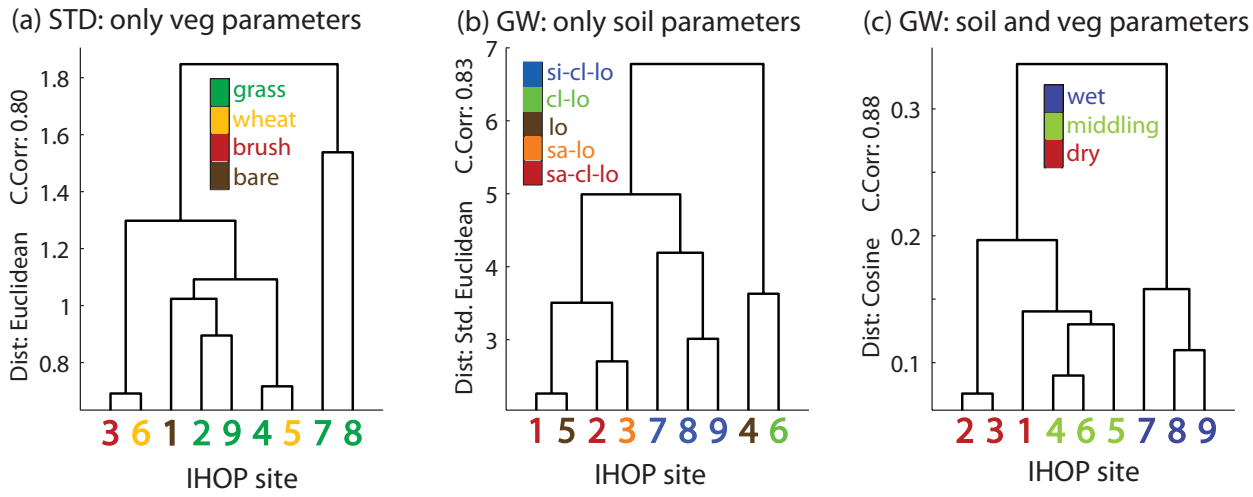

Figure 3.10. Clustering of sites using only the vegetation parameters of STD, only the soil parameters of GW, and both soil and vegetation parameters of GW.

The similarity between marginal distributions of behavioral parameters at all sites is compared using different distances. The plots report the distance that maximizes the cophenetic correlation coefficient of the linkage. Note that neither soil nor vegetation parameters render groups solely based on soil or vegetation type. The clusters of all parameters seem to have a strong relationship with the 3 climatic zones: (1-3) semi-arid, (4-6) middling, and (7-9) semi-humid.

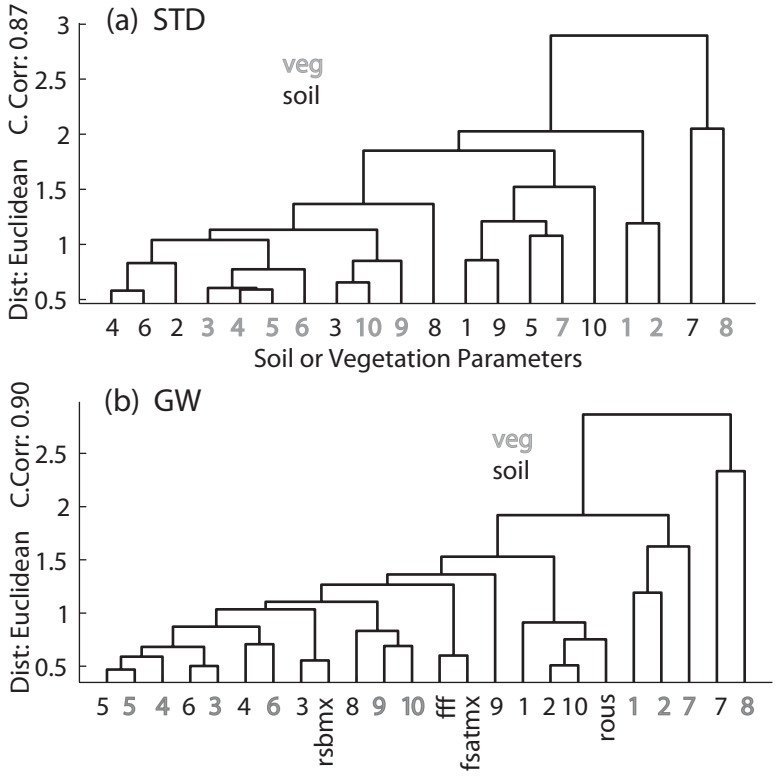

Figure 3.11. Clustering of soil, vegetation, and GW-only parameters for the behavioral, marginal posterior distributions STD and GW at all sites.

Soil parameters are shown in black; vegetation parameters are shown in gray; GW-only parameters are labeled with their names. The cophenetic correlation coefficient for the complete linkage for the parameters of STD and GW is 0.87 and 0.90, respectively. GW parameters seem to behave in a similar way as the soil parameters do.

Chapter 4: Partitioning of the water balance in high-resolution simulations over the Little Washita River Experimental Watershed

4.1. ABSTRACT

We evaluate the ability of two versions of the Noah LSM to simulate the water cycle of the Little Washita River experimental watershed (LWREW) in Oklahoma, USA, at high resolution. We compare Noah STD, which uses the standard hydrological parameterizations of release 2.7, to Noah GW, which uses a simple aquifer model and topography-related surface and subsurface runoff parameterizations in place of the STD subsurface hydrology. We ask: (1) Can STD simulate the high-temporal-resolution and long-term features of runoff when applied on a high-spatial-resolution grid? (2) Does GW improve on STD's ability to partition the water balance? We drive 125,000 (STD) and 200,000 (GW) realizations with NEXRAD Stage IV precipitation data on a 4-km grid, representing 1997−2007. Parameters important to runoff are varied: each realization uses a unique parameter set sampled within physically realistic bounds. Simulations are compared to observed daily-mean runoff, soil moisture, and latent heat. Despite extensive parameter variation, STD and GW overestimate the ratio of runoff to evapotranspiration. Behavioral ensembles of STD and GW overestimate the surface-to-subsurface runoff ratio; simulated streamflow is much flashier than observations. In its current formulation, GW extremely underestimates the contribution of baseflow to total runoff and requires a shallow water table to function realistically. In the LWREW (where the depth to water is >10 m), GW functions as a simple bucket model. We note that model parameters are

likely scale and site dependent, and we underscore the need for even 'physically based' models to be extensively calibrated for all domains on which they are applied.

4.2. INTRODUCTION

Runoff is an integral component of the water balance, and because it is a primary source of water for human use and consumption, it is of great importance for society. Yet, of the variables represented by land-surface models (LSMs), runoff, together with soil moisture, is in general poorly represented (Viterbo, 2002; Nijseen and Bastidas, 2005; Overgaard et al., 2006). Major uncertainty remains in LSMs' simulation of the surface water balance. Some of this uncertainty is governed by the parameterization of processes that drive runoff and the differences in the storage characteristics of LSMs (Pitman, 2003). The strong interaction between the surface water balance and energy balance means that systematic errors in the allocation of moisture to reservoirs and runoff lead to errors in the partitioning of turbulent heat fluxes (Chen et al., 1997; Koster and Milly, 1997; Liang et al., 1998; Wood et al., 1998; Dirmeyer, 2006). This cascade of errors affects the models' simulation of weather and climate (Pitman et al., 1999; Li et al, 2007). The Intergovernmental Panel on Climate Change (IPCC) identified freshwater resources as particularly vulnerable to climate change and highlighted the need for increased capacity to model runoff processes at high-resolution (catchment scale) within the LSMs that are linked to climate models (Bates et al., 2008). The IPCC asserts that such improvements, combined with more extensive, high-resolution runoff datasets, are

necessary for improved assessment of the feedbacks affecting humans' freshwater resources.

The complexity of the subsurface hydrology parameterizations of LSMs is relatively low when compared to the complexity of their parameterizations of above-ground processes (Stöckli et al., 2008). While most LSMs describe the canopy and root zone in great detail, the interactions between groundwater, the root zone, and surface water are normally neglected (Overgaard et al., 2006). Due to the lack of observations of the vadose zone flow, diverse representations of infiltration, drainage and interflow processes in LSMs stem mainly from their unconstrained development, which focused primarily on regional fluxes to the atmosphere (Wetzel et al., 1996). Many LSMs, like the Noah LSM (Ek et al., 2003), neglect topographic effects, assume spatially continuous soil moisture values, parameterize surface runoff with a simple infiltration-excess scheme, and treat baseflow as a linear function of bottom soil-layer drainage (Schaake et al., 1996). More complex in its subsurface-hydrology parameterizations than most LSMs, the multilevel reservoir Variable Infiltration Capacity (VIC) (Wood et al., 1992) family of models tends to perform relatively well in simulating runoff (e.g., Nijssen et al., 1997). VIC and its descendants (e.g., Liang et al., 1996) use a spatial probability distribution to represent subgrid heterogeneity in soil moisture and treat baseflow as a nonlinear recession curve. Other alternative LSM runoff schemes such as the Catchment model (Koster et al., 2000; Ducharne et al., 2000) have been used only in limited research applications (e.g., Reichle and Koster, 2005). Given sufficient data and computing power, lumped catchments may eventually replace rectilinear grid cells as the chosen method for

discretizing the land surface (Goteti et al., 2008); however, in current research and operational practice, LSMs are typically run using rectilinearly gridded domains. The parameterizations of LSMs continue to be developed. Recently, groundwater dynamics have been incorporated into LSMs (e.g., Gutowski et al., 2002; Liang and Huang, 2003; Yeh and Eltahir, 2005; Maxwell and Miller, 2005; Niu et al., 2007; Fan et al., 2007; Kollet and Maxwell, 2008). Advances in routing schemes for the high-resolution representation of the lateral transport of soil water are limited by the accuracy of their surface and subsurface runoff inputs (Gochis and Chen, 2003; Lyon et al., 2008).

Several major concerted efforts have evaluated the ability of multiple LSMs to simulate runoff at coarse scales in temperate regions. The PILPS 2(c) compared the simulations of the seasonal cycle of runoff and the mean annual runoff in the Arkansas-Red River basin (566,251 km^2) on 1×1° grids. The Rhone-AGG (86,000 km^2) addressed issues of domain resolution when comparing simulated land-surface states and fluxes, including runoff, at 8-km, ½°, and 1° aggregated grid. Such intercomparisons indicate that: [1] Bucket models are insufficiently complex to capture runoff processes (Wood et al., 1998; Lohmann et al., 1998). [2] Especially in semi-arid regions, most LSMs overpredict runoff: PILPS 2(c) showed that most LSMs overestimate mean annual runoff (and hence underestimate ET) and that the overestimation of runoff is especially pronounced during summer and in the drier portions of the Red-Arkansas River Basin (Wood et al., 1998; Lohmann et al., 1998), which is the region on which the study presented here focuses. [3] Models whose runoff schemes were dominated by subsurface runoff (baseflow) most accurately simulated summer-season runoff (Lohmann et al.,

1998). [4] Most LSMs can simulate monthly total river runoff relatively well, provided that the precipitation and other forcing input data are sufficiently accurate (Oki et al., 1999). Performance degrades significantly when evaluated at a daily timescale, although most LSMS are still able to slightly outperform the mean discharge (Boone et al. 2004). [5] Increasing model grid resolution tends to increase the volume of simulated runoff (Boone et al. 2004), which implies that there may be a need for the revision of modeling formulations as increasingly finely gridded models and/or at catchment-based models are used to meet societal demands for water-resource information. Model efficiency in the subbasins of the Rhone river, which are comparable in area to those of the MOPEX-basins (1,020 to 4,421 km^2), was found to be lower than the model efficiency for the entire watershed. [6] LSMs appear to be sensitive to subgrid runoff parameterizations (Stöckli et al., 2008) and model parameters (Wood et al., 1998); however, the multi-model intercomparisons' use of only one or a few model realizations has made it difficult to definitively attribute the sensitivity of runoff simulations to parameterization, to parameters, or to a combination of the two. For example: runoff ratios (runoff/precipitation) of single realizations of 16 different LSMs that were used as part of the PILPS 2(c) ranged from 0.02 to 0.41 (the observed runoff ratio was 0.15) (Wood et al., 1998).

The identification and evaluation of distributed hydrological models has been complicated by the large number of model parameters and the lack of sufficiently powerful methods that can be used to perform a truly distributed assesment of model performance (e.g., Beven, 1989; Beven, 2001; Beven, 2002; Konikow and Bredehoeft,

1992; Refsgaard, 1997; Refsgaard and Henriksen, 2004). Recently, Nasonova et al. (2009) showed that with appropriate automatic calibration of a large number of parameters and with the introduction of correction factors for the model forcing (precipitation and incoming radiation), LSMs can simulate runoff at 1/8° with accuracy comparable to that of the hydrological models participating in the MOPEX (Andreassian et al., 2006).

Research presented here evaluates the ability of two versions of the Noah LSM to simulate the water-cycle at high spatial and temporal resolution without the use of forcing-correction factors. The runoff parameterization of the standard version 2.71 of the Noah LSM (hereafter, STD), is relatively simple, as described above, but is still more complex than a bucket-model parameterization. Motivated by observations of other researchers regarding the overly simplistic hydrologic parameterizations in LSMs, we also evaluate a version of the Noah LSM that has been augmented with a lumped, unconfined aquifer model (hereafter, GW), which represents the vertical flow of water between the soil column and an aquifer according to a parameterization of Darcy's Law and which represents surface and subsurface runoff as a function of topography (Niu et al., 2007). In an effort to capture the subgrid heterogeneity in land surface properties that controls runoff generation, a TOPMODEL-based parameterization (Niu et al., 2005) replaces in GW the surface and subsurface runoff parameterizations of STD. Niu et al., (2007) reported improving a complex LSM's capacity to simulate monthly total runoff volume over continental-scale river basins on a 1×1° grid. We hypothesize that because of GW's increased complexity and conceptual realism when compared to STD, and

because of GW's previously reported good performance in reproducing near surface fluxes and states at single points in semi-humid regions of transition zones (Rosero et al, 2009a; Rosero et al., 2009b), GW will outperform STD when simulating runoff.

Our chosen modeling domain is the Little Washita River Experimental Watershed (LWREW) (Allen and Naney, 1991), which is a 611-km^2 basin in Oklahoma, USA (Fig. 4.1). The influence of frozen soil hydrology is negligible. The LWREW is slow draining: baseflow is a major component of overall runoff, which makes the basin an ideal location in which to test the parameterizations of Niu et al. (2005, 2007). Noting the community's call for increased spatial and temporal scales when predicting runoff, we run both versions of Noah LSM on a 4-km grid and evaluate daily river discharge. We further assess the models' abilities to simultaneously simulate runoff, soil moisture, and evapotranspiration. This analysis is at a finer temporal and spatial scale than has been done previously for LSMs.

We address the following broad questions: Can a medium-complexity LSM (i.e., Noah STD) simulate runoff at a fine spatial and temporal resolution in a zone of transition between humid and arid climates? Does the addition of a more complex, physically realistic parameterization of groundwater dynamics improve the model's capacity to simulate runoff? Note that we do not expect the LSMs to be able to provide highly accurate mean daily discharge predictions; rather, we evaluate them based on their capacity to reproduce the essential components and character of runoff generation and of the water balance of the LWREW. Following earlier work done by this research group (Rosero et al, 2009a; Rosero et al., 2009b), we use an extensive evaluation approach that

incorporates the models' typical behavior (i.e., their 'signatures') of ensembles that use realistic, near optimal sets of parameters. We focus extensively on a set of Monte-Carlo−derived behavioral runs that best reproduce the timing and the volume of streamflow.

This is a preliminary study aimed toward improved runoff simulation within LSMs. Because of the small scale of the basin (611 km^2), we assume routing is not necessary to predict daily total streamflow volumes (Fig. 4.2). We further assume that the meteorological input forcing data are accurate enough (i.e., not correction is required). We calibrate a subset of model parameters for seven groups of grid cells (that share the same soil and vegetation type) within the watershed; we leave more exhaustive calibrations (e.g., of parameters at each grid point) to future work.

Section 4.2 introduces the models, evaluation datasets, and Monte-Carlo–based methods. Section 4.3 presents the results of the intercomparison. Discussion is offered in section 4.4. Section 4.5 summarizes our conclusions.

4.3. DATA, MODELS, AND METHODS

We used two versions of the Noah LSM (Ek et al., 2003; Mitchell et al., 2004) to produce ensembles of LSM realizations of near-surface states and fluxes over the Little Washita basin in Oklahoma, USA from 1 January, 1997 to 31 December, 2007. The first five years of model output are treated as spin-up. We evaluate simulations of the period 1 January, 2002 to 31 December, 2007.

4.3.1. The Little Washita River Experimental Watershed (LWREW)

The LWREW (Fig. 4.1), a tributary of the Washita River, is just south of 35°N and is centered on -98°E. Grass, crops, and wooded grassland cover the 611-km^2 basin, which contains soil types ranging from fine sand to silty loam. The climate is temperate and continental: average annual rainfall is 760 mm. Most precipitation is received in the spring and fall. Summers are long, hot, and dry; winters are short, temperate, and dry. Mean annual temperature is 16°C. Daily mean maximum (minimum) temperature in July is 35°C (21°C); and daily mean maximum (minimum) temperature in January is 10°C (−4°C). The watershed is well drained, with gently rolling hills dominating the landscape. Maximum topographic relief is about 180 m. LWREW campaigns and datasets (e.g., Jackson et al., 1993) have been used to validate models (e.g., Wang et al., 2009). Additional description of the watershed can be found in Allen and Naney (1991).

4.3.2. The Noah LSM

Noah is a medium complexity LSM that takes meteorological forcing as input and uses physically based equations to simulate near-surface states and surface-to-atmosphere fluxes. Noah is used operationally by the National Centers for Environmental Prediction models and it is the land component of the Weather Research Forecasting model. Noah uses mass conservation and a diffusive form of the Richards equation to represent vertical water flow through its four-layer soil column (with lower boundaries at 0.1, 0.4, 1.0, and 2.0 m). The dependency of hydraulic conductivity and soil matric potential on soil moisture is parameterized according to Clapp and Hornberger (1978).

The two versions of Noah that we used are hydrologically distinct: (1) in STD, the standard hydrological parameterizations of Noah version 2.71 are used; (2) in GW, a simple aquifer model is coupled to the model's soil columns and the surface and subsurface runoff parameterizations of STD are replaced by the TOPMODEL-based parameterizations of Niu et al. (2005, 2007).

Because the maximum time-lag correlation of daily streamflow between gages upstream of the outlet is under 1 day, no routing scheme was used (Fig. 4.2).

4.3.2.1. The standard version of the Noah LSM (STD)

STD uses an infiltration-excess parameterization to represent surface runoff and a gravitational drainage parameterization to represent subsurface runoff (Schaake et al., 1996). Surface runoff (Q_s) is:

$$Q_s = P_d - Inf_{max} \tag{4.1}$$

where P_d is the rate at which water reaches the soil surface and Inf_{max} is the maximum rate of infiltration into the soil. Inf_{max} is calculated as function of the hydraulic conductivity of the first soil layer, according to the subgrid parameterization of the water balance deficit as:

$$Inf_{max} = Pd\,\frac{Dx[1-exp(-kdt\times\delta_t)]}{Pd+Dx[1-exp(-kdt\times\delta_t)]} \tag{4.2}$$

where *Dx* is the soil moisture (θ) deficit term integrated across soil layers (Δz_i) on time interval δ_t:

$$Dx = \sum_{i=1}^{4} \Delta z_{i(\theta_{sat}-\theta_i)} \tag{4.3}$$

and the variable *kdt* is calculated as a function of the ratio of the saturated hydraulic conductivity (K_{sat}) and its reference value (K_{ref}).

$$kdt = kdt_{ref} \times \frac{Ksat}{K_{ref}} \tag{4.4}$$

Subsurface runoff (Q_{sb}) is:

$$Q_{sb} = Slope \times K_{nsoil} \tag{4.5}$$

where $Slope$ is a scaling factor between 0 and 1 and K_{nsoil} is the hydraulic conductivity of the bottom layer.

4.3.2.2. The Noah LSM augmented with a groundwater parameterization (GW)

GW parameterizes both Q_s and Q_{sb} as a function of depth to water table (zwt). In GW:

$$Q_{sb} = \left(R_{sb_{max}}\right)e^{-(f)(zwt)} \tag{4.6}$$

where $R_{sb_{max}}$ is the maximum rate of subsurface runoff and the parameter f is the *e*-folding depth of saturated hydraulic conductivity, which, following Silvapalan et al. (1987), is assumed to exponentially decay with depth. GW uses a similar parameterization for Q_s:

$$Q_s = (P_d)\left(f_{sat_{max}}\right)e^{-0.5(f)(zwt)} \tag{4.7}$$

where P_d is the rate of precipitation reaching the ground and $f_{sat_{max}}$ is the maximum fraction of ground area that can be saturated.

4.3.3. Meteorological forcing inputs

We used hourly, 4-km NEXRAD stage IV as precipitation input for all model runs after 1 January, 2002. For all other meteorological forcing (longwave radiation, shortwave radiation, atmospheric pressure, wind speed, air temperature, and specific humidity), hourly North American Land Data Assimilation (NLDAS) meteorological forcing (Cosgrove et al., 2003) was used. The NLDAS forcing data were bilinearly interpolated from their native 12-km resolution to the 4-km grid used to represent the LWREW (Fig. 1). We chose to use the NEXRAD precipitation in place of the NLDAS precipitation because the timing of rainfall, the volume of precipitation in individual events, and the cumulative volume of precipitation specified by the NEXRAD data were more consistent with the characteristics of 24 single-point observations obtained by the USDA Agricultural Resource Service (ARS).

4.3.4. Initialization of model realizations

Each model realization was spun up between 1 January, 1997 and 31 December, 2001. All runs were initialized with snow-free ground, a dry canopy, and at the approximate multiannual mean temperature. Soil moisture was initialized as 50% of the realization's specified porosity. We used the equilibrium-water-table assumption of Niu et al. (2007) to initialize the water table for the Noah-GW realizations.

4.3.5. Evaluation data

4.3.5.1. USGS daily mean runoff

We evaluated model performance by comparing simulated daily mean discharge rate to observed data collected by the United States Geological Survey (USGS). We obtained data for the five gauging stations (73274406, 73274458, 7327442, 7327447, and 7327550) within the LWREW for which data was available for the model-evaluation period at http://waterdata.usgs.gov/nwis.

4.3.5.2. FLUXNET evapotranspiration data

We compared hourly simulated latent heat flux for 1 January, 1998 − 31 December, 1998 to the mean hourly observed latent heat flux (Meyers, 2001) obtained at the Ameriflux site at Little Washita (-97.9789°E, 34.9604°N). Data were accessed at http://public.ornl.gov/ameriflux/Site_Info/siteInfo.cfm?KEYID=us.little_washita.01. No latent heat flux observations within the LWREW were available for any other period of time.

