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

Maoyi Huang[1*], Jaideep Ray[2], Zhangshuan Hou[3],
Huiying Ren[3], Ying Liu[1], Laura Swiler[4]

[1]Earth System Analysis and Modeling Group, Pacific Northwest National Laboratory, P.O. Box 999, Richland, WA 99352, USA

[2]Sandia National Laboratories, P.O. Box 969, Livermore, CA 945551

[3]Hydrology Technical Group, Pacific Northwest National Laboratory, P.O. Box 999, Richland, WA 99352, USA

[4]Sandia 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

72 [*Bastidas et al.*, 2006; *Demaria et al.*, 2007; *Gan et al.*, 2015; *Gulden et al.*, 2008; *Henderson-*
73 *Sellers et al.*, 1995; *Hou et al.*, 2012; *Huang et al.*, 2013; *Liang and Guo*, 2003; *Rosero et al.*,
74 2010]. Williams et al. [2009] discussed how the FLUXNET database can be used to improve
75 forecasts of global biogeochemical and climate models. Some sensitivity studies were performed
76 at the flux tower sites [*Alton et al.*, 2006; *Baldocchi and Wilson*, 2001; *White et al.*, 2000; *Zobitz*
77 *et al.*, 2006], focusing on evaluating net primary production controls, biophysical parameters
78 governing light propagation, canopy photosynthesis, and carbon cycling. In the past few years, a
79 number of studies have documented the sensitivity of surface fluxes to model parameters in the
80 Community Land Model (CLM). *Göhler et al.* [2013] analyzed the sensitivity of latent heat
81 (LH), sensible heat (SH) and photosynthesis of the Community Land Model CLM version 3.5 to
82 its parameters. They found that photosynthesis is very sensitive to parameters associated with
83 plant functional types, whereas LH is sensitive to soil water parameters. *Bonan et al.* [2011]
84 investigated sensitivity of LH to photosynthetic parameters in CLM version 4.0 (CLM4) and
85 suggested that model structural errors in the model could be compensated by parameter
86 adjustment. *Hou et al.* [2012] performed sensitivity analyses at representative flux tower sites on
87 outputs of heat fluxes from CLM4 driven by satellite phenology and showed that they are most
88 sensitive to a subset of hydrological parameters. This raised the possibility that the parameters
89 could be estimated from measurements of heat fluxes, to improve the predictive skill of CLM.
90 The results also show that only a selected set of parameters, and not all hydrological parameters,
91 could be correctly estimated from the fluxes. More recently, *Huang et al.* [2013]; *Ren et al.*
92 [2015] extended the sensitivity analysis framework in *Hou et al.* [2012] to 431 relatively pristine
93 watersheds over the contiguous United States within minimal human perturbations. They
94 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

117 form of samples, each of which requires the forward model (e.g., CLM) to be run at least once. If
118 the model is computationally expensive, e.g., if it is a high fidelity model (HFM), it has to be
119 replaced by a fast-running proxy called a surrogate so that the inverse problem may be solved.

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

126 Surrogates have long been used in engineering and in the geosciences, for example in water
127 resources research. *Viana et al.* [2014], *Forrester and Keane* [2009] and *Razavi et al.* [2012] are
128 three recent review articles that describe those applications. The earliest surrogates were
129 polynomials or approximate neural networks fitted to data. Later, multivariate adaptive
130 regression splines [*Friedman*, 1991], Gaussian process (or kriging) and kernel methods such as
131 radial basis functions (RBF) [*Regis and Shoemaker*, 2007] were used to represent the HFM's
132 response surface. Some methods, such as Gaussian process models, can also provide an estimate
133 of the error in the surrogate model's prediction. Of the latter, there has been a move towards
134 mixtures of surrogates [*Goel et al.*, 2007]. Various types of "multifidelity surrogates",
135 constructed from a training set of HFM and low fidelity model (LFM) runs, have also been
136 investigated [*Eldred and Dunlavy*, 2006; *Pau et al.*, 2014; *Viana et al.*, 2014]. HFM responses
137 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

160 (not the HFM output) and estimated 11 parameters of the CLM4.5's methane module using a
161 global optimization method called DYnamic COordinate search using Response Surface models
162 (DYCORS) [*Regis and Shoemaker*, 2007]. *Sargsyan et al.* [2014] attempted to construct
163 surrogates for five variables of interests from CLM4 with prognostic carbon and nitrogen
164 modules turned on (i.e., CLM4-CN) using Bayesian compressive sensing (BCS) in combination
165 with polynomial chaos expansions (PCEs). They found that the input-output relationship in
166 CLM4-CN could be composed of qualitatively different regimes (i.e., live or dead vegetation
167 regimes associated with different regions in the parameter space), so that clustering- and
168 classification-based piecewise PCE construction is needed. In *Ray et al.* [2015], the authors used
169 polynomial and universal kriging surrogates to calibrate three hydrological parameters of CLM
170 4.0 using measurements of latent heat fluxes. Two competing models were used for the model –
171 data mismatch to estimate a composite of measurement error and (a crude estimate of) the
172 structural error of CLM. In *Gong et al.* [2015], the authors used adaptive surrogate-based
173 optimization to perform parameter estimation of the Common Land Model using six observables
174 jointly; 12 independent parameters were (deterministically) calibrated.

175 In this study we combine the advances in surrogate modeling described above with a
176 Bayesian model calibration framework as presented in *Ray et al.* [2015] to perform calibration of
177 CLM 4.0 at 12 selected flux tower sites using latent heat (LH) flux measurements. In Section 2,
178 we formulate the parameter estimation problem and describe the (Bayesian) parameter
179 estimation method. In Section 3, we estimate the joint PDFs of three hydrological parameters,
180 tabulate their modes and their credibility intervals for all the sites and correlate them with the site
181 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-NRl 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-

based inversion approach developed in *Ray et al.* [2015]. In sections 2.2 to 2.4, we will describe our inversion approach, including the choice of priors and the method of building the surrogates.

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

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

where $\boldsymbol{N}(0, \boldsymbol{\Gamma})$ denotes a zero-mean multivariate normal distribution with covariance $\boldsymbol{\Gamma}$. Neither the form nor the distribution of $\boldsymbol{\Gamma}$ 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

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

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

$$\mathrm{P}(\mathbf{p},\boldsymbol{\Gamma}|\mathbf{Y}^{(obs)}) \propto \Lambda\left(\mathbf{Y}^{(obs)}|\mathbf{p},\boldsymbol{\Gamma}\right)\pi_{prior}(\mathbf{p}) \quad \text{(Eq. 3)}$$

where $\pi_{prior}(\mathbf{p})$ is our prior belief regarding the distribution of **p**. The posterior distribution is arbitrary and is realized by a set of samples $\mathbf{p}^{(s)}$, $s = 1 \ldots N_{mcmc}$ 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

245 used to parameterize CLM4. Leaf area indices from MODIS (MODerate-resolution Imaging
246 Spectroradiometer) are retrieved from nine pixels surrounding the tower footprint and provided
247 by the NACP team as the satellite phenology to drive the CLM4SP simulation. Measured fluxes
248 of latent and sensible heat at the native time resolution of the observations (30- or 60-minute) are
249 provided and aggregated to monthly time step for calibration in this study. The data were gap-
250 filled following a standard protocol as well. Measurements were obtained using the eddy-
251 covariance (EC) method as part of the Ameriflux network. It has been widely recognized that
252 surface energy fluxes based on EC method are subject to energy closure problems [*Wilson et al.*,
253 2002] but unfortunately measurement uncertainty bounds are not reported at part of the NACP
254 site synthesis dataset and therefore are not addressed in this study. Rather, we treat the fluxes
255 provided as the "truth" in this study. We note that addressing the energy closure problem or
256 reporting errors from EC systems for modelers is out of the scope of the study but should be
257 addressed as a community effort as part of the FLUXNET network. Interested readers are
258 referred to *Schwalm et al.* [2010] on the NACP site synthsis dataset and the references listed in
259 Table 1 for detailed descriptions on the sites.

260 CLM4 was spun up by cycling the provided forcing for at least five times until all state variables
261 reached equilibrium. For the purpose of capturing first-order dynamics in the climate system, we
262 focus on evaluating CLM4's ability to simulate seasonal variability by deriving time series of
263 latent heat flux at monthly time steps from the raw datasets, consistent with our previous studies
264 [*Hou et al.*, 2012; *Ray et al.*, 2015]. As shown in *Ray et al.* [2015], climatologically averaging
265 smoothens CLM4 predictions and allows the surrogate to be fitted with an acceptable degree of
266 accuracy. A surrogate model is created for each of the 12 months. Parameter estimates obtained
267 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 $\boldsymbol{\Gamma}$. Consequently, we model it as a constant diagonal matrix i.e., $\boldsymbol{\Gamma} = \sigma^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. This is a very limiting assumption; when data and surrogate models allow, we can use more sophisticated representations for Γ as discussed in *Ray et al.* [2015]. Eq. 3 simplifies to

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

where $M = 12$ is the months of climatologically averaged data that we use in the calibration, $\| \cdot \|_2$ 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 σ^2 is an inverse Gamma, which, being a conjugate prior, allows us to sample σ^2 with a Gibbs sampler. We use the MCMC implementation in the `R` package `FME` to sample ($\mathbf{p}$, σ^2) from the posterior density distribution. `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 `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_{drai} and

287 $\ln(Q_{dm})$. The third parameter is the Clapp-Hornberger parameter B [*Clapp and Hornberger*,
288 1978] for US-ARM and US-Wlr, and specific yield S_y for the other study sites. Prior
289 distributions for each of these parameters are discussed in *Hou et al.* [2012]. They are
290 independent and usually (but not always) uniform distributions. However **p** sampled randomly
291 from $\mathcal{P}$ is not necessarily physically realistic, which causes complex (and non-physical)
292 behaviors of LH predictions and makes surrogate modeling difficult. Further, the LH predictions
293 generated bear little resemblance to $\mathbf{Y}^{(obs)}$ and the root-mean-square-error
294 $\mathrm{RMSE}(\mathbf{p}) = \left\|\mathbf{Y}^{(obs)} - \mathbf{Y}(\mathbf{p})\right\|_2$ is large. We would like to avoid sampling non-physical parts of the
295 parameter space with the MCMC method. Therefore, we re-define the hypercube encompassing
296 all parameter values to a more informative prior. We outline our approach in the following
297 paragraphs.

