SIAM/ASA J. Uncertainty Quantification
Vol. xx, pp. x

x–x

BAYESIAN CALIBRATION OF THE COMMUNITY LAND MODEL USING SURROGATES[*]

J. Ray[†][¶], Z. Hou[‡], M. Huang,[‡] K. Sargsyan[†] and L. Swiler[§]

Abstract. We present results from the Bayesian calibration of hydrological parameters of the Community Land Model (CLM), which is often used in climate simulations and Earth system models. A statistical inverse problem is formulated for three hydrological parameters, conditioned on observations of latent heat surface fluxes over 48 months. Our calibration method uses polynomial and Gaussian process surrogates of the CLM, and solves the parameter estimation problem using a Markov chain Monte Carlo sampler. Posterior probability densities for the parameters are developed for two sites with different soil and vegetation covers. Our method also allows us to examine the structural error in CLM under two error models. We find that accurate surrogate models could be created for CLM in three out of the four cases we investigated. The posterior distributions lead to better prediction than the default parameter values in CLM. Climatologically averaging the observations does not modify the parameters' distributions significantly. The structural error model reveals a correlation time-scale which can potentially be used to identify physical processes that could be contributing to it. While the calibrated CLM has a higher predictive skill, the calibration is under-dispersive.

Key words. Bayesian calibration, Community Land Model, surrogate models, structural error models, Markov chain Monte Carlo

AMS subject classifications. 62F15, 62F25, 62F86

1. Introduction.

The Community Land Model (CLM, [36]), the land component of the Community Earth System Model (CESM, [10]), is used to simulate terrestrial water, energy, and biogeochemical processes in offline and coupled climate simulations. The CLM contains a large number of parameters that govern its behavior, many of which are not directly measurable. They are estimated from indirect measurements, and are therefore subject to great uncertainty. Further, many parameters are site-dependent i.e., they vary within certain ranges [17, 22, 23]. In addition, due to difficulties in estimating such parameters at a global scale, CLM is released with default values for these parameters obtained by benchmarking its simulations against global datasets using simple statistics [11]. The predictive accuracy of CLM is, to a large degree, dependent on obtaining "correct" values of these parameters, and calibrating to site-specific observational data is the best means of doing so. Model calibration, to date, has meant optimizing parameter values to reduce the discrepancies between historical observations and their corresponding model predictions (e.g., from CLM). This leads to a number of practical challenges. For example, gradient-descent optimization methods e.g., L-BFGS-B [6] are sensitive to their starting guesses and can yield multiple "optimal" parameter combinations. More seriously, due to the limited amount of observational data, the measurement errors in observations, and the modeling shortcomings/simplifications in CLM, parameters cannot be estimated with a high degree of accuracy. As a result, the parameter estimates are uncertain, but such parametric uncertainty has not been well quantified. Consequently, CLM is not distributed with "error bounds" that reflect parametric uncertainty after calibration.

[*] This work was funded by the U.S. Department of Energy, via the Office of Advanced Scientific Computing Research (OASCR) and Biological and Environmental Research (BER). Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. Pacific Northwest National Laboratory is operated by Battelle Memorial Institute for the U.S. Department of Energy under contract DE-AC05-76RLO1830.

[†] Sandia National Laboratories, Livermore, CA

[‡] Pacific Northwest National Laboratory, Richland, WA

[§] Sandia National Laboratories, Albuquerque, NM

[¶] Corresponding author; jairay@sandia.gov

The problem of parametric uncertainty can be addressed using Bayesian calibration. It develops parameter estimates as probability density functions (PDFs). The PDFs can be general i.e. we do not have to stipulate a canonical family of distributions like Gaussian, log-normal etc. or make any approximations in the numerical scheme if the Bayesian calibration problem is solved using a Markov chain Monte Carlo (MCMC) method. The PDF captures parametric uncertainty and the correlation between parameter estimates concisely. Further, such a calibration also improves the predictive skill of CLM; instead of attempting to predict observations with one "optimal" parameter combination, one samples the PDF and constructs an ensemble of CLM predictions. Simple statistical measures [16, 15] can be used to summarize the "goodness of fit"; further, the statistical measures also reveal other aspects of the fit (e.g., over-/under-dispersive calibrations) that provide specific directions to pursue to improve CLM. However, Bayesian calibration poses two technical challenges. Firstly, like contemporary optimization methods, Bayesian calibration minimizes the model-observation discrepancy. In addition, it also requires one to specify a statistical model for the discrepancy (henceforth called the structural error model). The sensitivity of calibration to this choice then has to be gauged. Secondly, MCMC can require many ($O(10^4)$-$O(10^5)$) CLM evaluations to reach converged posterior estimates, which is prohibitive. Thus, while Bayesian calibration holds much promise for CLM calibration, its use has been rare [53, 59].

