On the Applicability of Surrogate-based MCMC-Bayesian Inversion to the Community Land Model: Case Studies at Flux Tower Sites

Maoyi Huang¹*, Jaideep Ray², Zhangshuan Hou³, Huiying Ren³, Ying Liu¹, Laura Swiler⁴

¹Earth System Analysis and Modeling Group, Pacific Northwest National Laboratory, P.O. Box 999, Richland, WA 99352, USA
²Sandia National Laboratories, P.O. Box 969, Livermore, CA 94551
³Hydrology Technical Group, Pacific Northwest National Laboratory, P.O. Box 999, Richland, WA 99352, USA
⁴Sandia National Laboratories, P.O. Box 5800, Albuquerque, NM 87123

*Corresponding author Tel.: +1 509 375-6827. E-mail address: maoyi.huang@pnnl.gov

To be considered for publication in Journal of Geophysical Research - Atmosphere.

12 June 2016
Key Points:

1. The feasibility of applying a Bayesian calibration technique to estimate CLM parameters is assessed;

2. CLM-simulated LH fluxes using the calibrated parameters are generally improved;

3. The parameter values are likely transferable within the plant functional type.
Abstract

The Community Land Model (CLM) has been widely used in climate and Earth system modeling. Accurate estimation of model parameters is needed for reliable model simulations and predictions under current and future conditions, respectively. In our previous work, a subset of hydrological parameters has been identified to have significant impact on surface energy fluxes at selected flux tower sites based on parameter screening and sensitivity analysis, which indicate that the parameters could potentially be estimated from surface flux observations at the towers. To date, such estimates do not exist.

In this paper, we assess the feasibility of applying a Bayesian model calibration technique to estimate CLM parameters at selected flux tower sites under various site conditions. The parameters are estimated as a joint probability density function (PDF) that provides estimates of uncertainty of the parameters being inverted, conditional on climatologically-average latent heat fluxes derived from observations. We find that the simulated mean latent heat fluxes from CLM using the calibrated parameters are generally improved at all sites when compared to those obtained with CLM simulations using default parameter sets. Further, our calibration method also results in credibility bounds around the simulated mean fluxes which bracket the measured data. The modes (or maximum a posteriori values) and 95% credibility intervals of the site-specific posterior PDFs are tabulated as suggested parameter values for each site. Analysis of relationships between the posterior PDFs and site conditions suggests that the parameter values are likely correlated with the plant functional type, which needs to be confirmed in future studies by extending the approach to more sites.

Keywords: Community Land Model; MCMC-Bayesian; Surrogate; Flux tower
1. Introduction

Land surface models (LSMs) are a critical component in Earth system models. Among essential LSM outputs are the heat fluxes, which drive important physical processes such as boundary layer processes, cloud formation, and precipitation (e.g., Qian et al., 2013). The inputs of an LSM include meteorological conditions/forcing, boundary conditions, and parameters introduced in various modules. These inputs, however, are all subject to certain levels of uncertainty, which are associated with data, model structure, and lack of knowledge about the model parameters.

Tremendous efforts have been made to evaluate and/or compare performances of various LSMs [Bastidas et al., 2006; Henderson-Sellers et al., 1995]. However, many LSM parameters are uncertain, and the default assignment of parameter values may be inappropriate (e.g., [Bastidas et al., 2006; Hou et al., 2012; Huang et al., 2013; Rosero et al., 2010]). Without calibration, better conceptual models do not warrant better match between model simulations and observations for variables of interest. Recently, as LSMs have become increasingly complex, their dimensionality (in terms of the parameter space) has increased dramatically and inverse problems that seek to estimate parameters have become very ill-posed. Therefore, dimensionality reduction is a pre-requisite for parameter estimation. In order to quantify the uncertainty in the model predictions, it is reasonable to adopt stochastic inversion (e.g., Bayesian) approaches rather than deterministic (e.g., least-square fitting). However, depending on the nonlinearity, non-uniqueness, and complexity of the inverse problem, these stochastic approaches could involve a large number of model simulations that are potentially computationally impractical.

Parametric dimensionality can be reduced via sensitivity analysis methods (Morris One at a Time, variance-based decomposition using Sobol’ indices, etc.) using ensembles of simulations
Williams et al. [2009] discussed how the FLUXNET database can be used to improve forecasts of global biogeochemical and climate models. Some sensitivity studies were performed at the flux tower sites [Alton et al., 2006; Baldocchi and Wilson, 2001; White et al., 2000; Zobitz et al., 2006], focusing on evaluating net primary production controls, biophysical parameters governing light propagation, canopy photosynthesis, and carbon cycling. In the past few years, a number of studies have documented the sensitivity of surface fluxes to model parameters in the Community Land Model (CLM). Göhler et al. [2013] analyzed the sensitivity of latent heat (LH), sensible heat (SH) and photosynthesis of the Community Land Model CLM version 3.5 to its parameters. They found that photosynthesis is very sensitive to parameters associated with plant functional types, whereas LH is sensitive to soil water parameters. Bonan et al. [2011] investigated sensitivity of LH to photosynthetic parameters in CLM version 4.0 (CLM4) and suggested that model structural errors in the model could be compensated by parameter adjustment. Hou et al. [2012] performed sensitivity analyses at representative flux tower sites on outputs of heat fluxes from CLM4 driven by satellite phenology and showed that they are most sensitive to a subset of hydrological parameters. This raised the possibility that the parameters could be estimated from measurements of heat fluxes, to improve the predictive skill of CLM. The results also show that only a selected set of parameters, and not all hydrological parameters, could be correctly estimated from the fluxes. More recently, Huang et al. [2013]; Ren et al. [2015] extended the sensitivity analysis framework in Hou et al. [2012] to 431 relatively pristine watersheds over the contiguous United States within minimal human perturbations. They confirmed the findings in Hou et al. [2012] that surface energy fluxes (i.e., latent and sensible
heat) and surface and subsurface runoff in CLM4 are highly sensitive to subsets of hydrological parameters depending on their hydrologic attributes, therefore the parameter values and/or inversion procedure is potentially transferrable among watersheds in similar hydrologic regimes.

Computational demand, though, still remains a great challenge, even when parameter dimensionality is reduced. In order to address this problem, surrogate models can be used as alternatives to the numerical simulators. Ensemble simulations, which are required to develop surrogate models, can be performed efficiently in a task-parallel manner on supercomputing facilities. Depending on the complexity/nonlinearity of the link between unknown input parameters and model outputs, the applicability (accuracy and consistency) of the developed surrogates need to be evaluated before they can be used in sensitivity analysis or calibration.

We define model calibration as the process of inferring uncertain/unknown model inputs (model parameters) from experimental or field observations. It is traditionally posed as a model-fitting problem, and the variables being calibrated (“calibration parameters”) are optimized to reproduce the observational data. Due to limitations of the observations and/or shortcomings of the model itself, it may be possible to infer the calibration parameters only with a large degree of uncertainty. In such a case, model outputs cannot be considered robust or predictive unless the uncertainties in the inputs are estimated and incorporated into model predictions. Inference of model parameters/inputs, along with their uncertainties, can be performed by posing a statistical or Bayesian inverse problem [Kaipio and Somersalo, 2006]. When solved using a Markov chain Monte Carlo (MCMC) method [Gilks et al., 1996], Bayesian inverse problem yield the calibration parameters in the form of a joint probability density function (PDF). The PDF succinctly captures the uncertainty in the estimates. MCMC methods construct the PDF in the
form of samples, each of which requires the forward model (e.g., CLM) to be run at least once. If the model is computationally expensive, e.g., if it is a high fidelity model (HFM), it has to be replaced by a fast-running proxy called a surrogate so that the inverse problem may be solved.