4.3.5.3. Soil moisture data

Daily volumetric soil moisture observations at 5 cm and 25 cm for the period 1 January, 2005 – 31 December, 2007 for 24 sites within the LWREW were obtained from the USDA's ARS Micronet website (http://ars.mesonet.org/). Time series for selected sites (A148 and A153) and statistics of soil moisture for all the sites are compared.

4.3.6. Land cover classification

We used 1-km University of Maryland vegetation data (Hansen et al. 2000) and 1-km FAO/STATSGO2 soil texture classifications (Soil Survey Staff, 2009), both of

which were aggregated (using the most predominant type) to the 4-km grid shown in Figure 4.1, to classify the basin according to seven unique soil-vegetation groups. Because it is unlikely that parameters vary solely as a function of soil type alone or of vegetation type alone (Rosero et al., 2009b), and to reduce the total number of parameters studied, we described the domain as a mosaic of soil-vegetation classes (Table 4.1; Figure 4.1d). For simplicity and to ease computational burden, when identifying the soil-vegetation groups, we treated crop and grass as the same vegetation class.

4.3.7. Parameter values

We vary a total of 61 parameters for STD and 64 for GW for the distributed run. In a given grid cell, for each of the STD realizations, 9 parameters deemed important to the simulation of soil hydrology (8 soil and vegetation parameters and 1 basin – topography-related– parameter) were randomly sampled from uniform distributions (see Table 4.2); for the GW realizations, 10 parameters (7 soil and vegetation parameters and 3 basin parameters) were allowed to vary. For each model run, a unique parameter set was assigned to each soil-vegetation class (Figure 4.1d) and to each of the five sub-basins in the watershed. That is, the parameters of each soil-vegetation class and each basin varied independently, and all the cells within a class (or basin) had the same soil-vegetation (or basin) parameters. Ranges in Table 4.2 were taken from the literature (e.g., Chen et al. 1996; Schaake et al. 1996; Bastidas et al. 2006; Hogue et al. 2006). Parameters that were held constant between realizations were set to the default value used

by Niu et al. (2009) for that vegetation or soil type or, in the case of Noah-GW parameters, to the default values set by Niu et al. (2007).

4.3.8. Methods

4.3.8.1. Latin Hypercube Monte Carlo model realizations

Using Monte Carlo simulation, we obtained ensemble predictions of watershed responses using samples of parameter sets drawn from within feasible parameter ranges (Table 4.2). We used uniform prior distributions independently defined for each parameter to sample 125,000 model realizations for STD and 200,000 for GW using a Latin Hypercube sampling algorithm. We used LH because it combines the strengths of stratified and random sampling to ensure that all regions of the parameter space are represented in the sample (McKay et al., 1979; Helton and Davis, 2003). We classified models as behavioral or as non-behavioral based on acceptable or unacceptable behavior (Hornberger and Spear, 1981). The behavioral sample fulfilled a subjective threshold (Beven and Binley 1992) for this classification: conservation of mass (Eq. 4.11). It also minimized two measures of performance: heteroscedastic maximum likelihood estimation (HMLE) of daily flows (at stations 07327447 and 07327550), which accounts for timing, and the Bias of monthly flows at the outlet gage 07327550, which accounts for volume:

$$HMLE = \sqrt[2]{\frac{1}{n}\sum_{i=1}^{n}\left(q_{sim,i}^{t} - q_{obs,i}^{t}\right)^{2}} \tag{4.8}$$

$$q_i^t = \frac{(Q_i - 1)^{\lambda}}{\lambda} \tag{4.9}$$

where q_i^t is the transformed flow (Box and Cox, 1964) with λ=0.3 (Sorooshian and Dracup, 1980).

$$Bias = \frac{1}{n}\sum_{i=1}^{n}\left(Q_{sim,i} - Q_{obs,i}\right) \quad (4.10)$$

$$\frac{E}{P} \leq 1 \quad (4.11)$$

The behavioral runs are those that are best able to reflect the timing and the volume of streamflow without violating the long term water balance. We used Monte Carlo filtering (Ratto et al., 2007) only as a screening tool after which further analysis of the behavioral ensemble was performed.

4.3.8.2. Sobol' sensitivity indexes

We used the variance-based method of Sobol' (Sobol', 1993; 2001) to efficiently identify the factors that contribute most to the variance of a model's response. The first-order sensitivity index ($S_{1,k}$) represents a measure of the sensitivity of the performance of a model realization that is evaluated against observations to variations in parameter x_k. $S_{1,k}$ is defined as the ratio of the variance conditioned on the k^{th} factor to the total unconditional variance of the performance measure (e.g., Eq. 4.8, 4.10). For details see Saltelli (2002). We used the Sobol' semi-random sampling sequence, as implemented in SimLab (Saltelli et al., 2004), to evaluate 8320 and 11008 runs for STD and GW, respectively. The number of realizations allowed us to use a sample size m>128 for each parameter.

4.3.8.3. Ensemble-based performance score

The performance of the behavioural ensemble at every time step *i* was quantified using the score of Gulden et al. (2008b):

$$\varsigma_i = \frac{CDF_{ens,i} - CDF_{obs,i}}{1 - CDF_{\overline{obs}}} \tag{4.12}$$

where CDF is the cumulative distribution function of the ensemble or the observed quantity. The score is lowest (i.e., best) when the ensemble brackets the observation, is highly skilled (observations centred on the ensemble mean), and has low spread. See Appendix 2.

4.4. RESULTS

4.4.1. Most frequent performance and selection of behavioral runs

The typical performance of the 125,000-member STD ensemble and the 200,000-member GW ensemble suggest a wet bias in the total amount of simulated discharge and the inability of both models to adequately capture the timing of the daily streamflow in the LWREW (Fig. 4.3). Both STD and GW tend to overpredict the ratio of runoff to total precipitation (Fig. 4.4b); however, the bias of total watershed discharge simulated by GW tends to be slightly lower than that simulated by STD (solid lines in Fig. 4.3a). The typical simulation of runoff by STD achieves an equally good HMLE as does that of GW (solid lines in Fig. 4.3b). GW tends to overestimate the evaporative flux; the RMSE of its simulated LE is significantly greater than STD (Fig. 4.3c). Dotted lines in the panels of Figure 4.3a and 4.3b show that calibration of model parameters leads to a significant reduction in the simulations' bias and HMLE; however, as reported in myriad other

studies (e.g Koster and Milly, 1997) there exists a tradeoff between a model's achieving better runoff performance and accurate simulation of evapotranspiration. Constraining the top 0.05% of model runs (behavioral) to better capture basic characteristics of the runoff does not yield improved simulations of latent heat flux. The tuning of model parameters significantly improves performance but is insufficient to overcome structural biases in model formulation.

4.4.2. Partitioning of the water cycle

The majority of simulations of STD and GW are unable to capture the fundamental features of the long-term hydrologic response of the basin (Fig. 4.4). We treat the 4-km NEXRAD stage IV precipitation data, used as meteorological input to the model cohorts, as observed precipitation and use it to compute evaporative (E/P) and runoff (Q/P) ratios. Noah's tendency to overestimate runoff volumes is shown in the positively (negatively) skewed E/P (Q/P). The Q/P ratio is overestimated by interquartile range of the STD and GW runs by a factor of 6 (Fig. 4.4b). Treating the evaporation observed at the AmeriFlux site to be approximately representative of the rates for the entire basin, we compute an estimated observed E/P ratio (solid line in Fig. 4.4a). Seventy-five percent of the runs of both models underestimate the evaporative ratio. We presume that this estimated observed E/P is itself likely an underestimate of the actual ET in the LWREW; therefore, the dry bias of the model-ensemble simulated ET is likely even greater than it appears in Figure 4.4. The subset of behavioral models (that achieve the lowest bias and best HMLE scores) do nearly conserve mass and are able to

reasonably accurately simulate the gross characteristics of the LWREW water balance (STD* and GW* in Fig. 4.4).

4.4.2. Ensemble-based evaluation of daily streamflow

Having established that neither STD nor GW is skilled in simulating the large-scale features of the water balance, we sharpened our focus to the daily timescale as a means for understanding why the two version of Noah fail to capture essential features of the water cycle in the LWREW. We examined the best-performing subset of models and examined in more detail the components of runoff simulation and the hydrologic cycle. Results presented in this section apply only to the behavioral (lowest-bias and best HMLE) subset of runs for both STD and GW.

4.4.2.1. Hydrographs and recession curves

The streamflow hydrographs suggest that the models are limited in their ability to capture the timing of daily runoff and have less skill with respect to the magnitude, especially during dry spells. During wet periods, such as the spring and summer of 2007, both STD and GW simulate runoff that is overly flashy: the models are too responsive to small inputs of precipitation, they overestimate the rate of discharge after precipitation events, and the simulated recession of discharge is too fast (Fig. 4.5). After dry-down, STD significantly outperforms GW; however, the difference in performance results from STD's larger baseflow (GW often has no baseflow at all; see further discussion below). During dry periods, such as the summer and early fall of 2005, STD again outperforms GW, especially when baseflow is the primary source of water in the channel (Fig. 4.5).

Both models overestimate post-precipitation increases in discharge and overestimate the speed at which channel flow recedes. Spurious peaks in the hydrograph may indicate that the precipitation forcing data contain errors. The time-mean performance score of the Box-Cox transformed runoff (over the period 1 January, 2002 – 31 December, 2007) at the outlet is 1.42 for GW and 0.99 for STD. (Table 4.3).

In Figure 4.8, it is evident that the models do not have the skill to reproduce the recession events observed on May, Jun and Aug 2007. We consequently do not try to fit a power/exponential-law−type model to the observations to better quantify the recession characteristics. Mismatch between the measured flow recession characteristics (Brutsaert and Nieber, 1977) and those of the modeled flow, is another clear indication that the subsurface flow dynamics of the model need to be investigated.

4.4.2.2. Flow duration curve

We use a flow exceedance probability curve (FEPC) (also known as the flow duration curve) (Vogel and Fennessey, 1994) to summarize the models' ability to simulate the long-term distribution of flows of different magnitudes, which in turn is indicative of the different contributions made by surface and subsurface runoff to total streamflow (Farmer et al., 2003; Yilmaz et al., 2008; van Werkhoven, et al., 2008) (Fig. 4.7). The FEPC represents the flow regime, and its steepness reflects the speed of watershed drainage, which is a result of the watershed functional behavior (Wagener et al., 2007). The gently sloping FEPC of the observed discharge, indicates that groundwater or 'slow' runoff is a significant contributor to the discharge (Smakhtin,

2001) of the Little Washita River system in both its upstream (Fig. 4.7a) and downstream (Fig. 4.7b) reaches. Neither STD nor GW is able to capture this essential baseflow-dominated character of the LWREW streamflow (Fig. 4.7).

Both STD and GW simulate too-frequent high and extreme flows and too-infrequent intermediate and low flows. The models' short, steep FEPCs indicate that the models exhibit significant flow variability and limited flow persistence. At the mid-catchment gauge (7327447), the entire GW behavioral ensemble and part of the STD cohort show that the model is much more flashy (i.e., with low water-storage capacity and overland-flow-dominated runoff) than the actual Little Washita River (Fig. 4.7a). STD is more sensitive to the choice of parameters. At the downstream gauge (7327550), the behavioral ensemble of STD obtains more baseflow from the lowlands of the watershed (likely because of a change in soil-vegetation group type in the downstream reaches). Although the probability of intermediate and low flows in STD is lower than the observed, at the downstream gauge, several realizations of STD do exhibit a distribution of flow volumes that somewhat resembles the slope of the observed FEPC, although STD's intermediate flows are dry-biased with respect to observations (Fig. 4.7b). Even at the downstream gauge, GW simulations remain much flashier than observations. The FEPC of GW provides evidence that, in the LWREW, GW behaves as a simple bucket model that does not parameterize groundwater flow (Farmer et al., 2003; Wagener et al., 2007) (see section 4.5 for a discussion of this dichotomy).

4.4.2.3. Spatial distribution of the runoff partitioning

Consistent with our foregoing observations, a spatial analysis of the ensemble-mean of the cumulative surface (Q_s) and subsurface (Q_{sb}) flow shows that the total runoff (Q_{Total}) estimated by the two versions of Noah LSM is composed mostly of surface, overland, fast runoff (Fig. 4.9). The dominance of Q_s is particularly pronounced for GW. That simulated Q_s/Q_{total} is high is inconsistent with the observed FEPC, which shows a more slowly responding watershed.

4.4.2.4. Sensitive parameters

Analysis of the parameters most responsible for the model's behavior (Fig. 4.10) shows that for STD more than 70% of the variance is controlled by the Clapp and Hornberger *b* of groups D-G, while for GW, less than 50% of the variance can be apportioned to *b* of groups D and E. A quarter of GW's variance corresponds to the porosity (*smcmax*), saturated soil matric potential (*psisat*), and aquifer specific yield (*rous*) of D-F. Despite that D-G correspond to the larger areas in the catchment, the fractions of the variance do not directly correspond to the area covered.

That the Clapp and Hornberger *b* exponent is important for both STD and GW is not surprising: [1] Parameter *b* controls the shape of the pedotransfer function from which the change of soil hydraulic conductivity with saturation is computed. [2] Paramter *b* is also used to provide physical consistency between parameters: multiple internal model parameters (e.g., the wilting point, the saturated soil diffusivity, etc.) are computed using *b* (Chen et al., 1996; Chen and Dudhia, 2001).

Parameter *b* plays a larger role in shaping the variance of runoff in STD than in GW (Fig. 4.10). In STD, *b* is used to compute the maximum rate of infiltration (which controls surface runoff); it is also used to compute the hydraulic conductivity of the bottom layer of soil, of which Q_{sb} is a linear function. In GW, although *b* still plays a role in determining the values of multiple soil hydraulic properties, it does not directly control surface runoff or subsurface runoff.

A comparison of Figure 4.1 and Figure 4.9 shows that in both models, but especially in GW, surface runoff is a function of soil-vegetation group and not of watershed. The only parameter indirectly used to compute surface runoff in GW that is also linked to land cover is the maximum canopy water content (*cmcmax*), which determines the amount of precipitation reaching the surface. Basin-linked parameters *fsatmx* and *f* are also used to compute surface runoff, but Figure 4.9 shows a clear dependence of surface runoff on soil-vegetation group. The variable groundwater table depth (*zwt*) is the only remaining aspect of the GW computation of surface runoff that is indirectly linked to land-cover group, and it clearly is controlled by parameters of each soil-vegetation group (Fig. 4.11). GW's method of calculation of *zwt* explains the contribution to model variance of *smcmax, psisat* and *rous.*

4.4.3. Ensemble-based evaluation of daily soil moisture

4.4.3.1. Soil moisture statistics

Point-based soil moisture measurements are difficult to compare with the spatially smoothed simulations of a model grid; however, statistical properties are often preserved

across scales (Famiglietti et al., 2008). We compare the first, second, and third moments of observed and modeled soil moisture across the LWREW (Fig. 4.12). Observed soil moisture observations reveal that the coefficient of variation (CV) exhibits an exponentially decreasing pattern with increasing mean moisture content. In the upper soil layer (5 cm), the skewness of observed moisture generally decreases, from positive to negative values, with increasing mean soil moisture, with most observations centered around zero. In the root zone (25 cm), observed skewness shows approximately the same pattern, but with more scatter, and is on average slightly positive. Of the behavioral subset of model realizations, neither STD nor GW captures the essential character of the soil moisture statistics. Skewness is far too positive at both depths, and the coefficient of variation of simulated moisture increases with mean soil moisture. The addition of the groundwater module to STD does not fundamentally change the character of simulated soil moisture (Fig. 4.12). Observed soil moisture is more normally distributed than is modeled. In both models, simulated soil moisture is especially positively skewed for the driest cells: the model soil columns saturate quickly and then dry quickly, favoring lower-than-mean moisture. At both depths, observed soil moisture is more variable than modeled, and is most variable in drier cells. Near the surface, lower-mean grid cells have less moisture variation than their wetter counterparts; at depth, lower-mean grid cells exhibit more variation than their wetter counterparts. For both STD and GW, model output is consistent with expectations but not with reality. We (and likely the model developers) expect that the mean state of the soil moisture profile will monotonically wet with depth; yet observations show that in some cases this is not the case.

4.4.3.2. Upper layer and root zone soil moisture

We use observations from two selected sites from within the basin (A148 and A153; see Fig. 4.1), each with distinct wetting profile and behavior, to evaluate model performance. The ensemble mean, time-mean soil moisture profile of GW and STD slowly wet with depth at both sites, which is not consistent with observations (Fig. 4.13). Simulated gradual wetting with depth is consistent across the basin; only the uppermost layer of soil varies consistently between soil-vegetation groups (Fig. 4.14).

Time series of simulated soil moisture are plausibly realistic at both 5 cm and in the rooting zone, although both STD and GW simulate soils that exhibit a dry bias in the top layer when compared to observations (Fig. 4.15, 4.16). Although the simulations exhibit little differentiation between sites and between regions of the catchment, the models tend to perform better in the root zone of site A148 (Fig. 4.16). The amount of time that it takes for the soil to dry down is consistent with observations, although the magnitude of the change in modeled soil moisture is normally much greater than what is observed. It appears that the model may have a (dry) equilibrium state that it strongly prefers, possibly in spite of local forcing (Fig. 4.16). Performance scores for both models at the sites and depths are very similar (Table 4.3).

4.4.4. Ensemble-based evaluation of daily evapotranspiration (ET)

In most parts of the basin, the time-averaged ensemble-mean ET rates are much larger in GW than in STD (Fig. 4.17); a qualitative examination of the spatial distribution of ET shows that ET rates are controlled by soil-vegetation group parameter choices, not

by basin-related parameters. Examination of the performance of the behavioral ensemble when simulating the time variation of daily ET at a single grid cell (where the FLUXNET tower is located) shows that both GW and STD are too variable in their ET simulation and show that both models, but especially GW, overestimate ET at the given site (Fig. 4.18). This result is consistent with the overly robust evapotranspiration pathway observed for GW in previous studies (Rosero et al., 2009a). We note that it is possible that the eddy-flux tower location from which the ET data was collected may not be representative of the ET flux averaged across the domain of the overlapping 4-km grid cell.

4.5. DISCUSSION

The failure of our implementations of STD and GW to realistically represent runoff in a small baseflow-dominated watershed appears to result in large measure from the models' inability to adequately represent the soil hydrology and a steady subsurface runoff. Consequently, both models significantly overestimate the fraction of total runoff (Q_T) that is rapid. Our results are consistent with the conclusions of Boone et al. (2004), who observed that, in general, higher ratios of surface runoff (Q_s) to total runoff (Q_s/Q_T >0.25) corresponded to less-realistic simulated discharge. Lower Q_s/Q_T values were especially important for obtaining good performance at a daily timescale. Boone et al. (2004) also observed that schemes with little water-storage capacity in their soils tend to overestimate runoff; both STD and GW can be characterized as having low water-storage capacity in their soils: they both wet and dry too quickly in response to precipitation

events. The flashy response of the model watersheds is in part a consequence of low water storage in the modeled soil column.

We note that in the current implementation of GW, surface runoff is needlessly increased by the scaling factor 0.5 in the exponential term used to scale the precipitation rate (Eq. 4.7). Given the observations of Boone et al. (2004) and others regarding improved simulations obtained with models that have a low Q_s/Q_T, we suggest that this factor need be either eliminated (thereby effectively increased to 1.0) or increased to force a decrease in surface runoff. However, a simple decrease in surface runoff is not sufficient to create a constant supply of baseflow. Modifying the groundwater formulation used here such that it provides a time-delayed second reservoir for flow and such that it is able to generate a steady subsurface flow, even in regions where the water table is low, will likely improve the Noah LSM's capacity to simulate more physically realistic streamflow in the LWREW.

The current GW parameterization does provide a constant reservoir that is a potential source of runoff, but in its current implementation, GW does not function effectively when the water table is low because modeled surface and subsurface runoff decrease exponentially with water table depth (Eq. 4.6 and 4.7). Given the current parameterization, when the water table falls below 10 meters beneath the land surface, little subsurface runoff is produced (Fig. 4.19). The modeled equilibrium groundwater tables for the LWREW in the behavioral GW runs range from 1 up to 80 m in some cells (Fig. 4.11), with most values being deeper than regional observations (10-25 m; USGS water data and D. Moriasi, personal communication). While previous work using the

same or similar implementations of the Niu et al. (2007) groundwater model have shown that the GW module performs realistically in simulating various aspects of the terrestrial water cycle (Niu and Yang, 2003; Niu et al., 2007; Gulden et al., 2007a; Lo et al., 2008: Rosero et al. 2009a, 2009b), it is necessary to point out that, in the other researchers' simulations, domain-average water tables have been shallow. GW seemed to degrade the simulation of near surface fluxes and states in regions of transition zones were the water table is believed to be deep (Rosero et al. 2009a, 2009b).

One potential, domain-specific solution is to set the tunable parameter *f* near zero such that there is only a very weak exponential dependency of runoff on depth to water (see Eq. 4.6 and 4.7). We investigated physically plausible values of *f* (Table 4.2). Niu and Yang (2003) provide a range from 1.5 to 5.2 of physically realistic values of *f* reported in the literature using similar topography-based runoff schemes in somewhat similar modeling environments (Famiglietti et al., 1992; Stieglitz et al., 1997; Chen and Kumar, 2001; Dai et al., 2003). The calibrated values adopted by Niu et al., (2005, 2007) are shown in Figure 4.19. However, in such studies, the resolution of the grid cell was coarser, and the depth to the (parameterized) water table was relatively shallow, which made the exponential component of the subsurface runoff significantly greater (see Eq. 4.7). The ideal value of *f* likely changes with grid cell size, with soil properties, with modeled equilibrium depth to water, and with host model (Rosero et al., 2009a). By comparing our runs with those of other researchers' results we clearly see that *f* must be treated as a scale dependent tunable parameter, not a physical quantity.

We also note that a potential explanation for the deeper-than-observed modeled water table is an overly robust parameterization of soil matric potential, which sucks water from the overly deep groundwater reservoir and contributes to significantly overestimated ET (Fig. 4.18).

Other potential explanations for the poor quality of simulated streamflow are that the soil hydrology representation of Noah is insufficiently complex and/or not realistic. This potential limitation is consistent with the poor-quality behavioral simulations of STD runoff, which appear to result from the model's low soil-water residence time. The increasing or constant CV with increasing mean soil moisture implies that the model does not have the capacity to retain or to buffer soil moisture. That is, model grid cells with high porosity likely have larger mean moisture because they occasionally are briefly saturated; however, all cells, including the cells that are wetter on average, dry quickly. For cells with higher porosity, this behavior increases the CV of soil moisture content. Such quick-to-wet, quick-to-dry behavior may be ameliorated by increasing the number of layers in the modeled soil column. However, such a change may be insufficient to fundamentally alter the statistics of the modeled soil moisture.

