Quantifying Parameter Sensitivity, Interaction and Transferability in Hydrologically Enhanced Versions of Noah-LSM over Transition Zones

ENRIQUE ROSENO,1 ZONG-LIANG YANG,1 THORSTEN WAGENER,2 LINDSEY E. GULDEN,1 SONI YATHEENDRADAS,3,4 and GUO-YUE NIU1

1Department of Geological Sciences, Jackson School of Geosciences, The University of Texas at Austin, Austin, Texas, USA
2Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park, Pennsylvania, USA
3NASA Goddard Space Flight Center, Hydrological Sciences Branch, Greenbelt, Maryland, USA
4Earth System Science Interdisciplinary Center, University of Maryland, College Park, Maryland, USA

Manuscript submitted 10 March 2009

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

We use sensitivity analysis to identify the parameters that are most responsible for shaping land surface model (LSM) simulations and to understand the complex 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 US 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 output. Significant parameter interaction occurs so that not only the optimal parameter values differ between models, but the relationships between parameters change. GW decreases parameter interaction and appears to improve model realism, especially at wetter sites. DV increases 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 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.

Key words: sensitivity analysis, land surface models, Noah LSM, model identification

Index terms: 1843 Land/atmosphere interactions, 1805 Computational hydrology, 0550 Model verification and validation, 1873 Uncertainty assessment, 1847 Modeling
1. 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), which 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 exploring high-dimensional parameter spaces of complex environmental models to better understand what controls model performance (Saltelli et al., 2008). Monte Carlo-based uncertainty and SA uses the results of 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). In this article, we inform LSM development by using sophisticated SA to guide the development of the commonly used Noah LSM (Ek et al., 2003).

SA has been an underutilized resource in LSM development. Approaches to quantify 'sensitivity' (the rate of change in model response with respect to a factor) have very frequently been restricted to a simple exploratory analysis of the effects of factors taken one-at-a-time
(OAT), without regard for interactions between factors. Such factors can be parameters (e.g., Pitman, 1994; Gao et al., 1996; Chen and Dudhia, 2001; Trier et al., 2008), meteorological forcing, or ancillary data sets (e.g., Kato et al., 2007; Gulden et al., 2008a). OAT is illicit and unjustified for non-linear models (Saltelli, 1999; Bastidas et al., 1999; Saltelli et al., 2006). More powerful and sophisticated approaches that account, explicitly or implicitly, for parameter interactions include the factorial method (e.g., Liang and Guo, 2003; Oleson et al., 2008) and regionalized sensitivity analysis (RSA) (e.g., Bastidas et al., 2006, Demaria et al., 2007; Prihodko et al., 2008). The factorial method is a global variance-based sensitivity analysis (VSA) in which a selected number of model runs whose parameters have been perturbed by an arbitrary percent from an arbitrary reference value (default) are evaluated to identify parameters that affect the variance of the model output and to calculate a rough estimate of the low-order interactions. RSA, on the other hand, representatively samples the entire parameter space and provides a more robust assessment of the way in which parameter distributions change between subjectively defined ‘good’ and ‘bad’ (i.e., behavioral and non-behavioral) model simulations. RSA does not explicitly account for interactions between parameters and is typically applied with the sole purpose of identifying sensitive parameters to reduce the dimensionality of the calibration problem (e.g., Bastidas et al., 1999). When RSA and VSA are used separately, both the lack of firm conclusions regarding the effect of dominant parameters (and their interactions) on the output (e.g., Bastidas et al., 2006) 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 model development. Our approach complements the information about the posterior distributions of behavioral parameters (RSA) with information about their contributions to model variance (VSA). The Monte Carlo-based VSA that we use (the method of
Sobol') is more robust than factorial analysis because it employs a representative sample of parameter space and bypasses the perceived 'complexities' often associated with powerful SA methods.

Our SA guides the investigation of 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). We evaluate the performance of Noah LSM in simulating 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), our work focuses on warm-season climates of the US Southern Great Plains. We document how parameter interaction and sensitivity varies with model, site, soil, vegetation, and climate. We investigate the ways in which alternate model structures affect the behavior of 'physically meaningful' LSM parameters, focusing on the dominant parameter interactions that dictate model response. We show that only a few parameters directly control model variance and that parameter interaction is significant.

We further compare the similarity of estimated multivariate posterior distributions of behavioral parameters and their sensitivity against those obtained at other sites. We show that the changes between sites are not solely controlled by soil or vegetation types but appear to be strongly related to the climatic gradient.

This paper is organized as follows. Datasets, models, and methods are described in section 2. Experimental design and driving questions are formulated in section 3. Section 4 presents the patterns of sensitivity obtained via RSA and the global variance-based method of Sobol'. Section 5 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 6 discusses
implications of the results for the transferability of parameters between locations with similar
physical characteristics. Conclusions are summarized in section 7.

2. Models, data and methods

2.1. Hydrologically enhanced versions of Noah LSM

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

2.1.1. 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.

2.1.2. 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.

2.1.3. 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). The model does not explicitly consider lateral transport of groundwater between grid cells.

2.2. IHOP_2002 sites and datasets

We used datasets available at http://www.rap.ucar.edu/research/land/observations/ihop/ from the IHOP_2002 field campaign (Weckwerth et al. 2004; 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. 1). The nine stations were sited to obtain a representative sample of the region, which spans a strong east–west rainfall gradient and 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 1.
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).

2.4. Evaluation datasets

To constrain and evaluate the models, we used sensible heat flux (H), latent heat flux (LE), ground heat flux (G), ground temperature (Tg), and first layer 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; Stockli et al., 2008). To score model performance, we used root mean square error (RMSE).

2.5. Parameters considered in the sensitivity analysis

We selected 10 soil and 10 vegetation parameters of STD that have been deemed important at similar locations (Demarty et al. 2004; Bastidas et al., 2006). We included eight parameters responsible for the phenology module and four that control the 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 2) for all parameters were taken from the literature (e.g., Chen and Dudhia, 2001; Hogue et al., 2006).
2.6. Methods for sensitivity analysis

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. Unlike RSA, 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., Ratto et al., 2007; Tang et al., 2006, 2007; Yatheendradas et al., 2008; van Werkhoven et al., 2008). To our knowledge, it has not yet been used for LSM SA.

Since variance-based SA (VSA) does not allow a mapping of the sensitivity back to the parameter space, we complemented our evaluation with regionalized sensitivity analysis (RSA).

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

Sobol' indices enable researchers to distinguish the subset of independent input factors \( X = \{x_1, x_2, \ldots, x_k\} \) 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 \( Y = f(x_1, x_2, \ldots, x_k) \) (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 \( (V) \):

\[
S_i = \frac{V_i}{V(Y)} = \frac{V(E(Y | x_i))}{V(Y)} = \frac{\hat{U}_i - \hat{E}^2(Y)}{\hat{V}(Y)} \tag{1}
\]

Where \[
\hat{U}_i = \frac{1}{n-1} \sum_{r=1}^{n} f(x_{r1}, x_{r2}, \ldots, x_{rk}) f(x'_{r1}, x'_{r2}, \ldots, x'_{r(i-1)}, x_{ri}, x'_{r(i+1)}, \ldots, x_{rk}) \tag{2}
\]
The estimates of the mean squared and the total variance are computed as:

\[
\hat{E}^2(Y) = \frac{1}{n} \sum_{r=1}^{n} f(x_{r1}, x_{r2}, \ldots, x_{rk}) f(x'_{r1}, x'_{r2}, \ldots, x'_{rk})
\]