In this study, we define a surrogate model as a response surface model i.e., a statistical “curve-fit” that relates the HFM model output of interest to the model inputs/parameters being varied. Surrogates are constructed by fitting a functional form to a training data corpus created by sampling the HFM model input/parameter space and obtaining the HFM’s response at those input combinations. The approximation inherent in surrogate models implies that the predictive skill of the parameters estimated using them has to be checked with the original HFM.

Surrogates have long been used in engineering and in the geosciences, for example in water resources research. Viana et al. [2014], Forrester and Keane [2009] and Razavi et al. [2012] are three recent review articles that describe those applications. The earliest surrogates were polynomials or approximate neural networks fitted to data. Later, multivariate adaptive regression splines [Friedman, 1991], Gaussian process (or kriging) and kernel methods such as radial basis functions (RBF) [Regis and Shoemaker, 2007] were used to represent the HFM’s response surface. Some methods, such as Gaussian process models, can also provide an estimate of the error in the surrogate model’s prediction. Of the latter, there has been a move towards mixtures of surrogates [Goel et al., 2007]. Various types of “multifidelity surrogates”, constructed from a training set of HFM and low fidelity model (LFM) runs, have also been investigated [Eldred and Dunlavy, 2006; Pau et al., 2014; Viana et al., 2014]. HFM responses display extreme nonlinearity when the model inputs are non-physical or infeasible.
Fitting surrogates in such cases is difficult and it may be worthwhile to excise the infeasible (or “nonsense”) regions of the parameter space [Giunta et al., 1995]. Thereafter, care has to be taken to ensure that the surrogate is never evaluated in the “nonsense” region, e.g., by using a classifier. Such classifier-response surface composites have been used in the surrogate modeling of CLM 4.0 [Sargsyan et al., 2014]. Another approach is to limit the surrogate to a “trust-region”, a small region in which local perturbations to the parameters are still valid [Alexandrov et al., 1998]. A variation of this is an Adaptive Response Surface Method, where portions of the input space that correspond to large objective function values are discarded at each iteration, gradually reducing the input space to the neighborhood of the global optimum [Wang et al., 2001].

Probabilistic methods, based on Monte Carlo simulations, have been used to calibrate LSMs. Lo et al. [2010] used Monte Carlo techniques to estimate hydrological parameters of Community Land Model (CLM) 3.0, while Prihodko et al. [2008] calibrated Simple Biosphere Model version 2.5. Sun et al. [2013] performed a MCMC calibration of 10 parameters in CLM version 4.0 without using surrogates. Järvinen et al. [2010]; Solonen et al. [2012] used multi-chain MCMC methods to address the formidable computational cost of calibrating the parameters of a climate model, while Zeng et al. [2013] used the same approach to calibrate the parameters of a crop module in CLM version 3.5. Bilionis et al. [2015] used a sequential Monte Carlo method to calibrate 10 parameters of the Crop module in CLM4.5. Tian and Xie [2008] used an unscented Kalman filter to calibrate CLM 2.0.

The use of surrogates in the calibration of climate models or LSMs is less common. In Müller et al. [2015], the authors used an RBF to create a surrogate of the data – model mismatch
(not the HFM output) and estimated 11 parameters of the CLM4.5’s methane module using a global optimization method called DYnamic COordinate search using Response Surface models (DYCORS) [Regis and Shoemaker, 2007]. Sargsyan et al. [2014] attempted to construct surrogates for five variables of interests from CLM4 with prognostic carbon and nitrogen modules turned on (i.e., CLM4-CN) using Bayesian compressive sensing (BCS) in combination with polynomial chaos expansions (PCEs). They found that the input-output relationship in CLM4-CN could be composed of qualitatively different regimes (i.e., live or dead vegetation regimes associated with different regions in the parameter space), so that clustering- and classification-based piecewise PCE construction is needed. In Ray et al. [2015], the authors used polynomial and universal kriging surrogates to calibrate three hydrological parameters of CLM 4.0 using measurements of latent heat fluxes. Two competing models were used for the model–data mismatch to estimate a composite of measurement error and (a crude estimate of) the structural error of CLM. In Gong et al. [2015], the authors used adaptive surrogate-based optimization to perform parameter estimation of the Common Land Model using six observables jointly; 12 independent parameters were (deterministically) calibrated.

In this study we combine the advances in surrogate modeling described above with a Bayesian model calibration framework as presented in Ray et al. [2015] to perform calibration of CLM 4.0 at 12 selected flux tower sites using latent heat (LH) flux measurements. In Section 2, we formulate the parameter estimation problem and describe the (Bayesian) parameter estimation method. In Section 3, we estimate the joint PDFs of three hydrological parameters, tabulate their modes and their credibility intervals for all the sites and correlate them with the site characteristics. In Section 4, we discuss our results and draw our conclusions.
2. Methodology and site information

2.1 Review of the previous study

Hou et al. [2012] applied an uncertainty quantification (UQ) framework to analyze the sensitivity of simulated surface fluxes to selected hydrologic parameters in the CLM 4.0 (henceforth CLM4) driven by Satellite Phenology (SP). We note that by choosing the SP mode, the biogeochemical modules of CLM4 are not activated so that the model is used as a standard land surface model focusing on water and energy budget simulations. The sensitivity analysis was conducted at thirteen flux towers that span a wide range of climate and site conditions. In this study, 12 of the sites studied in Hou et al. [2012] will be the subject of model calibration (see Table 1). The US-NRI site is not included in this study because Hou et al. [2012] showed that the heat fluxes at the site are insensitive to the selected hydrological parameters.

Simulations corresponding to sampled parameter sets were used to generate response curves and surfaces and statistical tests were used to rank the significance of the parameters for output responses including latent heat (LH) and sensible heat (SH) fluxes. Overall, CLM4-simulated LH and SH show the largest sensitivity to subsurface runoff generation parameters. However, study sites with deep root vegetation are also affected by surface runoff parameters, while sites with shallow root zones are sensitive to the vadose zone soil water parameters. Generally, sites with finer soil texture and shallower rooting systems tend to have larger sensitivity of outputs to the parameters. Their study suggests the necessity and possibility of parameter inversion/calibration using available measurements of latent/sensible heat fluxes. In this study, we attempt to invert the sensitive parameters identified in Hou et al. [2012], by applying and refining the surrogate-
2.2 Posing the parameter estimation problem

CLM4 contains a large number of parameterizations of biogeophysical and biogeochemical processes [Lawrence et al., 2011]. It is used to simulate global scale water, energy, carbon dynamics as the land component in the Community Earth System Model (CESM). By default, parameters are set at values that reproduce benchmark datasets globally [Y Q Luo et al., 2012]. When CLM4 is used to simulate processes at a site, it is used in its “single point” mode and its parameters have to be recalibrated to represent the site being modeled. The data used for calibration is often limited, spanning a few years. Further, due to model approximations, CLM4 cannot reproduce observations perfectly, even if the “optimal” parameters were known; this shortcoming is called the structural error. Consequently, CLM4 parameters can be estimated only with a large degree of uncertainty. Quantification of parametric uncertainty becomes an integral part of the calibration and hence Bayesian calibration, using MCMC to estimate the PDF of the parameters, becomes necessary for robust model predictions.