It is worth noting that Famiglietti et al. (2008), working in the same region, observed an overall increase in skewness with the scale of soil moisture measurements, which implies that positive skewness of the soil moisture distribution under dry surface conditions may be more pronounced at the larger end of a range of scales. This observation may help explain why our 4-km soil-moisture simulations have skewness values that are much larger than those of observations.

The runoff parameterizations within GW are related to topography but do not actually depend on the statistics of topography. In the development of the physically based parameterization of GW it was assumed that identification of parameters which, by derivation, are related to topography (*fsatmx* and *rsbmax*) has the potential to capture the heterogeneity of the land features and improve both simulated runoff and simulated soil moisture. Our sensitivity analysis showed that adjusting parameter *rsbmx* within GW to better reflect within-watershed variations of topography has little to no effect in improving both the statistics of soil moisture and the realism of the simulation of runoff. In the derivation in of the simplified model, Niu and Yang (2003) state: "it is attractive to develop a topography-related runoff parameterization which does not require the topographic index data set. In the simplest case, the topographic characteristics may be parameterized as constants for all land points, and the saturated fraction and subsurface runoffs are only determined by the soil moisture represented by the water table depth". It is evident that the dependency of the grid cell topography is largely lost when using the simplification embedded within the maximum rate of subsurface runoff (*rsbmx*). Similarly, the conceptual maximum saturated fraction (*fsatmx*) becomes a tunable parameter. Hence, GW's simplifications for surface and subsurface runoff used here are disconnected from the actual physics of topographic influence on groundwater discharge. It is therefore not surprising that, without extensive calibration, they do not yield significant improvement in the physical realism of model simulations.

In the LWREW, a combination approach may be warranted. GW is a modified saturation-excess runoff scheme, which is valid in humid regions, zones with large

infiltration capacity, and well-distributed precipitation. STD uses an infiltration-excess scheme, which is better suited to dry regions with sparse, localized rain or in humid areas where soils are impermeable.

In order to accurately predict streamflow or other hydrologic fluxes and states, choosing the appropriate model structure (and model parameters) is a crucial step in hydrologic modeling. The same can be said about understanding the dominant physical controls on the response of a watershed (Clark et al., 2008). In land-surface modeling, often a bottom-up approach is followed (as is done here). LSM are complex structures that generally require detailed information of the physical characteristics of the modeled watersheds and are often potentially over-parameterized. As pointed out by Jakeman and Hornberger (1993), model overparameterization is particularly acute when simulating streamflow. Instead, and in the context of hypothesis testing, a top-down approach to model development is advisable (e.g., Farmer et al., 2003; Schultz and Beven, 2003; Sivapalan et al., 2003; Bai et al., 2009). The aim should be to identify a model structure with the minimum level of complexity that is capable of reproducing the observed watershed response for the 'right reasons' (Kirchner, 2006). The gap between the simplified hydrological model components implemented in atmospheric models and the state-of-the-art integrated hydrological models (Overgaard et al., 2006) can only be bridged with approaches that systematically increase the complexity of the subsurface hydrologic parameterization in a framework that acknowledges explicitly the inherent uncertainty of the problem (Clark et al., 2008).

4.6. CONCLUSIONS

We conclude that, in their current formulations and on a 4-km grid, neither STD nor GW is able to capture the essential characteristics of runoff in the Little Washita River basin. A fundamental failure of the Noah STD soil parameterization is its inability to produce sustained baseflow for streams; the addition of the simple groundwater parameterization used here does not ameliorate this deficiency. In regions where the modeled water table is deep (> 10 m below the surface), GW does not simulate sufficient baseflow and instead causes the model to function as a simple bucket model. Both models have too high a ratio of surface to subsurface runoff and consequently simulate streamflow that is far too flashy. In both models, the soil column wets too quickly and dries too quickly. We note that parameters for both models are likely scale and site dependent, and we underscore the need for even 'physically based' models to be calibrated at all locations in which they are applied.

4.7. ACKNOWLEDGEMENTS

The author was supported by the Graduate Fellowship of the Hydrology Training Program of the OHD/NWS. The project was also funded by the NOAA grant NA07OAR4310216, NSF, and the Jackson School of Geosciences. B. Li, at NASA/GSFC, provided us with 1 km-LIS land-cover, monthly vegetation fraction and albedo climatology data. The NEXRAD stage IV was prepared by S. Hong at UT Austin. We acknowledge the USGS, ARS micronet and Ameriflux for the validation datasets. I

thank D. J. Gochis at NCAR and K. Mitchell at NCEP for their insight. We benefited from the computational resources at the Texas Advanced Computing Center (TACC).

Table 4.1. Soil-vegetation properties

Soil-vegetation group	Vegetation type	Vegetation index*	Soil type	Soil index*	Number of 4-km grid cells	Area km^2 (fraction %)
A	Wooded grassland	7	Sand	1	1	16 (2.63)
B	Wooded grassland	7	Loam	6	1	16 (2.63)
C	Grassland/Cropland	10, 11	Sand	1	2	32 (5.26)
D	Grassland/Cropland	10, 11	Sandy loam	3	9	144 (23.68)
E	Grassland/Cropland	10, 11	Silty loam	4	9	144 (23.68)
F	Grassland/Cropland	10, 11	Loam	6	10	160 (26.31)
G	Wooded grassland	7	Sandy loam	3	6	96 (15.78)

* See Figure 4.1

Table 4.2. Bounds of distributions of parameters allowed to vary between realizations

Name	Description	units	Feasible range
Soil-vegetation parameters[#]			
K_{ref}[§]	(*refdk*) Used with *kdtref* to compute runoff parameter *kdt*	−	0.05−3.0
kdt_{ref}[§]	(*refkdt*) Surface runoff parameter	−	0.1−10.0
rcmin	Minimum stomatal resistance	s m^{-1}	30−200
fxexp	Bare soil evaporation exponent	−	0.1−2.0
b	Clapp−Hornberger *b* exponent	−	2−12
θ_{max}	(*smcmax*) Porosity	m^3 m^{-3}	0.2−0.5
psisat	saturated soil matric potential	m m^{-1}	0.03−0.76
K_{sat}	(*satdk*) saturated soil hydraulic conductivity	m s^{-1}	0.1−10
rous[†]	Aquifer specific yield	m^3 m^{-3}	0.05−3.0
Basin parameters*			
rsbmax[†]	Maximum rate of subsurface runoff	m s^{-1}	$1.0E^{-6}$−$1.0E^{-3}$
f[†]	e-folding depth of saturated hydraulic conductivity	m^{-1}	0.5−10
fsatmx[†]	Maximum saturated fraction	%	0.1−90
slope[§]	Slope of bottom soil layer	−	0−1

[#] Assigned to all the cells within a soil-vegetation class (see Fig. 1d)

*Assigned to all the cells within a sub-basin to better capture the topographic relief of the catchment.

[§] Parameter is used by Noah-STD only.

[†] Parameter is used by Noah-GW only.

Table 4.3. Performance score of the behavioral ensembles

	Runoff (Q_{Total})			SMC_{5cm}		SMC_{25cm}	
Station	7327442	7327447	7327550	A148	A153	A148	A153
STD	1.66	1.42	0.99	0.58	0.57	0.61	0.67
GW	1.6	1.69	1.42	0.62	0.63	0.6	0.71

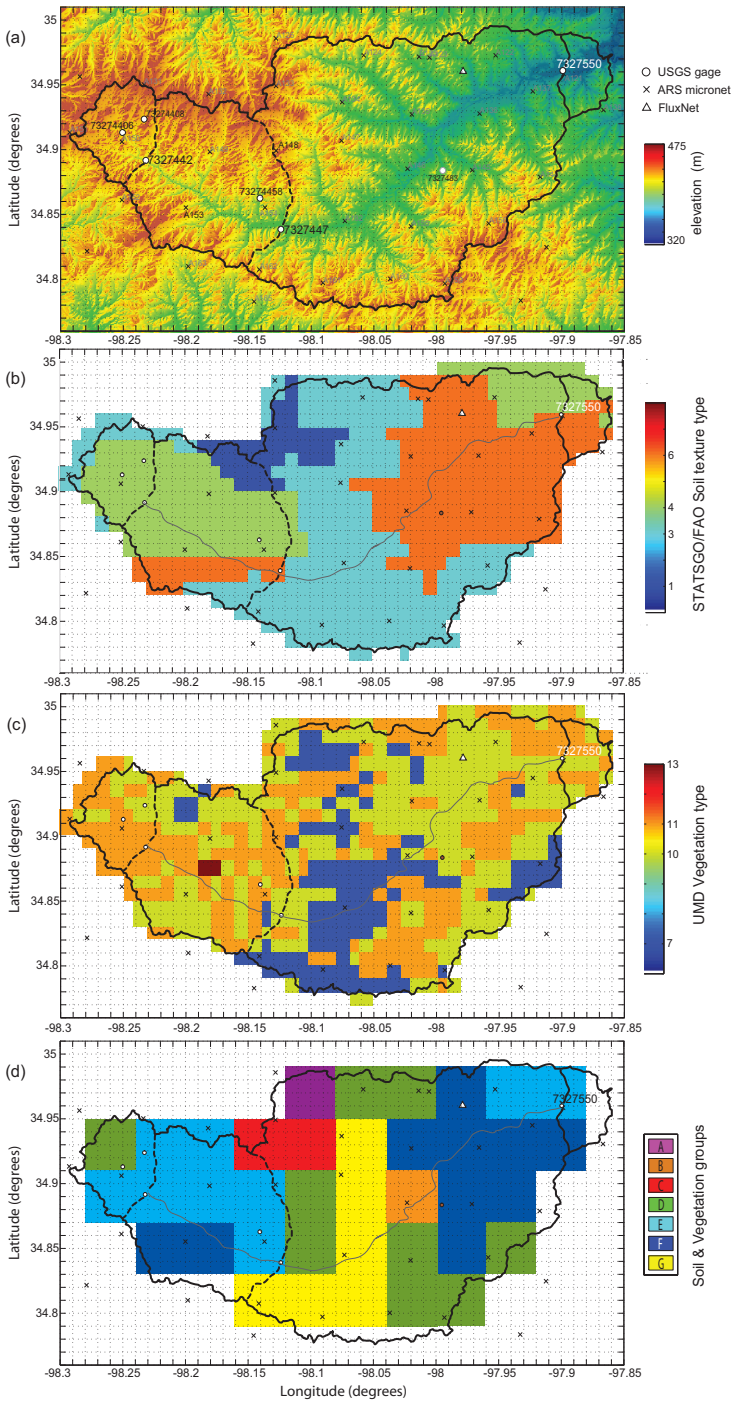

Figure 4.1. Little Washita River Experimental Watershed (LWREW) modeling domain.

(a) Hydrography and locations of the USGS streamflow gages, ARS soil moisture observation sites, and the FLUXNET tower. (b) 1-km FAO/STATSGO soil texture data. (c) 1-km UMD vegetation type data. (d) Groups A-G of cells with the same vegetation and soil types on the 4-km modeling domain used in all model realizations described here. Note the delineation of 3 sub-basins: upstream (7327442), mid-catchment (7327447) and downstream at the outlet (7327550). See Table 4.1 for soil and vegetation classification.

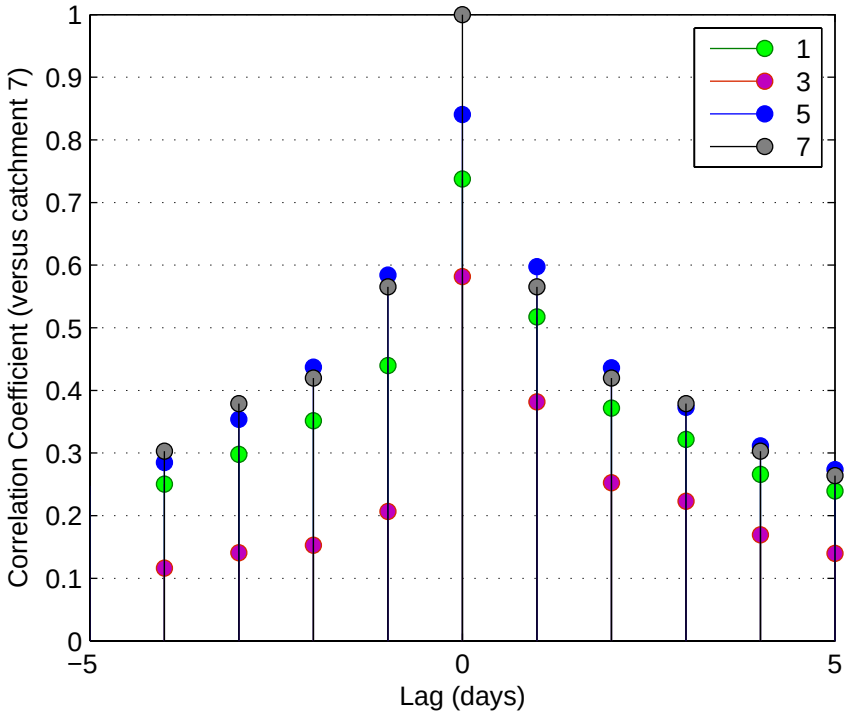

Figure 4.2. Lag-correlation coefficients between streamflow at the outlet gauge (07327550) and gages upstream.

The maximum correlation of the time series correlation has a time lag of 0 days. In the figure, 1 is USGS gauge 07327327; 3 is 07327442; 5 is 07327447; and 7 is the outlet 07327550.

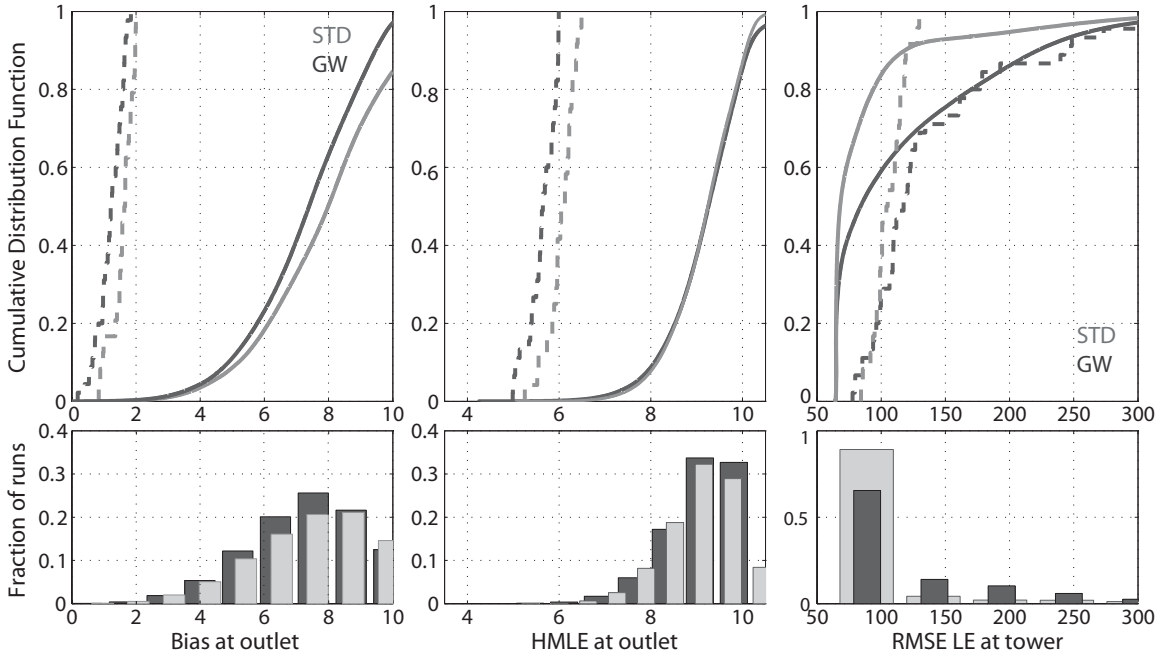

Figure 4.3. Performance of all realizations of STD and GW.

Cumulative distribution functions (CDF) and histograms are shown for (a) the Bias at the watershed outlet, (b) the HMLE at the watershed outlet, and (c) the RMSE of the 1998 latent heat flux. In all panels, CDFs with solid lines are those for all Monte Carlo realizations of STD (gray) and GW (black); dashed lines are the CDFs of the behavioral runs (for which both Bias and HMLE were minimized).

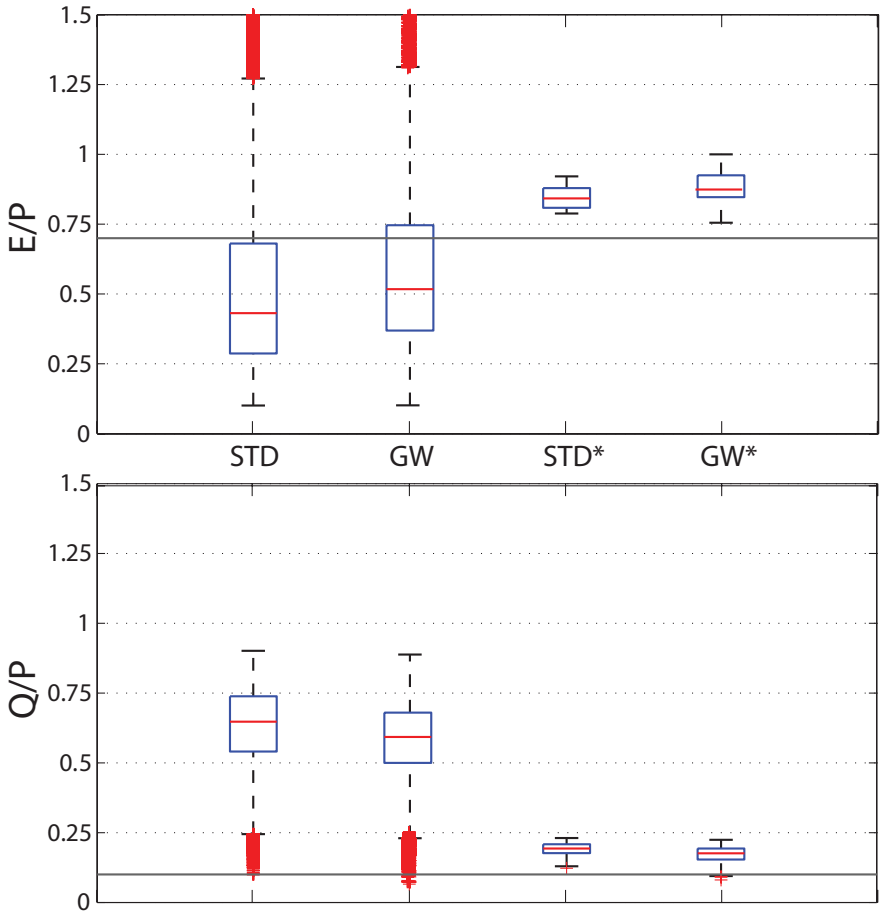

Figure 4.4. Box plots showing the 2002-2007 hydrologic response of the basin.

Hydrologic response is defined in terms of (a) evaporative (E/P) and (b) runoff (Q/P) ratios for are all Monte Carlo realizations and the behavioral subset of runs (*), which minimized Bias and HMLE. The box shows the interquartile range (i.e., the range between the first and the third quartiles) of the ratios and the length of the whiskers is 1.5 times the vertical scale of the boxes. Ratios outside of the whiskers are regarded as outliers and marked as crosses in the figure. For reference, the horizontal line in (a) stands for E/P=0.7 observed at the FLUXNET tower for 1997−1998. The line in (b) stands for Q/P=0.1 observed at the outlet for 1997−2007.

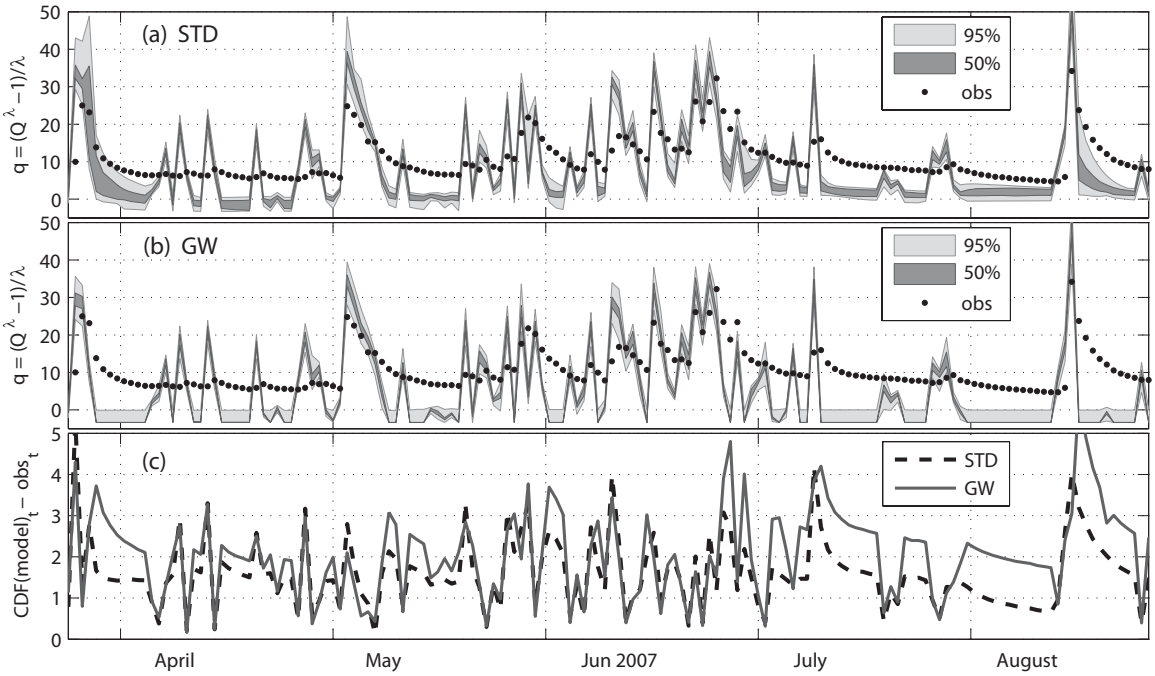

Figure 4.5. Daily streamflow hydrograph simulated at the outlet (7327550) by the behavioral ensemble of STD and GW during a wet period in 2007.

Transformed observed daily streamflow observations [cfs] are shown as black dots. Transformed runoff is used for improved visualization of both high and low flows. For both STD (a) and GW (b), the 50 and 95% confidence intervals are shown. (c) Performance score (lower is better) of both STD and GW shows that both are too flashy (too high peaks and too persistent low flows), but STD consistently outperforms GW, especially during dry-down periods.