298 As described in *Giunta et al.* [1995] and mentioned in Section 1, one may excise the
299 inappropriate portions of $\mathcal{P}$ to obtain $\mathcal{R}$, which contains physically realistic parameters. We do
300 so in this study. We draw N (=282) samples from $\mathcal{P}$ using a space-filling, quasi Monte Carlo
301 method and use them to generate an ensemble of LH predictions. The reason we use 282 samples
302 is described in Section. 2.4 on surrogate models. RMSE(**p**) are calculated for each realization
303 and we specify a threshold RMSE quartile Q_{RMSE} to identify **p** whose predictions are close to
304 observations at a given site. The selected samples of **p** discretely define $\mathcal{R}$. We define an
305 improper, informative prior $\pi_{prior}(\mathbf{p})$ with support $\mathcal{R}$ such that the prior density is one inside $\mathcal{R}$
306 and zero outside. We construct our surrogates using only parameter combinations **p** that reside
307 inside $\mathcal{R}$. Note that the use of a user-defined Q_{RMSE} makes $\pi_{prior}(\mathbf{p})$ somewhat subjective and we

will investigate its effect below. Usually setting Q_{RMSE} = 0.7 has allowed us remove the non-physical part of $\mathcal{P}$ and to construct accurate surrogates. Note that CLM4 surrogates that are valid only in a portion of $\mathcal{P}$ has been documented in earlier studies as well [*Sargsyan et al.*, 2014].

In order to use $\pi_{prior}(\mathbf{p})$ within MCMC, we require a precise definition of $\mathcal{R}$ so that we may unambiguously decide whether an arbitrary **p** resides within $\mathcal{R}$. The separation of the training set of runs into valid (i.e., $\mathbf{p} \in \mathcal{R}$) and invalid (i.e., ., $\mathbf{p} \in \mathcal{R}^*, \boldsymbol{\mathcal{R}}^* \in \mathcal{P}\backslash\mathcal{R}$)) 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 $\zeta(\mathbf{p})$

$$\zeta(p) = \begin{cases} +1, \mathbf{p} \in \mathcal{R} \\ -1, \mathbf{p} \in \mathcal{R}^* \end{cases}$$

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

The problem of approximation $\zeta(\mathbf{p})$ can be cast as a classification problem – we seek the separatrix $\partial\mathcal{R}$ that separates $\mathcal{R}$ from R^*. Once established, it can decide whether an arbitrary **p** lies within $\mathcal{R}$. The training set of 282 runs provides us with realizations of the binary function $\zeta(\mathbf{p})$ 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 $\zeta(\mathbf{p})$. The SVM identifies an optimal separating hyperplane i.e. $\partial\mathcal{R}$, in the three-dimensional **p** space. $\partial\mathcal{R}$ 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 `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_{RMSE} of the difference between observations and existing training data, i.e., $\mathrm{RMSE}(\mathbf{p}) = \left\|\mathbf{Y}^{(obs)} - \mathbf{Y}(\mathbf{p})\right\|_2$. $\mathcal{R}$ is chosen as a way of excluding the non-physical part of $\mathcal{P}$. Choosing Q_{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 $\mathbf{p}_{opt}$ that lead to the minimal value of RMSE. That is, if we use an optimization method to find the optimal parameters $\mathbf{p}_{opt}$, 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 $\mathbf{p}_{opt}$ 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_{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

348 achieved. As an example, for the site US-IB1, we used Q_{RMSE} = 0.7; the training data points in
349 the space spanned by (F_{drai}, ln(Q_{dm}), S_y) that lie in $\mathcal{R}$ are shown in Figure 1. Here we examine the
350 reason why Q_{RMSE} was set at that particular value.

351 A simple way of determining the suitability of $\mathcal{R}$ is to set Q_{RMSE} to a range of values and
352 compute $\mathbf{p}_{opt}$ for each. If all Q_{RMSE} lead to the same value of $\mathbf{p}_{opt}$, then none of the Q_{RMSE} values
353 are very restrictive. If the smallest Q_{RMSE} leads to a $\mathbf{p}_{opt}$ that is different from the others, then it is
354 too restrictive. We perform four analyses for Q_{RMSE} = 0.65, 0.7, 0.75 and 0.85. $\mathcal{R}$ computed
355 using a smaller Q_{RMSE} is a subset of an $\mathcal{R}$ corresponding to a larger Q_{RMSE}. $\mathcal{P}$ is defined by the
356 following bounds: $0.1 \leq F_{drai} \leq 5, \ln(10^{-6}) \leq \ln(Q_{dm}) \leq \ln(10^{-2}), 0.09 \leq S_y \leq 0.27$, obtained
357 from *Hou et al.* [2012]. At each site, we take N = 282 samples in $\mathcal{P}$, fit a classifier and quadratic
358 surrogates, using the approach described in Sections 2.2 and 2.3. LH is modeled as a function of
359 F_{drai}, ln(Q_{dm}) and S_y, for US-IB1. Note that we model LH and not the ln(LH) as in *Ray et al.*
360 [2015] to give more weights to goodness of fit in summer months when LH is high. For Q_{RMSE} =
361 0.65, surrogates for all months have errors less than 10%. For Q_{RMSE} = 0.7, the surrogates for
362 July and August have errors between 10% and 15%, whereas for Q_{RMSE} = 0.75, surrogates for
363 three months have errors between 10% and 20%. For Q_{RMSE} = 0.85, all surrogates have errors
364 above 10%. Thus as Q_{RMSE} increases and $\mathcal{R}$ encompasses an increasing fraction of $\mathcal{P}$, the
365 estimates of $\mathbf{p}_{opt}$ become less trustworthy (as the surrogates become increasingly poor
366 approximations of CLM4). Note that adding a kriging component to the surrogate i.e. $y_2(\mathbf{p}; \Theta_2)$
367 in Eq. 5 was not helpful.

368 To find $\mathbf{p}_{opt}$ within each region $\mathcal{R}$ defined by the different Q_{RMSE} thresholds, we used a
369 genetic algorithm (GA), as implemented in the `R` package `GA` [*Scrucca*, 2013]. GA algorithms
370 are described in *Yu and Gen* [2010]. The algorithm is started with an ensemble of 200 members
371 and run for 200 iterations. Surrogate models were used in this calibration.

372 **2.5 Surrogate models**

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

376 $$y_c(\mathbf{p}) = y_1(\mathbf{p}; \boldsymbol{\Theta_1}) + y_2(\mathbf{p}; \boldsymbol{\Theta_2}) + \delta, \quad \text{(Eq.5)}$$
377

378 where y_c is the CLM4 prediction of LH (unless specified otherwise), y_1 is the prediction due to a
379 polynomial surrogate, y_2 is the prediction due to a kriging surrogate which captures prediction
380 error of y_1 and δ is a residual. $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ are surrogate model parameters such as polynomial
381 coefficients and variogram parameters. The procedure for deciding N and estimating $y_1(\mathbf{p}, \Theta_1)$
382 and $y_2(\mathbf{p}; \Theta_2)$ are described in *Ray et al.* [2015] and a summary is provided here. The structure of
383 the model, i.e., the form of $y_1(\mathbf{p}, \Theta_1)$ and $y_2(\mathbf{p}; \Theta_2)$, are learnt from a training dataset of N CLM4
384 runs using samples of $\mathbf{p}$ drawn from $\mathcal{P}$. The polynomial surrogate is constructed first and error in
385 the fit is computed as

$$E = \left\| \frac{y(\mathbf{p})_c - y(\mathbf{p})_1}{y(\mathbf{p})_c} \right\|_2,$$

386 where the norm is taken over a uniformly distributed set of samples of $\mathbf{p}$ in the parameter space
387 (i.e., $\mathcal{R}$ introduced in section 2.3). The kriging surrogate is constructed only if E > 10%. The

388 construction of $y_1(\mathbf{p};\ \boldsymbol{\Theta_1})$ starts with a multivariate (i.e., three variables) fifth-order polynomial,
389 which is simplified using Bayesian compressive sensing and Akaike Information Criterion, as
390 described in *Ray et al.* [2015], in which for the two cases shown, this polynomial is simplified to
391 a quadratic form. In one case, the kriging surrogate $y_2(\mathbf{p};\ \boldsymbol{\Theta_2})$ is also required. The procedure for
392 choosing the set of samples to construct the training set for the surrogate models, the metrics for
393 assessing the accuracy of the surrogate, the steps taken to ensure that the surrogates do not
394 overfit the training CLM4 data are all described in *Ray et al.* [2015]. The size N of the training
395 data is decided iteratively. We attempted to construct surrogates based on N = 128 CLM4 runs
396 but failed to achieve the requisite predictive accuracy (relative error less than 15%) from the
397 surrogate models. Consequently we doubled the sampling density to obtain 256 samples of **p**,
398 and added the corners (8 samples of **p**), face-centers (6 samples) and edge-centers (12 samples)
399 of $\mathcal{P}$, for a total of 282 samples. This dataset allowed us to obtain acceptable surrogate models.
400 The amount of computation performed for calculating the optimal surrogate for each month at
401 each site was extensive: for each order of polynomial considered (one through five), we
402 constructed 200 training sets and performed 200-rounds of repeated random subsampling
403 validation (a type of cross-validation) to assess the goodness-of-fit and the information criterion
404 for selection of the optimal model. This amount of cross-validation and model selection to
405 obtain robust, accurate surrogates is not typically done in most studies.
406

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_{drai}, $\ln(Q_{dm})$ 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_{drai} lies at the upper limit of the prior, whereas the other values of Q_{RMSE} provide more realistic values of F_{drai}. Similar behavior can be observed in Figure 2 – the values of $\mathbf{p}_{opt}$ 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 $\mathcal{P}$ 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 $\{F_{drai}, \ln(Q_{dm}), S_y\}$, which is conducted for each site, as will be discussed in section 3.3.