(3)

\[
\hat{V}(Y) = \frac{1}{n} \sum_{r=1}^{n} f(x_{r1}, x_{r2}, \ldots, x_{rk})^2 - \hat{E}^2(Y)
\]

(4)

Instead of computing all \(2^k - 1\) terms of the variance decomposition:

\[
V(Y) = \sum_{i} V_i + \sum_{i,j} V_{ij} + \ldots + V_{12.k}
\]

(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 \((S_{T,i})\) associated with parameter \(x_i\). \(S_{T,i}\) 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. \(S_{T,i}\) 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:

\[
S_i = \frac{V(Y) - V(E(Y|x_i))}{V(Y)} = 1 - \frac{\hat{U}_{-i} - \hat{E}^2(Y)}{V(Y)}
\]

(6)

Where

\[
\hat{U}_{-i} = \frac{1}{n-1} \sum_{r=1}^{n} f(x_{r1}, x_{r2}, \ldots, x_{rk}) f(x_{r1}, x_{r2}, \ldots, x_{r(i-1)}, x'_{r(i)}, x_{r(i+1)}, \ldots, x_{rk})
\]

(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. To spare \(nk\) simulations for each model and each site (and to show that is possible to mine regular Monte Carlo runs, customary
used in land-surface modelling), we estimated the second part of this term by interpolating the
response surface $Y$ (as explained in Section 2.7.3).

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).

2.6.2. Kolmogorov-Smirnov test for regionalized sensitivity analysis (RSA)

In RSA (Hornberger and Spear, 1981; Spear et al., 1994; Bastidas et al., 1999, 2006), the
Kolmogorov-Smirnov (K-S) test helps identify which parameters differ when model
performance is 'behavioral' ($B$) ('good', 'acceptable') versus 'non-behavioral' ($\bar{B}$). Note that
RSA does not give any information regarding the extent to which a parameter affects the
variance of the model output. It instead focuses on which factors are believed to affect the
occurrence of behavioral model realizations.

The two-sample K-S test (two-sided version) is independently applied to the univariate
marginal cumulative density functions (CDF) of each parameter to test whether the distributions
of $B$ and $\bar{B}$ are different at a particular significance level. The maximum distance between the
two CDFs serves to evaluate the null hypothesis that the parameter values come from the same
probability distribution. It is intuitive that, if the CDFs are dissimilar from one another (as well
as different from the a priori marginal distribution of that parameter), then such parameter is
'sensitive' (Saltelli, 2006).
Generally, a set of subjective constraints or thresholds provides a qualitative definition for $B$, and model realizations are classified into either $B$ or $\bar{B}$. In the research presented here, we considered the calibrated parameter sets (which belong to the joint posterior probability distribution that maximizes the likelihood function) as representative of $B$. The $\bar{B}$ sample was defined by parameter sets whose RMSE is half a standard deviation or more away from the mode of the cost function. For robustness, for each parameter, several different samples from $\bar{B}$ were taken and compared to the one of $B$. We used 5% as the significance level for our RSA.

2.7 Sampling strategies for sensitivity analysis

We generated representative samples of model parameters using Latin Hypercube Sampling (LHS) and of the behavioral parameter sets through multi-objective calibration. We also sampled from the response surfaces obtained via LHS using inexpensive Support Vector Machines (SVM).

2.7.1 Latin Hypercube Monte Carlo sampling (LHS)

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 detailed representation of the range of model responses. We used LHS 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). LHS 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 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 $S_i$ index of the Sobol' method (Eq. 1), we used a modified LHS that enables replication. We
sequentially sampled the same parameter value twice from a given hypercube, while all other parameters in the set were chosen according to the LHS.

2.7.2 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 \(B\).

2.7.3 Response surface sampling (meta-modeling) using Support Vector Machines (SVM)
To inexpensively compute the second component of the $U_i$ term (Eq. 7) for the Sobol' method (Section 2.6.1), we approximated the response surfaces of the LSMs using SVM trained on the Monte Carlo samples. We used the SVMlight libraries of Joachims (1999) available at http://svmlight.joachims.org/.

SVM are a robust learning strategy that efficiently generalizes a non-linear functional dependence between a set of multi-dimensional inputs and a vector of outputs, $Y=f(X)$. SVM use kernels to map the complex data into high-dimensional feature spaces in which linear relationships can describe separation planes. Through a sparse set of training examples that 'support' the hyperplane that best fits the observed data, SVM create a model to approximate $Y$ (Vapnik, 1998). SVM have been successfully applied in a variety of hydrological applications (e.g., Asefa et al., 2004, 2005, 2006; Khalil et al., 2005; Kaheil et al., 2008a, 2008b). SVM perform better than artificial neural networks (Dibike et al., 2001; Gill et al., 2008), with lower complexity (only 3 parameters), less computational cost (faster training), and reduced sensitivity to noisy inputs and sparse training samples.

Based on statistical learning theory (Smola and Schölkopf, 2004), SVM are characterized by a mechanism that avoids overfitting and that results in good generalization. A coefficient $C$ determines the tradeoff between the complexity (or flatness) of the function $f$ and the amount to which deviations larger than empirical error $\varepsilon$ are tolerated. In the SVM training, prediction error and model complexity are simultaneously minimized in a quadratic optimization problem, which results in a uniquely global optimum (Vapnik, 1998). We used Gaussian radial basis function kernels and determined near-optimal values of the 3 SVM hyper-parameters (the capacity $C$, the soft margin $\varepsilon$ and the kernel width $\gamma$) using the SCE calibration algorithm (Duan et al., 1992).
For detailed explanation on the estimation procedure for SVM parameters, we refer interested readers to Khalil et al. (2006).

The error in the estimation of the \( U_t \) term is small. For example, an SVM \((C=97.12, \varepsilon=0.0025 \text{ and } \gamma=0.0323)\) trained on 5,000 random samples of GW \([Y = \text{RMSE}(\text{SMC}_{5cm})]\) at Site 9 (and calibrated against other 5,000 testing samples) predicted the remaining independent validation samples from the \( Y \) surface with a RMSE of 0.93 \%. The estimated variance of \( Y \) (8.857) is only 3\% different than the ‘true’ variance (8.584) obtained by the Monte Carlo runs.

2.8. Hierarchical clustering for comparisons of parameter distributions

Unsupervised classification of behavioral parameter distributions allowed us to gain understanding about the similarities of the data across locations, specifically about the relationships between types of parameters and sites. We used clustering methods to separate the collection of marginal posterior distributions of calibrated parameters sets into groups. 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 into a single cluster (Hair et al., 1995). Since 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. Driving questions and experimental design
We first ask: What are the dominant model parameters across the region? We ran 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 RSA and the method of Sobol'. Our SA guides our further investigation.

We then address the question: What are the dominant interactions between model parameters, and how do these 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 distributions according to types of soil and vegetation. Our classification provides insights into how parameters can be transferred to ungauged locations.

4. What parameters are sensitive?

We report the results of our regionalized and variance-based sensitivity analyses, which we use in the next section to help evaluate model behavior.

4.1. Regionalized sensitivity analysis (RSA)

The sensitivity of a given parameter changes between models and varies by location, without an easily recognizable pattern. Table 3 presents the model parameters deemed sensitive at the 5% significance level according to a Kolmogorov-Smirnov (K-S) test between samples.
that drive behavioral (multi-objectively calibrated) and non-behavioral simulations. A value of 1 means the parameter is sensitive, 0 means insensitive. The number of sensitive parameters is tabulated by site and by class. In all versions of Noah LSM, soil parameters tend to be more sensitive than vegetation parameters.