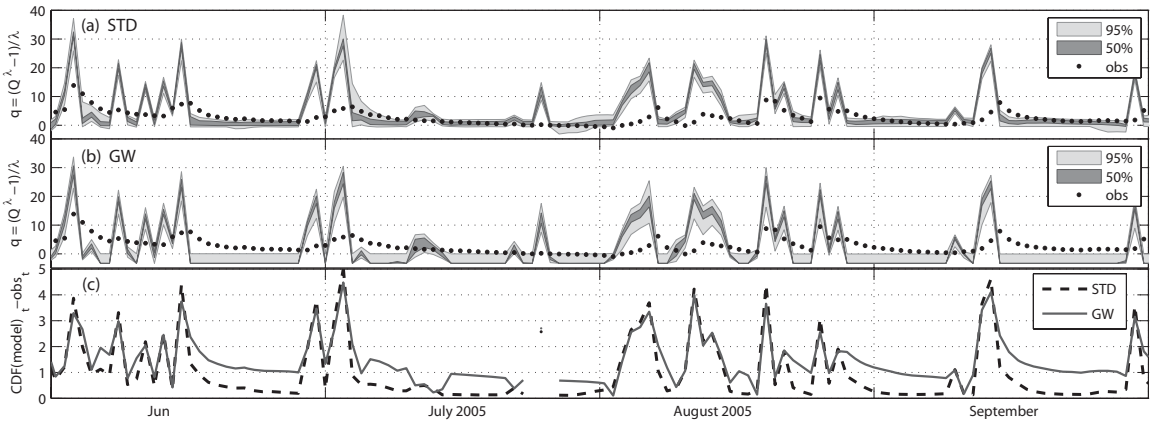

Figure 4.6. Daily streamflow hydrograph simulated at the outlet (7327550) by the behavioral ensemble of STD and GW during a dry period in 2005.

See Figure 4.5.

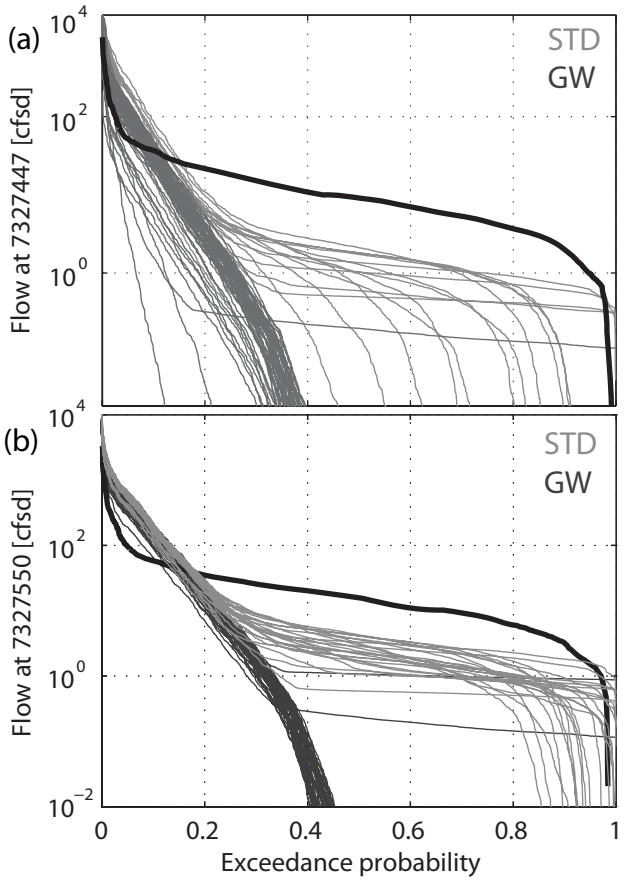

Figure 4.7. Flow exceedance probability curves (FEPC) of the Q_{total} simulated by the behavioral ensembles of STD and GW for 2002-2007.

FEPCs are shown at (a) the intermediate gage (7327447) and at (b) the outlet (7237550). The observed FEPC (black) is baseflow-dominated. The FEPCs of the GW cohort (dark gray) resemble those of a bucket model; the FEPCs of STD (light gray) show a distribution more similar in shape to the observed but underpredicts medium and high probability events. For both, low probability, high-flow events are overpredicted.

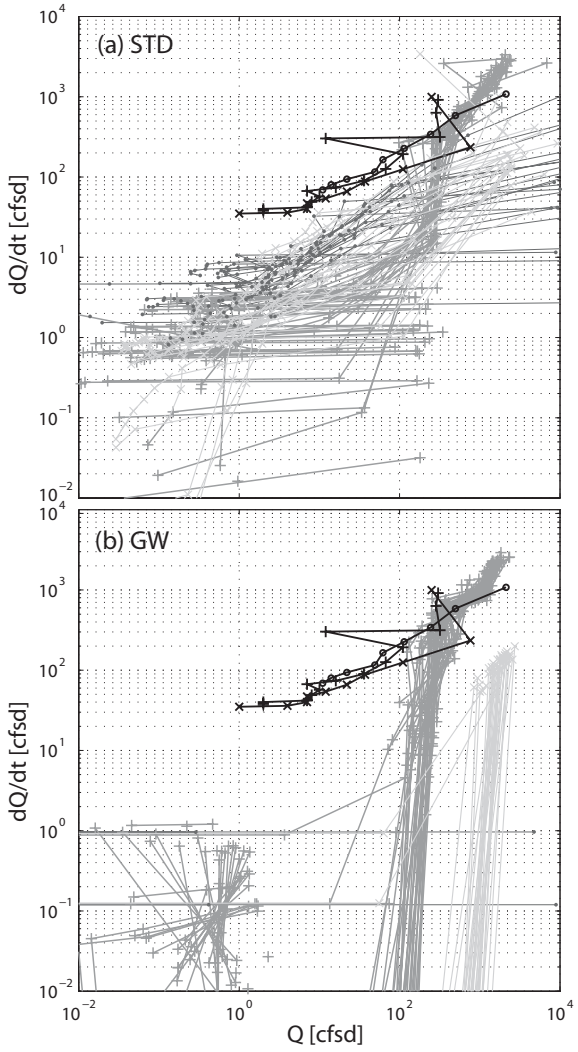

Figure 4.8. Flow recession curve for events in (x) April, (+) May and (o) August 2007. See Figure 4.5.

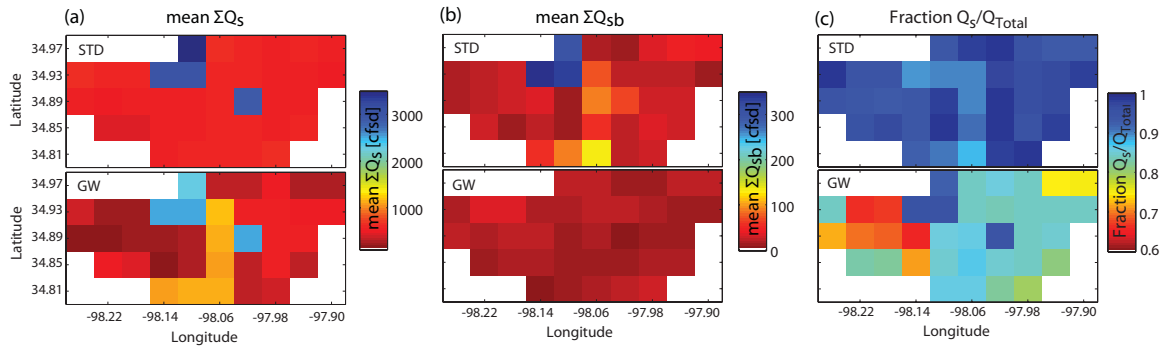

Figure 4.9. Spatial distribution of ensemble-mean cumulative surface and subsurface runoff.

Panel (a) shows surface runoff; (b) subsurface runoff; (c) the ratio of surface to total runoff. GW (lower panels) has a higher ensemble-mean Q_s/Q_{total} than STD (upper panels). In both STD and GW, surface runoff is controlled by soil-vegetation group distribution.

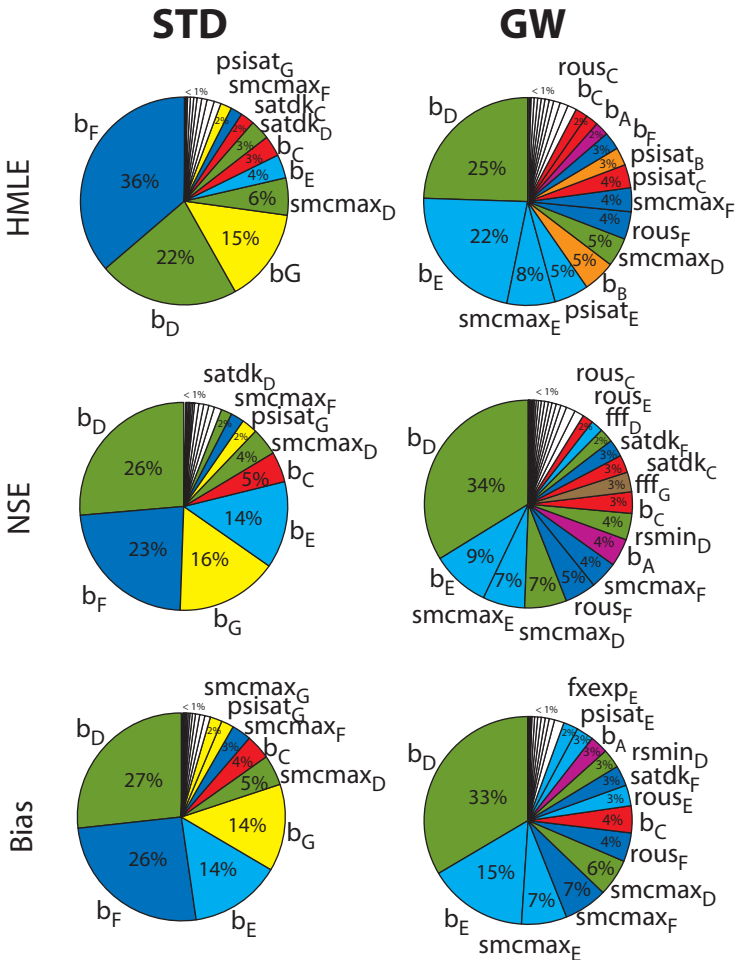

Figure 4.10. Relative contribution of parameters to variance of the HMLE, NSE, and Bias of simulated streamflow.

Sensitivity analysis for STD is shown in left column; that for GW is shown in right column. Parameters are color-coded by soil-vegetation group type (see also group colors in Fig. 4.1d). Group types that cover larger areas (e.g., soil-vegetation groups D, E, F, and G) tend to have more importance in shaping variance.

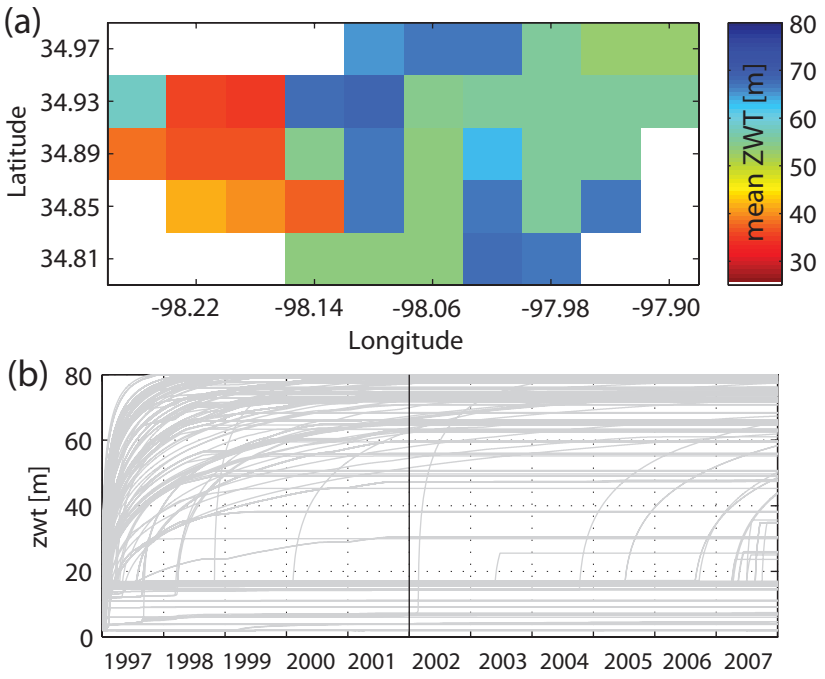

Figure 4.11. Depth to groundwater table (*zwt*) simulated by the behavioral ensemble of GW.

(a) Ensemble-mean depth to groundwater; (b) time series of water table depth for all grid cells of all behavioral ensemble members.

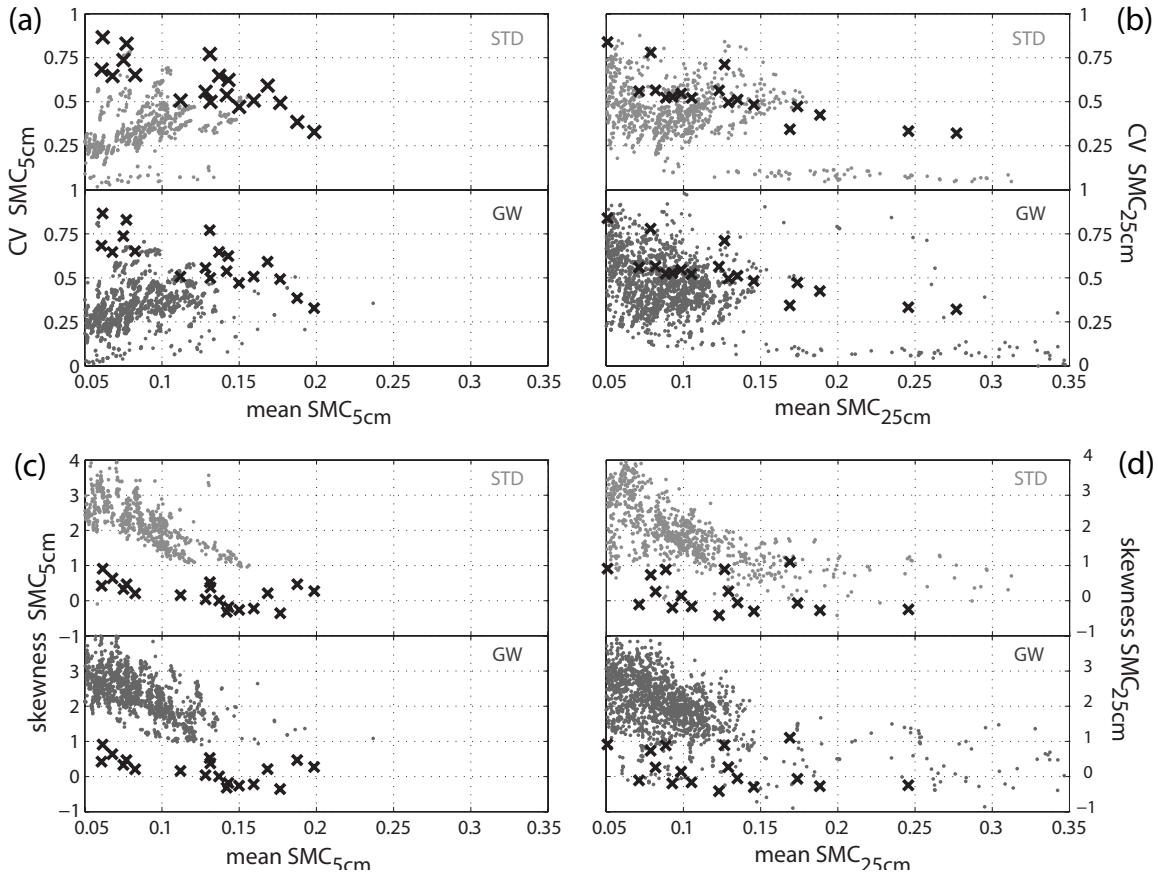

Figure 4.12. Scatter plots of soil moisture statistics for observed and simulated volumetric soil moisture content (SMC).

Mean SMC vs. the coefficient of variation (CV) of SMC are shown for (a) 5 cm and (b) 25 cm. Mean SMC vs. the skewness of SMC are shown for (c) 5 cm and (d) 25 cm. The subsets of the simulated soil moisture statistics (STD: light gray dots; GW: dark gray dots) tend not to follow the same patterns as ARS observations (black crosses).

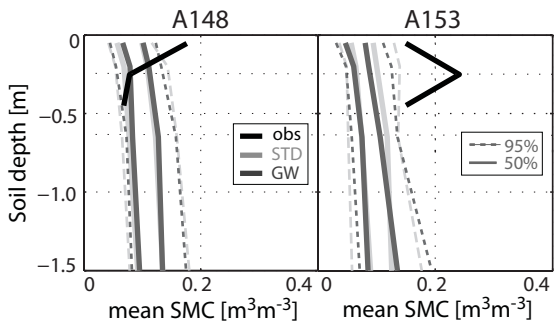

Figure 4.13. Ensemble-mean SMC profile compared with observations at ARS sites A148 (north upper catchment) and A153 (south upper catchment).

Ensemble mean SMC profiles are more consistent between behavioral models and between sites than they are with observations. See also Figure 4.1a.

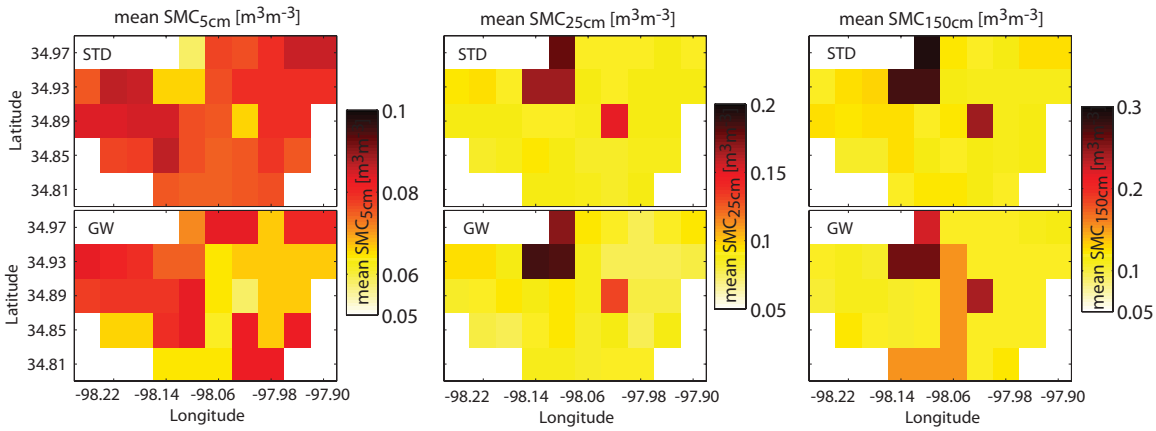

Figure 4.14. Spatial distribution of ensemble-mean average soil moisture content (SMC).

SMC is shown for STD (top panels) and GW (bottom panels) at depths of (a) 5, (b) 25, and (c) 150 cm. Note that the limits on the color-bar legends are not the same between panels. SMC at 5-cm appears to be strongly related to soil-vegetation group. Models in general experience slow wettening with depth.

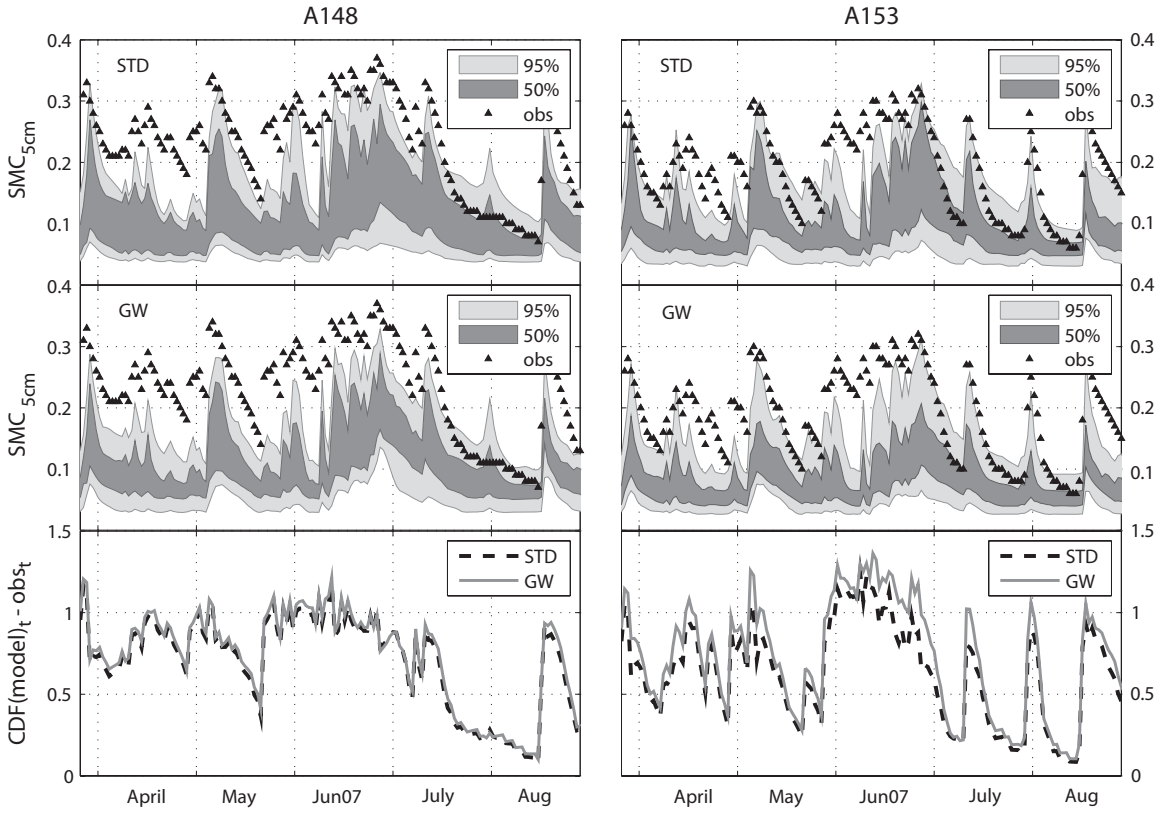

Figure 4.15. Time series of observed and modeled 5-cm soil moisture content (SMC) at ARS sites A148 and A153 for the spring and summer of 2007.

The 50% and 95% confidence intervals of the subset of STD (top row) and GW (middle row) are plotted with daily mean 5 cm-SMC (triangles). Time-varying performance scores (lower is better) of both models are shown in the bottom row. STD slightly outperforms GW at A148 and outperforms GW at A153.