3.2 Surrogate validation

The process of isolating $\mathcal{R}$ 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 Sectioin2.3, using the Q_{RMSE} described in Sec. 3.1, is used to restrict the calibration to the region $\mathcal{R}$ 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 $\{F_{drai}, \ln(Q_{dm}), S_y\}$ 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 σ, the model-data mismatch is also obtained. Figure 4 shows pairwise plots as well as marginalized PDFs for the parameters F_{drai}, $\ln(_{Qdm})$, and S_y, for US-IB2. The pairwise

450 plots show strong positive relationship between the posterior $\ln(Q_{dm})$ and S_y, and slightly
451 positive relationship between the posterior $\ln(Q_{dm})$ and F_{drai}. The marginalized PDFs are
452 constructed using kernel density estimation applied to the DRAM samples. Marginalized PDFs
453 for F_{drai}, $\ln(Q_{dm})$, S_y, and σ for the remaining sites are provided in the supporting information
454 (SI). As can be seen from the figures, parametric uncertainty has been reduced in two ways: (1)
455 the marginal PDFs (particularly that for S_y) are narrower than the prior bounds defining $\mathcal{P}$, and
456 (2) the correlation structure between various parameters (e.g., positive correlation between
457 $\ln(Q_{dm})$ and S_y) are exposed. Calibration indicates that the true value of S_y is close to 0.225, but
458 there is some probability that S_y might actually deviate from the value. The mode in the posterior
459 PDF of $\ln(Q_{dm})$ (also called the MAP or *maximum a posteriori* value of $\ln(Q_{dm})$) is not obvious,
460 but reduction in its uncertainty stems from the discovery of its positive correlation with S_y.
461 Knowledge of this correlation will help improve ensemble predictions of LH.

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

470 Next we investigate whether the estimates of hydrological parameters bear any correlation to
471 the soil and vegetation characteristics at the sites. The sites can be divided into different types

472 given their soil texture (sandy loam, sandy clay loam, loam, silty loam, silty clay loam, clay
473 loam, silty clay, and clay) and plant functional types (PFTs: deciduous broadleaf, croplands,
474 evergreen needleleaf, grasslands, and closed shrublands). In Figure 5 we overlay the PDFs of the
475 three parameters (F_{drai}, $\ln(Q_{dm})$, S_y) at all sites, color-coded by the PFT. It is clear that the two
476 sites with "evergreen needleleaf" PFT, US-Ho1 and US-Dk3, have very similar PDFs (plotted in
477 red) for all three parameters. It is worth mentioning that these two sites also have loamy soils.
478 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.
479 Sites classified as "croplands" (plotted in green) show similar PDFs for $\ln(Q_{dm})$ and S_y, with the
480 former at the lower end of the prior distribution and the latter at the upper end. This raises the
481 possibility that sites of a given PFT class may share parameters and developing a calibration for
482 one site might suffice for the others. The inverted parameters share some common features at
483 sites with finer soil (e.g., US-IB2 and US-Ne3), particularly in ln(Qdm) and Sy. This indicates a
484 certain level of soil texture control on the parameter values, but F_{drai} behaves slightly differently
485 at US-Ne3 compared to at US-IB2, probably due to the different PFTs.

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

494 using the default values of the parameters are in green, the mean prediction in red and the IQR is
495 plotted using error bars.

496 From Figure 6, we see that the credibility intervals (IQR and the outlier bounds) vary
497 significantly between sites. This reflects the fact that the precision (sharpness) of the PDFs for
498 different sites varies significantly (see SI for PDFs for other sites); wider PDFs lead to large
499 uncertainty bounds. We see that in many cases the median predictions are not too different from
500 the CLM4 predictions with default parameters (henceforth, "default predictions"). The main
501 contribution of the Bayesian calibration is the establishment of a predictive distribution of latent
502 heat fluxes, as summarized by the IQR and outlier bounds. We see that most of the observations,
503 over all the sites, lie within the IQR. The exceptions are US-IB1, US-Ne3 and US-Wlr where 4-6
504 observations lie outside the IQR bounds; this is not unusual since the IQR is expected to capture
505 50% of the observations (i.e., the IQR ranges between the 25^{th} and 75^{th} percentiles of the
506 predictions). Therefore, these exceptions do not indicate that the calibrations are particularly
507 deficient. Further, there are no observations that can be classified as outliers, which illustrates the
508 usefulness and effectiveness of Bayesian inversion. In certain cases, calibration rectifies CLM4's
509 shortcomings quite significantly. At the two loamy needleleaf sites, US-Dk3 and US-Ho1, the
510 default simulations systematically underestimate the LH for almost all the months, with up to
511 30% underestimates during summer; after calibration, the predictions are significantly improved.
512 This demonstrates the necessity of parameter estimation to improve CLM4's predictive skills. At
513 the croplands and grasslands sites, the mean predictions are close to the predictions generated
514 using the nominal/default values of the parameters, but Bayesian calibration allows us to define
515 the uncertainty bounds over the predictions.

516 To summarize, Bayesian model calibration improves CLM4's predictive skills, and provides
517 reliable quantification and reduction of the uncertainties. Although due to structural and
518 measurement errors, calibration will not enable CLM4 to reproduce latent heat fluxes exactly.
519 Rather it would provide a means to quantify parametric uncertainty as prediction intervals. These
520 are elements required for subsequent risk analysis and decision making.

521 **4. Discussion**

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

528 In this study, we assess the feasibility of calibrating CLM4 parameters at flux tower sites
529 with various soil and climate conditions using a surrogate-based Bayesian model calibration
530 procedure. The procedure starts with building surrogates using CLM4 simulations driven by
531 perturbed parameter sets using a space-filling quasi-MC sampling approach. The surrogates
532 provide simplified yet reliable relationships between dominant hydrological parameters (e.g,
533 F_{drai}, Q_{dm}, S_y, and B) and response variables such as latent heat fluxes. The surrogates, after
534 careful validation and selection, are then used as computationally efficient alternatives to the
535 CLM numerical simulator, for improving the estimates of the hydrological parameters, and
536 therefore LH predictions, with quantified uncertainties. This procedure had been demonstrated to
537 be effective at two of the 12 selected sites in a previous study [*Ray et al.*, 2015].

538 However, by extending the same technique to more sites, we acknowledge that there are
539 limitations in the previous version of the procedure [*Ray et al.*, 2015] in that the parameter space
540 confined within prescribed bounds contains non-physical parameter sets as demonstrated in other
541 studies [*Sargsyan et al.*, 2014]. Therefore a classifier is needed to separate the parameter space
542 into physical/non-physical portions, and serve as an informative prior (a joint PDF) before a
543 Bayesian calibration could be performed. The posterior distribution, again a joint PDF, is
544 obtained by inverting against climatologically-averaged latent heat fluxes derived from
545 observations. The posterior distribution provides a complete quantification of uncertainty in the
546 parameter estimates.

547 We find that the simulated mean latent heat fluxes from CLM4 using the calibrated
548 parameters are generally improved at all sites when compared to those from CLM4 simulations
549 using default parameter sets. Those sites with similar soil texture (e.g., loam) and PFTs share
550 similar posterior PDFs of the parameters, which indicate certain levels of parameter
551 transferability between these sites (i.e., as shown in Figure 5). Nevertheless, the number of sites
552 (i.e., 12) investigated is too small for evaluating model parameter transferability, and we would
553 like to leave it as a topic to be investigated in the future by applying the method to more flux
554 tower sites. On the other hand, it is worth mentioning that model parameter transferability among
555 431 watersheds in the United States has been investigated in a separate study, in which
556 sensitivity analysis results are used to classify the watersheds into different classes by grouping
557 basins with similar parameter significance patterns. Such a parameter-sensitivity-based
558 classification system helps reduce the complex climate/hydrologic system into subsets of more
559 homogeneous and smaller systems, and provides necessary information to setup the parameter

560 estimation or model calibration problem. Interested readers are referred to *Ren et al.* [2015] for
561 details.

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

574 In addition to the validation shown in Figure 6, validation is also done by checking if the
575 posterior PDFs are useful for predicting LH during testing time periods using calibrations from
576 training time periods. This validation is supplementary to the one shown in Figure 6; however, it
577 could be misleading as it depends on the reliability of inversion itself and also requires that the
578 favorable parameter values do not vary year to year. For example, Figure S12 shows the
579 calibration results for the loamy needleleaf site, US-Dk3, using data from the period 2002-2006.
580 The posteriors are very close to those inverted using data from the whole period 1997-2006 (see
581 Figure S3), and the posterior ranges of LH predictions are almost identical. This validation is not

582 preferred for sites with relatively short time periods of observations or with large LH variations
583 from year to year. Unfortunately, most sites investigated in this study have relatively short
584 observational periods when the NACP site synthesis datasets were collected. We expect to be
585 able to evaluate the influence of training/testing periods on model parameters when longer data
586 records from the Ameriflux network become available.