Let $Y^{(obs)}$ be measurements of the latent heat flux (LH) over a duration $T$, i.e., it is a time-series. Let $M(p; x)$ be CLM4 predictions due to a parameter setting $p$, and with external forcing e.g. meteorology $x$. We impose the relation

$$Y^{(obs)} = M(p; x) + \epsilon, \epsilon \sim N(0, \Gamma), \quad \text{(Eq. 1)}$$

where $N(0, \Gamma)$ denotes a zero-mean multivariate normal distribution with covariance $\Gamma$. Neither the form nor the distribution of $\Gamma$ is known. Thus, the error model is a choice and can
significantly affect calibration results. The likelihood of observing a particular parameter combination $p$ is given by

$$\Lambda(Y^{\text{obs}}|p, \Gamma) \propto \frac{1}{|\Gamma|^{1/2}} \exp \left( -\left( Y^{\text{obs}} - M(p) \right)^T \Gamma^{-1} \left( Y^{\text{obs}} - M(p) \right) \right),$$

(Eq. 2)

where we have omitted $x$ for brevity. Then by Bayes theorem, the posterior distribution (calibrated joint PDF) is given by

$$P(p, \Gamma | Y^{\text{obs}}) \propto \Lambda(Y^{\text{obs}}|p, \Gamma) \pi_{\text{prior}}(p)$$

(Eq. 3)

where $\pi_{\text{prior}}(p)$ is our prior belief regarding the distribution of $p$. The posterior distribution is arbitrary and is realized by a set of samples $p^{(s)}$, $s = 1\ldots N_{\text{mcmc}}$ drawn from it by an MCMC sampler. As described in Ray et al. [2015], $O(10^4) - O(10^5)$ samples are required to reach a stationary joint PDF and given the spin-up requirement and computational cost of CLM4, surrogates are required to perform the calibration.

The periods with available data for each site are listed in Table 1, during which the meteorological forcing and fluxes (e.g., LH) are measured at hourly or half-hourly time step. In this study, we keep inputs and simulations procedure to be identical to that in Hou et al. [2012]. That is, for each site, meteorological forcing, site information such as soil texture, vegetation cover, and satellite-derived phenology, as well as observational data sets (e.g., LH), are provided by the North American Carbon Program (NACP) site synthesis team [Schwalm et al., 2010]. Meteorological forcing to drive CLM4, including air temperature, specific humidity, wind speed, precipitation, surface pressure, surface incident short-wave radiation and surface incident long-wave radiation were gap-filled by the NACP team using the same protocol. Ancillary data and information describing tower location, soil and vegetation characteristics are also provided and
used to parameterize CLM4. Leaf area indices from MODIS (MODerate-resolution Imaging Spectroradiometer) are retrieved from nine pixels surrounding the tower footprint and provided by the NACP team as the satellite phenology to drive the CLM4SP simulation. Measured fluxes of latent and sensible heat at the native time resolution of the observations (30- or 60-minute) are provided and aggregated to monthly time step for calibration in this study. The data were gap-filled following a standard protocol as well. Measurements were obtained using the eddy-covariance (EC) method as part of the Ameriflux network. It has been widely recognized that surface energy fluxes based on EC method are subject to energy closure problems [Wilson et al., 2002] but unfortunately measurement uncertainty bounds are not reported at part of the NACP site synthesis dataset and therefore are not addressed in this study. Rather, we treat the fluxes provided as the “truth” in this study. We note that addressing the energy closure problem or reporting errors from EC systems for modelers is out of the scope of the study but should be addressed as a community effort as part of the FLUXNET network. Interested readers are referred to Schwalm et al. [2010] on the NACP site synthesis dataset and the references listed in Table 1 for detailed descriptions on the sites.

CLM4 was spun up by cycling the provided forcing for at least five times until all state variables reached equilibrium. For the purpose of capturing first-order dynamics in the climate system, we focus on evaluating CLM4’s ability to simulate seasonal variability by deriving time series of latent heat flux at monthly time steps from the raw datasets, consistent with our previous studies [Hou et al., 2012; Ray et al., 2015]. As shown in Ray et al. [2015], climatologically averaging smoothens CLM4 predictions and allows the surrogate to be fitted with an acceptable degree of accuracy. A surrogate model is created for each of the 12 months. Parameter estimates obtained in a test case where surrogates could be created without climatological averaging were not
substantially affected when re-estimated with climatological averaging. However, the limited
nature of the climatologically-averaged time-series does not allow the use of complex models for
\( \Gamma \). Consequently, we model it as a constant diagonal matrix i.e., \( \Gamma = \sigma^2 I \), where \( I \) is the identity
matrix. This is a very limiting assumption; when data and surrogate models allow, we can use
more sophisticated representations for \( \Gamma \) as discussed in Ray et al. [2015]. Eq. 3 simplifies to

\[
P(\mathbf{p}, \sigma^2 | Y^{(obs)}) \propto \frac{1}{\sigma^M} \exp \left( - \frac{\| Y^{(obs)} - \mathcal{M}^{(s)}(\mathbf{p}) \|^2_2}{2\sigma^2} \right) \pi_{\text{prior}}(\mathbf{p}). \tag{Eq. 4}
\]

where \( M = 12 \) is the months of climatologically averaged data that we use in the calibration, \( \| . \|_2 \)
is the \( l^2 \) norm and \( \mathcal{M}^{(s)}(\mathbf{p}) \) is a composite of the monthly surrogates of CLM4. That is, \( \mathcal{M}^{(s)}(\mathbf{p}) \) is
a 12-component vector, with 12 surrogates constructed separately for each month’s
climatological average over the four years. The use of a surrogate is a necessity since each
CLM4 invocation takes \(~30\) minutes on a CPU. The prior on \( \sigma^2 \) is an inverse Gamma, which,
being a conjugate prior, allows us to sample \( \sigma^2 \) with a Gibbs sampler. We use the MCMC
implementation in the \( R \) package \textsc{FME} to sample \( (\mathbf{p}, \sigma^2) \) from the posterior density distribution.

\textsc{FME} implements the Delayed Rejection Adaptive Metropolis (DRAM, [Haario et al., 2006])
MCMC sampler. The convergence of the Markov chain of samples is tested using the Raftery-
Lewis method [Raftery and Lewis, 1995], as implemented in the \( R \) package \textsc{mcgibbsit}
[mcgibbsit].

### 2.3 Designing an informative prior

The vector \( \mathbf{p} = \{p_1, p_2, p_3\} \) resides in a cuboidal \((p_1, p_2, p_3)\) space (henceforth \( \mathcal{P} \)). The
hydrological parameters that constitute the first two elements of \( \mathbf{p} \), for each site, are \( F_{\text{dai}} \) and
ln(Q_{dm}). The third parameter is the Clapp-Hornberger parameter B \cite{Clapp1978} for US-ARM and US-Wlr, and specific yield S_y for the other study sites. Prior distributions for each of these parameters are discussed in Hou et al. \cite{Hou2012}. They are independent and usually (but not always) uniform distributions. However p sampled randomly from \mathcal{P} is not necessarily physically realistic, which causes complex (and non-physical) behaviors of LH predictions and makes surrogate modeling difficult. Further, the LH predictions generated bear little resemblance to Y^{(obs)} and the root-mean-square-error

\[ \text{RMSE}(p) = \left\| Y^{(obs)} - Y(p) \right\|_2 \] is large. We would like to avoid sampling non-physical parts of the parameter space with the MCMC method. Therefore, we re-define the hypercube encompassing all parameter values to a more informative prior. We outline our approach in the following paragraphs.