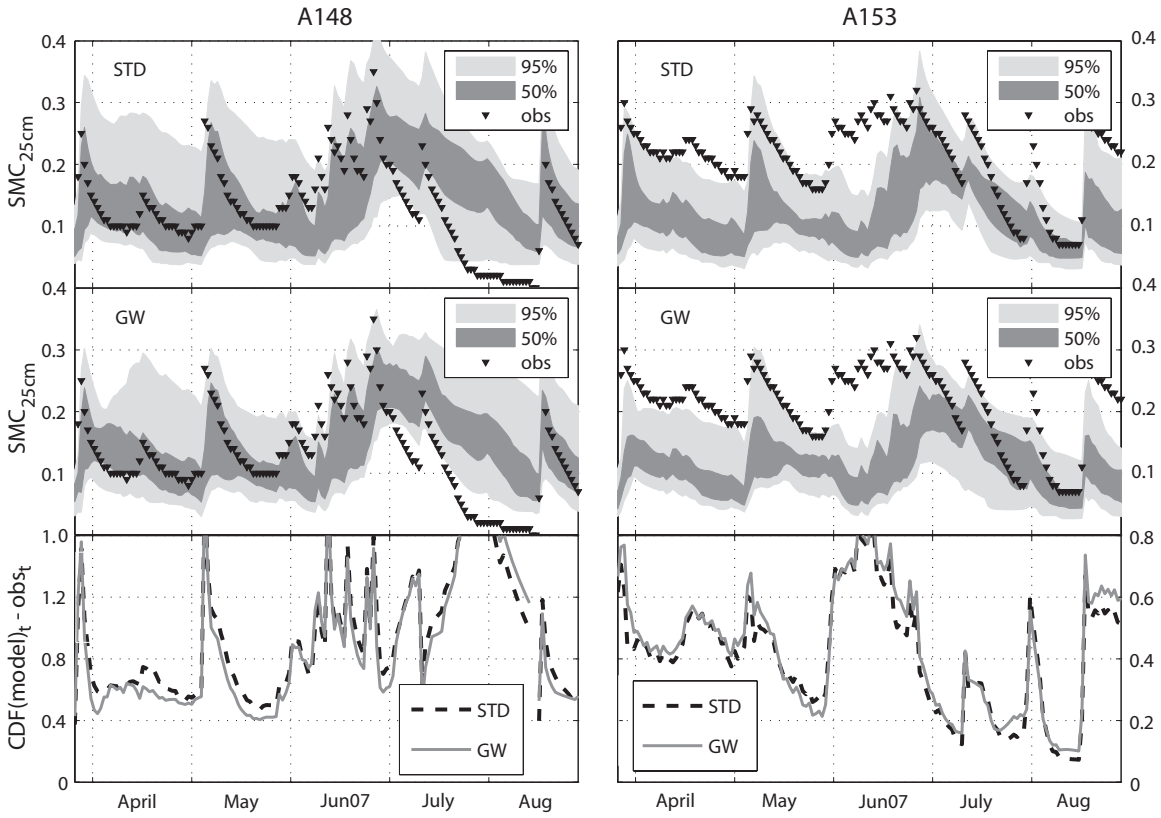

Figure 4.16. Time series of observed and modeled 25-cm soil moisture content (SMC) at ARS sites A148 and A153 for the spring and summer of 2007.

See legend of Figure 4.15. GW outperforms STD at site A148; the models are approximately equally well suited to simulate site A153.

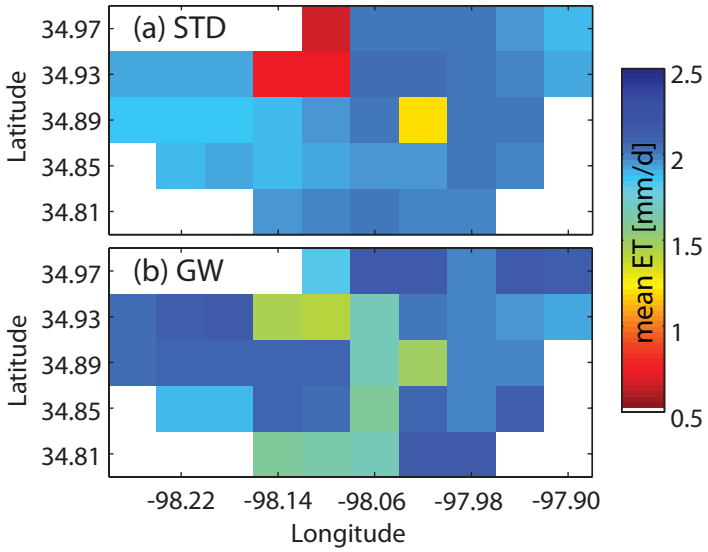

Figure 4.17. Spatial distribution of the simulated ensemble-mean, average evapotranspiration for 2002-2007.

Spatial patterns roughly correlate with soil-vegetation group.

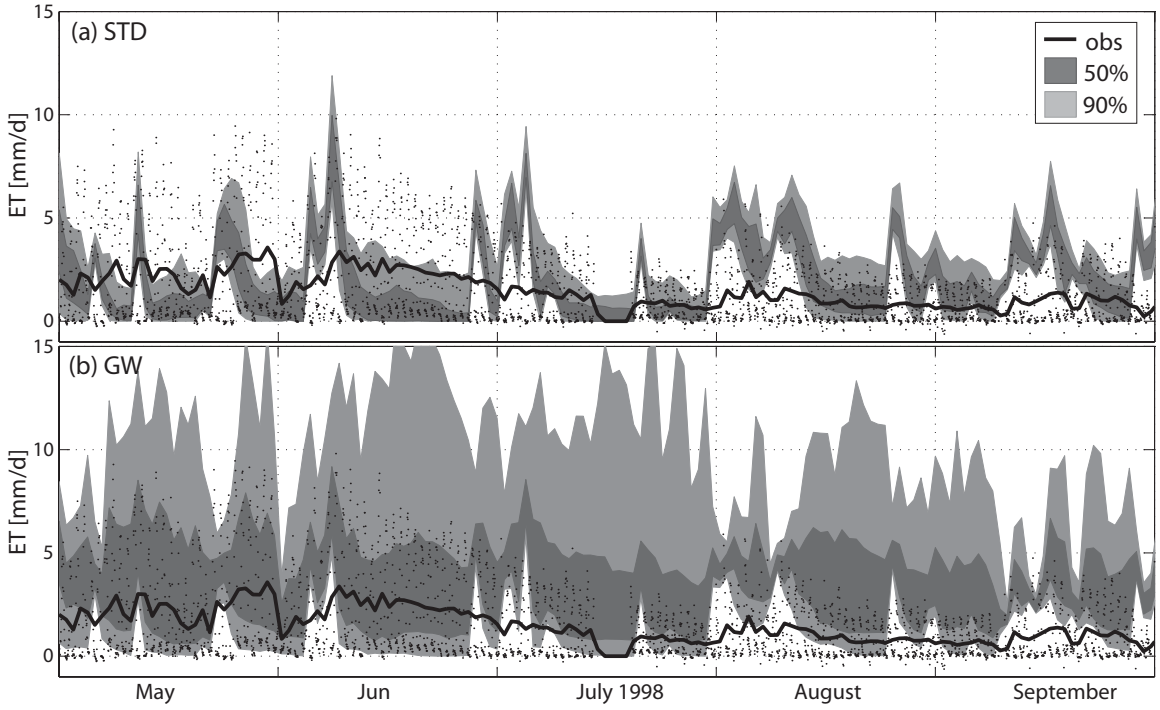

Figure 4.18. Time series of simulated and observed ET at FLUXNET site Little Washita for 1998.

Black dots are hourly ET observations. Daily mean observed values (black line) are significantly less variable and have a lower mean value than do the simulated daily mean values (a) STD (b) GW. Both the 50% and 90% intervals of the behavioral subset of GW realizations overestimate ET. See also Figure 4.1a.

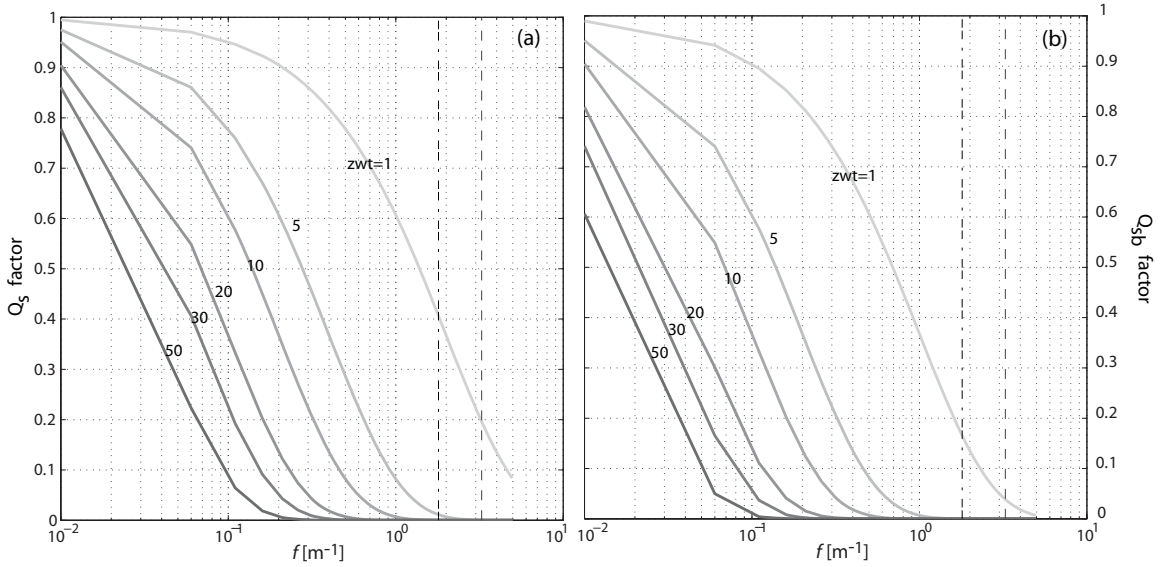

Figure 4.19. Sensitivity of GW's surface and subsurface runoff to depth to water table (*zwt*) and the *f* parameter.

Sensitivity of surface runoff to depth to water table (*zwt*) is shown in (a); subsurface is shown in (b). Parameter *f* is e-folding depth of the exponential decay of saturated hydraulic conductivity). Dashed lines are values of calibrated parameters used by Niu et al. (2005; 2007).

Chapter 5: Summary, conclusions, and contributions

5.1. Overview of work completed

The work presented here advances diagnosis of the hydrologic parameterizations of land-surface models (LSM) to encompass the assessment of characteristic model behavior ('signatures') in feature, cost-function, and parameter spaces. I exhaustively evaluated hypotheses underlying the implementation of new representations of land-memory mechanisms by comparing the performance of three versions of the medium-complexity Noah LSM when simulating high-resolution near-surface states and fluxes in zones of transition between humid and arid climates of the continental U.S. The first version of the LSM was the benchmark, standard Noah release 2.7 (STD), which is the land component of the model used in operational weather forecasting by the NOAA National Centers for Environmental Prediction and which is also widely used in weather and climate research. The two augmented versions, GW and DV, were developed as part of community efforts to increase the conceptual (and physical) realism of the model. GW couples a simple aquifer model to the soil column of Noah and implements topography-related parameterizations of surface and subsurface runoff. DV replaces prescribed biomass in Noah with a mechanistic representation of the vegetation response to short-term environmental variation. I used an ensemble-based framework for model diagnostics that allowed me to account for sources of uncertainty and to reach conclusions about the capacity of the models to accurately and reliably reproduce characteristics of the system that are independent of the choice of parameters.

In Chapters 2 and 3, I assessed typical model behavior at a point scale using high-temporal-resolution heat fluxes, soil states, and meteorological forcing data from the U.S. Southern Great Plains collected by the observational campaign of the International H_2O Project 2002 (IHOP). The dataset contains 45 days of 30-minute data for a set of three dry sites (MAP=550 mm), a set of three wet sites (MAP=900 mm), and a set of three semi-humid sites (MAP=750 mm). I quantified the models' ability to partition the energy balance and the evolution of moisture in the root zone, and I evaluated the parameters that the models require to produce accurate simulations. In Chapter 4, I used aggregated signatures of the land surface at the catchment scale to evaluate distributed simulations of 11 years of daily streamflow and high-spatial-resolution near-surface states in the Little Washita river basin (MAP=760 mm) on a 4-km grid. Using ensembles of models constrained to reproduce the timing and volume of streamflow, I evaluated the models' ability to partition the water balance, to simulate the long-term distribution of flow volumes, and to reproduce the characteristics of soil moisture in the upper and root zone.

The approach to LSM evaluation presented in this dissertation enabled me to identify shortcomings in the formulations of the parameterizations that hinder the models' capacity to simulate near-surface states and fluxes of water and energy even when the models employed optimized parameters; it facilitated my challenging of common practices and assumptions in LSM development; and it allowed me to present the community with new, stringent ways to test models that bridge the gap between the model identification and model development communities. Such an approach will help to

ensure continued improvement of our understanding and modeling of environmental processes.

5.2. MAJOR CONCLUSIONS AND CONTRIBUTIONS

Although some of the conclusions offered in Chapters 2, 3 and 4 are to a certain extent model and site specific, the associated implications for model development and evaluation are significant and broad-reaching.

The analysis presented in Chapter 2 showed that traditional, single realization model evaluation (using default or calibrated model parameters), as is typically performed in LSM intercomparison experiments and during the development of LSMs, is incomplete and uninformative. Traditional model evaluation does not have the capacity to distinguish between models under parametric uncertainty and therefore has little diagnostic power. Equifinality poses a significant problem for the *ad hoc* methods used to evaluate and compare models, and raising the models to optimal performance via calibration does not have the power to diagnose structural deficiencies in model formulation. Models are flexible, and the variation of parameters serves to mask errors. Indistinguishability of the calibrated models' goodness-of-fit occurred at all sites studied, even though multiple criteria parameter estimation was used. Such equifinality does not mean that models cannot be used for scientific inquiry; however, it indicates that single realization, aggregated goodness-of-fit comparisons provide very little information by which one can distinguish between models and very little evidence of the improvements gained via model development. More powerful methods, such as the ensemble-based,

hypotheses-testing-oriented framework presented in this dissertation, are required to evaluate models for scientific and engineering applications. Model diagnostics require the use of 'signatures' (typical patterns in the behavior of the system) to help attribute cause-effect relationships; explicit acknowledgement of the uncertainties in the simulation that come from structure, parameter, and data error; and the use of time-varying measures of performance or misfit that allow modelers to diagnose the potential reasons for shortcomings in the model's formulations.

Focusing on cost-function space and basing my hypothesis on an assumption that is widely held in the LSM development community, I asserted that: (1) Increasing a model's conceptual realism decreases the sensitivity of model output to parameter choices. My results do not provide definitive evidence regarding the role of conceptual realism in shaping model robustness. Hypothesis 1 was not supported by the simulations of GW and DV at the dry IHOP sites. Results did support the hypothesis for GW (but not DV) at the wet sites. Adding complexity to models (although crucial for research endeavors) frequently requires the use of immeasurable, uncertain parameters, which entails a significant risk of decreasing model robustness. Less robust models are less well suited for broad application in operational settings.

A significant byproduct of the work presented in Chapter 2 is the demonstration that calibration of only some, new model parameters ('partial calibration') of implemented modules (e.g., GW, DV) is insufficient to guarantee good results and significantly increases the chance of bad simulations. This part of my work suggests that parameter values may be model specific and that interactions between parameters are

necessary to accommodate new model structures. My work poses a caveat for simple 'plug and play' of functional modules between LSMs. 'Plug-and play' and partial calibration are currently standard practices in LSM development and application; my results provide evidence that both should be re-examined.

Focusing on model performance in the feature space and seeking to understand whether models are getting the right answer for the right reasons, I posed two additional hypotheses: (2) upward-flowing water from deep-soil stores is an important source of moisture in transition zones that supports latent heat flux and root zone soil moisture content during dry-downs or drought in the warm season (i.e., GW will perform better than STD in terms of simulated evaporative fraction and soil wetness); and (3) the rapid response of vegetation to changes in environmental conditions is an important control on evapotranspiration in transition zones during the warm season (i.e., DV will improve over STD on the simulation of observed evaporative fraction). The test of the hypotheses consisted of whether the models could reproduce a signature in the evolution of the observed evaporative fraction and root zone moisture: at dry sites the evaporation peaks and recedes immediately after rainfall, but at wet sites it does not peak until several days after precipitation. Results presented in Chapter 2 provide support for hypothesis 2 only in the case of soil moisture. However, it is likely that the improvement seen with GW is the result of bias correction in the mean moisture state, not improved soil moisture dynamics. Hypothesis 2 was not supported for evaporation. The current formulation of GW was unable to improve the partitioning of the energy cycle. My results provide support for hypothesis 3 at wet, vegetated sites where the partitioning between slow and

fast evaporation of DV improves over that of STD, but the results did not support hypothesis 3 at dry, sparsely vegetated sites. Note that I used ensembles of calibrated and uncalibrated models to quantify the 'real' performance of the models when simulating the signatures. Hence the identified shortcomings are likely structural and not the result of bad parameter choices.

Research presented in Chapter 3 primarily evaluated the augmented models in the multidimensional parameter space, a facet of model behavior that is often overlooked in the LSM development community. Because LSM developers have attempted to use physical principles when conceptualizing and parameterizing their models, parameters of such physically based models are assumed to be physically meaningful and to correspond to unchanging physical characteristics of the land surface that can be measured or inferred. I tested the following hypothesis: (4) Model parameters are physically meaningful characteristics of the system whose values do not change between models for a given site. My results provide evidence that the marginal distributions of behavioral physically meaningful parameters (of models constrained to reproduce high-resolution near-surface fluxes and states) differ between models at every site, which does not support the hypothesis. Furthermore, I showed that the relationships between parameters among models are not the same for a given site. I presented evidence that the preferred values of optimal parameters at a given site are not the same between models and that the covariance structure between the parameters is also different between models. My results are not consistent with hypothesis 4, but, confirming observations made in Chapter 2, they suggest that parameters of medium complexity LSMs are model dependent, effective

quantities. My work shows that although the identification of an optimal value for an effective parameter is useful information, the change in the functional relationship between parameters is more important for model development and hypothesis testing.

The corollary of the notion that in a hypothetical perfect model parameters correspond to observable, unchanging characteristics of a system is that the level of interaction between model parameters can serve as a measure of the model's physical realism. I tested the hypothesis: (5) More physically realistic models are better models, which have less –unwarranted– parameter interaction (i.e., the parameters of STD are more interactive than those of GW and DV). In general, very few model parameters directly control the variance of the model, and the interaction between parameters is a significant source of variability. Evidence of the cumulative amount of variance explained by a single parameter interacting with all the rest at all orders did not support hypothesis 5 for DV at any of the IHOP sites and for GW at all dry sites. Quantification of the level of interaction of the parameters in GW at semi-humid and wet sites seem to support hypothesis 5, suggesting that it is a better model. Because of its improved partitioning of the water balance (a robust connection of the soil with the water table, direct evaporation decoupled from baseflow, and an enhanced ratio of subsurface to surface runoff), GW appears to be a more realistic model in wet sites. The application of an efficient, quantitative, variance-based sensitivity analysis presented in Chapter 3 is innovative in the field of land-surface modeling and offers the community a suitable tool to test parameter behavior during model development.

The other implication of the assumption made in land-surface modeling about the physical nature of the parameters is that if they are physically meaningful quantities that do not change and that have strong relationships with physical characteristics of the land surface (e.g., vegetation type, soil texture), then parameters will be the same for sites within the same physical classification. This assumption has resulted in the common practice of *a priori* estimation of LSM parameters. With no other information, soil and vegetation parameters are assigned using look-up tables that are based on soil texture class and vegetation type. In Chapter 3, I tested the hypothesis: (6) Physically meaningful model parameters can be transferred between sites with physically similar characteristics. Using unsupervised classification of the similarity of the marginal posterior distributions of optimal parameters between sites, I found evidence that does not support hypothesis 6. Vegetation (soil) parameter distributions could not be grouped by similarity solely based on vegetation (soil) type but appear to be strongly related to the climatic gradient. These results are consistent with the quantified level of parameter interaction between soil and vegetation parameters and at least partially explain why transferability between sites (solely based on shared soil texture or vegetation type) and models is not straightforward. The implication of my results is that *a priori* assignment of medium complexity LSM parameters should, in addition to a land-cover classification that accounts for interactive pairs of soil and vegetation classes, also consider the climatic conditions of a study location.

In Chapter 4, I applied the ensemble-based, hypothesis-testing-oriented framework presented in Chapters 2 and 3 to diagnose shortcomings in the ability of STD

and GW to simulate the long-term distribution of streamflow and partitioning of the water cycle. I used ensembles of behavioral models to test two hypotheses: (7) The addition of a groundwater module provides sustained baseflow and improves the partitioning of surface to subsurface runoff; and (8) The topography-related parameterization of runoff and the explicit representation of the connection of the soil profile to an aquifer better represent the variability of upper and root-zone soil moisture. I used the flow duration curve of the baseflow-dominated watershed and the relationship between the first and second moments of the soil moisture observations as signatures of catchment hydrologic behavior. GW's enhanced representation of topographic effects on runoff and its augmentation of Noah with an aquifer model do not improve the distributed simulation of the timing and volume of streamflow.

In its current form, the deficient formulations of GW do not provide support for hypothesis 7. Streamflow simulated by STD and GW is too flashy: the models underestimate the persistency of low flows, and they fail to capture the flow recession curves properly. The dominance of surface over subsurface runoff in STD and GW was identified as the leading cause for the deficient performance. I showed that when the water table is deep, physically realistic values of GW model parameters make the model unable to simulate enough sustained baseflow and instead cause the model to function as a simple bucket. This unrealistic performance is consistent with the lack of sensitivity of the groundwater model parameters.

Results of simulated soil moisture provide evidence that does not support hypothesis 8. I showed in Chapter 4 that the soil profile in both models (STD and GW) is

too quick to wet and too quick to dry. The statistics of the simulated soil moisture are not improved by the implementation of GW. This evidence points toward structural deficiencies in STD's representation of the subsurface hydrologic processes that cannot be solved by the way in which GW represents groundwater.

The fundamental thesis of my work is that the use of innovative diagnostic-oriented methods is an approach to LSM evaluation that, unlike current practices, enables LSM developers to identify shortcomings in the formulations of the parameterizations that hinder the models' capacity to simulate high-frequency near-surface states and fluxes of water and energy even with optimized parameters. The use of an ensemble approach allows for the accounting of sources of uncertainty that are inherent to land-surface modeling and allows researchers to reach conclusions that are independent of the choice of parameters. As a result of the model intercomparison presented here, I have shown that only when the water table is shallow and vegetation is lush do the more conceptually realistic versions of the Noah LSM ameliorate biases in the estimation of the root zone moisture, latent heat flux, and (potentially) runoff during the warm season in transition zones. The inability of the (enhanced) models to reproduce particular hydrologic features, such as a characteristic temporal or spatial pattern summarized by a 'signature,' has pointed out specific aspects in the parameterizations that need to be modified by developers. My work has confirmed that models are flexible and that model parameters are effective, scale-, site- and model-dependent, which underscores the need to account for parameter uncertainty (via calibration) even when using 'physically based' models. My work challenges common practices and assumptions in LSM development and offers

other modeling groups new, stringent ways to test their models. By helping bridge the gap between the model identification and the model development communities, this work contributes to ensure the continued improvement of our understanding and modeling of environmental processes.