At five of the nine sites, GW has the fewest number of sensitive parameters of any of the three models, despite having more total parameters than STD. At the dry sites (1-3), the addition of GW reduces the number of vegetation parameters that are sensitive, perhaps because when GW processes are not included, the vegetation parameters are given more weight in controlling the water balance of the soil. At site 9, despite decreasing the number of sensitive vegetation and soil parameters with respect to STD, GW does not even make the parameters associated with GW sensitive.

DV most often has the highest number of sensitive parameters. The number of vegetation parameters at dry sites that are sensitive in DV is significantly higher than at wet sites. Not surprisingly, the conversion factor ($gl$) and the specific leaf area ($sla$) are sensitive everywhere. The fraction of carbon into growth ($fragr$) and the water stress parameter ($wtsr$) are sensitive at eight sites, while the coefficient for soil respiration ($rssoil$), the optical depth ($tauhf$) and the wood allocation parameter ($bf$) are the least often sensitive. The variability of parameter sensitivity between even similar sites perhaps points to a lack of constraint of DV with respect to the partitioning of the carbon budget or suggests that interactions with other parameters decreases the separation of the CDFs, consequently hindering the identification of sensitive parameters (Yatheendradas et al., 2008).

The widespread use of RSA in LSM development has been precluded by what has been RSA's lack of firm conclusions because of the relatively poor agreement between the
conclusions of RSA studies. Even though we and Bastidas et al. (2006) used roughly the same
version of Noah (STD) and allowed many of the same parameters to vary, we found that our
RSA is at best partially consistent with the assessment of Bastidas et al. (2006). One possible
reason for our divergent assessments is that we did not analyze exactly the same parameters and
did not sample from exactly the same ranges as did Bastidas et al. A single parameter’s
sensitivity, as defined in RSA approaches, is likely a function of the values of other model
parameters; alteration of the sampled range of a single parameter can influence the sensitivity of
many other parameters (Spear and Hornberger, 1980). Results presented elsewhere (Rosero et
al., 2009) have shown that adjusting only a partial set of parameters alters model structure (i.e.,
the relationships between parameters); it follows that, with changed model structure,
regionalized sensitivity would also vary.

Perhaps the most important reason why information obtained through RSA has not been
used to inform LSM development is that RSA provides neither information regarding the
magnitude of a parameter’s influence on model output nor information about the influence that
the parameter exerts through interactions. RSA only identifies a subset of parameter sets that
improves (if only slightly) model output. A K-S statistic that corresponds to ‘high sensitivity’
means the parameters used to generate the behavioral model realizations are ‘a lot different’ from
those used to generate the nonbehavioral realizations. ‘High sensitivity’ does not mean that a
parameter is highly important to (exerts significant control over) model output. Furthermore,
having invariant parameter distributions in the transition from \( \bar{B} \) to \( B \) is a necessary but not
sufficient condition for insensitivity (Spear and Hornberger, 1980). Quantifying regionalized
sensitivity has the potential to allow a mapping of the model sensitivity to parameters back to the
parameter space, but without information on the change in model's variance, RSA appears only useful for limiting the dimensionality of the calibration problem at a specific site.

RSA is distinct from the variance-based analysis of sensitivity in model output to parameter values. The method of Sobol' identifies parameters that exert significant control on the variance of model output. Sobol' indices are arguably a more useful tool for model developers and the model's operational users.

4.2 Global variance-based sensitivity analysis (VSA)

Our 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 2, 3, and 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). The easily recognizable pattern of sensitivity can be linked to physical characteristics of the sites.

4.2.1 First-order sensitivity ($S_i$)

The identified sensitive parameters (Figs. 2-4) are consistent with our expectations. When parameter influences changes as we would physically expect, we interpret these results as consistent with our hypothesis that to a first order, the models are adequately representing the site-to-site variation in these key components of the water and energy cycles. We observe that, for most sites and models, the greatest first-order control on simulated top-layer soil moisture is porosity (maxsmc) (Fig.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 ($f_{xexp}$) exerts the
most first-order control on soil moisture. The LE flux simulated by GW at dry sites is controlled
by $f_{xexp}$ and specific yield ($rous$), which controls depth to the water table. $lai$ directly exerts
control on transpiration and hence on the surface energy budget; at the most vegetated sites (7-9),
$lai$ consequently shapes the most the variance of $H$ and LE for both STD and GW (Fig 2a, 3a).
The initial value $lai$ is not important to determining DV’s simulated $H$ and LE because, unlike
the other two models, DV allows the initial value of $lai$ to change. Instead, minimum stomatal
resistance ($r_{cmin}$) exerts the most control on DV-simulated LE flux. Two new parameters
associated with DV, $gl$ and $sla$, which control the calculation of $lai$, also exert first-order control
on the simulated energy balance. In sparsely vegetated sites (1-3), the Zilintikevich coefficient
($c_{zil}$) plays a significant role in the variance of $H$.

The control of parameters on model variance changes between models and between sites.
The shift in dominant parameters gives clues about how the models are representing the water
and energy cycles from semi-arid (1-3) to semi-humid (7-9) sites. Looking only at STD, we see
that as the mean annual precipitation (MAP) at the sites increases, $f_{xexp}$ becomes less important
to determining the top-layer soil moisture, and $ref_{kdt}$, a parameter involved in determining
maximum rates of infiltration, becomes more important (Fig. 4a). Expanding our view to include
GW, in which surface runoff is de-emphasized and subsurface runoff is emphasized (see
discussion about GW’s preferred modes of operation, Section 5), we see that although $f_{xexp}$ is
still a key parameter exerting first-order control on SMC at the dry sites, $ref_{kdt}$ has little direct
influence on soil moisture at the wet sites. The most sensitive parameter for SMC$_{5cm}$ at sites 1-3
is the aquifer’s specific yield ($rous$), which effectively controls whether aquifer water will be
accessible to the near-surface soil. Consistent with our expectations, soil suction ($psi_{sat}$), which
in GW controls upward movement of water from the aquifer to the soil, has significant control on SMC within GW but not within STD, in which \textit{psisat} exerts less control over soil hydraulic behavior (Fig. 4a).

As sites get progressively wetter, surface exchange coefficient \textit{czil} exerts progressively less influence on simulated H, and \textit{rcmin} exerts progressively more influence on H (Fig. 2a). The shift in control 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, \textit{rcmin} and \textit{lai} have increasing control over simulated LE, while \textit{fxexp} becomes less influential (Fig. 3a). Even at dry sites (1-3), DV tends to favor larger values of vegetation fraction (\textit{shdfac}) than are prescribed by STD and GW. As a consequence, DV stands apart from GW and STD in that \textit{fxexp} does not directly contribute to variance of any objective at the three driest sites (with the exception of the unvegetated, bare-soil site 1, where LE is controlled by \textit{fxexp}).

Examinations of \textit{\$i} that are not in line with expectations may be used to help modelers diagnose problems with forcing data and/or model structure. For instance, \textit{fxexp} has the highest \textit{\$i} of simulated H and LE at site 6. Site 6 receives 800 mm of precipitation each year. We do not expect direct evaporation to be a 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 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.

In contrast to what was observed with RSA, the VSA shows a clear pattern of sensitivity across the sites. 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 climatic gradient.