As described in Giunta et al. \cite{Giunta1995} and mentioned in Section 1, one may excise the inappropriate portions of \mathcal{P} to obtain \mathcal{R}, which contains physically realistic parameters. We do so in this study. We draw N (=282) samples from \mathcal{P} using a space-filling, quasi Monte Carlo method and use them to generate an ensemble of LH predictions. The reason we use 282 samples is described in Section 2.4 on surrogate models. RMSE(p) are calculated for each realization and we specify a threshold RMSE quartile \text{Q}_{\text{RMSE}} to identify p whose predictions are close to observations at a given site. The selected samples of p discretely define \mathcal{R}. We define an improper, informative prior \pi_{\text{prior}}(p) with support \mathcal{R} such that the prior density is one inside \mathcal{R} and zero outside. We construct our surrogates using only parameter combinations p that reside inside \mathcal{R}. Note that the use of a user-defined \text{Q}_{\text{RMSE}} makes \pi_{\text{prior}}(p) somewhat subjective and we
will investigate its effect below. Usually setting \( Q_{\text{RMSE}} = 0.7 \) has allowed us remove the non-
physical part of \( P \) and to construct accurate surrogates. Note that CLM4 surrogates that are valid
only in a portion of \( P \) has been documented in earlier studies as well [Sargsyan et al., 2014].

In order to use \( \pi_{\text{prior}}(p) \) within MCMC, we require a precise definition of \( R \) so that we may
unambiguously decide whether an arbitrary \( p \) resides within \( R \). The separation of the training set
of runs into valid (i.e., \( p \in R \)) and invalid (i.e., \( p \in R^*, R^* \in P \setminus R \)) ones is used to train a
classifier (similar to the approach in Sargsyan et al. [2014]). The problem is posed as follows:

We define a function \( \zeta(p) \)

\[
\zeta(p) = \begin{cases} 
+1, & p \in R \\
-1, & p \in R^* 
\end{cases}
\]

where the level set \( \zeta(p) = 0 \) defines \( \partial R \), the boundary of \( R \). All that remains is to approximate
the function \( \zeta(p) \) using the training set defined over \( P \).

The problem of approximation \( \zeta(p) \) can be cast as a classification problem – we seek the
separatrix \( \partial R \) that separates \( R \) from \( R^* \). Once established, it can decide whether an arbitrary \( p \)
lies within \( R \). The training set of 282 runs provides us with realizations of the binary function \( \zeta(p) \) which we use to construct a binary classifier using Support Vector Machines (SVM). The
points in the training data are first distributed into two opposing classes based on the binary
values of \( \zeta(p) \). The SVM identifies an optimal separating hyperplane i.e. \( \partial R \), in the three-
dimensional \( p \) space. \( \partial R \) Details of the theory of SVMs can be found in Hastie et al. [2009].

In this work we use the SVM implementation in the \( R \) package \texttt{e1071}[Meyer et al., 2014].
The training data were randomly split into a Learning Set (LS) consisting of 85% of the points
and a Testing Set (TS) with the remainder. The SVM was trained on the LS, with a five-fold cross-validation to find optimize the hyperparameters of the SVM classifier. The resulting SVM was tested on the TS and a misclassification rate is computed. In order to check the sensitivity of the SVM on the LS, the whole process was repeated 50 times with different LS/TS pairs. The average misclassification rate (over the 50 rounds of testing) was reported to be between 6% and 9%, depending upon the site. This is similar to the 15% misclassification rate achieved for the classifier associated with a CLM4 surrogate in Sargsyan et al. [2014]. At the end of the SVM training process, we have a classifier that can be used to determine if a point in the 3-D original hypercube is located in the informative prior. If so, this point is used in the MCMC process.

2.4. Checking the effect of $\mathcal{R}$

The procedure in section 2.3 aims to choose a subset $\mathcal{R}$ of the parameter space $\mathcal{P}$, based on a quantile $Q_{\text{RMSE}}$ of the difference between observations and existing training data, i.e., $\text{RMSE}(\mathbf{p}) = \| \mathbf{Y}^{(\text{obs})} - \mathbf{Y}(\mathbf{p}) \|_2$. $\mathcal{R}$ is chosen as a way of excluding the non-physical part of $\mathcal{P}$. Choosing $Q_{\text{RMSE}}$ is somewhat subjective. The limited amount of training data points can influence the RMSE values and the resulting quantiles. Ideally, we would like $\mathcal{R}$ to include the “true” optimal parameters $\mathbf{p}_{\text{opt}}$ that lead to the minimal value of RMSE. That is, if we use an optimization method to find the optimal parameters $\mathbf{p}_{\text{opt}}$, these are the parameters which will lead to a minimal RMSE value of the latent heat predictions. If $\mathcal{R}$ is too restrictive, it may exclude $\mathbf{p}_{\text{opt}}$ altogether or lead to a case where there are not enough parameter samples to result in a good surrogate fit. On the other hand, if $Q_{\text{RMSE}}$ is too large, it may include regions with complex non-physical CLM4 responses, rendering our quadratic surrogates inaccurate. Thus a balance needs to be
achieved. As an example, for the site US-IB1, we used $Q_{\text{RMSE}} = 0.7$; the training data points in the space spanned by ($F_{\text{drai}}$, $\ln(Q_{\text{dm}})$, $S_y$) that lie in $\mathcal{R}$ are shown in Figure 1. Here we examine the reason why $Q_{\text{RMSE}}$ was set at that particular value.

A simple way of determining the suitability of $\mathcal{R}$ is to set $Q_{\text{RMSE}}$ to a range of values and compute $p_{\text{opt}}$ for each. If all $Q_{\text{RMSE}}$ lead to the same value of $p_{\text{opt}}$, then none of the $Q_{\text{RMSE}}$ values are very restrictive. If the smallest $Q_{\text{RMSE}}$ leads to a $p_{\text{opt}}$ that is different from the others, then it is too restrictive. We perform four analyses for $Q_{\text{RMSE}} = 0.65$, 0.7, 0.75 and 0.85. $\mathcal{R}$ computed using a smaller $Q_{\text{RMSE}}$ is a subset of an $\mathcal{R}$ corresponding to a larger $Q_{\text{RMSE}}$. $\mathcal{P}$ is defined by the following bounds: $0.1 \leq F_{\text{drai}} \leq 5, \ln(10^{-6}) \leq \ln(Q_{\text{dm}}) \leq \ln(10^{-2}), 0.09 \leq S_y \leq 0.27$, obtained from Hou et al. [2012]. At each site, we take $N = 282$ samples in $\mathcal{P}$, fit a classifier and quadratic surrogates, using the approach described in Sections 2.2 and 2.3. LH is modeled as a function of $F_{\text{drai}}$, $\ln(Q_{\text{dm}})$ and $S_y$, for US-IB1. Note that we model LH and not the $\ln(\text{LH})$ as in Ray et al. [2015] to give more weights to goodness of fit in summer months when LH is high. For $Q_{\text{RMSE}} = 0.65$, surrogates for all months have errors less than 10%. For $Q_{\text{RMSE}} = 0.7$, the surrogates for July and August have errors between 10% and 15%, whereas for $Q_{\text{RMSE}} = 0.75$, surrogates for three months have errors between 10% and 20%. For $Q_{\text{RMSE}} = 0.85$, all surrogates have errors above 10%. Thus as $Q_{\text{RMSE}}$ increases and $\mathcal{R}$ encompasses an increasing fraction of $\mathcal{P}$, the estimates of $p_{\text{opt}}$ become less trustworthy (as the surrogates become increasingly poor approximations of CLM4). Note that adding a kriging component to the surrogate i.e. $y_2(p; \Theta_2)$ in Eq. 5 was not helpful.
To find $p_{opt}$ within each region $R$ defined by the different $Q_{RMSE}$ thresholds, we used a genetic algorithm (GA), as implemented in the R package GA [Scrucca, 2013]. GA algorithms are described in Yu and Gen [2010]. The algorithm is started with an ensemble of 200 members and run for 200 iterations. Surrogate models were used in this calibration.