5.3. FUTURE WORK

Immediate future work consists of implementing the suggested modifications to the augmented models identified in Chapters 2, 3 and 4. As a result of the discussions presented in Chapter 2 and 4, I advocate for a top-down approach to model development in which changes (additional complexity) are sequentially implemented.

I caution that, as with any modeling endeavor, it is possible that the results presented in this dissertation might change if the same (or similar) models are used to represent other biomes (e.g., temperate forest) or locations where other hydrologic processes (e.g., snow accumulation and ablation) are dominant. Immediate future work consists of applying the ensemble-based diagnostic framework, in a similar fashion as in Chapters 2 and 3, to evaluate at a point scale the augmented versions of Noah (or similar variants) to locations throughout the world. Currently there are almost 30 FLUXNET stations worldwide, which collect high-frequency land-to-atmosphere fluxes. They present the opportunity to benchmark models. We need to know: How do results differ when models are used to represent other biomes? Where do models work? Are *in-situ* observations sufficient to capture the temporal and spatial dynamics of the energy and water balances? How can we use model diagnostics to better identify sites to collect

measurements? For which biomes are model process representations best suited? What parts of the model require the most attention in terms of improved process representation? How do model parameters vary across dominant biomes? How can we bridge the gap between our understanding of LSM parameters and *in situ* characteristics of the land surface for better *a priori* parameter identification?

As we move away from *in-situ* to distributed evaluation of models (Chapter 4), other questions become more relevant. In Chapter 4, I used statistics of soil moisture that appear to be scale invariant. Are there other ways to bridge potential scale disconnects between models and observations? It would be useful to know: What is the role of remote sensing in providing meaningful information to rate models in the distributed setting? Specifically whether incorporating brightness temperature, estimates of moisture or evaporation can help constrain the models in the distributed setting.

Although the findings of this work are relevant for coupled simulations, I can only be confident that my results on the value of the models hold in offline settings. Future research should focus on: What is the value of land-memory mechanisms for the predictability of precipitation? What is the impact of offline calibrated model structures for the online prediction of precipitation? How does the scatter of LSM-predicted fluxes and states propagate to the atmosphere? How can the ensemble-based signature-oriented framework be extended to evaluate simulations of precipitation? What are useful signatures of boundary layer meteorological processes?

The analysis presented in the chapters of this thesis neglected uncertainty in the forcing and evaluation data. Assuming we can estimate their magnitudes, the framework

for model evaluation presented here can accommodate both sources of uncertainty. To that end it is important to be able to describe uncertainty *a priori*: How can we define prior uncertainties in data, model parameters, and model structure? It also is important to understand how different sources of uncertainty propagate through feature, cost-function, and parameter spaces: How can we estimate the contribution of these sources to the overall uncertainty? What does the inclusion of sources of uncertainty mean for decision making and scenario analysis?

For operational prediction and forecasting, future work relates to addressing questions on the use of the performance metrics used in Chapters 2 and 4 to correct for systematic errors in the models: How can estimates of model structural error be used to correct for systematic biases? What is a meaningful way to combine multi-model ensemble predictions and their relative performance for an improved product? What is the relationship between the performance of behavioral ensembles conditioned on observed data and weights used in Bayesian Model Averaging?

More important for the field of model development is the design of meaningful experiments to test whether energy and water cycle parameterizations in LSMs are sufficiently accurate to be used with confidence in land-use/land-cover and climate change attribution. In the diagnostic framework, signatures hold the key of identifying cause-and-effect relationships; therefore it is fundamental to select appropriate, informative diagnostic criteria. We need to know: Which signatures are relevant for the representation of land-surface fluxes and states? What is the nature of the system function or pattern summarized in the signature? Does it change with time and in space? Ideally,

choosing a signature reflects knowledge about a function of the natural system that is known in advance to be relevant. In Chapter 2 we identified the evolution of evaporative fraction and the depletion of root zone wetness. In Chapter 4 we identified the long-term distribution of flows and the relationship between the statistics of the near surface soil moisture as meaningful signatures. The timing of the partitioning of water and energy budgets was the essential characteristic of land-atmosphere interactions to be captured by the LSMs. They allowed me to provide explanations to the questions: What does it mean when models fail to reproduce the pattern contained in the signature? The fact that signatures need to be tailored according to the problem means that the community has yet to focus on: What does each signature tell the model developer about deficiencies in the parameterizations or the understanding of the system? Our inability to identify signatures hinders our capacity to pose relevant tests for evaluation and development of models. Choice of signature is not trivial and requires attention from the community, but it has the potential to help synthesize the knowledge of observationalists and modelers.

If other modeling groups implement the LSM evaluation approach presented in this dissertation, it will help the community to understand the relationship between model complexity and predictive uncertainty. It would also be advisable for the community to revisit whether the conclusions reached by other model intercomparison experiments hold when accounting for uncertainty.

Appendices

1. 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. k is the number of free parameters in the model (Legates and McCabe, 1999; Akaike, 1974; Schwarz, 1978).

Observation mean: $$\overline{O} = \frac{1}{T}\sum_{t=1}^{T}\left(O_t\right) \tag{A.1.1}$$

Model mean: $$\overline{P} = \frac{1}{T}\sum_{t=1}^{T}\left(P_t\right) \tag{A.1.2}$$

Observation Standard Deviation: $$\text{StdDev} = \left[\frac{1}{T}\sum_{t=1}^{T}\left(O_t - \overline{O}\right)^2\right]^{0.5} \tag{A.1.3}$$

Model Standard Deviation: $$\text{StdDev} = \left[\frac{1}{T}\sum_{t=1}^{T}\left(P_t - \overline{P}\right)^2\right]^{0.5} \tag{A.1.4}$$

Root mean squared error: $$RMSE = \left[\frac{1}{T}\sum_{t=1}^{T}\left(O_t - P_t\right)^2\right]^{0.5} \tag{A.1.5}$$

Coefficient of determination: $$r^2 = \left[\frac{\sum_{t=1}^{T}\left(O_t - \overline{O}\right)\left(P_t - \overline{P}\right)}{\left[\sum_{t=1}^{T}\left(O_t - \overline{O}\right)^2\right]^{0.5}\left[\sum_{t=1}^{T}\left(P_t - \overline{P}\right)^2\right]^{0.5}}\right]^2 \tag{A.1.6}$$

Bias $$bias = \frac{1}{T}\sum_{i=1}^{T}\left(P_i - O_i\right) \tag{A.1.7}$$

Nash-Sutcliffe Efficiency $$NSE = 1 - \frac{\sum_{t=1}^{T}(O_t - P_t)^2}{\sum_{t=1}^{T}(O_t - \overline{O})^2} \quad \text{(A.1.8)}$$

Akaike Information Criteria $$AIC = 2k + T\ln\left[\frac{RMSE}{T} + 1\right] + \frac{2k(k+1)}{T-k-1} \quad \text{(A.1.9)}$$

Bayesian Information Criteria $$BIC = T\ln\left[\frac{RMSE}{T}\right] + k\ln(T) \quad \text{(A.1.10)}$$

2. 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, 1997).

Ensemble mean: $$\bar{x}_t = \frac{\sum_{i=1}^{N_{ens}} x_{i,t}}{N_{ens}} \quad \text{(A.2.1)}$$

Ensemble bias: $$\beta_t = \bar{x}_t - o_t \quad \text{(A.2.2)}$$

Ensemble skill score: $$\kappa_t = \left(\bar{x}_t - o_t\right)^2 \quad \text{(A.2.3)}$$

Ensemble spread: $$\pi_t = \frac{\sum_{i=1}^{N_{ens}} \left(x_{i,t} - \bar{x}_t\right)^2}{N_{ens} - 1} \quad \text{(A.2.4)}$$

2.1. Metrics for model evaluation

2.1.1. Model performance (ς_t)

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

$$\varsigma_t = \frac{CDF_{ens,t} - CDF_{obs,t}}{1 - CDF_{\overline{obs}}} \quad \text{(A.2.5)}$$

where $CDF_{ens,t}$ is the cumulative distribution function (CDF) of the ensemble at time *t*, $CDF_{obs,t}$ is the CDF of the observations at time *t,* and $CDF_{\overline{obs}}$ is the CDF of the time mean of observation time series. As ς_t decreases, model performance at time *t* increases. Inspired by ensemble verification metrics, model performance score ς_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, $CDF_{obs,t}$ is a step function. Denominator $1-CDF_{\overline{obs}}$ 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: e.g., using a denominator of $1-CDF_{obs,t}$ would increase the stringency of the score more when observations are low than when observations are high.

2.1.2. 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:

$$\rho = \frac{\left|\bar{\varsigma}_{ps} - \bar{\varsigma}_{mf}\right|}{\bar{\varsigma}_{ps} + \bar{\varsigma}_{mf}} \tag{A.2.6}$$

where $\bar{\varsigma}_{p}$ is the time median performance score of the Pareto set (PS) ensemble. $\bar{\varsigma}_{mf}$ is the time median performance score of the most-frequent performing (MF) ensemble.

2.1.3. Model fitness (φ)

The ς-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:

$$\phi = \rho\bar{\varsigma}_{ps} \tag{A.2.7}$$

where ρ is the robustness score for a given model where $\bar{\varsigma}_p$ 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.

3. SIMPLE GROUNDWATER MODEL AND TOPOGRAPHY-RELATED RUNOFF PARAMETERIZATION

Following Niu et al. (2007), the temporal variation in water stored in the aquifer is determined by the residual of recharge rate, Q, minus discharge rate (baseflow or subsurface runoff), R_{sb}.

$$\frac{dW_a}{dt} = Q \tag{A.3.1}$$

Q is then parameterized following Darcy's law (to balance gravitational and capillary forces) and is positive when water enters the aquifer:

$$Q = -K_a \frac{-z_{wt}-(\psi_{bot}-z_{bot})}{z_{wt}-\varphi_{bot}} \tag{A.3.2}$$

where z_{wt} is the water table depth, ψ_{bot} is the matric potential of the bottom soil layer, z_{bot} (1.5 m in Noah) is the midpoint of the bottom soil layer, calculated according to Clapp and Hornberger (1978) as:

$$\psi_{bot} = \psi_{sat,bot}\left(\frac{\theta_{bot}}{\theta_{sat,bot}}\right)^{-b} \tag{A.3.3}$$

and K_a is the and hydraulic conductivity of the aquifer, obtained by integrating the hydraulic conductivity below the soil column (which is assumed to decay exponentially with depth at rate f), as:

$$K_a = \frac{\int_{z_{bot}}^{z_{wt}} k_{bot} e^{-f(z-z_{bot})} dz}{z-z_{bot}} = \frac{k_{bot}\left(1-e^{-f(z-z_{bot})}\right)}{f(z_{wt}-z_{bot})} \tag{A.3.4}$$

The water table depth is related to the aquifer water storage through the specific yield of the aquifer, S_y:

$$z_{wt} = \frac{W_a}{S_y} \tag{A.3.5}$$

With the recharge rate, the volumetric soil moisture of the bottom layer θ_{bot} is updated using the Richards' equation with zero flux lower boundary condition as:

$$-Q = \rho_w \Delta z_{bot} \frac{d\theta_{bot}}{dt} \tag{A.3.6}$$

When the water table is within the soil column, equation A3.2 is expressed as:

$$Q_i = -K_{i,wt} \frac{(\psi_{sat} - z_{wt}) - (\psi_i - z_i)}{z_{wt} - z_i} \tag{A.3.7}$$

where z_i and ψ_i are node depth and the matric potential of the i^{th} layer right above the layer where the water table is. $K_{i,wt}$ is the hydraulic conductivity between layer i and the water table.

Niu et al. 2005 use a simple TOPMODEL-based runoff model to compute surface runoff and groundwater discharge, which are both parameterized as exponential functions of the depth to water table. Surface runoff is mainly saturation-excess (Dunne) runoff, i.e., the water (sum of rainfall, dew, and snowmelt) incident (P_{in}) on the fractional saturated area of a model grid-cell (F_{sat}), or

$$R_s = (P_{in} F_{sat}) + (1 - F_{sat}) \max(0, P_{in} - I_{max}) \tag{A.3.8}$$

where I_{max} is the maximum infiltration capacity and the fractional saturated area, F_{sat}, is parameterized as:

$$F_{sat} = \left(1 - F_{frz}\right) F_{sat_{max}} e^{-0.5 f(z_{wt})} + F_{frz} \tag{A.3.9}$$

where the potential or maximum saturated fraction of a gridcell is F_{satmax} and the impermeable fraction is F_{frz}

Analogously, the groundwater discharge (baseflow or subsurface runoff) rate is parameterized as:

$$R_{sb} = R_{sb_{max}} e^{-f(z_{wt})} \tag{A.3.10}$$

where $R_{sb_{max}}$ is the maximum rate of subsurface runoff and f is the e-folding depth of saturated hydraulic conductivity, which, following Silvapalan et al. (1987), is assumed to exponentially decay with depth.

4. DYNAMIC VEGETATION MODEL

The dynamic leaf model (Dickinson et al., 1998) describes the carbon budget of vegetation (leaf, wood, and root) and soil carbon pools (fast and slow). The model represents various processes including carbon assimilation through photosynthesis, allocation of the assimilated carbon to various carbon pools (leaf, stem, wood, root, and soil), and respiration from each of the carbon pools.

The leaf carbon mass, C_{leaf}, (g m^{-2}) balance is calculated according to:

$$\frac{\partial C_{leaf}}{\partial t} = F_{leaf} A - (S_{cd} + T_{leaf} + R_{leaf}) C_{leaf} \quad \text{(A.4.1)}$$

where A is the total carbon assimilation rate of the sunlit and shaded leaves (g m^{-2} s^{-1}), S_{cd} is death rate due to cold and drought stresses, and T_{leaf} is the rate of leaf turnover due to senescence, herbivory, or mechanical loss [see *Dickinson et al*., 1998 for details]. R_{leaf} is leaf respiration rate including maintenance and growth respiration and F_{leaf} is the fraction of the assimilated carbon allocated to leaf and parameterized as an exponential function of LAI:

$$F_{leaf} = e^{(0.01 * LAI(1-\exp(LAI))} \quad \text{(A.4.2)}$$

LAI is converted from C_{leaf} using specific leaf area (m^2 g^{-1}), a vegetation-type–dependent parameter. Greness vegetation fraction (F_{veg}) is then simply converted from LAI:

$$F_{veg} = 1 - e^{-0.52 LAI} \quad \text{(A.4.3)}$$

The rate of photosynthesis per unit LAI of shaded and sunlit leaves, A_i (A_{shd} and A_{sun}), depends on the Ball-Berry stomatal resistance per unit LAI of shaded and sunlit leaves, $r_{s,i}$ ($r_{s,shd}$ and $r_{s,sun}$),

$$\frac{1}{r_{s,i}} = m \frac{A_i}{c_{air}} \frac{e_{air}}{e_{sat}(T_v)} P_{air} + g_{\min} \quad \text{(A.4.4)}$$

where c_{air} is the CO_2 concentration at leaf surface ($355 \times 10^{-6} \times P_{air}$ in the unit of pa), P_{air} surface air pressure (pa), e_{air} vapor pressure at the leaf surface (pa), $e_{sat}(T_v)$ saturation vapor pressure inside leaf (pa), g_{min} minimum stomatal conductance ($\mu molm^{-2}s^{-1}$), m is an empirical parameter to relate transpiration with CO_2 flux (a larger m indicates the leaf consumes more water, i.e., greater transpiration, to produce the same carbon mass).

The total carbon assimilation (or photosynthesis) rate (g m^{-2} s^{-1}),

$$A = 12 \times 10^{-6}(A_{sun}L_{sun} + A_{shd}L_{shd}) \tag{A.4.5}$$

where A_{sun} and A_{shd} are photosynthesis rates (μmol m^{-2} s^{-1}) per unit LAI of sunlit and shaded leaves, and L_{sun} and L_{sha} are sunlit and shaded leaf area indices, respectively. L_{sun} and L_{sha} are respectively proportional to sunlit and shaded fractions of the canopy, which are computed from the two-stream radiation transfer scheme. The factor 12×10^{-6} is to transform the unit μmol $m^{-2}s^{-1}$ to g m^{-2} s^{-1}.

$$A_i = I_{gs} \min(A_C, A_{L,i}, A_S) \qquad i \text{ for sunlit and shaded leaves} \tag{A.4.6}$$

where I_{gs} is a growing season index depending on leaf temperature, A_C, $A_{L,i}$, and A_S are carboxylase-limited (Rubisco-limited), light-limited, and export-limited (for C3 plants) photosynthesis rates per unit LAI, respectively.

A_C, $A_{L,i}$, and A_S are respectively,

$$A_c = \frac{(c_i - c_{cp})V_{\max}}{c_i + K_c(1 + o_i / K_o)} \tag{A.4.7}$$

$$A_{L.i} = \frac{(c_i - c_{cp})4.6\alpha PAR_i}{c_i + 2c_{cp}} \tag{A.4.8}$$

$$A_s = 0.5V_{\max} \tag{A.4.9}$$

where c_i is the CO_2 concentration inside leaf cavity, which is about 0.7 times of the atmospheric CO_2 concentration, c_{air}, (pa), and o_i are the atmospheric O_2 concentration (pa). PAR_i (i for shaded and sunlit leaves) is photosynthetically active radiation (Wm^{-2}) per unit shaded and sunlit LAI. The factor 4.6 (μmol photons J^{-1}) is used to convert Wm^{-2} to μmol photons m^{-2} s^{-1}. c_{cp} is the CO_2 compensation point and equals to $0.5\frac{K_c}{K_o}0.21o_i$ (pa), where K_c and K_o are the Michaelis-Menton constants (pa) for CO_2 and O_2, respectively, varying with vegetation temperature T_v [Collatz et al., 1991]. α is the quantum efficiency (μmol CO_2 per μmol photon).

The maximum rate of carboxylation varies with temperature, foliage nitrogen, and soil water,

$$V_{\max} = V_{\max 25} a_{v\max}^{\frac{T_v - 25}{10}} f(N) f(T_v) \beta \qquad \text{(A.4.10)}$$

where $V_{\max 25}$ is maximum carboxylation rate at 25˚C (μmol CO_2 m^{-2} s^{-1}) and $\alpha_{v\max}$ is a temperature sensitive parameter. The $f(T_v)$ is a function that mimics thermal breakdown of metabolic processes [Collatz et al., 1991]. The $f(N) \le 1$ is a foliage nitrogen factor, and $f(N) = 1$ assumes saturation. The β factor is the soil moisture controlling factor, and it is parameterized as a function of soil moisture:

$$\beta = \sum_{i=1}^{N_{root}} \frac{\Delta z_i}{z_{root}} \min(1.0, \frac{\theta_{liq,i} - \theta_{wilt}}{\theta_{ref} - \theta_{wilt}}) \qquad \text{(A.4.11)}$$

where θ_{wilt} and θ_{ref} are soil moisture at witling point (m^{-3} m^{-3}) and a reference soil moisture (m^{-3} m^{-3}) (close to field capacity), respectively. Both depend on soil type. N_{root} and z_{root} are total number of soil layers containing roots and total depth of root-zone, respectively

5. MULTIOBJECTIVE SHUFFLED COMPLEX EVOLUTION METROPOLIS

The Multi-Objective Shuffled Complex Evolution Metropolis (MOSCEM) algorithm (for details see Vrugt et al., 2003) used in this dissertation is a multi-criteria extension of the Shuffled Complex Evolution Metropolis (SCEM) algorithm.

In contrast to local optimization methods, the SCEM is a general purpose global optimization algorithm that provides an estimate of the most likely parameter set and its underlying posterior probability distribution. SCEM is basically is an approximate Markov Chain Monte Carlo (MCMC) sampler, which generates a number of sequences of parameter sets that converges to the stationary posterior distribution for a large enough number of simulations. SCEM is only related to the Shuffled Complex Evolution (SCE) of Duan et al. (1992) global optimization method but uses the Metropolis-Hastings instead of the Downhill Simplex method for population evolution. The SCEM algorithm starts by sampling an initial population of parameter sets randomly distributed within the given feasible parameter ranges. The hydrologic model is run for each parameter set θ. The posterior density $p(\theta|y)$ (or the chance of θ being the optimal parameter set given the information from measurements y) is computed from the likelihood of the model score and the prior information using a Bayesian inference scheme:

$$p\left(\theta^{(t)} \mid y\right) \propto \frac{p\left(\theta^{(t)}\right) L\left(\theta^{(t)} \mid y\right)}{p(y)} \tag{A.5.1}$$

Assuming that the residuals between model prediction and observation are mutually independent, Gaussian distributed, with constant variance σ^2, the likelihood of a parameter set $\theta^{(t)}$ for describing the observed data y is:

$$L\left(\theta^{(t)} \mid y\right) = \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{e\left(\theta^{(t)}\right)_i}{\sigma}\right)^2\right] \tag{A.5.2}$$

where e is a vector of error terms (to be minimized).

Box and Tiao (1973) showed that, assuming a noninformative prior of the form:

$$p\left(\theta^{(t)}\right) \propto \frac{1}{\sigma} \tag{A.5.3}$$

the influence of σ can be integrated out, leading to the following form of the posterior density of $\theta^{(t)}$:

$$p\left(\theta^{(t)} \mid y\right) \propto \left[\sum_{i=1}^{N} e\left(\theta^{(t)}\right)_i^2\right]^{-N/2} \tag{A.5.4}$$

The classical approximation to obtain the $p(\theta|y)$ is to use a first-order Taylor series expansion of the nonlinear model evaluated at the globally optimal parameter set estimate θ_{opt}. The estimated multivariate posterior joint probability density function of θ is then expressed as:

$$p\left(\theta^{(t)} \mid y\right) \propto \exp\left[-\frac{1}{2\sigma^2}\left(\theta-\theta_{opt}\right)^T X^T X\left(\theta-\theta_{opt}\right)\right] \tag{A.5.5}$$

where X is the Jacobian evaluated at $\theta_{opt.}$ This means that $p(\theta^{(t)}|y)$ is approximated by a normal distribution, $N_i(\theta_{opt}, \sigma^2\Sigma_{ii})$, where Σ_{ii} is the i^{th} diagonal element of the covariance matrix computed as:

$$\sqrt{\left(X^T X\right)^{-1}} \tag{A.5.6}$$

For nonlinear models (e.g., hydrologic models), this approximation can be quite poor. $p(\theta|y)$ exhibits strong and nonlinear parameter interdependence, and can deviate

significantly from the multinormal distribution. In this case, an explicit expression of the joint and marginal probability density functions is often not possible and Markov Chain schemes are general approach for sampling from the $p(\theta|y)$.