4.2.2 Sensitivity through interactions ($S_i-S_T$)

Figures 2b, 3b, and 4b show that interactions between parameters are responsible for most of the variance in the models’ predicted $H$, LE, and SMC. If we assume that the parameterizations are correct, then the significant parameter interaction indicates model overparameterization (Saltelli et al., 2008). Overparameterization causes increased parameter interaction (observed here) and decreases the percentage of parameters that are deemed sensitive by RSA (Bastidas et al., 2006; Yatheendradas et al., 2008) (observed in the RSA described above). VSA enables the quantification of the significance of parameter interaction and the identification of the most interactive parameters. 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), especially when there are no known functional relationships between the physical quantities that developers believe two parameters represent, parameter interaction is likely to be indication that the model is working in ways that are 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, GW reduces parameter interaction at the middling moisture (4-6) and semi-humid sites (7-9) (e.g., Fig.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 is consistent with observations on the robustness of GW (Gulden et al., 2007). Conversely, GW appears to produce
significant parameter interaction at semiarid sites 1-3, indicating a need for an improved parameterization of groundwater processes at semi-arid sites. DV parameters are much more interactive than those of STD and GW, especially at the wettest sites when simulating LE and SMC. 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., 2009).

There is significant interaction between the vegetation parameters and the sensitivity of soil parameters. These results fully support the findings of Demaria et al. (2007) and Liang and Guo (2003). The significant interplay between soil and vegetation parameters implies that parameter interaction between different model parts changes both the optimal value of a given parameter and the relationships between parameters. In the next section, we use the multivariate posterior distribution of behavioral parameters to characterize the impact of adding new modules on parameter interactions.

Looked at in full, we conclude that the models best represent the surface water and energy balance 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.

5. **How do sensitive parameters interact and shape model behavior?**

We present a case study in which sensitivity analysis (SA) links model identification and model development. SA identifies parameters whose behavior merits further examination: VSA
indicates which parameters are most influential in determining model output; RSA helps us identify modes in parameter distributions that yield distinctive model behavior.

5.1 A case study at Site 7

At site 7, STD, DV, and GW show nearly equivalent performance when using their behavioral parameter sets (Figs. 5, 6). Such ‘equifinality’ is well documented in the hydrologic modeling literature (e.g. Beven and Freer, 2001; Rosero et al., 2009). 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 but to fully 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 conform to our physical understanding of the systems.

5.2 Focus on sensitive parameters to better understand model function

We follow the impact of shifted preferred values of three ‘physically meaningful’ parameters that made considerable contributions to first-order variance: porosity \((smc_{\text{max}})\), the muting factor for vegetation’s effect on thermal conductivity \((sbeta)\), and minimum stomatal resistance \((rcmin)\). 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 functional relationship between the parameters (the information contained in the interactions) is most relevant for purposes of model development.

Figure 7 shows the marginal CDF of the posterior multivariate distribution for selected sensitive parameters in the behavioral range at Site 7. Along with the CDFs, box plots show the median and interquartile ranges of the skewed parameter distributions. The scatterplots show the variation in RMSE of LE and \(SMC_{5cm}\) along the range of plausible values of the parameter. The
trend is emphasized by fitting the points with a polynomial of minimal complexity. Even a
cursory analysis of Figure 7 shows that the direction of ‘sensitivity’ (understood as the rate of
change of score with parameter value) changes between models (e.g., Fig. 7a). The simulation of
SMC$_{5cm}$ by STD and DV degrades as porosity increases, while GW improves. We also note that,
along the plausible range, the response can be enhanced (Fig. 7d) or become relatively
insensitive to changes in parameter value (Fig. 7c). The identifiability of parameters (understood
as having 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 ($cmc_{max}$), but STD and GW have less of a preference (Fig. 7c). The
interquartile range of $rc_{min}$ of STD is smaller than that of GW or DV (Fig. 7b). We can clearly
see that different models have distinct optimal parameter values for the same physical parameter,
implying not only that the parameters cannot be transferred between models but that the
relationships between them are different.

5.2.1 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 $maxsmc$:

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

(8)

$SMC_{dry}$ is the lowest volumetric water content that can exist in the top soil layer, and $fx_{exp}$ is a
parameter ranging from 0.2 to 4.
In STD and DV, the error in simulating LE tends to be relatively small when \( \text{maxsmc} \) is low and relatively large when \( \text{maxsmc} \) is high (Fig. 7a). However, GW shows slight improvement in simulating LE as \( \text{maxsmc} \) increases. The tendency of STD and DV to simulate LE well when \( \text{maxsmc} \) 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. Given the same \( \text{maxsmc} \) value, GW more easily simulates sufficient direct evaporation, perhaps because of wetter soil (Rosero et al., 2009).

In STD and DV, \( \text{maxsmc} \) controls both surface and subsurface runoff. Hydraulic conductivity of the soil layer (\( \text{wcnd} \)) is computed using Clapp and Hornberger’s method, which scales saturated hydraulic conductivity (\( \text{dksat} \)) by wetness (\( \text{SMC}/\text{maxsmc} \)), raised to an exponent determined by the Clapp and Hornberger parameter (\( b \)):

\[
\text{wcnd} = \text{dksat} \left( \frac{\text{SMC}}{\text{maxsmc}} \right)^{2b-3}
\]

(9)

Lower \( \text{maxsmc} \) yields higher \( \text{wcnd} \), which means water moves through the soil more quickly. Note that, in the case of parameterizing subsurface runoff, hydraulic conductivity is a representation of the speed at which the water can move laterally through the soil. In STD and DV, subsurface runoff (\( \text{Runoff2} \)) is \( \text{wcnd} \) times the slope. Consequently, higher values of \( \text{maxsmc} \) decrease \( \text{Runoff2} \). Higher values of \( \text{maxsmc} \) also decrease surface runoff (\( \text{Runoff1} \)) by increasing the maximum rate of infiltration. Both changes increase the wetness of the surface soil.

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

\[
\text{Runoff2} = \text{rsbmx} e^{-0.76x_{wtr}}
\]

(10)
Where $rsbmx$ is the maximum rate of subsurface runoff, $fff$ is the e-folding depth of saturated hydraulic conductivity, and $Z_{WT}$ is a variable that represents the depth to the water table. Runoff is a modified version of the same function used to compute subsurface runoff (Niu et al., 2005):

$$\text{Runoff} = pcpdrp \times \left( f_{satmx} e^{-0.5 \times fff \times Z_{WT}} \right)$$

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, virtually eliminates the tradeoff between good simulation of LE and SMC5cm. GW is, in this regard, a better model for simulating fluxes at site 7.

The question remains – why does GW poorly simulate SMC when $maxsmc$ increases? $maxsmc$ is used to compute vertical hydraulic conductivity (using the same function described above), which GW uses to control 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 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 recommend use of runoff constraints before drawing firm conclusions regarding the physical reality of the runoff-related processes in GW.

5.2.2 The role of SBETA

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

\[ G = DF_1 \frac{STC_1 - T_1}{0.5 \times ZSOIL_{(1)}} \] (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. It uses parameter \( sbeta \) and the vegetated fraction (\( shdfac \)) to mute the thermal conductivity of the top layer of the soil (\( DF_1 \)) by:

\[ DF_1 = DF_1 \times e^{sbeta \times shdfac} \] (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. 7d). The preference for near-bound values is more pronounced in DV, which at site 7 tends to have \( shdfac \) values near 1.0, which puts downward pressure on the value of \( sbeta \). The skewed posterior parameter distribution for all suggests that an even-less-negative value of \( sbeta \) would 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
which 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 described by Eq.13.

5.2.3 The role of minimum stomatal resistance ($rcmin$)

$rcmin$ 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 LE.

At site 7, all three models prefer a low $rcmin$ (Fig. 7b), which increases LE for a given
potential evapotranspiration; however, GW and DV show decreased identifiability of $rcmin$. The
mode of the $rcmin$ distribution is higher for GW than for STD, perhaps because GW tends to
have a wetter soil and more robust LE. The spread of the posterior 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. 2]). Because DV simulations include a wider spread of $lai$ states, they also
have a wider spread of ‘good’ $rcmin$ values.