In this paper, we will describe a method that can allow MCMC calibration of CLM. The method is based on surrogates of CLM - inexpensive polynomial or Gaussian process representations of the mapping between CLM parameters being calibrated and the CLM outputs for which we have measurements. We therefore build on, and extend, recent developments on the use of surrogates to calibrate computationally expensive models [30, 28] and MCMC calibration of complex (e.g., those based on partial differential equations) models including structural errors (i.e., the fundamental inability of the model to reproduce observations due to modeling simplifications) [8, 9]. Our method is general, but we will demonstrate it in the estimation of three hydrological parameters using observations from two sites, US-ARM, located in Oklahoma, and US-MOz, located in Missouri. The method will also yield an approximation of CLM's structural error. Our method is dependent on an accurate surrogate model; in its absence, our calibration method does not work. We will also present an example of this shortcoming.

The novel contributions of this paper are:

1. *Procedure for building CLM surrogates:* While the idea of building surrogates for computationally expensive models is not new [29], the particular form chosen for the surrogate is problem dependent. We describe the practical details of sampling the space of calibration parameters, performing the runs (which, in our case, produce a time-series of outputs), and the process of constructing surrogates while simultaneously simplifying them using sparsity. In particular, we will exploit a sparse reconstruction method, Bayesian compressive sensing [4], to perform model simplification.
2. *Choice of error model and their ramifications:* Bayesian calibration requires one to specify an error model. If competing models exist (as they do in our case), there has to be a systematic way of selecting one. We present an illustration of how to select an error model.
3. *Gauging the post-calibration predictive skill of CLM:* When one has a "point" estimate of parameters (the defaults or optimal values obtained from deterministic optimization), the predictive skill of a model is estimated by calculating bias and root-mean-square-error (RMSE) with respect to observations. When parameters are estimated as PDFs, a different set of error metrics can be used. Further, some of them can reveal how the model needs to be improved. We will compute these error metrics as a demonstration of the usefulness of Bayesian calibration beyond just parameter-estimation-with-uncertainty-quantification.

The paper is organized as follows. In § 2, we review background literature on surrogate models, sparse reconstruction, kriging and MCMC methods. We also review our previous work, based on sensitivity analysis of CLM at the two chosen sites, which underlie the selection of calibration parameters given the observational dataset at hand. In addition, we provide a brief description of the hydrologic modules and their parameters in version 4 of CLM. We also describe the observational dataset used to drive, parameterize and calibrate the model used in this study. In § 3 we construct surrogate models. In § 4, we use them to perform the calibration and discuss the implications of the results. We conclude in § 5.

2. Background.

2.1. Probabilistic calibration of climate models. The implications of parametric uncertainty in climate models (or their submodels) have long been appreciated and there have been efforts to estimate them as PDFs [25]. Due to the computational cost of such models, these methods have sought to reduce the number of model invocations necessary, largely via approximations in the numerical formulation of the estimation problem. Variants of the Very Fast Simulated Annealing Method (VFSA, [24, 25]) have been used to tune parameters of the CAM5 Zhang-McFarlane convection scheme [57]. VFSA leverages simulated annealing to reduce CAM5 (Community Atmosphere Model, version 5) runs, whereas multiple starting points allowed an efficient search in a high-dimensional parameter space. The same method was used to tune 6 parameters in the Weather Research and Forecasting (WRF, [48]) model in [58]. In order to address the high-dimensionality of the problem, the authors used three separate starting points and a total of 150 WRF runs. PDFs of parameters that had higher predictive skill than the default parameter settings were plotted but the quality of the calibration was checked only using an optimal parameter estimate from the calibration i.e., the accuracy of a point summary, rather than the full probabilistic calibration was checked. The ensemble Kalman filter (EnKF, [12]) provides a scalable Bayesian calibration technique, under the assumption that the calibrated PDFs of the parameters are Gaussian. In [2], the authors calibrated a coupled AOGCM (atmospheric ocean coupled general circulation model) of intermediate complexity using EnKFs. In [37], the authors used EnKFs to optimize a hydrology-crop model using data from central Belgium.

Of late, due to advances in computational resources, there have been attempts to perform the calibration without any approximations i.e., to solve the Bayesian calibration problem using MCMC. In [53], 10 hydrological parameters of the CLM version 4 (CLM4) were calibrated using latent heat flux measurements from the flux tower sites at US-ARM and US-MOz. Parameter samples from the posterior PDF (the post-calibration PDFs of the parameters) provided better predictions compared to the default CLM4 settings when their predictions were model averaged. In [59], the authors present a MCMC calibration of 6 parameters of a CLM crop model. The convergence of the MCMC chain was checked via the Brooks-Gelman-Rubin statistic [5]. The paper does not contain any plots of the parameter PDFs or any discussion on estimates of structural error of the model. The improved ability of the calibrated PDFs to predict observations is shown. In [55], the authors applied Bayesian uncertainty analysis to 12 parameters of the Bern2.5D climate model. They first defined a nonparametric set of prior distributions for climate sensitivity and then updated the entire set using MCMC. Motivated by practical needs in estimating parameters of climate and Earth system models, the authors in [50] evaluate the computational gains attainable through parallel adaptive MCMC and Early Rejection using a realistic climate model. In [26] the authors use an adaptive MCMC method (DRAM, [20]) to estimate four parameters of a general circulation model, ECHAM5, using measurements of radiative fluxes at the top of the atmosphere (TOA). The study develops joint posterior PDFs of the four parameters and investigates multiple likelihood formulations which differ in the type and number of summary statistics (of the TOA radiative fluxes) which are included in the likelihood function. The runs were performed with ECHAM5 (and not its surrogates) and the longest run involved 5600 ECHAM5 evaluations. Rather than using a conventional method such as the Gelman-Rubin-Brooks statistic or the Raftery-Lewis test as the stopping criteria for the MCMC run, the sampling was stopped when the tenth, fiftieth and ninetieth percentiles of all four parameters were deemed to have reached a stationary value. The posterior distributions were used to predict global annual mean radiative fluxes and compared with measurements to identify the best likelihood formulation (alternatively, the best set of summary statistics of the TOA radiative fluxes to be used for parameter calibrations).