587 **5. Conclusion and future work**

588 This work demonstrates a generalizable procedure that can be adopted for calibrating CLM4
589 under various site and climate conditions using a Bayesian inversion technique integrated with
590 surrogate model development. Surrogate models, as computationally-economic alternatives to
591 the direct CLM simulator, can be successfully developed with 15% threshold for training and
592 testing errors in the climatologically averaged heat fluxes. At all selected flux tower sites, most
593 of the latent heat flux observations lie within the IQR ranges of predictions based on parameter
594 values drawn from the posterior distribution. The procedure can be applied to other models
595 including newer versions of CLM and other components of an Earth system model, given that
596 the metrics for measuring model performance and for defining objective functions can be well-
597 defined. Further, since the calibration is performed using surrogates, the computational cost of
598 the Earth system model (or its component) ceases to be an issue. As demonstrated in this study,
599 the Bayesian calibration procedure could be used as a tool for parameter estimation with
600 uncertainty bounds, as well as for identifying potential model structural errors by extensively
601 exploring the parameter space and comparing discrepancies between model predictions and
602 observations. Such a tool will be valuable for model applications for quantifying prediction
603 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.

627 **References**

628

629 Alexandrov, N. M., J. E. Dennis Jr, R. M. Lewis, and V. Torczon (1998), A trust-region framework for
630 managing the use of approximation models in optimization, *Structural Optimization*, *15*(1), 16-23.
631 Allison, V. J., R. M. Miller, J. D. Jastrow, R. Matamala, and D. R. Zak (2005), Changes in Soil Microbial
632 Community Structure in a Tallgrass Prairie Chronosequence, *Soil Science Society of America*
633 *Journal*, *69*(5), 1412-1421, doi:10.2136/sssaj2004.0252.
634 Alton, P., L. Mercado, and P. North (2006), A sensitivity analysis of the land-surface scheme JULES
635 conducted for three forest biomes: Biophysical parameters, model processes, and meteorological
636 driving data, *Global Biogeochemical Cycles*, *20*(1), n/a-n/a, doi:10.1029/2005GB002653.
637 Baldocchi, D. D., and K. B. Wilson (2001), Modeling CO2 and water vapor exchange of a temperate
638 broadleaved forest across hourly to decadal time scales, *Ecological Modelling*, *142*(1–2), 155-184,
639 doi:http://dx.doi.org/10.1016/S0304-3800(01)00287-3.
640 Bastidas, L. A., T. S. Hogue, S. Sorooshian, H. V. Gupta, and W. J. Shuttleworth (2006), Parameter
641 sensitivity analysis for different complexity land surface models using multicriteria methods, *Journal*
642 *of Geophysical Research: Atmospheres*, *111*(D20), n/a-n/a, doi:10.1029/2005JD006377.
643 Bilionis, I., B. Drewniak, and E. Constantinescu (2015), Crop physiology calibration in the CLM,
644 *Geoscientific Model Development*, *8*(4), 1071-1083.
645 Bonan, G. B., P. J. Lawrence, K. W. Oleson, S. Levis, M. Jung, M. Reichstein, D. M. Lawrence, and S. C.
646 Swenson (2011), Improving canopy processes in the Community Land Model version 4 (CLM4)
647 using global flux fields empirically inferred from FLUXNET data, *Journal of Geophysical Research:*
648 *Biogeosciences*, *116*(G2), n/a-n/a, doi:10.1029/2010JG001593.
649 Clapp, R. B., and G. M. Hornberger (1978), Empirical equations for some soil hydraulic properties, *Water*
650 *Resources Research*, *14*(4), 601-604, doi:10.1029/WR014i004p00601.
651 Cowles, M. K., and B. P. Carlin (1996), Markov Chain Monte Carlo Convergence Diagnostics: A
652 Comparative Review, *Journal of the American Statistical Association*, *91*(434), 883-904,
653 doi:10.2307/2291683.
654 Demaria, E. M., B. Nijssen, and T. Wagener (2007), Monte Carlo sensitivity analysis of land surface
655 parameters using the Variable Infiltration Capacity model, *Journal of Geophysical Research:*
656 *Atmospheres*, *112*(D11), n/a-n/a, doi:10.1029/2006JD007534.
657 Eldred, M., and D. Dunlavy (2006), Formulations for surrogate-based optimization with data fit,
658 multifidelity, and reduced-order models, paper presented at Proceedings of the 11th AIAA/ISSMO
659 Multidisciplinary Analysis and Optimization Conference, number AIAA-2006-7117, Portsmouth,
660 VA.
661 Fischer, M. L., D. P. Billesbach, J. A. Berry, W. J. Riley, and M. S. Torn (2007), Spatiotemporal
662 Variations in Growing Season Exchanges of CO2, H2O, and Sensible Heat in Agricultural Fields of
663 the Southern Great Plains, *Earth Interactions*, *11*(17), 1-21, doi:10.1175/EI231.1.
664 Forrester, A. I., and A. J. Keane (2009), Recent advances in surrogate-based optimization, *Progress in*
665 *Aerospace Sciences*, *45*(1), 50-79.
666 Friedman, J. H. (1991), Multivariate adaptive regression splines, *The annals of statistics*, 1-67.
667 Gan, Y., X.-Z. Liang, Q. Duan, H. I. Choi, Y. Dai, and H. Wu (2015), Stepwise sensitivity analysis from
668 qualitative to quantitative: Application to the terrestrial hydrological modeling of a Conjunctive
669 Surface-Subsurface Process (CSSP) land surface model, *Journal of Advances in Modeling Earth*
670 *Systems*, n/a-n/a, doi:10.1002/2014MS000406.
671 Gao, W., R. L. Coulter, B. M. Lesht, J. Qiu, and M. L. Wesely (1998), Estimating Clear-Sky Regional
672 Surface Fluxes in the Southern Great Plains Atmospheric Radiation Measurement Site with Ground
673 Measurements and Satellite Observations, *Journal of Applied Meteorology*, *37*(1), 5-22,
674 doi:10.1175/1520-0450(1998)037<0005:ECSRSF>2.0.CO;2.

675 Gilks, W., S. Richardson, and D. Spiegelhalter (1996), *Markov chain Monte Carlo in practice*, Chapman
676 & Hall.
677 Giunta, A. A., V. Balabanov, S. Burgee, B. Grossman, R. T. Haftka, W. H. Mason, and L. T. Watson
678 (1995), Variable-complexity multidisciplinary design optimization using parallel computers, in
679 *Computational Mechanics' 95*, edited, pp. 489-494, Springer.
680 Goel, T., R. T. Haftka, W. Shyy, and N. V. Queipo (2007), Ensemble of surrogates, *Structural and*
681 *Multidisciplinary Optimization*, *33*(3), 199-216.
682 Göhler, M., J. Mai, and M. Cuntz (2013), Use of eigendecomposition in a parameter sensitivity analysis
683 of the Community Land Model, *Journal of Geophysical Research: Biogeosciences*, *118*(2), 904-921,
684 doi:10.1002/jgrg.20072.
685 Gong, W., Q. Duan, J. Li, C. Wang, Z. Di, Y. Dai, A. Ye, and C. Miao (2015), Multi-objective parameter
686 optimization of common land model using adaptive surrogate modeling, *Hydrology and Earth System*
687 *Sciences*, *19*(5), 2409-2425.
688 Gu, L., T. Meyers, S. G. Pallardy, P. J. Hanson, B. Yang, M. Heuer, K. P. Hosman, J. S. Riggs, D. Sluss,
689 and S. D. Wullschleger (2006), Direct and indirect effects of atmospheric conditions and soil moisture
690 on surface energy partitioning revealed by a prolonged drought at a temperate forest site, *Journal of*
691 *Geophysical Research: Atmospheres*, *111*(D16), n/a-n/a, doi:10.1029/2006JD007161.
692 Gulden, L. E., E. Rosero, Z.-L. Yang, T. Wagener, and G.-Y. Niu (2008), Model performance, model
693 robustness, and model fitness scores: A new method for identifying good land-surface models,
694 *Geophysical Research Letters*, *35*(11), n/a-n/a, doi:10.1029/2008GL033721.
695 Haario, H., M. Laine, A. Mira, and E. Saksman (2006), DRAM: Efficient adaptive MCMC, *Stat Comput*,
696 *16*(4), 339-354, doi:10.1007/s11222-006-9438-0.
697 Hastie, T., R. Tibshirani, and J. Friedman (2009), *The elements of statistical learning – Data mining,*
698 *inference and prediction*, Springer.
699 Henderson-Sellers, A., A. J. Pitman, P. K. Love, P. Irannejad, and T. H. Chen (1995), The Project for
700 Intercomparison of Land Surface Parameterization Schemes (PILPS): Phases 2 and 3, *Bulletin of the*
701 *American Meteorological Society*, *76*(4), 489-503, doi:10.1175/1520-
702 0477(1995)076<0489:TPFIOL>2.0.CO;2.
703 Hollinger, D. Y., S. M. Goltz, E. A. Davidson, J. T. Lee, K. Tu, and H. T. Valentine (1999), Seasonal
704 patterns and environmental control of carbon dioxide and water vapour exchange in an ecotonal
705 boreal forest, *Global Change Biology*, *5*(8), 891-902, doi:10.1046/j.1365-2486.1999.00281.x.
706 Hou, Z., M. Huang, L. R. Leung, G. Lin, and D. M. Ricciuto (2012), Sensitivity of surface flux
707 simulations to hydrologic parameters based on an uncertainty quantification framework applied to the
708 Community Land Model, *Journal of Geophysical Research: Atmospheres*, *117*(D15), n/a-n/a,
709 doi:10.1029/2012JD017521.
710 Huang, M., Z. Hou, L. R. Leung, Y. Ke, Y. Liu, Z. Fang, and Y. Sun (2013), Uncertainty Analysis of
711 Runoff Simulations and Parameter Identifiability in the Community Land Model – Evidence from
712 MOPEX Basins, *Journal of Hydrometeorology*.
713 Kaipio, J., and E. Somersalo (2006), *Statistical and Computational Inverse Problems*, Springer New
714 York.
715 Lawrence, D. M., et al. (2011), Parameterization improvements and functional and structural advances in
716 Version 4 of the Community Land Model, *Journal of Advances in Modeling Earth Systems*, *3*(3), n/a-
717 n/a, doi:10.1029/2011MS000045.
718 Liang, X., and J. Guo (2003), Intercomparison of land-surface parameterization schemes: sensitivity of
719 surface energy and water fluxes to model parameters, *Journal of Hydrology*, *279*(1–4), 182-209,
720 doi:http://dx.doi.org/10.1016/S0022-1694(03)00168-9.
721 Lo, M. H., J. S. Famiglietti, P. F. Yeh, and T. Syed (2010), Improving parameter estimation and water
722 table depth simulation in a land surface model using GRACE water storage and estimated base flow
723 data, *Water Resources Research*, *46*(5).