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

Figure 8 shows the bivariate posterior distribution of selected behavioral parameters of
STD and GW. The response surface of RMSE $SMC_{5cm}$ changes between STD and GW (e.g., see
$maxsmc$ vs. $psisat$). For GW, the shape of the posterior distributions of soil parameters that are
shared with STD is significantly different because of interaction with the GW parameters and module. Such shifts in model function affect the model covariance structure (Table 4).

After multi-objective parameter estimation, at site 7, GW functions in one of two preferred modes (Fig. 8b). In the first, slightly preferred 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 functions with parameters that render GW nonfunctional) (Fig. 8b). Relatively high values of the reciprocal of the e-folding depth of hydraulic conductivity (fff) effectively seal the bottom of the soil column, limiting communication between the aquifer and the soil column; high porosity (maxsmc) decreases the vertical conductivity, further inhibiting the already poor communication between the soil and aquifer. High porosity favors decreased direct evaporation. Surface runoff is augmented by a relatively high surface runoff scaling factor fsatmx; subsurface runoff is lessened by the relatively low subsurface runoff scaling factor (rsbmx).

These alternative behaviors are a possible explanation for the issue identified by Rosero et al. (2009), who showed that despite very good performance of calibrated GW, the model
suffered from low robustness (high sensitivity to unmeasurable, errant parameters). The bimodal behavior of GW highlights the need to constrain GW with runoff observations, which will help identify a more consistent, better model.

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

DV and STD differ functionally in two ways: 1) DV predicts (rather than prescribes) shdfac as a function of environmental variation in moisture and radiation availability, and 2) rather than use a temporally constant, prescribed lai, DV uses shdfac to predict lai variation using a functional relationship:

\[ lai = \max \left( x_{lai_{\text{min}}}, \frac{-1}{gl} \log(1 - \text{shdfac}) \right) \] (14)

The vegetated fraction affects all three components of LE: 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}). Because of shdfac’s role in scaling all components of evapotranspiration, a high value of conversion parameter gl 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}. When there is little vegetation (e.g., at sites 1-3), the coupling to DV may be failing to consider special water use features associated with the semi-arid vegetation (Unland et al., 1996).

STD keeps shdfac fixed at monthly climatological values (~0.7 at site 7) and does not relate shdfac to lai. 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 $E_{\text{transp}}$) 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. 9). The identifiability in the response surface of RMSE LE has changed (e.g. $lai$ vs. $rcmin$). 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. 9). We see additional evidence for increased interaction between parameters in DV when we note that the models’ covariance structure has been altered (Table 5). For example, $rcmin$ and $maxsme$ 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 time scales 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 may need to be constrained with observations of carbon fluxes and plant growth. The function of the DV module may be hindered by Noah’s lack of a separate canopy layer (Rosero et al., 2009) or the absence of a more complex Ball-Berry type of stomatal conductance formulation (Niu et al., 2009).

6. What does sensitivity analysis tell us about transferability?
LSM parameters are frequently assumed to be physical quantities, which can be measured and which have strong relationships with physical properties of the system. These assumptions imply that optimal values should not vary between models and that relationships between model parameters should be constant between models and between sites with similar physical properties. By making sets of vegetation-related parameters functions of vegetation type, LSMs contain the implicit assumption that vegetation type solely determines the ideal values of vegetation parameters. A priori assignment of parameter values based on a site’s physical characteristics (‘parameter transferability’) depends on the above assumption. Our VSA shows that interaction between parameters contributes most to a LSM’s total variance causing optimal parameter values to change significantly between models and sites. If parameters are transferable between similar sites, then the relationships between parameters should also be transferable between similar sites. The joint multivariate posterior distribution summarizes much of the information regarding the relationships between model parameters (model structure) at a particular location given observed datasets. We use the stable posterior distributions of the behavioral parameter sets to test the assumption that parameters and parameter relationships relate directly to physical characteristics. We also evaluate the extent to which climate determines the similarity of parameters between locations.

6.1 Testing parameter transferability between sites using soil and vegetation types

If parameters were readily ‘transferable’ between sites solely based on the vegetation class, 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). Using evidence from the marginal posterior parameter distributions of
behavioral, sensitive, ‘physically meaningful’ vegetation parameters (rcmin, lai, and sbeta),

Figure 10 shows that this expectation does not hold. Figure 10 contrasts the vegetation
parameters’ posterior distributions for sites 2 and 8 (grassland) against the posterior distributions
for contiguous sites 1 and 2 (dry). The distributions of rcmin, rmax, lai and z0 are more similar
between sites with similar climate than they are between sites with the same vegetation (Fig. 10a,
10b). 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. 10c). Parameter cfactr does show transferability, but only for STD.
rgl could be considered ‘transferable,’ but only for DV and GW.

Using the IHOP dataset, we cannot test parameter transferability using two sites with the
same soil type but different climatology. We instead tested whether the distributions of soil
parameters transfer along with soil or vegetation characteristics between these sites. Only fxexp
appears ‘transferable’ between dry sites (1 and 2) with the same soil type (sandy clay loam) in
STD and GW (Fig. 11b). Some parameters have more similar distributions between sites 2 and 8
than between sites 1 and 2: psisat and satdk only for GW, Clapp and Hornberger’s b only for
STD, maxsmc for STD and GW (Fig. 11a), and czil for STD and DV (Fig. 11a). Perhaps this is
because, despite sharing the same soil type, site 1 received more than twice the amount of
precipitation observed at site 2. We also hypothesize that transferability based on soil or
vegetation type is not uniformly viable because of the interactions between soil and vegetation
parameters, which condition the shape of the distributions. The interactions between the soil and
vegetation parameters at grassland site 2 are significantly different than the interactions at
unvegetated site 1.
We recognize that 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 land-surface model parameters are more sensitive to climatic forcing than to a specific soil or land-cover classification. Such results provide support to similar observations made about other hydrologic models (Demaria et al., 2007; van Werkhoven et al., 2008) and made about the Noah LSM and using single optimal parameter sets (Hogue et al., 2005; Rosero and Bastidas, 2007).

6.2 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 have obtained a stable, multivariate probability distribution $\chi$ of behavioral parameter sets $X=\{x_1, x_2, \ldots, x_i, \ldots x_k\}$ using multi-objective parameter estimation. The marginal probability distribution for the $i^{th}$ parameter is $\chi_i$. To circumvent comparing between sites against each other (two at a time, as done above), 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 between the cumulative distribution of the ‘optimal’ values of $x_i$ (i.e, the marginal $\chi_i$) 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:
\[ S = \begin{bmatrix}
\chi_{11} - D_{11} & \cdots & \chi_{k1} - D_{k1} \\
\vdots & \ddots & \vdots \\
\chi_{k1} - D_{k1} & \cdots & \chi_{kn} - D_{kn}
\end{bmatrix} \] (15)

\[
S \text{ can be used to identify groups of parameters that are similar between locations or to identify}
\]
locations where groups of parameters behave alike. We use the unsupervised, agglomerative
hierarchical clustering algorithm (described in Section 2.8) to find these groups without making
any further assumptions about the number of groups.