2.2. The Community Land Model. Community Land Model, Version 4 (CLM4), was released by the National Center for Atmospheric Research to serve as the land component of the Community Earth System Model [36, 31]. It simulates biogeophysical processes such as energy and water fluxes from canopy and soil, heat transfer in soil and snow, hydrology of soil, canopy and snow and stomatal physiology and photosynthesis. Even though most of its applications are conducted at continental or global scales [33, 32], CLM4 can be run at any resolution such as flux tower sites [22] or small watersheds [23].

In CLM4, soil water up to a depth of 3.8 meters from the surface is simulated using the one-dimensional Richards equation

$$\frac{\partial\theta}{\partial t} = -\frac{\partial q}{\partial z} - S \tag{2.1}$$

where θ [mm^3/mm^3] is the volumetric soil water content, z is the height above some datum in the soil column, t is time and S is a soil moisture sink (e.g., extraction by roots, or subsurface drainage). q is the moisture flux through the soil [kg m^{-2} s^{-1}]. The moisture flux q is driven by the soil matric potential Ψ[mm] by the equation

$$q = -K\frac{\partial(\Psi - \Psi_E)}{\partial z}$$

where K is the hydraulic conductivity [mm s^{-1}] and Ψ_E the equilibrium potential. Both K and Ψ_E depend on the local moisture content. This dependence is modeled with an exponential e.g.,

$$\frac{\Psi_E}{\Psi_{E,sat}} = \left(\frac{\theta}{\theta_{sat}}\right)^{-b},$$

where b is the Clapp-Hornberger exponent and θ_{sat} and $\Psi_{E,sat}$ constants that depend on local soil composition. The dependence of K is more complex (as it includes the effect of ice), but in its absence, a similar exponential model holds (and, correspondingly, introduces yet another parameter K_{sat}). b also appears in that expression (see the CLM4 technical note [36] for details).

The upper boundary condition is the infiltration flux q_{infl}, [kg m^{-2} s^{-1}] into the top soil layer given by

$$q_{infl} = q_{liq,grnd} - q_{over} - q_{evp}$$

where q_{evp} is the evaporation from the top soil layer, $q_{liq,grnd}$ is the liquid precipitation reaching the ground plus any snow melt and q_{over} is the surface runoff, parameterized as [34, 35]

$$q_{over} = f_{sat}q_{liq,grnd} + (1 - f_{sat})\max\left\{0, (q_{liq,grnd} - q_{infl,max})\right\}$$

where $q_{infl,max}$ is the maximum soil infiltration capacity, $f_{sat} = f_{max}\exp(-C_s f_{over} z_\nabla)$ is the saturated fraction of the location, z_∇ is the water table depth and f_{max}, C_s, f_{over} are model parameters.

The lower boundary condition, parameterized as the recharge to the subsurface aquifer ($q_{recharge}$, [kg m^{-2} s^{-1}]) is given by

$$q_{recharge} = \frac{\Delta\theta_{liq,N} + \Delta z_N}{\Delta t}$$

where $\Delta\theta_{liq,N}$ and Δz_N are changes in liquid water content at the bottom of the soil column, obtained by solving equation (2.1) numerically. The aquifer recharge, alternatively, the subsurface runoff into the aquifer (q_{drai}) is modeled as distributed sinks S in (2.1) and is expressed as

$$q_{recharge} = q_{drai} = Q_{dm} \exp\left(-F_{drai} z_{\nabla}\right),$$

where $F_{drai}[m^{-1}]$, the runoff decay factor and $Q_{dm}[kgm^{-2}s^{-1}$, the maximum subsurface drainage, are model parameters. The aquifer, which lies beneath the soil column, and exchanges water with it, has an average specific yield of S_y.

Thus the hydrological dynamics in CLM4 are governed by 10 parameters - $f_{max}, C_s, f_{over}, F_{drai}$, $Q_{dm}, S_y, b, \Psi_{E,sat}, \theta_{sat}$ and K_{sat}. The sensitivity of latent heat fluxes to these 10 parameters were investigated in [22]. For the US-ARM site, $\{F_{drai}, Q_{dm}, b\}$ were found to be most important; their counterparts for US-MOz were $\{F_{drai}, Q_{dm}, S_y\}$.