### 2.5 Surrogate models

As described in Section 1, there are many ways of constructing surrogates and we will use a mixture of polynomial and kriging surrogates in this work. As described in Ray et al. [2015], we set

$$y_c(p) = y_1(p; \Theta_1) + y_2(p; \Theta_2) + \delta, \quad \text{(Eq.5)}$$

where $y_c$ is the CLM4 prediction of LH (unless specified otherwise), $y_1$ is the prediction due to a polynomial surrogate, $y_2$ is the prediction due to a kriging surrogate which captures prediction error of $y_1$ and $\delta$ is a residual. $\Theta_1$ and $\Theta_2$ are surrogate model parameters such as polynomial coefficients and variogram parameters. The procedure for deciding $N$ and estimating $y_1(p, \Theta_1)$ and $y_2(p; \Theta_2)$ are described in Ray et al. [2015] and a summary is provided here. The structure of the model, i.e., the form of $y_1(p, \Theta_1)$ and $y_2(p; \Theta_2)$, are learnt from a training dataset of $N$ CLM4 runs using samples of $p$ drawn from $P$. The polynomial surrogate is constructed first and error in the fit is computed as

$$E = \left\| \frac{y_p(p) - y(c)}{y(c)} \right\|_2,$$

where the norm is taken over a uniformly distributed set of samples of $p$ in the parameter space (i.e., $R$ introduced in section 2.3). The kriging surrogate is constructed only if $E > 10\%$. The
construction of $y_1(p; \Theta_1)$ starts with a multivariate (i.e., three variables) fifth-order polynomial, which is simplified using Bayesian compressive sensing and Akaike Information Criterion, as described in Ray et al. [2015], in which for the two cases shown, this polynomial is simplified to a quadratic form. In one case, the kriging surrogate $y_2(p; \Theta_2)$ is also required. The procedure for choosing the set of samples to construct the training set for the surrogate models, the metrics for assessing the accuracy of the surrogate, the steps taken to ensure that the surrogates do not overfit the training CLM4 data are all described in Ray et al. [2015]. The size N of the training data is decided iteratively. We attempted to construct surrogates based on N = 128 CLM4 runs but failed to achieve the requisite predictive accuracy (relative error less than 15%) from the surrogate models. Consequently we doubled the sampling density to obtain 256 samples of $p$, and added the corners (8 samples of $p$), face-centers (6 samples) and edge-centers (12 samples) of $P$, for a total of 282 samples. This dataset allowed us to obtain acceptable surrogate models. The amount of computation performed for calculating the optimal surrogate for each month at each site was extensive: for each order of polynomial considered (one through five), we constructed 200 training sets and performed 200-rounds of repeated random subsampling validation (a type of cross-validation) to assess the goodness-of-fit and the information criterion for selection of the optimal model. This amount of cross-validation and model selection to obtain robust, accurate surrogates is not typically done in most studies.
3. Results

3.1. Determining the feasible parameter space

Results from the GA calibration, described in section 2.4, are summarized in Table 2. They are plotted as vertical lines in Figure 2, along with the PDFs of $F_{dri}$, $\ln(Q_{dm})$ and $S_y$ developed with the same values of $Q_{RMSE}$. In Table 2, the estimates of $\ln(Q_{dm})$ and $S_y$, are independent of the various values of $Q_{RMSE}$. However, for $Q_{RMSE} = 0.65$ the estimate of $F_{dri}$ lies at the upper limit of the prior, whereas the other values of $Q_{RMSE}$ provide more realistic values of $F_{dri}$. Similar behavior can be observed in Figure 2 – the values of $p_{opt}$ are close to the MAP values of the parameters for all values of $Q_{RMSE}$, except $Q_{RMSE} = 0.65$. This is because as we reduce $Q_{RMSE}$, we retain a smaller portion of the original training set to construct the SVM classifier. This leads to classifiers of low accuracy which can remove promising parts of the parameter space and which can lead to wrong results. Consequently, we reject $Q_{RMSE} = 0.65$ as being too restrictive, and select $Q_{RMSE} = 0.7$ as a compromise between a large coverage of $P$ and an acceptable surrogate accuracy. However, the true evaluation of a calibration is its ability to improve LH predictions when compared to that based on the default values of $\{F_{dri}, \ln(Q_{dm}), S_y\}$, which is conducted for each site, as will be discussed in section 3.3.

3.2 Surrogate validation

The process of isolating $R$ and developing surrogate models is an iterative one, as it involves finding a good $Q_{RMSE}$ and surrogate models of acceptable accuracy. A large $Q_{RMSE}$ is first chosen, and we fit surrogate models (one for each month) by partitioning our training data into learning and testing sets (as described in Section 2.4). Surrogate models are expected to pass a
validation test with two criteria: (1) both the training errors and testing errors should be comparable in magnitude and (2) they should be below 15%. Testing-set errors being larger than learning-set errors indicates overfitting. Figure 3 shows examples of validation checks at US-IB1 and US-IB2 sites. We see that learning and testing errors are similar for all months and for both sites. Further, they do not breach 15%. The surrogate models are mostly quadratic polynomials, but the model simplification step (using Akaike Information Criterion) sometimes removes a few second-order terms too. Identical tests are performed to validate surrogate models for all other sites, but their corresponding plots are omitted for brevity.

### 3.3. Bayesian inversion with surrogate models

The surrogate models, once constructed for all sites, are used to solve Eq. 4 using DRAM. The SVM-based classifier described in Section 2.3, using the $Q_{RMSE}$ described in Sec. 3.1, is used to restrict the calibration to the region $R$ in the parameter space. Model calibrations are performed for each flux tower site separately, to identify whether soil properties and plant functional types (PFT) affect the estimated parameters. Each calibration results in $O(10^4)$ samples of $\{F_{dri},\ln(Q_{dm}), S_y\}$ collected by DRAM. As mentioned in Section 2.2, the convergence of the MCMC chain is checked using the Raftery-Lewis procedure, with the median of the distribution checked with an error bound of 0.025. The median is the most stringent test of convergence for this procedure as discussed in Cowles and Carlin [1996]. Each DRAM run is repeated thrice, starting from an over-dispersed set of points. The samples are used to develop pair-wise correlation plots as well as PDFs for each of the parameters, obtained by marginalizing over all other parameters. An estimate of $\sigma$, the model-data mismatch is also obtained. Figure 4 shows pairwise plots as well as marginalized PDFs for the parameters $F_{dri}$, $\ln(Q_{dm})$, and $S_y$, for US-IB2. The pairw
plots show strong positive relationship between the posterior $\ln(Q_{dm})$ and $S_y$, and slightly positive relationship between the posterior $\ln(Q_{dm})$ and $F_{drai}$. The marginalized PDFs are constructed using kernel density estimation applied to the DRAM samples. Marginalized PDFs for $F_{drai}$, $\ln(Q_{dm})$, $S_y$, and $\sigma$ for the remaining sites are provided in the supporting information (SI). As can be seen from the figures, parametric uncertainty has been reduced in two ways: (1) the marginal PDFs (particularly that for $S_y$) are narrower than the prior bounds defining $P$, and (2) the correlation structure between various parameters (e.g., positive correlation between $\ln(Q_{dm})$ and $S_y$) are exposed. Calibration indicates that the true value of $S_y$ is close to 0.225, but there is some probability that $S_y$ might actually deviate from the value. The mode in the posterior PDF of $\ln(Q_{dm})$ (also called the MAP or maximum a posteriori value of $\ln(Q_{dm})$) is not obvious, but reduction in its uncertainty stems from the discovery of its positive correlation with $S_y$. Knowledge of this correlation will help improve ensemble predictions of LH.