Luo, H., W. C. Oechel, S. J. Hastings, R. Zulueta, Y. Qian, and H. Kwon (2007), Mature semiarid chaparral ecosystems can be a significant sink for atmospheric carbon dioxide, *Global Change Biology*, *13*(2), 386-396, doi:10.1111/j.1365-2486.2006.01299.x.

Luo, Y. Q., et al. (2012), A framework for benchmarking land models, *Biogeosciences*, *9*(10), 3857-3874, doi:10.5194/bg-9-3857-2012.

Meyer, D., E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch (2014), e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-3, edited.

Müller, J., R. Paudel, C. Shoemaker, J. Woodbury, Y. Wang, and N. Mahowald (2015), CH 4 parameter estimation in CLM4. 5bgc using surrogate global optimization, *Geoscientific Model Development Discussions*, *8*(1), 141-207.

Oishi, A. C., R. Oren, and P. C. Stoy (2008), Estimating components of forest evapotranspiration: A footprint approach for scaling sap flux measurements, *Agricultural and Forest Meteorology*, *148*(11), 1719-1732, doi:http://dx.doi.org/10.1016/j.agrformet.2008.06.013.

Oren, R. A. M., C.-I. Hsieh, P. Stoy, J. Albertson, H. R. McCarthy, P. Harrell, and G. G. Katul (2006), Estimating the uncertainty in annual net ecosystem carbon exchange: spatial variation in turbulent fluxes and sampling errors in eddy-covariance measurements, *Global Change Biology*, *12*(5), 883-896, doi:10.1111/j.1365-2486.2006.01131.x.

Pau, G. S. H., G. Bisht, and W. J. Riley (2014), A reduced-order modeling approach to represent subgrid-scale hydrological dynamics for land-surface simulations: application in a polygonal tundra landscape, *Geosci. Model Dev.*, *7*(5), 2091-2105, doi:10.5194/gmd-7-2091-2014.

Prihodko, L., A. Denning, N. Hanan, I. Baker, and K. Davis (2008), Sensitivity, uncertainty and time dependence of parameters in a complex land surface model, *agricultural and forest meteorology*, *148*(2), 268-287.

Qian, Y., M. Huang, B. Yang, and L. K. Berg (2013), A Modeling Study of Irrigation Effects on Surface Fluxes and Land–Air–Cloud Interactions in the Southern Great Plains, *Journal of Hydrometeorology*, *14*(3), 700-721, doi:10.1175/JHM-D-12-0134.1.

Raftery, A., and S. Lewis (1995), The number of iterations, convergence diagnostics and generic Metropolis algorithms, in *Practical Markov Chain Monte Carlo (W.R. Gilks, D.J. Spiegelhalter and S. Richardson, eds.)*, edited, pp. 115-130, Chapman and Hall, London, U.K.

Ray, J., Z. Hou, M. Huang, K. Sargsyan, and L. Swiler (2015), Bayesian Calibration of the Community Land Model Using Surrogates, *SIAM/ASA Journal on Uncertainty Quantification*, 199-233, doi:10.1137/140957998.

Razavi, S., B. A. Tolson, and D. H. Burn (2012), Review of surrogate modeling in water resources, *Water Resources Research*, *48*(7).

Regis, R. G., and C. A. Shoemaker (2007), A stochastic radial basis function method for the global optimization of expensive functions, *INFORMS Journal on Computing*, *19*(4), 497-509.

Ren, H., Z. Hou, M. Huang, J. Bao, Y. Sun, T. Tesfa, and L. R. Leung (2015), Classification of Hydrological Parameter Sensitivity and Evaluation of Parameter Transferability across 431 US MOPEX, *Journal of Hydrology*, *Accepted*.

Rosero, E., Z.-L. Yang, T. Wagener, L. E. Gulden, S. Yatheendradas, and G.-Y. Niu (2010), Quantifying parameter sensitivity, interaction, and transferability in hydrologically enhanced versions of the Noah land surface model over transition zones during the warm season, *Journal of Geophysical Research: Atmospheres*, *115*(D3), n/a-n/a, doi:10.1029/2009JD012035.

Sargsyan, K., C. Safta, H. N. Najm, B. J. Debusschere, D. Ricciuto, and P. Thornton (2014), DIMENSIONALITY REDUCTION FOR COMPLEX MODELS VIA BAYESIAN COMPRESSIVE SENSING, *4*(1), 63-93, doi:10.1615/Int.J.UncertaintyQuantification.2013006821.

Schwalm, C. R., et al. (2010), A model-data intercomparison of CO2 exchange across North America: Results from the North American Carbon Program site synthesis, *Journal of Geophysical Research: Biogeosciences*, *115*(G3), G00H05, doi:10.1029/2009JG001229.

773 Scrucca, L. (2013), GA: A Package for Genetic Algorithms in R, *Journal of Statistical Software*, *53*(4), 1-
774 37.
775 Stoy, P. C., G. G. Katul, M. B. S. Siqueira, J.-Y. Juang, K. A. Novick, H. R. McCarthy, A. C. Oishi, and
776 R. A. M. Oren (2008), Role of vegetation in determining carbon sequestration along ecological
777 succession in the southeastern United States, *Global Change Biology*, *14*(6), 1409-1427,
778 doi:10.1111/j.1365-2486.2008.01587.x.
779 Stoy, P. C., G. G. Katul, M. B. S. Siqueira, J.-Y. Juang, K. A. Novick, J. M. Uebelherr, and R. Oren
780 (2006), An evaluation of models for partitioning eddy covariance-measured net ecosystem exchange
781 into photosynthesis and respiration, *Agricultural and Forest Meteorology*, *141*(1), 2-18,
782 doi:http://dx.doi.org/10.1016/j.agrformet.2006.09.001.
783 Sun, Y., Z. Hou, M. Huang, F. Tian, and L. Ruby Leung (2013), Inverse modeling of hydrologic
784 parameters using surface flux and runoff observations in the Community Land Model, *Hydrology and*
785 *Earth System Sciences*, *17*(12), 4995-5011.
786 Suyker, A. E., and S. B. Verma (2010), Coupling of carbon dioxide and water vapor exchanges of
787 irrigated and rainfed maize–soybean cropping systems and water productivity, *Agricultural and*
788 *Forest Meteorology*, *150*(4), 553-563, doi:http://dx.doi.org/10.1016/j.agrformet.2010.01.020.
789 Suyker, A. E., S. B. Verma, and G. G. Burba (2003), Interannual variability in net CO2 exchange of a
790 native tallgrass prairie, *Global Change Biology*, *9*(2), 255-265, doi:10.1046/j.1365-
791 2486.2003.00567.x.
792 Tian, X., and Z. Xie (2008), A land surface soil moisture data assimilation framework in consideration of
793 the model subgrid-scale heterogeneity and soil water thawing and freezing, *Science in China Series*
794 *D: Earth Sciences*, *51*(7), 992-1000.
795 Urbanski, S., C. Barford, S. Wofsy, C. Kucharik, E. Pyle, J. Budney, K. McKain, D. Fitzjarrald, M.
796 Czikowsky, and J. W. Munger (2007), Factors controlling CO2 exchange on timescales from hourly
797 to decadal at Harvard Forest, *Journal of Geophysical Research: Biogeosciences*, *112*(G2), n/a-n/a,
798 doi:10.1029/2006JG000293.
799 Verma, S. B., et al. (2005), Annual carbon dioxide exchange in irrigated and rainfed maize-based
800 agroecosystems, *Agricultural and Forest Meteorology*, *131*(1–2), 77-96,
801 doi:http://dx.doi.org/10.1016/j.agrformet.2005.05.003.
802 Viana, F. A., T. W. Simpson, V. Balabanov, and V. Toropov (2014), Metamodeling in multidisciplinary
803 design optimization: how far have we really come?, *AIAA Journal*, *52*(4), 670-690.
804 Wang, G. G., Z. Dong, and P. Aitchison (2001), Adaptive response surface method--a global optimization
805 scheme for approximation-based design problems, *Engineering Optimization*, *33*(6), 707-734.
806 White, M. A., P. E. Thornton, S. W. Running, and R. R. Nemani (2000), Parameterization and Sensitivity
807 Analysis of the BIOME–BGC Terrestrial Ecosystem Model: Net Primary Production Controls, *Earth*
808 *Interactions*, *4*(3), 1-85, doi:10.1175/1087-3562(2000)004<0003:PASAOT>2.0.CO;2.
809 Williams, M., et al. (2009), Improving land surface models with FLUXNET data, *Biogeosciences*, *6*(7),
810 1341-1359, doi:10.5194/bg-6-1341-2009.
811 Wilson, K., et al. (2002), Energy balance closure at FLUXNET sites, *Agricultural and Forest*
812 *Meteorology*, *113*(1–4), 223-243, doi:http://dx.doi.org/10.1016/S0168-1923(02)00109-0.
813 Yang, B. A. I., S. G. Pallardy, T. P. Meyers, L.-H. Gu, P. J. Hanson, S. D. Wullschleger, M. Heuer, K. P.
814 Hosman, J. S. Riggs, and D. W. Sluss (2010), Environmental controls on water use efficiency during
815 severe drought in an Ozark Forest in Missouri, USA, *Global Change Biology*, *16*(8), 2252-2271,
816 doi:10.1111/j.1365-2486.2009.02138.x.
817 Yu, X., and M. Gen (2010), *Introduction to Evolutionary Algorithms*, Springer London.
818 Zeng, X., B. A. Drewniak, and E. M. Constantinescu (2013), Calibration of the Crop model in the
819 Community Land Model, *Geoscientific Model Development Discussions*, *6*(1), 379-398.
820 Zobitz, J. M., J. P. Keener, H. Schnyder, and D. R. Bowling (2006), Sensitivity analysis and
821 quantification of uncertainty for isotopic mixing relationships in carbon cycle research, *Agricultural*
822 *and Forest Meteorology*, *136*(1–2), 56-75, doi:http://dx.doi.org/10.1016/j.agrformet.2006.01.003.
823