2.3. Surrogate models. The task of calibrating computationally expensive models can be considerably eased if one can devise a computationally inexpensive surrogate. A surrogate model approximately captures the input-output mapping of the true (computationally expensive) model. It can prove to be an efficient solution to problems in sensitivity analysis and optimization that require multiple model invocations; see [28, 30] for some examples of their use in aerodynamics. Frequently surrogates are lower-fidelity or statistical models (e.g., regression models) obtained by fitting to a limited number of sample runs of the true model (also called the training data). In [51, 52], the authors compare various smoothing predictors and non-parametric approaches that can act as surrogate models. In [47] the authors provide an overview of statistical surrogates and lower-fidelity models that can be used as proxies for computationally expensive models.

Polynomials and kriging (also called Gaussian process or GP models) are two very common surrogates, and they can also used together (called regression kriging models). Polynomial surrogates are called trend functions when used together with GP models. Polynomials are very efficient in capturing large-scale variations/trends in the parameters space. A multivariate polynomial form is postulated (with unknown coefficients multiplying the terms) and their values are estimated from the training data via regression. The orders of the polynomial and the terms to be retained are dictated by the training data. One can incrementally simplify (remove terms from) the polynomial expression, refit to data and gauge the improvement in fit using the Akaike Information Criterion [56]. Alternatively, one may use shrinkage regression methods like Bayesian compressive sensing (BCS, [4]) to simplify an overly complex model; see [46] for an example of its use to make a polynomial surrogate for CLM4. Note that the terms retained in the polynomial are dependent on the training data. K-fold cross-validation [21] of the model is recommended.

Stationary smooth Gaussian processes [40, 44, 45] are the approach we adopted for some of our surrogate models. They embody the input-output mapping via a set of multivariate normal random variables. A parametric covariance function (alternatively, a semi-variogram) is then constructed as a function of the inputs. The covariance function is based on the idea that when the inputs are close, the correlation between the outputs will be high. As a result, the uncertainty associated with the model's predictions is small for input values that are close to the training points, and large for input values that are further away. Gaussian processes are popular surrogate models because they (1) typically interpolate the data from which they are built, (2) provide a spatially varying estimate of the variance of the error in their predictions, and (3) do not require a specific type of input sample design. As mentioned above, they are often used in conjunction with simple polynomial models (linear or quadratic), which model the large-scale trends whereas the GP represents short-range deviations from the polynomial predictions. A Bayesian perspective on such models is in [29].

2.4. Bayesian inverse problems and their MCMC solution. Estimation of parameters from observations can be cast as a Bayesian inverse problem. Let $\mathbf{y} = \mathbf{m}(\mathbf{p})$ be a model with parameters $\mathbf{p}$. The model outputs are related to observations $\mathbf{y}^{(obs)}$ as

$$\mathbf{y}^{(obs)} = \mathbf{y} + \varepsilon = \mathbf{m}(\mathbf{p}) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \Gamma) \tag{2.2}$$

where ε is a combination of measurement and structural error and $\mathcal{N}(0, \Gamma)$ denotes a multivariate Gaussian distribution with zero mean and Γ as the covariance matrix. Let $\pi(\mathbf{p}, \Gamma)$ be the prior belief regarding the distribution of the parameters and the structural error. By Bayes' theorem, the posterior PDF $P(\mathbf{p}, \Gamma | \mathbf{y}^{(obs)})$ of the parameters, conditioned on observations, can be given by

$$P(\mathbf{p}, \Gamma | \mathbf{y}^{(obs)}) \propto \underbrace{|\Gamma|^{-\frac{1}{2}} \exp\left[-\frac{1}{2} \left(\mathbf{y}^{(obs)} - \mathbf{m}(\mathbf{p}) \right)^T \Gamma^{-1} \left(\mathbf{y}^{(obs)} - \mathbf{m}(\mathbf{p}) \right) \right]}_{Likelihood, \mathcal{L}(\mathbf{y}^{(obs)} | \mathbf{p}, \Gamma)} \pi(\mathbf{p}, \Gamma) \tag{2.3}$$