The Noah LSM works with three classes of parameters: parameters associated with soil
type (soil parameters), parameters associated with vegetation cover type (vegetation parameters),
and general parameters. 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 1). Similarly, clustering of \( S(x_{\text{soil},1..n}) \) would result in sites
grouped according to the IHOP soil type classification (Table 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 12a 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 \( S(x_{\text{veg},1..n}) \) by location
in a way that matched the IHOP vegetation classifications. Given the subset \( S(x_{\text{soil},1..n}) \), 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. 12b). Using signature vectors for
STD, GW, and DV, some (but not all) of the distances related 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 together; sites 5 and 6 (wheat crops) are never classified together according to
vegetation (Fig 12a). Sites 1 and 2 (sandy clay loam) and sites 4 and 5 (loam) do not cluster
together using soil parameters (Fig 12b). These results support the earlier findings presented
here, which suggest that interaction between soil and vegetation parameters is significant, to the
point that it shapes the posterior parameter distributions. These results also suggest that soil and
vegetation type are not, by themselves, good physical characteristics by which to transfer
parameters.

Next, to account for the interdependence between soil and vegetation parameters, we
classified the entire matrix S(x\text{soil},x\text{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 12c 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 type and the same soil type.

If we instead cluster S looking for groups of parameters, the expectation is that x\text{soil} will
behave as a whole 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 13 shows representative groupings of the behavioral, marginal
posterior distributions of STD and GW parameters at all sites. Again, 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 relate
primarily to soil or vegetation, but to a combination of both. The new GW parameters do seem to
behave in a way that is similar to other soil parameters (Fig. 13b), which informs the estimation
of these parameters for distributed applications.

We conclude that the primary site-to-site control on the parameters identified as
‘vegetation’ or ‘soil’ parameters is not a site’s type of soil or vegetation used in isolation. This is
consistent with the notion that given that parameters used by LSMs must represent physical
processes across multiple scales, the parameters become “effective” rather than physically
derived quantities (Wagener and Gupta, 2005) and that interaction between classes of parameters
are very important. Our clustering analysis suggests that climate is a major control of site-to-site
variation in parameter values and supports recommendations that mean climate be considered
when transferring parameter values between sites (Liang and Guo, 2003; Demaria et al., 2007;
vanWerkhoven et al., 2008).

7. Summary and Conclusions

We combine two powerful sensitivity analysis methods, regionalized sensitivity analysis
(RSA) and global variance-based sensitivity analysis (VSA), to inform model identification and
development. We draw conclusions regarding 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 VSA complements RSA for the purposes of model development. We perform VSA and RSA on a single set of Monte Carlo runs. When used together, VSA and RSA help developers understand model behavior (i.e., how and when different modules interact, how a model varies in performance between climate regimes, how a model performs differently between sites, how different models perform at the same site, etc.). SA identifies parameters whose behavior merits further examination: VSA identifies parameters responsible for model uncertainty; RSA identifies modes in parameter distributions that yield distinctive behavior.

We show that the clear patterns of parameter importance identified by 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, and we show that the interactions can be traced using RSA. 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. We assert that to advance model development, VSA should be used in conjunction with RSA.

Transfer of parameters based solely on shared vegetation type or on shared soil type is not a viable method for a priori parameter assignment. The work presented here shows that vegetation type and soil type 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 or soil type 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 and soil type 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 type 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., 2009) 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.

Acknowledgements We thank Pedro Restrepo at OHD/NWS, Dave Gochis at NCAR and Ken Mitchell at NCEP for their insight. We appreciated suggestions by M. Bayani Cardenas, Charles S. Jackson and Yasir H. Kaheil. We acknowledge the International H₂O Project for the datasets. We benefited from the computational resources at the Texas Advanced Computing Center (TACC). This project was funded by the Graduate Fellowship of the Hydrology Training Program of the OHD/NWS, the NOAA grant no. NA07OAR4310216, NSF, and the Jackson School of Geosciences.
REFERENCES


42 of 67


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


Figure 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 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 2. (a) First-order Sobol' sensitivity indices for the parameters of STD, GW and DV at all sites. $S_i$ stands for 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. See Table 2 for 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. Same as Figure 2 but for LE.
Figure 4. Same as Figure 2 but for SMC_{3cm}.
Figure 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 6. Uncertainty ranges of the 150 ensemble members of behavioral runs of STD (black), GW (dark gray) and DV (light gray) at Site 7. (a) Average diurnal LE. (b) Hourly SMC_{5cm}. Plots show the interquartile range IQR (50% of the runs) in continuous lines and the 90% confidence bounds (5% to 95% quantile) in dashed lines. Observations are shown with symbols. The spread of LE by STD is larger than that of DV, GW. However, IQR shows a very similar LE envelope. The ensemble of GW shows wetter SMC_{5cm} than the rest. The IQR of STD and DV are very similar, but the 90% confidence of STD has lower spread than DV.
Figure 7. 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\textsubscript{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 8. 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 $\text{SMC}_{5\text{cm}}$ 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 9. 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 10. Comparison of the marginal posterior parameter distributions of selected, sensitive vegetation parameters: (a) rcmin, (b) lai, and (c) sbeta. The total difference between parameter distributions at sites with the same vegetation cover type (Site 2 and 8)—grassland—(continuous, bright lines) is not smaller than the difference of distributions of the same parameters between contiguous sites (Site 2 and 1) (dashed, dark lines), which share the same soil type (sandy clay loam).
Figure 11. Comparison of the marginal posterior parameter distributions of selected, sensitive soil parameters: (a) maxsmc, (b) fxexp, and (c) czil. The total difference between parameter distributions at contiguous sites with the same soil type (Site 1 and 2) –sandy clay loam– (dashed, dark lines) is only sometimes smaller than the difference of distributions of the same parameters between sites with different meteorology (Site 2 and 8) (continuous, bright lines), which share the same vegetation cover type (grassland).
Figure 12. Clustering of sites using: (a) only the vegetation parameters of STD, (b) only the soil parameters of GW, and (c) 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 13. Clustering of soil (black), vegetation (gray) and GW-only parameters for the behavioral, marginal posterior distributions of (a) STD and (b) GW at all sites. 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.
Table 1. Average meteorology, near-surface states and turbulent fluxes observed during the calibration period (13 May - 25 Jun) at the nine IHOP_2002 sites. Indices of vegetation 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), sensible (H), latent (LE) and ground (G) heat flux, ground temperature (Tg) and soil moisture content at 5-cm (SMC$_{5cm}$).