824 **Tables:**

825 Table 1. Site characteristics of the 12 flux towers.

	Site	Longitude	Latitude	Elev	Soil texture	Plant functional type	Period	Reference
1	US-Ha1	-72.1715	42.5378	343.00	sandy loam	Deciduous Broadleaf	1991-2006	*Urbanski et al.* [2007]
2	US-Dk2	-79.1004	35.9736	163.00	sandy clay loam	Deciduous Broadleaf	2003-2005	[*Oishi et al.* [2008]; *Stoy et al.* [2008]]
3	US-Dk3	-79.0942	35.9782	163.00	loam	Evergreen Needleleaf	1998-2005	[*Oren et al.* [2006]; *Stoy et al.* [2006]]
4	US-IB1	-88.2227	41.8593	225.00	silty clay loam	Croplands	2005-2007	*Allison et al.* [2005]
5	US-IB2	-88.2410	41.8406	226.00	silty clay loam	Grasslands	2004-2007	*Allison et al.* [2005]
6	US-Shd	-96.6827	36.9601	350.00	silty clay loam	Grasslands	1997-2000	*Suyker et al.* [2003]
7	US-SO2	-116.6230	33.3739	1406.00	loam	Closed Shrublands	1998-2006	*H Luo et al.* [2007]
8	US-Ne3	-96.4397	41.1797	363.00	clay loam	Croplands	2001-2006	[*Suyker and Verma* [2010]; *Verma et al.* [2005]]
9	US-Ho1	-68.7403	45.2041	79.00	loam	Evergreen Needleleaf	2004-2007	*Hollinger et al.* [1999]
10	US-MOz	-92.2000	38.7441	219.00	loam	Deciduous Broadleaf	2004-2007	[*Gu et al.* [2006]; *Yang et al.* [2010]]
11	US-ARM	-97.4884	36.6050	311.00	clay	Croplands	2000-2007	*Fischer et al.* [2007]
12	US-Wlr	-96.8550	37.5208	408.00	silty clay	Grasslands	2001-2004	*Gao et al.* [1998]

826

827 Table 2. p_{opt} computed using different QRMSE and surrogate models of different qualities.

Q_{RMSE}	Surrogate model accuracy	$\mathbf{p}_{opt} = \{F_{drai}, \ln(Q_{dm}), S_y\}$
0.65	All 12 surrogates have < 10% error	{5.00, -13.46, 0.27}
0.70	2 out of 12 surrogates have errors between 10% and 15%	{0.80, -13.77, 0.27}
0.75	3 out of 12 surrogates have errors between 10% and 20%. Surrogates for summer months have the largest errors	{1.33, -13.78, 0.27}
0.85	All surrogates have errors above 10%	{1.09, -13.77, 0.27}

828

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.

CLM parameter	**F_{drai}**				**$\ln(Q_{dm})$**				**S_y or B**				**σ**			
Sites	MAP	Q2.5	median	Q97.5	MAP	Q2.5	median	Q97.5	MAP	Q2.5	median	Q97.5	MAP	Q2.5	median	Q97.5
US-Ha1	4.62	0.19	2.70	4.93	-12.70	-13.64	-9.57	-4.90	0.15	0.14	0.21	0.40	9.55	5.68	11.82	29.77
US-DK2	4.62	0.19	2.80	4.93	-10.88	-13.54	-9.49	-4.96	0.11	0.09	0.17	0.26	47.91	29.02	60.26	159.08
US-DK3	4.78	2.35	4.40	4.97	-13.53	-13.78	-12.92	-9.10	0.09	0.09	0.11	0.19	49.87	28.08	64.93	195.98
US-IB1	0.45	0.14	1.56	4.87	-13.15	-13.73	-10.88	-4.90	0.26	0.11	0.22	0.27	257.10	149.05	324.96	854.62
US-IB2	1.76	0.75	2.74	4.84	-12.97	-13.64	-9.59	-4.88	0.23	0.11	0.21	0.26	24.38	14.91	31.38	81.85
US-Shd	2.90	1.26	3.13	4.88	-8.30	-9.74	-7.38	-4.74	0.11	0.09	0.12	0.15	25.45	14.76	31.69	83.95
US-SO2	0.66	0.21	2.27	4.85	-11.08	-13.38	-9.38	-4.90	0.19	0.11	0.18	0.26	14.71	9.25	19.07	50.28
US-Ne3	4.28	0.34	3.33	4.91	-13.04	-13.70	-9.97	-4.89	0.27	0.17	0.25	0.27	372.85	222.77	459.19	1194.72
US-Ho1	4.56	3.12	4.26	4.96	-13.55	-13.78	-12.99	-11.23	0.11	0.09	0.11	0.14	10.33	6.09	12.98	35.54
US-MOz	0.34	0.12	0.82	4.75	-9.87	-13.46	-9.52	-5.20	0.21	0.11	0.21	0.27	41.31	23.83	51.38	143.37
US-ARM	0.20	0.11	0.40	1.57	-4.95	-9.38	-5.62	-4.64	1.05	1.00	1.12	1.86	130.89	95.09	140.11	215.80
US-Wlr	0.32	0.12	0.60	4.90	-5.52	-13.08	-7.70	-4.77	1.25	1.02	1.64	14.40	175.98	88.22	198.66	547.14

Note: the third parameter is B for US-ARM and US-Wlr, and Sy otherwise.

Figure captions:

Figure 1. CLM4 parameters (F_{drai}, $\ln(Q_{dm})$, 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_{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 $\mathbf{p}_{opt}$ 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 Sy.

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:

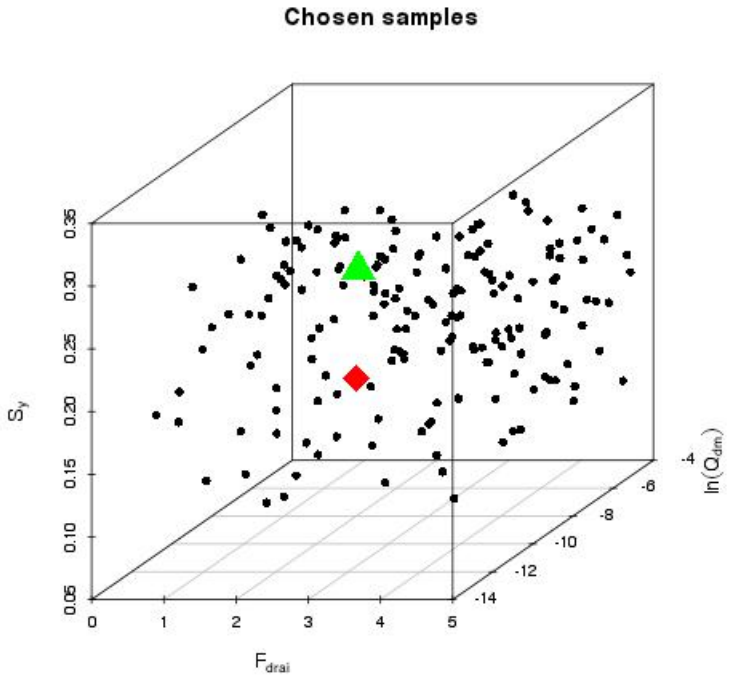

Figure 2. CLM4 parameters (F_{drai}, $\ln(Q_{dm})$, S_y), from the training data, that lie inside R. The red diamond plots the nominal value, the green triangle the parameter combination in the training set for US-IB1 with the best agreement with observations.