This is the post-calibration or posterior distribution of the parameters $\mathbf{p}$. It can be constructed by sampling from the right hand side of (2.3) and generating a histogram of the samples. Markov chain Monte Carlo (MCMC) methods [14] allow the sampling to be performed efficiently. In MCMC, one starts with a guess of the parameter $\mathbf{p}_0$. Using this as the base, a proposal $\mathbf{p}'$ is chosen from a proposal PDF (often, but not necessarily, a multivariate Gaussian) $Q(\mathbf{p}' | \mathbf{p}_0)$. $\mathbf{p}'$ is retained according to certain acceptance criteria, which ensure that the chain is ergodic (so that a chain of infinite length visits all parts of the parameter space) and satisfies detailed balance (i.e., high-probability parameters are visited more often than the low probability ones). The mixing of the MCMC chain in the parameter space is largely dependent on $Q(:)$. Adaptive MCMC methods [20] seek to tune an optimal Q i.e., estimate its covariance periodically using samples $\mathbf{p}_i$ that have already been collected by the MCMC chain. Multichain MCMC methods [50, 7] that use multiple concurrent chains to explore the parameter space have been used in the estimation of climate model parameters [26]. The MCMC chain is stopped when the samples it collects results in a stationary posterior distribution $P(\mathbf{p}, \Gamma | \mathbf{y}^{(obs)})$. An efficient MCMC method can require $O(10^4)$ samples to represent a posterior distribution for 3-4 parameters; for complex-shaped distributions, far more samples may be required. The convergence of a MCMC chain can be judged using the Raftery-Lewis [39] or Brooks-Gelman-Rubin [5] statistics. An unconverged MCMC chain usually leads to parameter PDFs that are too narrow i.e., it underestimates parametric uncertainty, and provides erroneous estimates of high-order moments of the distribution such as inter-parameter correlations. The quality of a Bayesian calibration is gauged by posterior predictive tests (PPTs; chapter on “Model Checking and Improvement” in [13]). Samples of $(\mathbf{p}, \Gamma)$ are drawn from the posterior distribution and used to replicate observations via an ensemble of model simulations using (2.2). The predictive skill of the ensemble is gauged by metrics such as the cumulative rank predictive score (CRPS), verification rank histogram (VRH), mean absolute error (MAE) etc. [16, 15]. The significance of these metrics will be discussed in § 4 where we use them to test our calibration.

Note that the formulation used for Bayesian calibration in (2.2) and (2.3) is a standard one. However, the model $\mathbf{m}(\mathbf{p})$ used in the calculation of the likelihood is not. $\mathbf{m}(\mathbf{p})$ provides monthly averaged predictions of latent heat fluxes and consists of a set of surrogates of CLM4, one for each month. In the studies presented here, the set consisted of 12 or 48 surrogates, depending on the length of the observational datastream. The surrogate models are novel. Each model consists of a polynomial trend function paired, in some cases, with a Gaussian Process model. When constructing each model, we investigate polynomials of orders 1 to 5, using a rigorous process of shrinkage regression and cross-validation to simplify them (i.e., remove superfluous polynomial terms). This is done to ensure that the surrogate models do not overfit the training data, and thus display a spurious degree of accuracy. If the resultant polynomial surrogate fails to achieve a minimum acceptable level of fidelity (with respect to

CLM4 predictions), it is augmented with a Gaussian Process model. In addition, we consider two forms for ε in (2.2), to account for uncorrelated and correlated errors. Details are in § 3 and § 4. This degree of rigor is necessary because (1) the simplifying assumptions underlying surrogate models hold only approximately for complex, nonlinear models such as CLM4 and (2) the data-model discrepancy has the potential to vary drastically as the Markov chain explores the entire parameter space i.e., the behavior of the data-model discrepancy in the vicinity of an optimal point may not be very representative of the entire parameter space. In return, the rigor confers accuracy and robustness to the MCMC results.

2.5. Observational data. The dataset used in the calibration was provided by the North American Carbon Program (NACP) Site-Level Synthesis. The synthesis provides various measurements for 47 sites (called flux towers), including US-ARM and US-MOz. The spatial extent of each site is modest; the largest, US-ARM, contains instruments arrayed over 143,000 square kilometers. The following data are available:

1. Meteorology data including air temperature, specific humidity, wind speed, precipitation, surface pressure, surface incident shortwave radiation, surface incident long-wave radiation, and CO2 concentration. The data were gap-filled by the NACP teams using the same protocol. This data was used to drive CLM4 in our study.
2. Measured fluxes of latent and sensible heat at the native time resolution of the observations (30 or 60-minute) as well as the diurnal, seasonal, and annual time scales. The data were gap-filled following a standard protocol as well. Measurements were performed using the eddy-covariance method [3] i.e., the fluxes were not directly measured, unlike, for example, temperature or wind speed. The observed data are not provided with measurement uncertainties.
3. Remotely sensed NDVI (Normalized Difference Vegetation Index), LAI (Leaf Area Index), and fPAR (fraction of Photosynthetically Active Radiation) phenology data derived from MODIS (MODerate-resolution Imaging Spectroradiometer; an instrument orbiting on NASA's Earth Observing System program of satellites). MODIS-based LAI were used to parameterize CLM4 in this study.
4. Ancillary data and information describe tower location and physical characteristics, disturbance history, and biological and ecological attributes of the vegetation, litter, and soil. These data were also used to parameterize CLM4.

Measurement details and the data themselves can be obtained from [1].