The MAP estimates of the three parameters ($F_{drai}$, $\ln(Q_{dm})$, $S_y$ or $B$), along with the 95% credibility intervals, for the 12 sites are summarized in Table 3. Using $\ln(Q_{dm})$ as an example, we can see the MAP estimates vary dramatically from site to site: from around -13 for US-Dk3, US-IB1, US-IB2, US-Ne3, up to around -5 for US-ARM and US-Wlr. It shows that a simple constant default value is inadequate and unrealistic in modeling heat fluxes at various flux tower sites, not to mention to be used globally. The values in Table 3 provide recommended values/ranges for more accurate and realistic CLM simulations for future studies at the corresponding sites.

Next we investigate whether the estimates of hydrological parameters bear any correlation to the soil and vegetation characteristics at the sites. The sites can be divided into different types
given their soil texture (sandy loam, sandy clay loam, loam, silty loam, silty clay loam, clay loam, silty clay, and clay) and plant functional types (PFTs: deciduous broadleaf, croplands, evergreen needleleaf, grasslands, and closed shrublands). In Figure 5 we overlay the PDFs of the three parameters ($F_{drai}$, $\ln(Q_{dm})$, $S_Y$) at all sites, color-coded by the PFT. It is clear that the two sites with “evergreen needleleaf” PFT, US-Ho1 and US-Dk3, have very similar PDFs (plotted in red) for all three parameters. It is worth mentioning that these two sites also have loamy soils. $F_{drai}$ for both sites lies at the upper end of the range, while $\ln(Q_{dm})$ and $S_Y$ are at the lower end. Sites classified as “croplands” (plotted in green) show similar PDFs for $\ln(Q_{dm})$ and $S_Y$, with the former at the lower end of the prior distribution and the latter at the upper end. This raises the possibility that sites of a given PFT class may share parameters and developing a calibration for one site might suffice for the others. The inverted parameters share some common features at sites with finer soil (e.g., US-IB2 and US-Ne3), particularly in $\ln(Q_{dm})$ and $S_Y$. This indicates a certain level of soil texture control on the parameter values, but $F_{drai}$ behaves slightly differently at US-Ne3 compared to at US-IB2, probably due to the different PFTs.

Finally, we validate the calibration results by checking whether the estimated PDFs can reproduce the calibration data and provide better predictions than the default parameter values. The validations are done with direct CLM4 simulations (i.e., not the surrogate models) for constructing the PDFs. For any given site, we draw 100 samples from the posterior sample sets and use them to seed an ensemble of CLM4 runs. This results in 100 LH predictions for each month, from which we compute the monthly mean, the interquartile range (IQR) and bounds to denote outliers (defined as 1.5 IQR, from the first and third quartiles of the predictions). This is repeated for all the sites. The results are summarized in Figure 6. The LH predictions obtained
using the default values of the parameters are in green, the mean prediction in red and the IQR is plotted using error bars.

From Figure 6, we see that the credibility intervals (IQR and the outlier bounds) vary significantly between sites. This reflects the fact that the precision (sharpness) of the PDFs for different sites varies significantly (see SI for PDFs for other sites); wider PDFs lead to large uncertainty bounds. We see that in many cases the median predictions are not too different from the CLM4 predictions with default parameters (henceforth, “default predictions”). The main contribution of the Bayesian calibration is the establishment of a predictive distribution of latent heat fluxes, as summarized by the IQR and outlier bounds. We see that most of the observations, over all the sites, lie within the IQR. The exceptions are US-IB1, US-Ne3 and US-Wlr where 4-6 observations lie outside the IQR bounds; this is not unusual since the IQR is expected to capture 50% of the observations (i.e., the IQR ranges between the 25th and 75th percentiles of the predictions). Therefore, these exceptions do not indicate that the calibrations are particularly deficient. Further, there are no observations that can be classified as outliers, which illustrates the usefulness and effectiveness of Bayesian inversion. In certain cases, calibration rectifies CLM4’s shortcomings quite significantly. At the two loamy needleleaf sites, US-Dk3 and US-Ho1, the default simulations systematically underestimate the LH for almost all the months, with up to 30% underestimates during summer; after calibration, the predictions are significantly improved. This demonstrates the necessity of parameter estimation to improve CLM4’s predictive skills. At the croplands and grasslands sites, the mean predictions are close to the predictions generated using the nominal/default values of the parameters, but Bayesian calibration allows us to define the uncertainty bounds over the predictions.
To summarize, Bayesian model calibration improves CLM4’s predictive skills, and provides reliable quantification and reduction of the uncertainties. Although due to structural and measurement errors, calibration will not enable CLM4 to reproduce latent heat fluxes exactly. Rather it would provide a means to quantify parametric uncertainty as prediction intervals. These are elements required for subsequent risk analysis and decision making.

4. Discussion

CLM has been widely used in climate and Earth system modeling. Accurate estimation of model parameters is needed for reliable model simulations and predictions under current and future conditions, respectively. In our previous work, a subset of hydrological parameters in CLM4 has been shown to have significant impact on surface energy fluxes based on parameter screening and sensitivity analysis, and therefore could potentially be inverted at the selected flux tower sites using observed surface fluxes.

In this study, we assess the feasibility of calibrating CLM4 parameters at flux tower sites with various soil and climate conditions using a surrogate-based Bayesian model calibration procedure. The procedure starts with building surrogates using CLM4 simulations driven by perturbed parameter sets using a space-filling quasi-MC sampling approach. The surrogates provide simplified yet reliable relationships between dominant hydrological parameters (e.g, $F_{\text{drai}}$, $Q_{\text{dms}}$, $S_y$, and $B$) and response variables such as latent heat fluxes. The surrogates, after careful validation and selection, are then used as computationally efficient alternatives to the CLM numerical simulator, for improving the estimates of the hydrological parameters, and therefore LH predictions, with quantified uncertainties. This procedure had been demonstrated to be effective at two of the 12 selected sites in a previous study [Ray et al., 2015].
However, by extending the same technique to more sites, we acknowledge that there are limitations in the previous version of the procedure [Ray et al., 2015] in that the parameter space confined within prescribed bounds contains non-physical parameter sets as demonstrated in other studies [Sargsyan et al., 2014]. Therefore a classifier is needed to separate the parameter space into physical/non-physical portions, and serve as an informative prior (a joint PDF) before a Bayesian calibration could be performed. The posterior distribution, again a joint PDF, is obtained by inverting against climatologically-averaged latent heat fluxes derived from observations. The posterior distribution provides a complete quantification of uncertainty in the parameter estimates.