<table>
<thead>
<tr>
<th>Site</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lat (°N)</td>
<td>36.4728</td>
<td>36.6221</td>
<td>36.8610</td>
<td>37.3579</td>
<td>37.3781</td>
<td>37.3545</td>
<td>37.3132</td>
<td>37.4070</td>
<td>37.4103</td>
</tr>
<tr>
<td>Lon (°W)</td>
<td>100.6179</td>
<td>100.6270</td>
<td>100.5945</td>
<td>98.2447</td>
<td>98.1636</td>
<td>97.6533</td>
<td>96.9387</td>
<td>96.7656</td>
<td>96.5671</td>
</tr>
<tr>
<td>Vegetation type</td>
<td>bare ground (1)</td>
<td>grassland (7)</td>
<td>sagebrush (9)</td>
<td>pasture (7)</td>
<td>wheat (12)</td>
<td>wheat (12)</td>
<td>pasture (7)</td>
<td>grassland pasture (7)</td>
<td></td>
</tr>
<tr>
<td>Soil type</td>
<td>sandy clay loam (7)</td>
<td>sandy clay loam (7)</td>
<td>sandy clay loam (4)</td>
<td>loam (8)</td>
<td>loam (8)</td>
<td>clay loam (6)</td>
<td>silty clay loam (2)</td>
<td>silty clay loam (2)</td>
<td></td>
</tr>
<tr>
<td>Rain (mm)</td>
<td>154.5</td>
<td>69.1</td>
<td>72.4</td>
<td>164.5</td>
<td>173.6</td>
<td>203.6</td>
<td>175.4</td>
<td>296.6</td>
<td>250.8</td>
</tr>
<tr>
<td>MAP (mm)</td>
<td>530</td>
<td>540</td>
<td>560</td>
<td>740</td>
<td>750</td>
<td>800</td>
<td>900</td>
<td>880</td>
<td>900</td>
</tr>
<tr>
<td>Ta (°C)</td>
<td>21.4</td>
<td>21.7</td>
<td>22.5</td>
<td>20.7</td>
<td>20.7</td>
<td>21.0</td>
<td>20.7</td>
<td>20.1</td>
<td>19.9</td>
</tr>
<tr>
<td>H (Wm$^{-2}$)</td>
<td>70.5</td>
<td>70.7</td>
<td>75.7</td>
<td>43.9</td>
<td>51.9</td>
<td>61.4</td>
<td>25.9</td>
<td>17.1</td>
<td>27.9</td>
</tr>
<tr>
<td>LE (Wm$^{-2}$)</td>
<td>65.1</td>
<td>76.1</td>
<td>68.2</td>
<td>106.2</td>
<td>111.2</td>
<td>97.1</td>
<td>126.4</td>
<td>122.8</td>
<td>115.3</td>
</tr>
<tr>
<td>G (Wm$^{-2}$)</td>
<td>-10.4</td>
<td>-6.4</td>
<td>-9.3</td>
<td>-2.7</td>
<td>-5.1</td>
<td>-7.5</td>
<td>-5.6</td>
<td>-12.1</td>
<td>-10.5</td>
</tr>
<tr>
<td>Tg (°C)</td>
<td>24.1</td>
<td>24.1</td>
<td>25.8</td>
<td>23.2</td>
<td>21.9</td>
<td>22.9</td>
<td>22.3</td>
<td>22.4</td>
<td>22.7</td>
</tr>
<tr>
<td>SMC$_{5cm}$ (%)</td>
<td>15.4</td>
<td>18.0</td>
<td>7.0</td>
<td>18.0</td>
<td>18.1</td>
<td>19.0</td>
<td>33.2</td>
<td>32.8</td>
<td>34.0</td>
</tr>
</tbody>
</table>
Table 2. Feasible ranges of Noah-LSM parameters considered in the sensitivity analysis.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>units</th>
<th>min</th>
<th>max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soil parameters</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>maxsmc</td>
<td>Maximum volumetric soil moisture</td>
<td>m$^3$m$^{-3}$</td>
<td>0.35</td>
<td>0.55</td>
</tr>
<tr>
<td>psisat</td>
<td>Saturated soil matric potential</td>
<td>m m$^{-1}$</td>
<td>0.1</td>
<td>0.65</td>
</tr>
<tr>
<td>satdk</td>
<td>Saturated soil hydraulic conductivity</td>
<td>m s$^{-1}$</td>
<td>1E-6</td>
<td>1E-5</td>
</tr>
<tr>
<td>b</td>
<td>Clapp-Hornberger b parameter</td>
<td></td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>quartz</td>
<td>Quartz content</td>
<td></td>
<td>0.1</td>
<td>0.82</td>
</tr>
<tr>
<td>refkdt</td>
<td>Used with refkd to compute runoff parameter kdt</td>
<td></td>
<td>0.05</td>
<td>3</td>
</tr>
<tr>
<td>fxexp</td>
<td>Bare soil evaporation exponent</td>
<td></td>
<td>0.2</td>
<td>4</td>
</tr>
<tr>
<td>refkd</td>
<td>Surface runoff parameter</td>
<td></td>
<td>0.1</td>
<td>10</td>
</tr>
<tr>
<td>czil</td>
<td>Zilintikevich parameter</td>
<td></td>
<td>0.05</td>
<td>8</td>
</tr>
<tr>
<td>csoil</td>
<td>Soil heat capacity</td>
<td>Jm$^3$K$^{-1}$</td>
<td>1.26</td>
<td>3.5</td>
</tr>
<tr>
<td>Vegetation parameters</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rcmin</td>
<td>Minimal stomatal resistance</td>
<td>s m$^{-1}$</td>
<td>40</td>
<td>400</td>
</tr>
<tr>
<td>rgl</td>
<td>Radiation stress parameter used in F1 term of canopy resistance</td>
<td>30</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>hs</td>
<td>Coefficient of vapor pressure deficit term F2 in canopy resistance</td>
<td>36</td>
<td>47</td>
<td></td>
</tr>
<tr>
<td>z0</td>
<td>Roughness length</td>
<td>m</td>
<td>0.01</td>
<td>0.1</td>
</tr>
<tr>
<td>lai</td>
<td>Leaf area index</td>
<td></td>
<td>0.1</td>
<td>5</td>
</tr>
<tr>
<td>cfactr</td>
<td>Exponent in canopy water evaporation function</td>
<td></td>
<td>0.4</td>
<td>0.95</td>
</tr>
<tr>
<td>cmcmax</td>
<td>Maximum canopy water capacity used in canopy evaporation</td>
<td>m</td>
<td>0.1</td>
<td>2.0</td>
</tr>
<tr>
<td>sbeta</td>
<td>Used to compute canopy effect on ground heat flux</td>
<td></td>
<td>-4</td>
<td>-1</td>
</tr>
<tr>
<td>rmax</td>
<td>Maximum stomatal resistance</td>
<td>s m$^{-1}$</td>
<td>2,000</td>
<td>10,000</td>
</tr>
<tr>
<td>topt</td>
<td>Optimum air temperature for transpiration</td>
<td>K</td>
<td>293</td>
<td>303</td>
</tr>
<tr>
<td>Dynamic Phenology parameters (Noah-DV only)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fragr</td>
<td>Fraction of carbon into growth respiration</td>
<td></td>
<td>0.1</td>
<td>0.5</td>
</tr>
<tr>
<td>gl</td>
<td>Conversion between greenness fraction and LAI</td>
<td></td>
<td>0.1</td>
<td>1.0</td>
</tr>
<tr>
<td>rssoil</td>
<td>Soil respiration coefficient</td>
<td>s$^{-1}$ x1E-6</td>
<td>0.005</td>
<td>0.5</td>
</tr>
<tr>
<td>tauhf</td>
<td>Average inverse optical depth for 1/e decay of light</td>
<td></td>
<td>0.1</td>
<td>0.4</td>
</tr>
<tr>
<td>bf</td>
<td>Parameter for present wood allocation</td>
<td></td>
<td>0.4</td>
<td>1.3</td>
</tr>
<tr>
<td>wstrec</td>
<td>Water stress parameter</td>
<td></td>
<td>10</td>
<td>400</td>
</tr>
<tr>
<td>xlaimin</td>
<td>Minimum leaf area index</td>
<td></td>
<td>0.05</td>
<td>0.5</td>
</tr>
<tr>
<td>sla</td>
<td>Specific leaf area</td>
<td></td>
<td>5</td>
<td>70</td>
</tr>
<tr>
<td>Groundwater parameters (Noah-GW only)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rous</td>
<td>Specific yield</td>
<td>m$^3$m$^{-3}$</td>
<td>0.01</td>
<td>0.5</td>
</tr>
<tr>
<td>fff</td>
<td>e-folding depth of saturated hydraulic capacity</td>
<td>m$^{-1}$</td>
<td>0.5</td>
<td>10</td>
</tr>
<tr>
<td>fsatmx</td>
<td>Maximum saturated fraction</td>
<td>%</td>
<td>0</td>
<td>90</td>
</tr>
<tr>
<td>rsbmx</td>
<td>Maximum rate of subsurface runoff</td>
<td>ms$^{-1}$</td>
<td>1E-3</td>
<td>0.01</td>
</tr>
</tbody>
</table>
Table 3. Sensitive parameters according to a Kolmogorov-Smirnov test between samples that drive behavioral (multi-objectively calibrated) and non-behavioral simulations. 1 stands for sensitive, 0 for insensitive. The number of sensitive parameter is tabulated by class (soil, vegetation, and new GW or DV). See Table 1 for parameter names. No clear regional pattern of sensitivity can be readily discerned.