2.6. Calibration parameters and sites. This paper is one in a series of studies focused on CLM calibration using data from Ameriflux towers: US-ARM (US Atmospheric Radiation Measurement Climate Research Facility, Southern Great Plains site, `http://www.arm.gov/sites/sgp`) and US-MOz (the Missouri Ozark tower, `http://ameriflux.lbl.gov/SitePages/siteInfo.aspx?US-MOz`). US-ARM, located in Oklahoma, has clay soils and a vegetation cover of shallow-rooted grasses [43, 54]. US-MOz has loamy soil and deciduous broadleaf vegetation [19, 18]. The energy closure in these Eddy Covariance measurements ranges between 75%—90% (i.e., the 10%—25% error is split between latent, sensible and ground heat measurements) with an uncertainty of around 5% for Latent Heat. Significant differences from the measurements have been observed in energy fluxes and runoff at these sites when simulated using default CLM4 parameters, making them attractive calibration test cases. Due to the high-dimensionality of input parameter space, and the complexity in model behavior, sensitivity analyses have to be performed first, to identify a subset of parameters that could be optimized with the available observational data. In [22, 23] the authors examined the sensitivity of Latent Heat (LH) fluxes and runoff computed using CLM4 to 10 hydrological parameters with a view of ranking the important parameters. In [53], the authors leveraged the sensitivity analysis to calibrate all 10 parameters, using LH and runoff observations. It was found that LH was more informative than runoff for calibration purposes. The study also identified parameters whose posterior distributions were appreciably different from the prior. The study used daily and monthly observations, collected over 2003-2006, for US-ARM; for US-MOz, the duration was 2004-2007. Meteorological forcing, site information (vegetation, soil type etc.), satellite-derived phenology, and validation data (water and energy fluxes) were obtained from the North American Carbon Program; see [22, 53] for details on site information and calibration data. In [53], a slightly modified version of MCMC, with a user-defined "reference acceptance probability" input, was integrated with CLM4 and used to calibrate its parameters. The study does not present any of the statistical details such as convergence analysis and posterior predictive tests. However, the authors did present results on the improvement in predictive skill, post-calibration. Issues related to structural error were not investigated.

Our study is an extension of the calibration performed in [53]. We limit ourselves to the top three parameters that could be calibrated from observations; these parameters were identified via the sensitivity analysis in [22] and the preliminary tuning that was performed in [53]. These parameters are $\{F_{drai}, \log(Q_{dm}), b\}$ for US-ARM and $\{F_{drai}, \log(Q_{dm}), S_y\}$ for US-MOz. F_{drai} represents the reciprocal of the effective storage capacity of the subsurface aquifer used in subsurface runoff generation and is positively correlated to LH. Small values of F_{drai} lead to quick drainage of water away from the shallow root zone, reducing evapotranspiration and LH fluxes. Beyond $F_{drai} \approx 2$, the sensitivity of LH to F_{drai} decreases. Q_{dm} is the maximum subsurface drainage rate and its high values lead to water depletion in the shallow root zone i.e., it is negatively correlated with LH. S_y is the drainable porosity (i.e., average specific yield) under gravity and it is positively correlated with LH. b is the Clapp-Hornberger exponent [27] describing the characteristic curves that relates the soil potential to the volumetric water content. These parameters control the seasonality of heat fluxes (i.e., processes far slower than diurnal variations) but whose accurate prediction is a fundamental requirement of any climate model. In this study, we perform our calibration using quick running surrogates of CLM4 so that the MCMC scheme can be run to convergence. The surrogates also introduce an approximation error (the inability of the surrogate to reproduce CLM4 outputs) motivating us to model and estimate structural error. We will perform our calibration using both monthly and climatologically averaged observations, such that daily / stochastic variability in observations can be averaged out, and structural error models other than i.i.d. Gaussians can be examined. PPTs and metrics such as CRPS etc., discussed in § 2.4 are used to gauge the quality of the calibration and also identify shortcomings in the model (surrogate or CLM4). Thus the aim of this study is to investigate, in detail, the preliminary calibration performed for US-ARM and US-MOz in [53], with emphasis on statistical rigor of the calibration e.g., structural errors, predictive skill, and the effect of climatological averaging.

3. Surrogate models.

In this section we will develop polynomial and GP surrogates for $y_c(\mathbf{p}) = \log(\text{LH})$ where LH are the monthly-averaged latent heat fluxes predicted by CLM4 for parameter setting $\mathbf{p} = \{p_1, p_2, p_3\} = \{F_{drai}, \log(Q_{dm}), b\}$ for US-ARM and $\{F_{drai}, \log(Q_{dm}), S_y\}$ for US-MOz. The fluxes are averaged over a month. The surface fluxes are log-transformed to reduce the dynamic range of LH, which spans an order of magnitude. The prior distributions [22] are:

$$\begin{aligned} F_{drai} &= \mathcal{U}(0.1,\ 5.0) \\ \log(Q_{dm}) &= \mathcal{U}(\log(10^{-6}),\ \log(10^{-2})) \\ S_y &= \mathcal{U}(0.09,\ 0.27) \\ b &= \mathcal{U}(1,\ 15) \end{aligned} \tag{3.1}$$

where $\mathcal{U}(a,\ b)$ denotes a uniform distribution with (a,b) as the lower and upper limits. The parameter space (p_1, p_2, p_3) is thus a cuboid, which is also the domain of applicability of our surrogate models. The default values of the parameters are $F_{drai} = 2.5$, $\log(Q_{dm}) = \log(5.5 \times 10^{-3})$, $S_y = 0.18$ and

$b = 9.76$. The bounds in the uniform priors were obtained from literature as well as via discussions with CLM4 developers and generally reflect physically realistic values; the precise sources and rationale are in [22].