We find that the simulated mean latent heat fluxes from CLM4 using the calibrated parameters are generally improved at all sites when compared to those from CLM4 simulations using default parameter sets. Those sites with similar soil texture (e.g., loam) and PFTs share similar posterior PDFs of the parameters, which indicate certain levels of parameter transferability between these sites (i.e., as shown in Figure 5). Nevertheless, the number of sites (i.e., 12) investigated is too small for evaluating model parameter transferability, and we would like to leave it as a topic to be investigated in the future by applying the method to more flux tower sites. On the other hand, it is worth mentioning that model parameter transferability among 431 watersheds in the United States has been investigated in a separate study, in which sensitivity analysis results are used to classify the watersheds into different classes by grouping basins with similar parameter significance patterns. Such a parameter-sensitivity-based classification system helps reduce the complex climate/hydrologic system into subsets of more homogeneous and smaller systems, and provides necessary information to setup the parameter
estimation or model calibration problem. Interested readers are referred to Ren et al. [2015] for details.

Further, our calibration method also results in credibility bounds around the simulated median fluxes which bracket the measured LH data. Sites with large measurement errors, and potential large model structural errors (e.g., ignoring snow melting where the process may be critical) would result in large prediction intervals in model predictions, as shown in Figure 6. This demonstrates that Bayesian calibration could be useful for (1) parameter estimation at sites where model structural assumption is sound; and (2) identifying model structural uncertainty at sites where the current model parametrizations might fail. In this paper, however, we have not attempted to isolate the structural error in CLM4. The PDFs of the estimated parameters are sufficiently wide that the IQR of CLM4 predictions contain the observations. If we had more data, and could create surrogates for them, an isolation of the structural error of CLM4 for each of the sites could be possible. Interested readers are referred to Ray et al. [2015] for an example for US-ARM/Southern Great Plains site.

In addition to the validation shown in Figure 6, validation is also done by checking if the posterior PDFs are useful for predicting LH during testing time periods using calibrations from training time periods. This validation is supplementary to the one shown in Figure 6; however, it could be misleading as it depends on the reliability of inversion itself and also requires that the favorable parameter values do not vary year to year. For example, Figure S12 shows the calibration results for the loamy needleleaf site, US-Dk3, using data from the period 2002-2006. The posteriors are very close to those inverted using data from the whole period 1997-2006 (see Figure S3), and the posterior ranges of LH predictions are almost identical. This validation is not
preferred for sites with relatively short time periods of observations or with large LH variations from year to year. Unfortunately, most sites investigated in this study have relatively short observational periods when the NACP site synthesis datasets were collected. We expect to be able to evaluate the influence of training/testing periods on model parameters when longer data records from the Ameriflux network become available.

5. Conclusion and future work

This work demonstrates a generalizable procedure that can be adopted for calibrating CLM4 under various site and climate conditions using a Bayesian inversion technique integrated with surrogate model development. Surrogate models, as computationally-economic alternatives to the direct CLM simulator, can be successfully developed with 15% threshold for training and testing errors in the climatologically averaged heat fluxes. At all selected flux tower sites, most of the latent heat flux observations lie within the IQR ranges of predictions based on parameter values drawn from the posterior distribution. The procedure can be applied to other models including newer versions of CLM and other components of an Earth system model, given that the metrics for measuring model performance and for defining objective functions can be well-defined. Further, since the calibration is performed using surrogates, the computational cost of the Earth system model (or its component) ceases to be an issue. As demonstrated in this study, the Bayesian calibration procedure could be used as a tool for parameter estimation with uncertainty bounds, as well as for identifying potential model structural errors by extensively exploring the parameter space and comparing discrepancies between model predictions and observations. Such a tool will be valuable for model applications for quantifying prediction intervals, as well as to model developers to better understand model structural uncertainties by
comparing the model uncertainty range to observations, and identify ways to improve the model. We will explore ways to integrate the procedure with model benchmark systems such as the International Land Model Benchmarking Project (http://www.ilamb.org) to accelerate such a process.

On the other hand, a surrogate-based calibration procedure is intrinsically subject to errors as a result of approximating a complex model using simplified functions, not to mention the potential risk of failures in building the surrogates due to the complex relationships between model parameters and outputs of interest. To address this limitation, we are testing a scalable MCMC algorithm that features multiple chains on high performance computing facilities that could be integrated with any real ESMs to avoid the issues rooted from surrogates. We will report our progress towards that direction in the near future.
Acknowledgement

This work is supported by US Department of Energy Office of Science’s Advanced Scientific Computing Research through the Applied Mathematics program. PNNL is operated for the US Department of Energy by Battelle Memorial Institute under Contract DE-AC05-76RLO1830. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the US Department of Energy’s National Nuclear Security Administration under contract DE-AC04-94AL85000. The authors would like to thank the site PIs, research teams, and their funders and the NACP site synthesis team for providing the flux tower data sets, which make this study possible. The NACP site synthesis dataset can be found at http://nacp.ornl.gov/mast-dc/int_synth_site.shtml.
References


**Tables:**

Table 1. Site characteristics of the 12 flux towers.

<table>
<thead>
<tr>
<th>Site</th>
<th>Longitude</th>
<th>Latitude</th>
<th>Elev</th>
<th>Soil texture</th>
<th>Plant functional type</th>
<th>Period</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>US-Ha1</td>
<td>-72.1715</td>
<td>42.5378</td>
<td>343.00</td>
<td>sandy loam</td>
<td>Deciduous Broadleaf</td>
<td>1991-2006</td>
<td>Urbanski et al. [2007]</td>
</tr>
<tr>
<td>US-Dk2</td>
<td>-79.1004</td>
<td>35.9736</td>
<td>163.00</td>
<td>sandy clay loam</td>
<td>Deciduous Broadleaf</td>
<td>2003-2005</td>
<td>Oishi et al. [2008]; Stoy et al. [2008]</td>
</tr>
<tr>
<td>US-Dk3</td>
<td>-79.0942</td>
<td>35.9782</td>
<td>163.00</td>
<td>loam</td>
<td>Evergreen Needleleaf</td>
<td>1998-2005</td>
<td>Oren et al. [2006]; Stoy et al. [2006]</td>
</tr>
<tr>
<td>US-IB1</td>
<td>-88.2227</td>
<td>41.8593</td>
<td>225.00</td>
<td>silty clay loam</td>
<td>Croplands</td>
<td>2005-2007</td>
<td>Allison et al. [2005]</td>
</tr>
<tr>
<td>US-IB2</td>
<td>-88.2410</td>
<td>41.8406</td>
<td>226.00</td>
<td>silty clay loam</td>
<td>Grasslands</td>
<td>2004-2007</td>
<td>Allison et al. [2005]</td>
</tr>
<tr>
<td>US-Shd</td>
<td>-96.6827</td>
<td>36.9601</td>
<td>350.00</td>
<td>silty clay loam</td>
<td>Grasslands</td>
<td>1997-2000</td>
<td>Suyker et al. [2003]</td>
</tr>
<tr>
<td>US-Ne3</td>
<td>-96.4397</td>
<td>41.1797</td>
<td>363.00</td>
<td>clay loam</td>
<td>Croplands</td>
<td>2001-2006</td>
<td>Suyker and Verma [2010]; Verma et al. [2005]</td>
</tr>
<tr>
<td>US-Ho1</td>
<td>-68.7403</td>
<td>45.2041</td>
<td>79.00</td>
<td>loam</td>
<td>Evergreen Needleleaf</td>
<td>2004-2007</td>
<td>Hollinger et al. [1999]</td>
</tr>
<tr>
<td>US-Wlr</td>
<td>-96.8550</td>
<td>37.5208</td>
<td>408.00</td>
<td>silty clay</td>
<td>Grasslands</td>
<td>2001-2004</td>
<td>Gao et al. [1998]</td>
</tr>
</tbody>
</table>
Table 2. $p_{opt}$ computed using different QRMSE and surrogate models of different qualities.