<table>
<thead>
<tr>
<th>Site No.</th>
<th>Par. 1</th>
<th>Par. 2</th>
<th>Par. 3</th>
<th>Par. 4</th>
<th>Par. 5</th>
<th>Par. 6</th>
<th>Par. 7</th>
<th>Par. 8</th>
<th>Par. 9</th>
<th>Par. 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Total 16 14 20 12 23 13 12 25 16 21 25 13 20 24 11 22 25 16 13 19 13 16 18 15 10 22
Table 4. Spearman rank correlation coefficients between parameter sets belonging to the behavioral set for STD (up the diagonal) and GW (below the diagonal). Note the change in the covariance structure in Fig. 8. See Table 1 for abbreviations of parameter names.

<table>
<thead>
<tr>
<th>GW</th>
<th>maxsmc</th>
<th>psisat</th>
<th>satdk</th>
<th>fxexp</th>
<th>rous</th>
<th>fff</th>
<th>fsatmx</th>
</tr>
</thead>
<tbody>
<tr>
<td>maxsmc</td>
<td></td>
<td>-0.10</td>
<td>-0.40</td>
<td>0.29</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>psisat</td>
<td>-0.33</td>
<td></td>
<td>-0.14</td>
<td>-0.32</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>satdk</td>
<td>-0.09</td>
<td>0.49</td>
<td></td>
<td>0.22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fxexp</td>
<td>-0.26</td>
<td>0.41</td>
<td>0.23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rous</td>
<td>-0.01</td>
<td>0.26</td>
<td>0.24</td>
<td>0.14</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fff</td>
<td>0.11</td>
<td>-0.46</td>
<td>-0.45</td>
<td>-0.49</td>
<td>-0.37</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fsatmx</td>
<td>-0.22</td>
<td>-0.04</td>
<td>-0.17</td>
<td>0.09</td>
<td>-0.37</td>
<td>0.17</td>
<td></td>
</tr>
<tr>
<td>rsbmx</td>
<td>0.11</td>
<td>-0.25</td>
<td>-0.13</td>
<td>-0.21</td>
<td>0.32</td>
<td>0.08</td>
<td>-0.24</td>
</tr>
</tbody>
</table>
Table 5. Spearman rank correlation coefficients between parameter sets belonging to the behavioral set for STD (up the diagonal) and DV (below the diagonal). Note the change in the covariance structure in Fig 9. See Table 1 for abbreviations of parameter names.

<table>
<thead>
<tr>
<th></th>
<th>STD</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>rcmin</td>
<td>hs</td>
<td>maxsmc</td>
<td>psisat</td>
<td>fragr</td>
<td>bf</td>
</tr>
<tr>
<td>rcmin</td>
<td>-0.35</td>
<td>0.44</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hs</td>
<td>0.30</td>
<td>-0.15</td>
<td>-0.36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>maxsmc</td>
<td>-0.16</td>
<td>-0.29</td>
<td>-0.10</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>psisat</td>
<td>0.50</td>
<td>0.36</td>
<td>-0.21</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fragr</td>
<td>0.58</td>
<td>0.24</td>
<td>-0.02</td>
<td>0.10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bf</td>
<td>0.61</td>
<td>0.30</td>
<td>-0.19</td>
<td>0.59</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>xclaimin</td>
<td>-0.72</td>
<td>-0.31</td>
<td>0.10</td>
<td>-0.31</td>
<td>-0.62</td>
<td>-0.54</td>
</tr>
<tr>
<td>sla</td>
<td>0.80</td>
<td>0.21</td>
<td>-0.15</td>
<td>0.35</td>
<td>0.66</td>
<td>0.45</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

FIG 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 sited in representative land covers (see Table 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.

FIG 2. (a) First-order Sobol' sensitivity indices for the parameters of STD, GW and DV at all sites. $S_i$ stands for 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. 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.

FIG 3. Same as Figure 2 but for LE.

FIG 4. Same as Figure 2 but for SMC$_{5cm}$.

FIG 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).

FIG 6. Uncertainty ranges of the 150 ensemble members of behavioral runs of STD (black), GW (dark gray) and DV (light gray) at Site 7. (a) Average diurnal LE. (b) Hourly SMC$_{5cm}$. Plots show the interquartile range (50% of the runs) in continuous lines and the 90% confidence bounds (5% to 95% quantile) in dashed lines. Observations are shown with symbols. The spread of LE by STD is larger than that of DV, GW. The ensemble of GW shows wetter SMC$_{5cm}$ than the rest.

FIG 7. 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 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.

FIG 8. 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 $m_1$ and $m_2$. See text for explanation.

FIG 9. 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.

FIG 10. Comparison of the marginal posterior parameter distributions of selected, sensitive vegetation parameters: (a) rcmin, (b) lai, and (c) sbeta. The total difference between parameter distributions at sites with the same vegetation cover type (Site 2 and 8) (continuous, bright lines) is not smaller than the difference of distributions of the same parameters between contiguous sites (Site 2 and 1) (dashed, dark lines), which share the same soil type.

FIG 11. Comparison of the marginal posterior parameter distributions of selected, sensitive soil parameters: (a) maxsmc, (b) fxexp, and (c) czil. The total difference between parameter distributions at sites with the same soil type (Site 5 and 7) (continuous, bright lines) is in general not smaller than the difference of distributions of the same parameters between contiguous sites (Site 5 and 6) (dashed, dark lines), which share the same vegetation cover type.

FIG 12. Clustering of sites using: (a) only the vegetation parameters of STD, (b) only the soil parameters of GW, and (c) 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.

FIG 13. Clustering of soil (black), vegetation (gray) and GW parameters for the behavioral, marginal posterior distributions of (a) STD and (b) GW at all sites. 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.
LIST OF TABLES

TABLE 1. Average meteorology, near-surface states and turbulent fluxes observed during the calibration period (13 May - 25 Jun) at the nine IHOP_2002 sites. Indices of vegetation 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), sensible (H), latent (LE) and ground (G) heat flux, ground temperature (Tg) and soil moisture content at 5-cm (SMC5cm).

TABLE 2. Feasible ranges of Noah-LSM parameters considered in the sensitivity analysis.

TABLE 3. Sensitive parameters according to a Kolmogorov-Smirnov test between samples that drive behavioral (multi-objectively calibrated) and non-behavioral simulations. 1 stands for sensitive, 0 for insensitive. The number of sensitive parameter is tabulated by class (soil, vegetation, and new GW or DV). No clear regional pattern of sensitivity can be readily discerned.

TABLE 4. Spearman rank correlation coefficients between parameter sets belonging to the behavioral set for STD (up the diagonal) and GW (below the diagonal). Note the change in the covariance structure in Fig. 8.

TABLE 5. Spearman rank correlation coefficients between parameter sets belonging to the behavioral set for STD (up the diagonal) and DV (below the diagonal). Note the change in the covariance structure in Fig. 9.