In order to construct surrogate models, we generate a training set of CLM4 runs. We draw 256 samples from the (p_1, p_2, p_3) cuboid via quasi Monte Carlo sampling; see [22] for details. This training set is augmented with the 8 corners, 6 face-centers and 12 edge-centers of the parameter cuboid, leading to $N = 282$ parameter samples where CLM4 is evaluated. For each parameter set, at each site, the model is spun up by cycling the forcing at least five times (i.e., 282 × 5 for the entire set of parameter samples) until all state variables reach equilibrium. Using the initial conditions generated by the spin-up, CLM4 simulates hourly latent heat (LH) fluxes over 2003-2006 for US-ARM and 2004-2007 for US-MOz. These are archived and averaged over each month to generate a monthly time-series of LH predictions. The training set consists of $\{\mathbf{p}_l, y_c^{(l,m)}\}, l = 1 \dots N, m = 1 \dots N_m$, N_m being the number of months in the time-series. $y_c^{(l,m)}$ is the CLM4 output for the l^{th} parameter combination $\mathbf{p}_l$, for month m.

Our interest in developing a new set of surrogate models arises from the challenges we faced when trying to leverage “conventional” surrogate models based on polynomial chaos expansions (PCE) and kriging (Gaussian Process or GP surrogates). PCE models consist of a series of weighted orthogonal polynomial terms of increasing order; the weights are estimated from a training set of CLM4 runs (CLM4 predictions at a set of points in the (p_1, p_2, p_3) cuboid). Conventionally, the calculation of the weights is performed using Galerkin projection (which leverages the orthogonal nature of the terms in the series). These points may be chosen randomly in a space-filling manner (as we have done in this study) or at a very precise set of quadrature points which (1) simplifies the computation of the Galerkin projection and (2) minimizes the size of the training set. The latter approach (i.e., quadrature) is the conventional one, and requires that CLM4 predictions be available at *all* quadrature points. Frequently, the quadrature points involve non-physical parameter combinations where a complex, physics-based simulator such as the CLM4 could potentially show (numerically) unstable behavior or outright crash. This brittleness of the quadrature approach led us to generate the training set of runs via random sampling. In such a situation, the weights, computed via Galerkin projection, are approximate and have to be checked for statistical significance (since not all terms in the PCE model need be retained). Such issues of the significance of model coefficients are routinely faced when fitting models to data via regression, and the theory (and techniques) of model simplification and tests of model robustness are well developed in that context. Consequently, we decided to dispense with projection-based estimation of PCE model coefficients, and developed a method, based on shrinkage regression, of fitting the model to data (described in § 3.1). As an added benefit (over conventional, projection-based construction of PCE models), our new method does not require us to guess the order of the PCE model, but rather “discovers” it in a data-driven manner.

GP modeling is another conventional way of developing surrogate models. The training data is de-trended by fitting a linear model and the residuals are modeled as a multivariate Gaussian with a stationary covariance. Challenges arise when the residuals are large and non-stationary, as was the case in this study. The approximation introduced by the stationary assumption led to a model that did not meet the minimum accuracy that we required of the models (described below). The simplest solution to this problem is to reduce the magnitude of the residuals, for example, by using a more sophisticated model for de-trending, so that the impact of the stationary assumption is muted. In our new approach, we use a GP model paired with high-order PCE, which serves as the sophisticated de-trending function. This is quite apparent in the structure of (3.2).

3.1. Formulation. For a given month, we will represent the monthly-averaged, log-transformed surface fluxes as

$$y_c(\mathbf{p}) = y_1(\mathbf{p};\mathbf{\Theta}_1) + y_2(\mathbf{p};\mathbf{\Theta}_2) + \delta \tag{3.2}$$

where $y_1(\mathbf{p};\mathbf{\Theta}_1)$ is a polynomial surrogate, $y_2(\mathbf{p};\mathbf{\Theta}_2)$ is a GP surrogate and δ is a surrogate model error. $y_c(\mathbf{p})$ is the CLM4 output for a month, for parameter setting $\mathbf{p}$. In our model, we will aim for $\|\delta\|_2/\|y_c(\mathbf{p})\|_2 < 0.1$. $\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$ are parameters of the surrogate models that are estimated from the training set. We postulate a polynomial surrogate model, of order M, as

$$y_1(\mathbf{p};\mathbf{\Theta}_1) = \sum_{i=0}^{M}\sum_{j=0}^{M}\sum_{k=0}^{M} c_{ijk} p_1^i p_2^j p_3^k, \qquad c_{ijk} \in \mathbf{\Theta}_1, \qquad i+j+k \le M. \tag{3.3}$$

Not all c_{ijk} (or terms in the polynomial) are required to model $y_c(\mathbf{p})$. Also, since they have to be estimated from a limited training set, some of the estimates may have significant uncertainty, especially if $y_c(\mathbf{p})$ is not very sensitive to them. Consequently, we estimate them using shrinkage regression, specifically BCS. Separate surrogate models are made for each month.