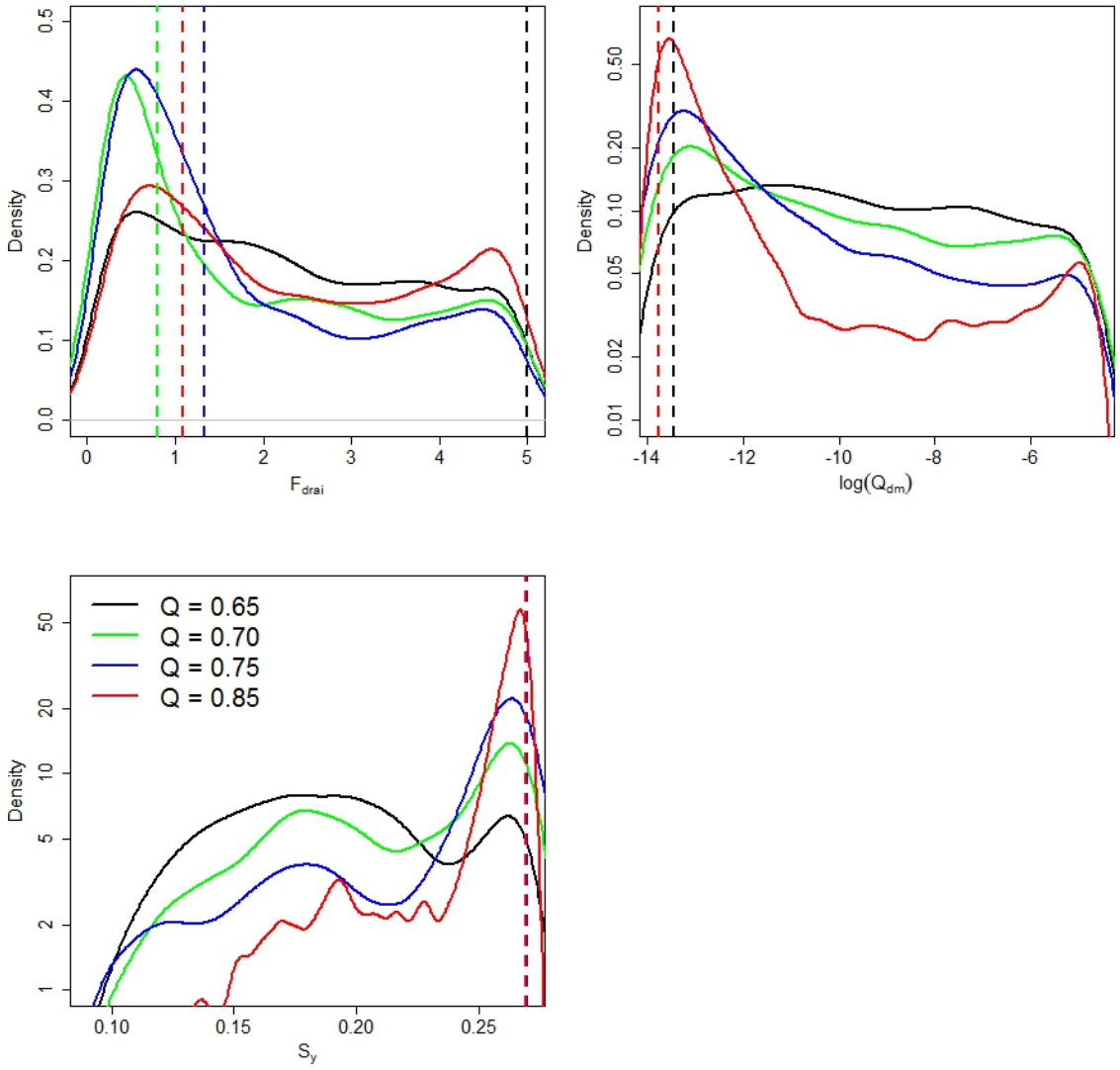

Figure 2. PDFs of the three parameters for each Q_{RMSE} with the GA estimate plotted as vertical lines. In the top right and bottom left sub-figures, the vertical lines showing the values of $\mathbf{p}_{opt}$ coincide and are thus obscured.

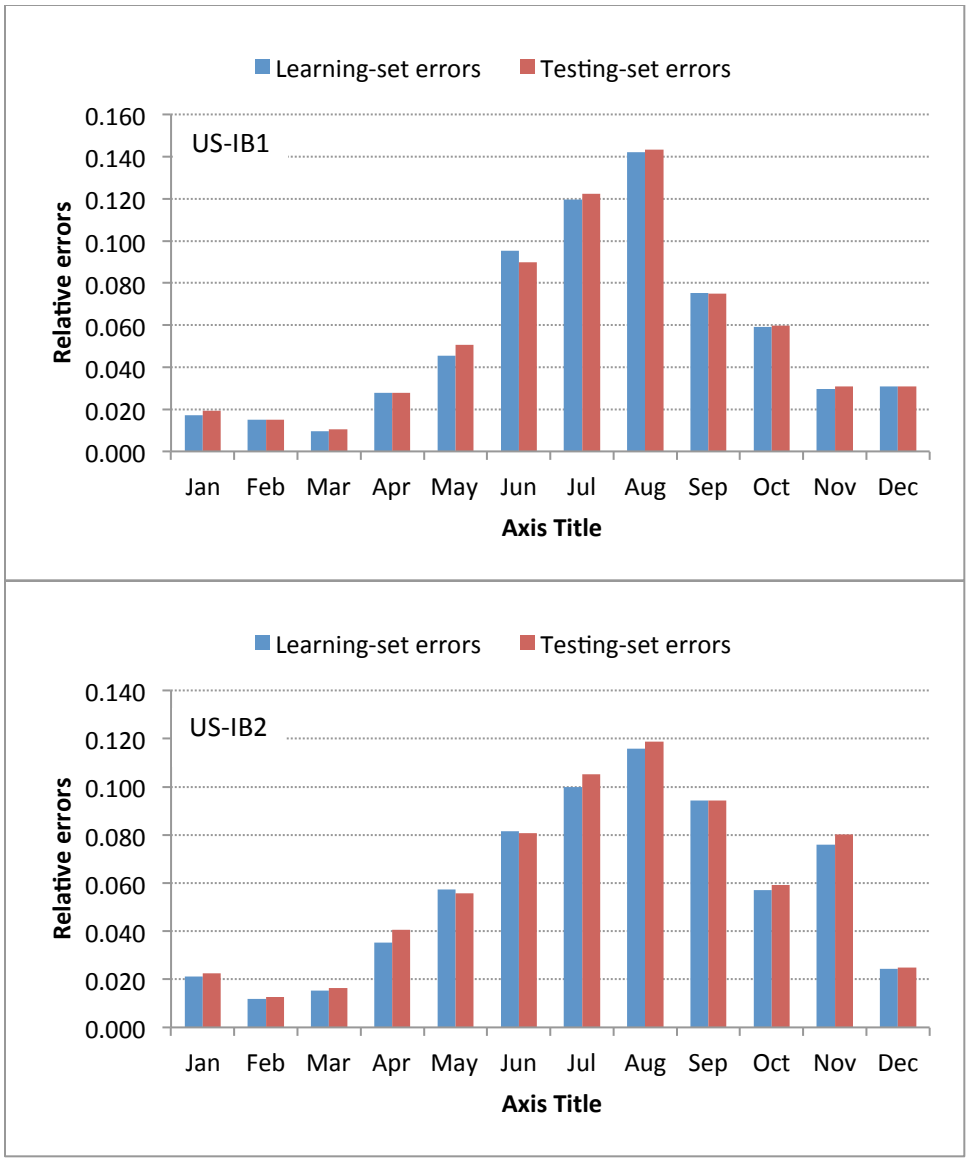

Figure 3. Learning-set and testing-set relative predictive errors for surrogate validation at two selected example sites (US-IB1 and IB2).