A Markov Chain is generated by sampling $\theta^{(t+1)} \sim z(\theta|\ \theta^{(t)})$ using a transition kernel z or proposal distribution of the Markov Chain. The most general Markov Chain Monte Carlo (MCMC) algorithm is the Metropolis-Hastings algorithm, summarized as follows:

1. Randomly start at a location in the feasible parameter space, $\theta^{(t)}$, and compute the posterior density, $p(\theta^{(t)}|y)$, relevant to this point according to Eq. *A.5.2 or A.5.4.*

2. Generate a new candidate point $\theta^{(t+1)}$ from $z(\theta|\ \theta^{(t)})$, where $z(.)$ is called the proposal distribution.

3. Evaluate , $p(\theta^{(t)}|y)$, using Eq. *A.5.2 or A.5.4.* and compute $\Omega = p(\theta^{(t+1)}|y)/ p(\theta^{(t)}|y)$.

4. Randomly sample a uniform label Z over the interval 0 to 1.

5. If $Z \leq \Omega$, then accept the new configuration. However, if $Z > \Omega$, then reject the candidate point and remain at the current position, that is, $\theta^{(t+1)} = \theta^{(t)}$.

6. Increment t. If t is less than a prespecified number of draws, then return to step 2.

In the SCEM, to increase information exchange between the sampled candidate points the population of parameter sets is partitioned into q complexes, and in each C^k complex k (k=1,2,...,q) a parallel sequence S^k is launched by the SEM (Sequence Evolution Metropolis) algorithm from the point that exhibits the highest posterior density.

SEM evolves each sequence and complex. The SEM algorithm produces new candidate points in each of the parallel sequences S^k by generating draws from an adaptive multivariate normal proposal distribution either centered around the current draw of the sequence (k) or the mean of the points in complex (k) extended with the covariance structure induced between the points in complex k by using the information induced in the *m* samples of C^k. The Metropolis-annealing criterion is used to test whether the candidate point should be added to the current sequence. The steps are summarized as follows:

I. Compute the mean μ^k, and covariance structure Σ^k of the parameters of C^k. Sort the *m* point in complex C^k in order of decreasing posterior density and compute Γ^k, the ratio of the posterior density of the first ("best") to the posterior density of the last ("worst") member of C^k.

II. Compute α^k, the ratio of the mean posterior density of the *m* points in C^k to the mean posterior density of the last *m* generated points in S^k.

III. If α^k is smaller than a predefined likelihood ratio, *T*, generate a candidate point, $\theta^{(t+1)}$, by using a multinormal distribution centered on the last draw, $\theta^{(t)}$, of the sequence S^k, and covariance structure $c_n^2 \Sigma^k$, where c_n is a predefined jumprate. Go to step V, otherwise continue with step IV.

IV. Generate offspring, $\theta^{(t+1)}$, by using a multinormal distribution with mean μ^k and covariance structure $c_n^2\Sigma^k$, and go to step V.

V. Compute the posterior density, $p(\theta^{(t+1)}|y)$, of $\theta^{(t+1)}$ using Eq. A.5.2 or A.5.4. If the generated candidate point is outside the feasible parameter space, set $p(\theta^{(t+1)}|y)$ to zero.

VI. Compute the ratio $\Omega = p(\theta^{(t+1)}|y)/p(\theta^{(t)}|y)$ and randomly sample a uniform label Z over the interval 0 to 1.

VII. If $Z \leq \Omega$, then accept the new candidate point. However, if $Z > \Omega$, reject the candidate point and remain at the current position in the sequence, that is, $\theta^{(t+1)} = \theta^{(t)}$.

VIII. Add the point $\theta^{(t+1)}$ to the sequence S^k.

IX. If the candidate point is accepted, replace the best member of C^k with $\theta^{(t+1)}$, and go to step X; otherwise replace the worst member (m) of C^k with $\theta^{(t+1)}$, provided that Γ^k is larger than the predefined likelihood ratio, T, and $p(\theta^{(t+1)}|y)$ is higher than the posterior density of the worst member of C^k.

X. Repeat the steps I–VIII L times, where L is the number of evolution steps taken by each sequence before complexes are shuffled.

The SEM routine passes the new candidate point back to SCEM and subsequently the new candidate point randomly replaces an existing member of the complex. Finally, after a certain number of iterations (q*L) new complexes are formed through a process of shuffling. L is the number of evolution steps taken by each sequence before complexes are shuffled. The Gelman and Rubin convergence statistic is calculated on the generated posterior densities to check whether convergence to a stationary target distribution has

been achieved. SCEM stops the search when the convergence criterion is met or when the maximum number of iterations is reached.

The Multi-Objective Shuffled Complex Evolution Metropolis (MOSCEM) algorithm (Vrugt et al., 2003) is capable of generating a fairly uniform approximation of the Pareto frontier within a single optimization run using a newly developed, improved concept of Pareto dominance.

Note that MOSCEM is different from the Multi-Objective Complex evolution MOCOM algorithm (Yapo et al., 1998), an extension of SCE that merged the strengths of controlled random search with a competitive evolution, Pareto ranking, and multi-objective downhill Simplex strategy. During the course of hydrologic investigations, it became apparent that MOCOM showed serious weaknesses typical of the evolutionary algorithms, which are currently available for solving the multiobjective optimization problem. The first shortcoming of MOCOM is that it does not consistently generate a uniform approximation to the Pareto front, but tends to cluster the solutions in the compromise region among the objectives, thereby leaving the ends of the Pareto frontier unrepresented. Consequently, the Pareto set of solutions does not contain the individual single-criterion solutions, which represent the theoretical extreme ends of the Pareto frontier. The second, perhaps more important, failure is the inability of the evolution strategy in the MOCOM algorithm to converge to solutions within the "true" Pareto set for case studies involving large numbers of parameters and highly correlated performance criteria (e.g., typical of land-surface models (LSMs)). The algorithm tends, instead, to converge to a fuzzy region surrounding the Pareto set and, in some cases, does not

converge at all. The phenomenon of genetic drift, where the members of the population drift to a single solution, is a characteristic typical of many evolutionary search algorithms.

MOSCEM differs from MOCOM in three essential ways:

First, to prevent the collapse of the search to a single region of highest attraction, the MOSCEM incorporates a strategy that preserves the diversity of the sampled population by using an improved fitness assignment method, whereas the MOCOM algorithm uses the standard Pareto ranking concept. The rank fitness assignment procedure begins by identifying all of the nondominated individuals in the population and assigning them rank "one". While the original Pareto ranking concept now proceeds by peeling off these points and identifies the nondominated points of the remaining population (assigned rank "two"), the proposed fitness assignment as follows:

a. Store all of the rank "one" points in an external nondominated set P^1 and the remaining dominated points of the population in a set entitled P.

b. Each solution $i \in P^1$ is assigned a real value $r_i \in [0, 1)$, called strength. The strength is proportional to the number of population members $j \in P$ for which $i \geq j$. Let N be the number of individuals in P that are covered by i and s is the population size ($P + P^1$). The strength is now defined as, $r_i = N/S$. For each member i of P^1, the fitness (f_i) is identical to its computed strength (r_i).

c. The fitness of the remaining dominated individuals $j \in P$ is calculated by summing the strengths of all external nondominated solutions $i \in P^1$ that cover j:

$$f_i = 1 + \sum_{i=1}^{i \leq j} r_i \tag{A.5.7}$$

where, $f_i \in [1,s)$.

To ensure that the members of P have a lower fitness than the members of P^1, the number one is added to the total sum. The closer the computed f value is to zero, the higher the fitness of the sampled point.

Second, the multi-objective downhill simplex method used by the MOCOM algorithm is replaced with a probabilistic covariance-annealing search method (SCEM), which (as discussed above) is well-suited to deal with the strong correlation structures between the parameters in the Pareto set that are typically encountered in hydrologic modeling. Moreover, the stochastic nature of the annealing scheme prevents the collapse of MOSCEM into a relatively small region of some single "best" parameter set, thereby further preserving diversity of the sampled population and enabling the algorithm to generate a fairly uniform approximation of the Pareto front.

Finally, the MOSCEM algorithm uses the strengths of the shuffling procedure and complex partitioning employed in the single-objective SCE algorithm to conduct an efficient search of the parameter space.

References

Abramowitz, G., H. V. 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.

Abramowitz, G., A. J. Pitman, . H. V. Gupta, E. Kowalczyk, and Y. Wang (2009), On the need for a biophysically based benchmark for land surface models, J. Hydrometeor.,: In Press.

Akaike, H. (1974), A new look at the statistical model identification. IEEE Trans. Automat. Contr., 19 (6): 716–723.

Allen, P. B., and J. W. Naney (1991), Hydrology of the Little Washita River Watershed, Oklahoma: Data and Analyses, U.S. Department of Agriculture, Agricultural Research Service, ARS-90, 74 pp.

Andreassian, V., et al. (2006), Catalogue of the Models Used in MOPEX 2004/2005, IAHS Publ., 307, 41–93.

Bai, Y., T. Wagener, and P. Reed (2009), A top-down framework for watershed model evaluation and selection under uncertainty, Environ. Modell. Softw., 24, 901-916. doi: 10.1016/j.envsoft.2008.12.012.

Baldocchi, D., et al. (2001), FLUXNET: A New Tool to Study the Temporal and Spatial Variability of Ecosystem–Scale Carbon Dioxide, Water Vapor, and Energy Flux Densities. Bull. Amer. Meteor. Soc., 82, 2415–2434.

Bastidas, L.A., H. V. Gupta, S. Sorooshian., W. J. Shuttleworth, and Z.-L. Yang (1999), Sensitivity analysis of a land surface scheme using multicriteria methods, J. Geophys. Res., 104(D16), 19, 481–19, 490.

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 et al. (Ed.), Water Science and Application, Vol.3, AGU, 65–76.

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

Bastidas, L. A., E. Rosero, and B. Nijssen (2006b), The PILPS Semi-Arid Experiment - Preliminary Results., 20th Conference on Hydrology, AMS Annual Meeting, Atlanta, Georgia, January 28-31.

Bastidas, L. A., E. Rosero, and B. Njissen (2007), Results from the PILPS semiarid experiment (PILPS San Pedro). GEWEX Newsletter, May 2007.

Bates, B. C., Z. W. Kundzewicz, S. Wu, and J. P. Palutikof, Eds. (2008), Climate Change and Water. Technical Paper of the Intergovernmental Panel on Climate Change, IPCC Secretariat, Geneva, 210 pp.

Beck, M. B. (1987), Water Quality Modeling: A Review of the Analysis of Uncertainty, Water Resour. Res., 23(8), 1393–1442.

Beck, M. B., Ed. (2002), Environmental Foresight and Models: A Manifesto, Elsevier Science, Oxford, UK, 473 pp.

Beven, K. J. (1989), Changing ideas of hydrology: The case of physically based models, J. Hydrol., 105, 157–172.

Beven, K. and A. Binley (1992), The future of distributed models: Model calibration and uncertainty prediction, Hydrol. Process., 6, 279–298.

Beven, K. J. (2001), How far can we go in distributed hydrological modelling?, Hydrol. Earth Syst. Sci., 5, 1–12.

Beven, K. J., and J. Freer (2001), Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems, J. Hydrol., 249, 11–29.

Beven K. J., (2002), Toward an alternative blueprint for a physically based digitally simulated hydrologic response modelling system, Hydrol. Process., 16, 189–206.

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

Bois, B. et al. (2007), Using remotely sensed solar radiation data for reference evapotranspiration estimation at a daily time step, Agric. Forest Meteorol.,doi:10.1016/j.agrformet.2007.11.005.

Boone, A., et al. (2004), The Rhone-Aggregation Land Surface Scheme Intercomparison Project: An Overview. J. Climate, 17, 187-208.

Bowling, L. C., et al. (2003), Simulation of high latitude hydrological processes in the Torne-Kalix basin: PILPS Phase 2e. 1: Experimental description and summary intercomparisons, Glob. Planet. Change, 38, 1-30.

Box, G. E. P. and D. R. Cox (1964). An analysis of transformations. Journal of the Royal Statistical Society, Series B 26: 211–252.

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(12), 3663–3674.

Brutsaert, W., and J. L. Nieber (1977). Regionalized drought flow hydrographs from a mature glaciated plateau. Water Resour. Res., 13, 637-643.

Carlson, J. M. and J. Doyle (2002), Complexity and robustness, PNAS, 99:2538-2545.

Chen, F., K. E. Mitchell, J. Schaake, Y. Xue, H.-L. Pan, V. Koren, Q. Y. Duan, M. Ek, and A. Betts (1996), Modeling of land surface evaporation by four schemes and comparison with FIFE observations, J. Geophys. Res., 101(D3), 7251–7268.

Chen, F., and J. Dudhia (2001), Coupling an advanced land surface hydrology model with the Penn State/NCAR MM5 modeling system. Part 1: Model description and implementation. Mon. Wea. Rev 129, 569-586.

Chen, F., et al. (2007), Description and Evaluation of the Characteristics of the NCAR High-Resolution Land Data Assimilation System, J. Appl. Meteor. Climatol., 46, 694-713, doi: 10.1175/JAM2463.1.

Chen, J. and Kumar, P. (2001), Topographic influence on the seasonal and inter-annual variation of water and energy balance of basins in North America. J. Climate, 14, 1989– 2014.

Chen, T. H., et al. (1997), Cabauw experimental results from the Project for Intercomparison of Land-Surface Parameterization Schemes. J. Climate, 10, 1194-1215, doi:10.1175/1520-0442.

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.

Clapp, R. B., and G. M. Hornberger (1978), Empirical Equations for Some Soil Hydraulic Properties, Water Resour. Res., 14(4), 601–604.

Clark, M. P., A. G. Slater, D. E. Rupp, R. A. Woods, J. A. Vrugt, H. V. Gupta, T. Wagener, and L. E. Hay (2008), FUSE: A modular framework to diagnose differences between hydrological models. Water Resour. Res., 44, W00B02. doi:10.1029/2007WR006735.

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., et al. (2003), Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) project, J. Geophys. Res., 108(D22), 8842, doi:10.1029/2002JD003118.

Dai, Y., X. Zeng, R. E. Dickinson, I. Baker, G. Bonan, M. Bosilovich, S. Denning, P. A. Dirmeyer, P. Houser, G.-Y. Niu, K. Oleson, C. A. Schlosser, and Z.-L. Yang (2003), The common land model (CLM), Bull. Amer. Meteor. Soc., 84, 1013-1023

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.

de Goncalves, L. G., N. Restrepo-Coupe, H. da Rocha., S. Saleska, R. Stöckli (2008), The Large Scale Biosphere-Atmosphere Experiment in Amazônia, Model Intercomparison Project (LBA-MIP) protocol, http://www.climatemodeling.org/lba-mip/.

Demaria, E. M., B. Nijssen, and T. Wagener (2007), Monte Carlo sensitivity analysis of land surface parameters using the Variable Infiltration Capacity model, J. Geophys. Res., 112, D11113, doi:10.1029/2006JD00.

Demarty, J., C. Ottlé, I. Braud, A. Olioso, J. P. Frangi, L. A. 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., A. Henderson-Sellers, P. J. Kennedy, and M. F.Wilson, (1986), Biosphere-Atmosphere Transfer Scheme (BATS) for the NCAR Community Climate Model. NCAR Technical Note, NCAR/TN–275+STR, National Center for Atmospheric Research, Boulder, CO, AAP, 69 pp.

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.

Dirmeyer, P. A. (2006), The hydrologic feedback pathway for land-climate coupling, J. Hydrometeor., 7, 857–867.

Dirmeyer, P. A., R. D. Koster, and Z. C. Guo (2006), Do global models properly represent the feedback between land and atmosphere? J. Hydrometeor., 7(6), 1177-1198.

Duan, Q., S. Sorooshian, and V. Gupta (1992), Effective and Efficient Global Optimization for Conceptual Rainfall-Runoff Models, Water Resour. Res., 28(4), 1015 1031.

Duan, Q., et al. (2006), Model Parameter Estimation Experiment (MOPEX): An overview of science strategy and major results from the second and third workshops. J. Hydrol., 320, 3-17.

Ducharne, A., R. D. Koster, M. J. Suarez, M. Stieglitz, and P. Kumar (2000), A catchment-based approach to modeling land surface processes in a general circulation model 2. Parameter estimation and model demonstration, J. Geophys. Res., 105(D20), 24, 823–24, 838.

Ek, M.B., et al. (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.

Famiglietti, J. S., E. F. Wood, M. Sivapalan, and D, J. Thongs (1992), A catchment scale water balance model for FIFE, J. Geophys. Res., 97, 18997–19007.

Famiglietti, J. S., D. Ryu, A. A. Berg, M. Rodell, and T. J. Jackson (2008), Field observations of soil moisture variability across scales, Water Resour. Res., 44, W01423, doi:10.1029/2006WR005804.

Fan, Y., G. Miguez-Macho, C. P. Weaver, R. Walko, and A. Robock (2007), Incorporating water table dynamics in climate modeling: 1. Water table observations and the equilibrium water table, J. Geophys. Res., 112, D10125, doi:10.1029/2006JD008111.

Farmer, D., M. Sivapalan, and C. Jothityangkoon (2003), Climate, soil, and vegetation controls upon the variability of water balance in temperate and semiarid landscapes: Downward approach to water balance analysis, Water Resour. Res., 39(2), 1035, doi: 10.1029/2001WR000328.

Fox, S., A. J. Pitman, A. Boone, and F. Habets (2006), The relationship between intermodel differences and surface energy balance complexity in the Rhone-Aggregation Intercomparison Project, J. Hydrometeor., 7 (1): 81-100.

Gao, X., S. Sorooshian, and H. V. Gupta (1996), Sensitivity analysis of the biosphere-atmosphere transfer scheme, J. Geophys. Res., 101(D3), 7279–7289.

Gochis, D. J. and F. Chen (2003), Exploration of subgrid routing responses in Noah router. 18th Conference on Hydrology, AMS Annual Meeting (Seattle, WA).

Goteti, G., J. S. Famiglietti, and K. Asante (2008), A Catchment-Based Hydrologic and Routing Modeling System with explicit river channels, J. Geophys. Res., 113, D14116, doi:10.1029/2007JD009691.

Gulden, L. E., E. Rosero, Z.-L. Yang, M. Rodell, C. S. Jackson, G.-Y. Niu, P. J.-F. Yeh, and J. S. 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.

Gulden, L.E., Z.-L. Yang, and G.-Y. Niu (2007b), Interannual variation in biogenic emissions on a regional scale, J. Geophys. Res., 112 (D14), D14103, doi: 10.1029/2006JD008231.

Gulden, L. E., Z.-L. Yang, and G.-Y. Niu (2008a), Sensitivity of biogenic emissions simulated by a land-surface model to land-cover representations, Atmos. Env., doi: 10.1016/j.atmosenv.2008.01.045

Gulden, L. E., E. Rosero, Z.-L. Yang, T. Wagener, and G.-Y. Niu (2008b), 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 non-commensurable measures of information, Water Resour. Res., 34(4), 751-763.

Gupta, H. V., 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), 19491-19504.

Gupta, H. V., K. J. Beven, and T. Wagener (2005), Model calibration and uncertainty estimation, in Encyclopedia of Hydrological Sciences, edited by M. G. Anderson, John Wiley, Hoboken, N.J

Gupta, H. V., T. Wagener, and Y. Liu (2008), Reconciling theory with observations: elements of a diagnostic approach to model evaluation, Hydrol. Process, 22, doi:10.1002/hyp.6989.

Gutmann, E. D., and E. E. Small (2007), A comparison of land surface model soil hydraulic properties estimated by inverse modeling and pedotransfer functions, Water Resour. Res., 43, W05418, doi: 10.1029/2006WR005135.

Gutowski, W. J., Jr., C. J. Vörösmarty, M. Person, Z. Ötles, B. Fekete, and J. York (2002), A Coupled Land-Atmosphere Simulation Program (CLASP): Calibration and validation, J. Geophys. Res., 107(D16), 4283, doi:10.1029/2001JD000392.

Hair, J. F., R. E. Anderson, R. L. Tatham, and W. C. Black (1995), Multivariate Data Analysis with Readings. Prentice-Hall: Upper Saddle River, NJ.

Hansen, M.C., R. S. DeFries, J. R. G. Townshend, and R. Sohlberg, R. (2000), Global land cover classification at 1km spatial resolution using a classification tree approach, Int. J. Remote Sens., 21(6-7), 1389-1414.

Helton, J., and F. Davis (2003), Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems, Reliab. Eng. Syst. Safety, 81(1), 23– 69.

Henderson-Sellers, A. (1993), A Factorial Assessment of the Sensitivity of the BATS Land-Surface Parameterization Scheme, J. Climate, 6, 227–247.

Henderson-Sellers, A., Z.-L. Yang, and R. E. Dickinson (1993), The Project for Intercomparision of Land-surface Parameterizatoin Schemes, Bull. Amer. Meteor. Soc., 74 (7): 1335-1349.

Henderson-Sellers, A., A. J. Pitman, P. K. Love, P. Irannejad, and T. H. Chen (1995), The Project for Intercomparision of Land-surface Parameterization Schemes (PILPS)—Phase 2 and Phase 3, Bull. Amer. Meteor. Soc., 76 (4): 489-503.

Henderson-Sellers, A., K. McGuffie, and A. J. Pitman (1996), The Project for Intercomparison of Land-surface Parametrization Schemes (PILPS): 1992 to 1995, Clim. Dynamics 12 (12): 849-859.

Hogue, T. S., L. A. Bastidas, H. V. Gupta, S. Sorooshian, K. E. Mitchell, and W. Emmerich, (2005), Evaluation and Transferability of the Noah Land-surface Model in Semi-arid Environments, J. Hydrometeor., 6(1), 68–84.

Hogue, T. S., L. A. Bastidas, H. V. Gupta, and S. Sorooshian (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.

Hornberger, G. M., and R. C. Spear (1981), An approach to the preliminary analysis of environmental systems, J. Environ. Manage., 12, 7–18.

Jackson T. J., and F. R. Schiebe, Editors (1993), Hydrology data report: WASHITA '92. NAWQL-93-1, National Agricultural Water Quality Laboratory, USDA Agricultural Research Service, Durant (OK).

Jakeman, A. J., and G. M. Hornberger (1993), How Much Complexity Is Warranted in a Rainfall-Runoff Model? Water Resour. Res., 29(8), 2637–2649.

Jakeman, A.J., R. A. Lecter and J. P. Norton (2006), Ten iterative steps in development and evaluation of environmental models, Environ. Modell. Softw., 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, Nutrient cycling in agrosystems, 50: 247-253.