<table>
<thead>
<tr>
<th>$Q_{RMSE}$</th>
<th>Surrogate model accuracy</th>
<th>$p_{opt} = { F_{drai}, \ln(Q_{dm}), S_y }$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.65</td>
<td>All 12 surrogates have &lt; 10% error</td>
<td>{5.00, -13.46, 0.27}</td>
</tr>
<tr>
<td>0.70</td>
<td>2 out of 12 surrogates have errors between 10% and 15%</td>
<td>{0.80, -13.77, 0.27}</td>
</tr>
<tr>
<td>0.75</td>
<td>3 out of 12 surrogates have errors between 10% and 20%.</td>
<td>{1.33, -13.78, 0.27}</td>
</tr>
<tr>
<td></td>
<td>Surrogates for summer months have the largest errors</td>
<td></td>
</tr>
<tr>
<td>0.85</td>
<td>All surrogates have errors above 10%</td>
<td>{1.09, -13.77, 0.27}</td>
</tr>
</tbody>
</table>
Table 3. Summary of posterior PDFs for the 12 sites. For each site, we tabulate the maximum a posteriori (MAP) and median estimates of each of the parameters and the 2.5th and 97.5th percentiles i.e., the 95% credibility bounds of the estimates. The original PDFs are in the Appendix.

<table>
<thead>
<tr>
<th>CLM parameter</th>
<th>$F_{\text{drai}}$</th>
<th>$\ln(Q_{\text{dim}})$</th>
<th>$S_y$ or $B$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sites</td>
<td>MAP</td>
<td>Q2.5</td>
<td>median</td>
<td>Q97.5</td>
</tr>
<tr>
<td>US-Ha1</td>
<td>4.62</td>
<td>0.19</td>
<td>2.70</td>
<td>4.93</td>
</tr>
<tr>
<td>US-DK2</td>
<td>4.62</td>
<td>0.19</td>
<td>2.80</td>
<td>4.93</td>
</tr>
<tr>
<td>US-DK3</td>
<td>4.78</td>
<td>2.35</td>
<td>4.40</td>
<td>4.97</td>
</tr>
<tr>
<td>US-IB1</td>
<td>0.45</td>
<td>0.14</td>
<td>1.56</td>
<td>4.87</td>
</tr>
<tr>
<td>US-IB2</td>
<td>1.76</td>
<td>0.75</td>
<td>2.74</td>
<td>4.84</td>
</tr>
<tr>
<td>US-Shd</td>
<td>2.90</td>
<td>1.26</td>
<td>3.13</td>
<td>4.88</td>
</tr>
<tr>
<td>US-SO2</td>
<td>0.66</td>
<td>0.21</td>
<td>2.27</td>
<td>4.85</td>
</tr>
<tr>
<td>US-Ne3</td>
<td>4.28</td>
<td>0.34</td>
<td>3.33</td>
<td>4.91</td>
</tr>
<tr>
<td>US-Ho1</td>
<td>4.56</td>
<td>3.12</td>
<td>4.26</td>
<td>4.96</td>
</tr>
<tr>
<td>US-MOz</td>
<td>0.34</td>
<td>0.12</td>
<td>0.82</td>
<td>4.75</td>
</tr>
<tr>
<td>US-ARM</td>
<td>0.20</td>
<td>0.11</td>
<td>0.40</td>
<td>1.57</td>
</tr>
<tr>
<td>US-Wlr</td>
<td>0.32</td>
<td>0.12</td>
<td>0.60</td>
<td>4.90</td>
</tr>
</tbody>
</table>

Note: the third parameter is $B$ for US-ARM and US-Wlr, and $S_y$ otherwise.
Figure captions:

Figure 1. CLM4 parameters ($F_{\text{d}}$, $\ln(Q_{\text{d}})$, $S_y$), from the training data, that lie inside $R$. The red diamond plots the nominal value, the green triangle the parameter combination in the training set for US-IB1 with the best agreement with observations.

Figure 2. PDFs of the three parameters for each $Q_{\text{RMSE}}$ with the GA estimate plotted as vertical lines. In the top right and bottom left sub-figures, the vertical lines showing the values of $p_{\text{opt}}$ coincide and are thus obscured.

Figure 3. Learning-set and testing-set relative predictive errors for surrogate validation at two selected example sites (US-IB1 and IB2).

Figure 4. Marginal distribution of the joint prior (dashed), posterior (solid) PDFs, the default values (dashed vertical line), and the maximum a posteriori (MAP) values (solid vertical line), and paired scatters of posterior samples of the four parameters for inversion at US-IB2.

Figure 5. Posterior distributions of inverted parameters color-coded by Plant Functional Types for 10 out of 12 sites that share the same parameters. It is evident that the two Evergreen Needleleaf sites have very similar PDFs for all three parameters. Croplands share similar estimates for $S_y$.

Figure 6. Validation of posterior parameter using by CLM4. The symbols are the monthly-mean observed LH fluxes, climatologically averaged over the durations tabulated in Table 1. The line with the error bound is the median prediction from the ensemble of runs seeded with samples from the posterior distribution. The error bars are the first and third quartiles of the predictions. The green dashed line is the prediction using nominal parameter values. The dashed blue and purple lines denote outlier bounds.
Figures:

Figure 2. CLM4 parameters ($F_{\text{dual}}$, $\ln(Q_{\text{dual}})$, $S_y$), from the training data, that lie inside R. The red diamond plots the nominal value, the green triangle the parameter combination in the training set for US-IB1 with the best agreement with observations.
Figure 2. PDFs of the three parameters for each $Q_{\text{RMSE}}$ with the GA estimate plotted as vertical lines. In the top right and bottom left sub-figures, the vertical lines showing the values of $p_{\text{opt}}$ coincide and are thus obscured.
Figure 3. Learning-set and testing-set relative predictive errors for surrogate validation at two selected example sites (US-IB1 and IB2).
Figure 4. Marginal distribution of the joint prior (dashed), posterior (solid) PDFs, the default values (dashed vertical line), and the maximum a posteriori (MAP) values (solid vertical line), and paired scatters of posterior samples of the four parameters for inversion at US-IB2.
Figure 5. Posterior distributions of inverted parameters color-coded by Plant Functional Types for 10 out of 12 sites that share the same parameters. It is evident that the two Evergreen Needleleaf sites have very similar PDFs for all three parameters. Croplands share similar estimates for $S_y$. 
Figure 6. Validation of posterior parameter using by CLM4. The symbols are the monthly-mean observed LH fluxes, climatologically averaged over the durations tabulated in Table 1. The line with the error bound is the median prediction from the ensemble of runs seeded with samples from the posterior distribution. The error bars are the first and third quartiles of the predictions. The green dashed line is the prediction using nominal parameter values. The dashed blue and purple lines denote outlier bounds.