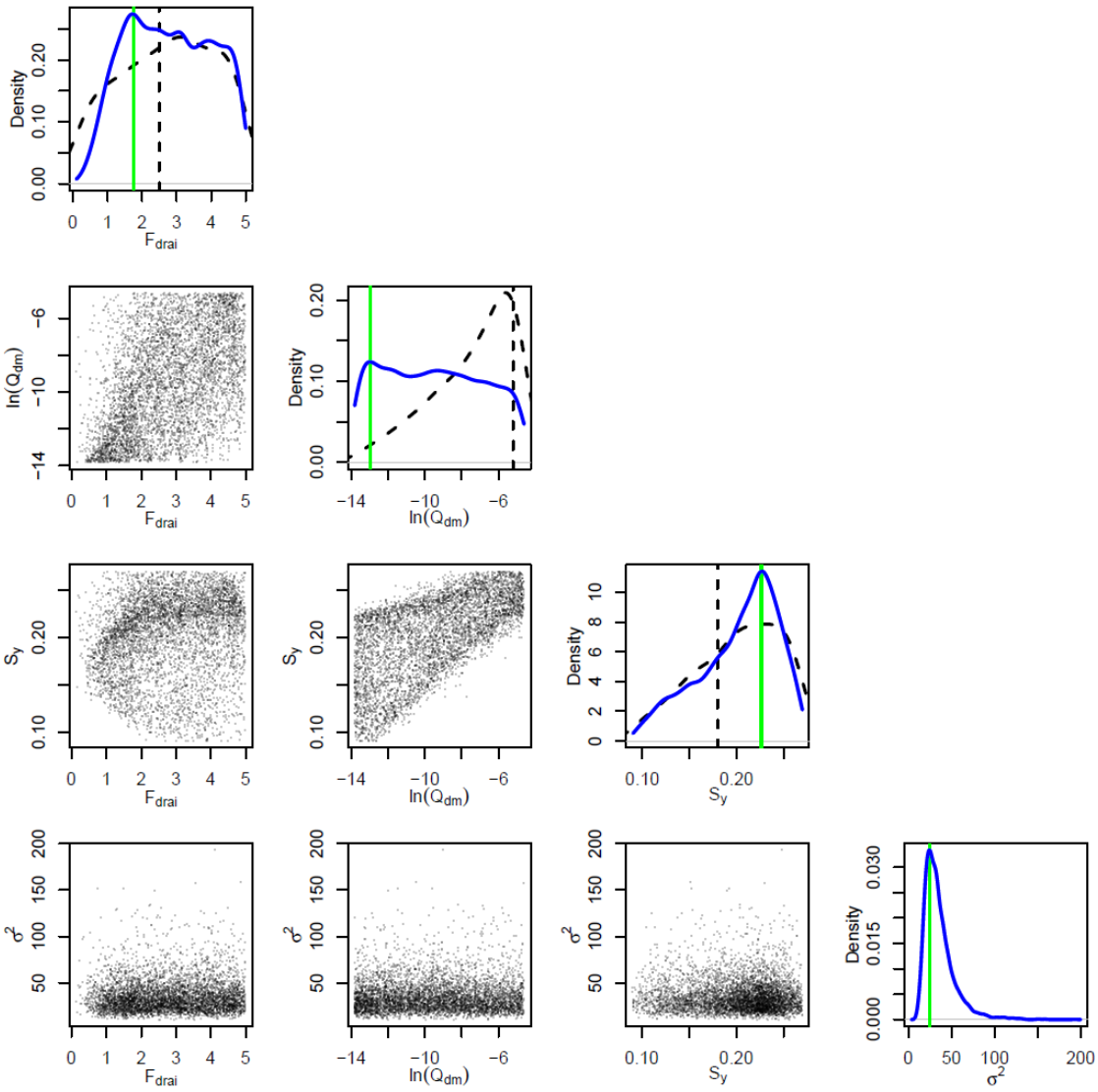

Figure 4. Marginal distribution of the joint prior (dashed), posterior (solid) PDFs, the default values (dashed vertical line), and the maximum a posteriori (MAP) values (solid vertical line), and paired scatters of posterior samples of the four parameters for inversion at US-IB2.

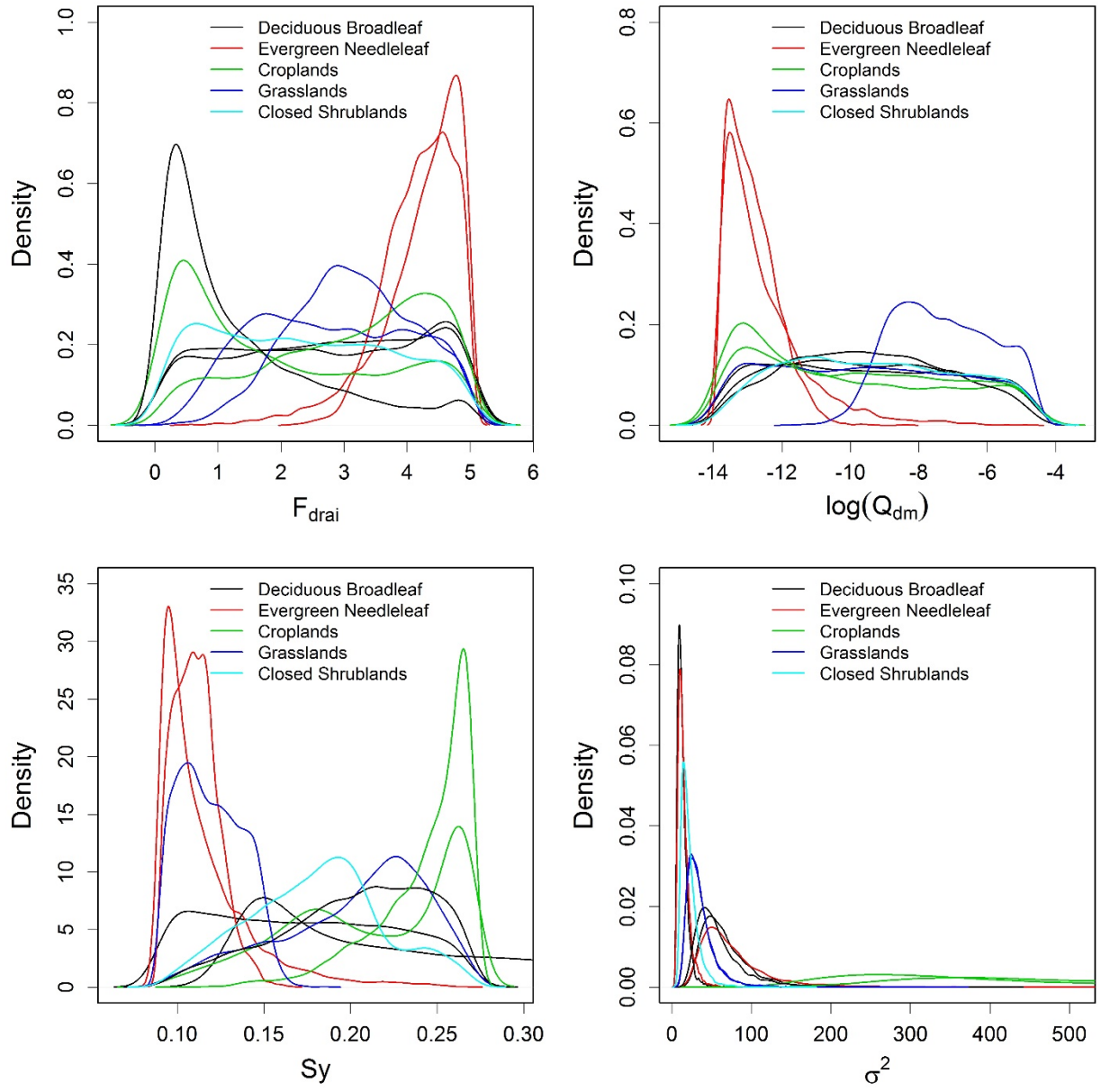

Figure 5. Posterior distributions of inverted parameters color-coded by Plant Functional Types for 10 out of 12 sites that share the same parameters. It is evident that the two Evergreen Needleleaf sites have very similar PDFs for all three parameters. Croplands share similar estimates for S_y.

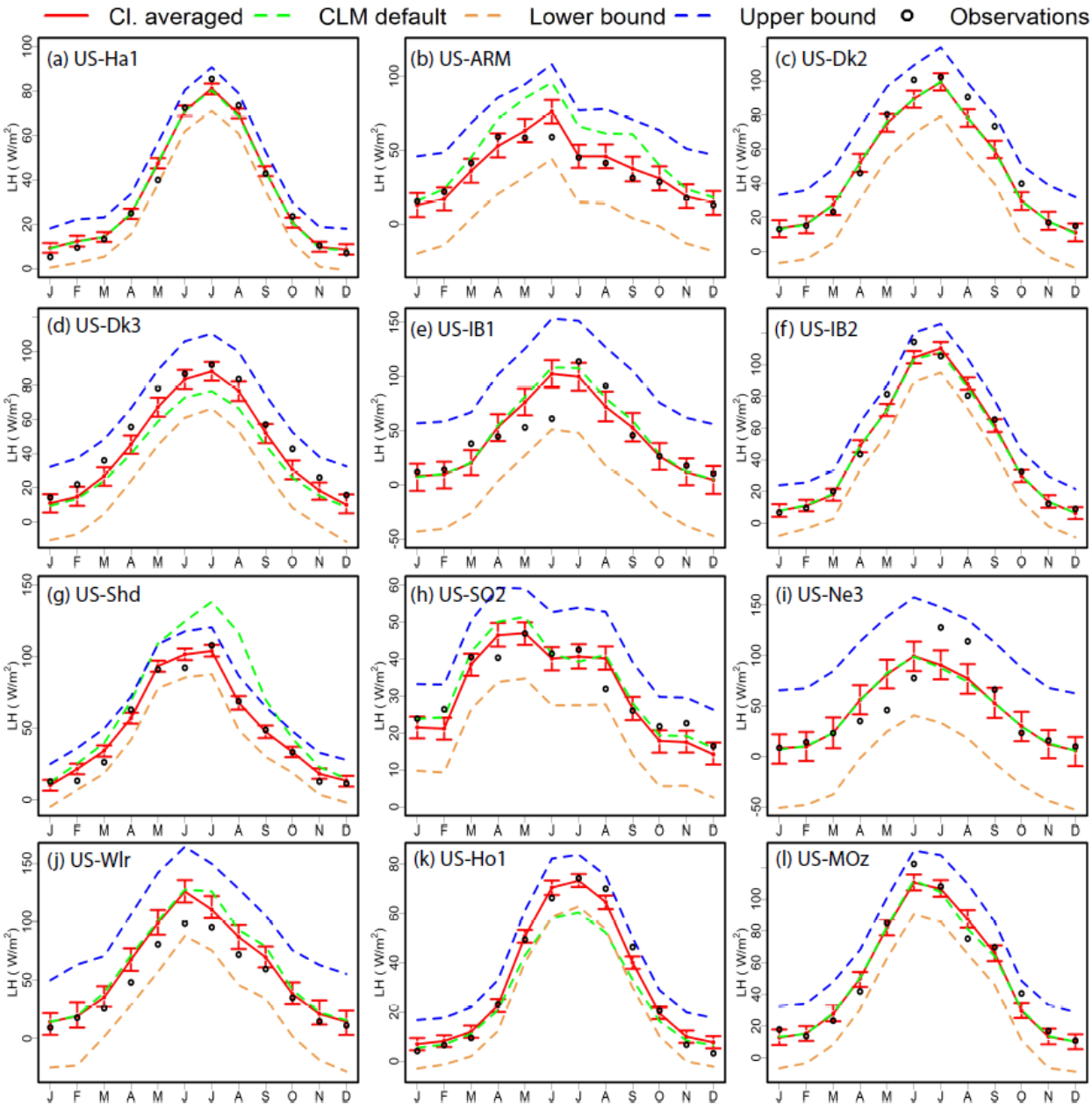

Figure 6. Validation of posterior parameter using by CLM4. The symbols are the monthly-mean observed LH fluxes, climatologically averaged over the durations tabulated in Table 1. The line with the error bound is the median prediction from the ensemble of runs seeded with samples from the posterior distribution. 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.