Kato, H., M. Rodell, F. Beyrich, H. Cleugh, E. van Gorsel, H. Liu, and T. P. Meyers, (2007), Sensitivity of Land Surface Simulations to Model Physics, Land Characteristics, and Forcings, at Four CEOP Sites, J. Meteor. Soc. Japan, 85A, 187-204.

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.

Kirchner, J. W. (2006), Getting the right answers for the right reasons: linking measurements, analyses, and models to advance the science of hydrology, Water Resour. Res., 42, W03S04, doi:10.1029/2005WR004362.

Klemes, V. (1986), Dilettantism in hydrology -- transition or destiny?, Water Resour. Res., 22, S177-S18.

Kollet, S. J., and R. M. Maxwell (2008), Capturing the influence of groundwater dynamics on land surface processes using an integrated, distributed watershed model, Water Resour. Res., 44, W02402, doi:10.1029/2007WR006004.

Konikow, L. F. and J. D. Bredehoeft (1992), Groundwater models cannot be validated, Adv. Water Res., 15, 13–24.

Koster, R. D., and P. C. D. Milly (1997), The Interplay between Transpiration and Runoff Formulations in Land Surface Schemes Used with Atmospheric Models. J. Climate, 10, 1578–1591.

Koster, R. D., M. J. Suarez, A. Ducharne, M. Stieglitz, and P. Kumar (2000), A catchment-based approach to modeling land surface processes in a general circulation model 1. Model structure, J. Geophys. Res., 105(D20), 24, 809–24, 822.

Koster, R. D., et al. (2004), Regions of Strong Coupling Between Soil Moisture and Precipitation, Science, 305 (5687), 1138.

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(1) 233-241.

LeMone, M. A., et al. (2007), NCAR/CU surface, soil, and vegetation observations during the International H2O 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(D20), 4443, doi:10.1029/2001JD000931.

Li, H., A. Robock, and M. Wild (2007), Evaluation of Intergovernmental Panel on Climate Change Fourth Assessment soil moisture simulations for the second half of the twentieth century, J. Geophys. Res., 112, D06106, doi:10.1029/2006JD007455.

Liang, X., E. F. Wood, and D. P. Lettenmaier (1996), Surface soil moisture parameterization of the VIC-2L model: Evaluation and modification, Global Planet. Change, 13, 195-206 .

Liang, X., et al. (1998), The Project for Intercomparison of Land-surface Parameterization Schemes (PILPS) phase 2(c) Red-Arkansas River basin

experiment: 2. Spatial and temporal analysis of energy fluxes, Global Planet. Change, 19(1-4): 137-159.

Liang, X., and J. Guo (2003), Intercomparison of land surface parameterization schemes: Sensitivity of surface energy and water fluxes to model parameters, J. Hydrol., 279, 182-209.

Liang, X., Z. Xie, and M. Huang (2003), A new parameterization for surface and groundwater interactions and its impact on water budgets with the variable infiltration capacity (VIC) land surface model, J. Geophys. Res., 108(D16), 8613, doi:10.1029/2002JD003090.

Lo, M. H., P. J.-F. Yeh and J. S. Famiglietti (2008), Constraining water table depth simulations in a land surface model using estimated baseflow, Adv. Water Resources, 31(12), 1552-1564.

Lohmann, D., et al. (1998), The Project for Intercomparison of Land-surface Parameterization Schemes (PILPS) phase 2(c) Red-Arkansas River basin experiment: 3. Spatial and temporal analysis of water fluxes, Global Planet. Change, 19(1-4) 161-179.

Luo, L. F., et al. (2003), Effects of frozen soil on soil temperature, spring infiltration, and runoff: Results from the PILPS 2(d) experiment at Valdai, Russia. J. Hydromet 4 (2), 334-351.

Lyon, S. W., et al. (2008), Coupling Terrestrial and Atmospheric Water Dynamics to Improve Prediction in a Changing Environment, Bull. Amer. Meteor. Soc., 89, 1275–1279.

Martinez, W and A. Martinez (2002), Computational Statistics Handbook with MATLAB.Chapman and Hall/CRD. ISBN:1584885661 2nd ed. - Boca Raton, Fla. : Chapman & Hall/CRC, 767 p

Maxwell, R. M., and N. L. Miller (2005), Development of a coupled land surface and groundwater model, J. Hydrometeor., 6, 233– 247.

McKay, M., R. Beckman, and W. Conover (1979), A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, Technometrics, 21(2), 239–245 doi:10.2307/1268522.

Meyers, T. P. (2001), A comparison of summertime water and CO2 fluxes over rangeland for well watered and drought conditions, Agric. Forest Meteorol., 106(3), 205-214.

Mitchell, K. E., et al. (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.

Nasonova O. N., Y. M. Gusev, and Y. E. Kovalev (2009), Investigating the ability of a land surface model to simulate streamflow with the accuracy of hydrological models: A case study using MOPEX materials. J. Hydrometeor.: In Press.

Nijssen, B., D. P. Lettenmaier, X. Liang, S. W. Wetzel, and E. F. Wood (1997), Streamflow Simulation for Continental-Scale River Basins, Water Resour. Res., 33(4), 711–724.

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, pg. 3089-3102, John Wiley, Hoboken, N.J.

Niu, G.-Y., and Z.-L. Yang (2003), The versatile integrator of surface and atmosphere processes (VISA). Part II: Evaluation of three topography-based runoff schemes. Global Planet. Change, 38, 191–208.

Niu, G.-Y., Z.-L. Yang, R.E. Dickinson, and L.E. Gulden (2005), A simple TOPMODEL-based runoff parameterization (SIMTOP) for use in GCMs, 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.

Niu, G.-Y., Z.-L. Yang, K. E. Mitchell, F. Chen, M. B. Ek, M. Barlage, L. Longuevergne, A. Kumar, K. Manning, D. Niyogi, E. Rosero, M. Tewari, and Y.-L. Xia (2009), The community Noah land surface model with multi-physics options, J. Geophys. Res.: In Review.

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.

Niyogi, D., K. Alapaty, S. Raman, and F. Chen (2006) Development and evaluation of a coupled photosynthesis-based gas exchange evapotranspiration model (GEM), J. Appl. Meteorol. Clim,: In Review.

Oki, T., T. Nishimura, and P. A. Dirmeyer (1999), Assessment of land surface models by runoff in major river basins of the globe using Total Runoff Integrating Pathways (TRIP). J. Met. Soc. Japan 77, 235–255.

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

Oleson, K. W., G. B. Bonan, J. Feddema, M. Vertenstein (2008b), An urban parameterization for a global climate model. Part II: Sensitivity to input parameters and the simulated urban heat island in offline simulations. J. Appl. Meteor. Climat., 47, 1061-1076, doi: 10.1175/2007JAMC1598.1.2

Overgaard, J., Rosbjerg, D. and Butts, M. B. (2006), Land-surface modelling in hydrological perspective—a review, Biogeosciences, 3(2), 229-241

Peters-Lidard, C. D., D. M. Mocko, M. Garcia, J. A. Santanello Jr., M. A. Tischler, M. S. Moran, and Y. Wu (2008), Role of precipitation uncertainty in the estimation of hydrologic soil properties using remotely sensed soil moisture in a semiarid environment, Water Resour. Res., 44, W05S18, doi:10.1029/2007WR005884.

Pielke Sr., R. A. (2001), Influence of the spatial distribution of vegetation and soils on the prediction of cumulus convective rainfall, Reviews of Geophysics, 39(2), 151-177.

Pitman, A. (1994), Assessing the Sensitivity of a Land-Surface Scheme to the Parameter Values Using a Single Column Model. J. Climate, 7, 1856–1869.

Pitman, A. J., A. Henderson-Sellers (1995), Simulating the diurnal temperature range – results from Phase-1(a) of the Project for Intercomparision of Land-surface Parameterization Schemes (PILPS), Atm. Res. 37 (1-3): 229-245.

Pitman, A. J. et al. (1999), Key results and implications from phase 1(c) of the Project for Intercomparison of Land-Surface Parametrization Schemes. Clim. Dynamics, 15 (9): 673-684.

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

Prihodko L., A. S. Denning, N. P. Hanan, I. Baker, and K. Davis (2008), Sensitivity, uncertainty and time dependence of parameters in a complex land surface model, Agric. Forest Meteorol., 148(2), 268-287, doi: 10.1016/j.agrformet.2007.08.006.

Randall, D.A., R.A. Wood, S. Bony, R. Colman, T. Fichefet, J. Fyfe, V. Kattsov, A. Pitman, J. Shukla, J. Srinivasan, R.J. Stouffer, A. Sumi and K.E. Taylor (2007), Climate Models and Their Evaluation. In: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M.Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Ratto, M., P. C. Young, R. Romanowicz, F. Pappenberger, A. Saltelli, and A. Pagano (2007), Uncertainty, sensitivity analysis and the role of data based mechanistic modeling in hydrology. Hydrol. Earth Syst. Sci., 11, 1249–1266.

Refsgaard, J. C. (1997), Parameterization, calibration and validation of distributed hydrological models, J. Hydrol., 198, 69–97.

Refsgaard, J. C. and Henriksen, H. J. (2004), Modelling guidelines terminology and guiding principles, Adv. Water Res., 27, 71–82.

Refsgaard, J. C., H. J. Henriksen, W. G. Harrar, H. Scholten, A. Kassahun (2005), Quality assurance in model based water management - review of existing practice and outline of new approaches, Environ. Modell. Softw., 20(10), 1201-1215.

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(11):1586-1597.

Refsgaard, J.C., J.P. van der Sluijs, A.L. Højberg, P. A. Vanrolleghem (2007), Uncertainty in the environmental modeling process - A framework and guidance, Environ. Modell. Softw., 11, 1543-1556.

Reichle, R. H., and R. D. Koster (2005), Global assimilation of satellite surface soil moisture retrievals into the NASA Catchment land surface model, Geophys. Res. Lett., 32, L02404, doi:10.1029/2004GL021700.

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.

Rosero, E., and L. A. Bastidas (2007), Evaluation of LSM Parameter Transferability Across Semi-Arid Environments, in Proceedings of the 21st Conference on Hydrology, AMS Meeting, San Antonio, Texas, January 14-18. http://ams.confex.com/ams/pdfpapers/117116.pdf

Rosero, E., L. E. Gulden, Z.-L. Yang, G.-Y. Niu (2007), When different LSMs drive the same phenology module, which better simulates surface to atmosphere fluxes?, 11th LBA-ECO Science Team Meeting, Salvador, Brazil. Sept. 25-28.

Rosero, E., Z.-L. Yang, L. E. Gulden, G.-Y. Niu, and D. J. Gochis (2009a), Evaluating Enhanced Hydrological Representations in Noah LSM over Transition Zones: Implications for Model Development. J. Hydrometeor., 10(3), 600-622. doi: 10.1175/2009JHM1029.1

Rosero E., Z.-L. Yang, T. Wagener, L. E. Gulden, S. Yatheendradas, and G.-Y. Niu (2009b), Quantifiying parameter sensitivity, interaction and transferability in hydrologically enhanced versions of Noah-LSM over transition zones. J. Geophys. Res., In Review.

Saleska, S., L. G. Goncalves, I. Baker, M. Costa, B. Poulter, B. Christoffersen, H. R. da Rocha, K. Didan, A. Huete, H. Imbuziero, B. Kruijt, A. Manzi, C. von Randow, N. Restrepo-Coupe, R. Silva, J. Tota, S. Denning, L. Gulden, E. Rosero, X. Zeng (2008), Effects of seasonality and land-use change on carbon and water fluxes across the Amazon basin: synthesizing results from satellite-based remote sensing, towers, and models, Eos Trans. AGU, 89(53), Fall Meet. Suppl., B54A-07.

Saltelli, A. (1999), Sensitivity analysis: Could better methods be used?, J. Geophys. Res., 104(D3), 3789–3793

Saltelli, A. (2002), Making best use of model evaluations to compute sensitivity indices. Computer Physics Communication, 145, 580–297.

Saltelli, A., A. Tarantola, F. Campolongo, and M. Ratto (2004), Sensitivity Analysis in Practice-A Guide to Assessing Scientific Models, John Wiley and Sons, Chichester.

Saltelli, A., M. Ratto, S. Tarantola, F. Campolongo (2006), Sensitivity analysis practices: Strategies for model-based inference, Reliability Engineering & System Safety, 91(10-11), 1109-1125.

Saltelli, A., M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, S. Tarantola, (2008), Global sensitivity analysis: The primer. Wiley-Interscience. ISBN-13: 978-0470059975

Scanlon, B. R., R. C. Reedy, D. A. Stonestrom, and D. E. Prudic (2005), Impact of land use and land cover change on groundwater recharge and quantity in the southwestern USA, Global Change Biology, 11, 1577–1593, doi: 10.1111/j.1365-2486.2005.01026.x

Schaake, J. C., V. I. Koren, Q.-Y. Duan, K. Mitchell, and F. Chen (1996), Simple water balance model for estimating runoff at different spatial and temporal scales, J. Geophys. Res., 101(D3), 7461–7475.

Schultz, K. and K. Beven (2003), Data-supported robust parameterizations in land surface-atmosphere flux predictions: toward a top-down approach. Hydrol. Process., 17, 2259-2277.

Schwarz, G., (1978), Estimating the dimension of a model, Annals of Statistics, 6(2):461-464.

Sellers, P., S. Los, C. Tucker, C. Justice, D. Dazlich, G. Collatz, and D. Randall (1996), A revised land surface parameterization (SiB2) for atmospheric GCMs part II: The generation of global fields of terrestrial biophysical parameters from satellite data, J. Clim., 9, 706–737.

Shuttleworth, W. J. (2007), Putting the vap into evaporation, Hydrol. Earth Syst. Sci., 11, 210-244.

Sivapalan, M., K. Beven, and E. F. Wood (1987), On hydrologic similarity: 2. A scaled model of storm runoff production, Water Resour. Res., 23, 2266– 2278.

Sivapalan, M., G. Blöschl, L. Zhang, and R. Vertessy (2003), Downward approach to hydrological prediction. Hydrological Processes, 17: 2101-2111, doi: 10.1002/hyp.1425.

Smakhtin, U. (2001) Low flow hydrology: a review, J. Hydrol., 240, 147–186.

Smith, L.A. (2002), What might we learn from climate forecasts? PNAS 99: 2487-2492. doi:10.1073/pnas.012580599.

Sobol', I. M. (1993), Sensitivity analysis for non-linear mathematical models, Math. Modelling Comput. Exp., 1, 407–414.

Sobol', I. M. (2001), Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates, Math. Comput. Simul., 55, 271–280, doi:10.1016/S0378-4754(00)00270-6

Soil Survey Staff, Natural Resources Conservation Service, United States Department of Agriculture. U.S. General Soil Map (STATSGO2) for Oklahoma Available online at http://soildatamart.nrcs.usda.gov accessed 01/05/2009.

Sorooshian, S., and J. A. Dracup (1980), Stochastic Parameter Estimation Procedures for Hydrologic Rainfall-Runoff Models: Correlated and Heteroscedastic Error Cases, Water Resour. Res., 16(2), 430–442

Spear R. C. and G. M. Hornberger (1980), Eutrophication in Peel Inlet-II: Identification of critical uncertainties via generalized sensitivity analysis, Water Resour. Res., 14, pp. 43–99.

Spear, R. C., T. M. Grieb, and N. Shang, (1994), Parameter uncertainty and interaction in complex environmental models, Water Resour. Res., 30(11), 3159-3169.

Stieglitz, M., D. Rind, J. S. Famiglietti, and C. Rosenzweig (1997), An efficient approach to modeling the topographic control of surface hydrology for regional and global modeling. J. Climate 10, 118–137.

Stöckli, R., D. M. Lawrence, G.-Y. Niu, K. W. Oleson, P. E. Thornton, Z.-L. Yang, G. B. Bonan, A. S. Denning, and S. W. Running (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. Proceedings, ECMWF Workshop on Predictability.

Tang, Y., P. Reed, T. Wagener, and K. van Werkhoven (2006), Comparing sensitivity analysis methods to advance lumped watershed model identification and evaluation, Hydrol. Earth Syst. Sci. Discuss., 3, 3333–3395.

Tang, Y., P. Reed, K. van Werkhoven, and T. Wagener (2007), Advancing the identification and evaluation of distributed rainfall-runoff models using global sensitivity analysis, Water Resour. Res., 43, W06415, doi:10.1029/2006WR005813.

Trenberth, K. E., A. Dai, R.M. Rasmussen, and D.B. Parsons (2003), The Changing Character of Precipitation. Bull. Amer. Meteor. Soc., 84, 1205–1217.

Trier, S. B., F. Chen, K. W. Manning, M. A. LeMone, and C. A. Davis (2008), Sensitivity of the PBL and Precipitation in 12-Day Simulations of Warm-Season Convection Using Different Land Surface Models and Soil Wetness Conditions. Mon. Wea. Rev., 136, 2321–2343.

Unland, H., P. Houser, W. J. Shuttleworth, and Z.-L. Yang (1996), Surface flux measurement and modeling at a semi-arid Sonoran Desert site, Agric. For. Meteorol., 82, 119–153.

van Werkhoven, K., T. Wagener, P. Reed, and Y. Tang (2008), Characterization of watershed model behavior across a hydroclimatic gradient, Water Resour. Res., 44, W01429, doi:10.1029/2007WR006271.

van Werkhoven, K., T. Wagener, P. Reed, and Y. Tang (2009), Sensitivity-guided reduction of parametric dimensionality for multiobjective calibration of watershed models, Adv. Water Resour., in press

Viterbo, P., (2002): A review of parametrization schemes for land surface processes. Meteorological Training Course Lecture Series, ECMWF, Shinfield Park, Reading, England, 1-49

Vogel, R.M., and Fennessey, N.M. (1994), Flow duration curves. I. A new interpretation and confidence intervals. J. Water Resour. Plan. Manag. 120 (4), 485–504.

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., Boyle, D. P., Lees, M. J., Wheater, H. S., Gupta, H. V. and Sorooshian, S. (2001), A framework for development and application of hydrological models, Hydrol. Earth Syst. Sci., 5, 13-26.

Wagener, T., and H. V. Gupta (2005), Model identification for hydrological forecasting under uncertainty, Stoch. Environ. Res. Risk Assess., 19(6), 378–387, doi:10.1007/s00477-005-0006-5.

Wagener, T., et al. (2006), The Model Parameter Estimation Experiment (MOPEX): Its structure, connection to other international initiatives and future directions. IAHS Pub. 307.

Wagener, T., and J. Kollat (2007), Numerical and visual evaluation of hydrological and environmental models using the Monte Carlo Analysis Toolbox, Environ. Modell. Softw., 2, 1021–1033.

Wagener, T., Sivapalan, M., Troch, P., Woods, R. (2007), Catchment classification and hydrologic similarity. Geography Compass, 1, 901-931. doi:10.1111/j.1749-8198.2007.00039.x.

Wagener, T., Reed, P., Van Werkhoven, K., Tang, Y. and Zhang, Z. (2009), Advances in the identification and evaluation of complex environmental system models. Journal of Hydroinformatics, 11(3-4), 266-281. doi:10.2166/hydro.2009.040.

Wang, L., T. Koike, K. Yang, T. J. Jackson, R. Bindlish, and D. Yang (2009), Development of a distributed biosphere hydrological model and its evaluation with the Southern Great Plains Experiments (SGP97 and SGP99), J. Geophys. Res., 114, D08107, doi:10.1029/2008JD010800.

Weckwerth, T. M. et al. (2004), An overview of the International H2O Project (IHOP_2002) and some preliminary highlights. Bull. Amer. Meteor. Soc., 85, 253–277.

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

Wetzel, P., X. Liang, P. Irannejad, A. Boone, J. Noilhan, Y. Shao, C. Skelly, Y. Xue and Z.-L. Yang, 1996: Modeling vadose zone liquid water fluxes: Infiltration, runoff, drainage, interflow, Global Planet. Change, 13, 57-71

Wood, E. F., D. P. Lettenmaier, and V. G. Zartarian (1992), A Land-Surface Hydrology Parameterization With Subgrid Variability for General Circulation Models, J. Geophys. Res., 97(D3), 2717–2728.

Wood, E. F. et al. (1998), The Project for Intercomparison of Land-surface Parameterization Schemes (PILPS) phase 2(c) Red-Arkansas River basin experiment: 1. experiment 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 multi-criteria methods and the Cabauw data set, J. Hydrometeor., V3(2) pp. 181-194.

Yang, Z.-L., and G.-Y. Niu (2003), Versatile integrator of surface and atmosphere processes (VISA) Part 1: Model description. Glob. Planet. Change 38, 175–189.

Yang, Z.-L., (2004), Modeling land surface processes in short-term weather and climate studies, in Observation, Theory and Modeling of Atmospheric Variability, edited by X. Zhu, X. Li, M. Cai, S. Zhou, Y.Zhu, F.-F. Jin, X. Zou, and M. Zhang, World Scientific Series on Meteorology of East Asia, World Scientific, New Jersey, pp. 288-313.

Yatheendradas, S., T. Wagener, H. Gupta, C. Unkrich, D. Goodrich, M. Schaffner, and A. Stewart (2008), Understanding uncertainty in distributed flash flood forecasting for semiarid regions, Water Resour. Res., 44, W05S19, doi: 10.1029/2007WR005940.

Yeh, P.J.-F., and E. A. B. Eltahir (2005), Representation of water table dynamics in a land-surface scheme. Part I: Model development, J. Clim., 18, 1861– 1880.

Yilmaz, K. K., H. V. Gupta, and T. Wagener (2008), A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resour. Res., 44, W09417, doi:10.1029/2007WR006716

Vita

Enrique Rosero was awarded the NOAA National Weather Service, Office of Hydrologic Development Graduate Fellowship. He graduated cum laude with a B.S. in Civil Engineering and a minor in Environmental Engineering from the Escuela Politécnica Nacional in Quito, Ecuador. He then went on to earn an M.S. in Environmental Fluid Mechanics at the University of Karlsruhe in Karlsruhe, Germany, and an M.E. in Civil and Environmental Engineering at Utah State University in Logan, Utah. He has worked as Junior Hydrologist and Modeler at Caminos y Canales Consulting Co. and as Graduate Research Assistant at the Institute for Hydromechanics at the University of Karlsruhe, the Utah Water Research Laboratory, and the Department of Geological Sciences at the University of Texas at Austin. He was a visiting scientist at the Hydrological Sciences Branch at NASA's Goddard Space Flight Center. A complete list of published peer-reviewed articles and conference proceedings can be found at https://webspace.utexas.edu/err449/erosero.html. He joined the Integrated Reservoir Performance Prediction Division of ExxonMobil Upstream Research Company in Houston, Texas, as a Senior Research Geoscientist.

Permanent address: 3133 Buffalo Speedway Apt. 8310, Houston, TX 77098-1870

This dissertation was typed by the author.