The form of (3.2) is deliberate. The polynomial component $y_1(\mathbf{p};\mathbf{\Theta}_1)$ will be constructed first since it is conceptually simple and the computational complexity of evaluating it (e.g., inside a MCMC loop) is proportional to the number of terms in the polynomial. This is far smaller than the size of the training set. If we fail to achieve an acceptable degree of accuracy with $y_1(\mathbf{p};\mathbf{\Theta}_1)$, we will progress to constructing $y_2(\mathbf{p};\mathbf{\Theta}_2)$ which requires more sophisticated modeling (e.g., form of the correlation function / variogram etc.). Further, evaluating $y_2(\mathbf{p};\mathbf{\Theta}_2)$ involves operations with the covariance matrix of the GP and the computational complexity scales as the square of the size of training set. Thus, GP models $y_2(\mathbf{p};\mathbf{\Theta}_2)$ are far more expensive than $y_1(\mathbf{p};\mathbf{\Theta}_1)$.

Modeling with polynomial chaos expansions: The $p_1^i p_2^j p_3^k$ terms in (3.3) can be collated into orthonormal polynomial chaos expansions. We normalize $p_i = C_i + D_i\xi_i$, where $\xi_i \sim \mathcal{U}(0,1,)$, ξ_i are independently and identically distributed, $i = 1\ldots3$. (3.3) can then be written as

$$y_1(\mathbf{p};\mathbf{\Theta}_1) = \sum_{m=0}^{M} \beta_m \mathbf{\Psi}_m(\boldsymbol{\xi}), \tag{3.4}$$

where $\mathbf{\Psi}_m(\boldsymbol{\xi})$ is an orthonormal polynomial basis and $\boldsymbol{\xi} = \{\xi_i\}, i = 1\ldots3$ i.e., we normalize each parameter between the upper and lower bounds of its prior. Each index m corresponds to a multi-index vector $\mathbf{r}(m) = \{r_1^{(m)}, r_2^{(m)}, r_3^{(m)}\}$ such that

$$\mathbf{\Psi}_m(\boldsymbol{\xi}) = \mathbf{\Psi}_{\mathbf{r}}(\boldsymbol{\xi}) = \Psi_{r_1}(\xi_1)\Psi_{r_2}(\xi_2)\Psi_{r_3}(\xi_3), \qquad r_i \in \{1\ldots M\}, \qquad \sum_{i=1}^{3} r_i = M. \tag{3.5}$$

In our case, $\Psi_{r_i}(\xi_i)$ are obtained from univariate Legendre polynomials $L_n(\zeta)$

$$L_0(\zeta) = 1, \quad L_1(\zeta) = \zeta, \quad L_2(\zeta) = \frac{1}{2}\left(3\zeta^2 - 1\right) \text{ and } L_{n+1}(\zeta) = \frac{2n+1}{n+1}\zeta L_n(\zeta) - \frac{n}{n+1}L_{n-1}(\zeta). \tag{3.6}$$

We will work with normalized Legendre polynomials i.e., $\Psi_n(\zeta) = \sqrt{2n+1}L_n(\zeta)$. Note that the RHS of (3.5) and (3.3) are formally identical. The choice of Legendre polynomials as an orthogonal basis set is a matter of convenience, since they have been used before to model CLM4 outputs [46]. We formulate the shrinkage regression problem for β_m next.

Shrinkage regression: For a given month, we divide our N-member training set into a learning set (LS) with 85% of the runs and a testing set (TS) with the remaining 15%. The set $\{\mathbf{p}_l, y_c^{(l,m)}\}, l \in LS$

and $m = 1 \ldots N_m$ are used to set up a shrinkage regression problem. We write the likelihood $\mathcal{L}(\mathbf{y}_c^{(LS)}|\boldsymbol{\beta})$, $\boldsymbol{\beta} = \{\beta_m\}$, as

$$\mathcal{L}\left(\mathbf{y}_c^{(LS)}|\boldsymbol{\beta}\right) \propto \left(2\pi\varsigma^2\right)^{-\frac{|LS|}{2}} \exp\left(-\frac{\|\mathbf{y}_c^{(LS)} - \sum_m \beta_m \boldsymbol{\Psi}_m(\boldsymbol{\xi}^{(LS)})\|_2^2}{2\varsigma^2}\right) \tag{3.7}$$

where $\mathbf{y}_c^{(LS)}$ is the vector of CLM4 predictions from the LS runs, $\boldsymbol{\xi}^{(LS)}$ are the corresponding (normalized) CLM4 parameters and the discrepancy between the CLM4 and polynomial surrogate model predictions is modeled using i.i.d. normals $\mathcal{N}(0,\varsigma^2)$. In order to estimate the sparsest model conditional on the data, we impose a Laplace prior

$$\pi(\boldsymbol{\beta}|\lambda) = \left(\frac{\lambda}{2}\right)^{M+1} \exp\left(-\lambda \sum_{m=0}^{M} |\beta_m|\right)$$

and solve the deterministic optimization problem to obtain the maximum *a posteriori* (MAP) values of β_m

$$\arg\max_{\boldsymbol{\beta}} \left[\log\left(\mathcal{L}\left(\mathbf{y}_c^{(LS)}|\boldsymbol{\beta}\right)\right) - \lambda\|\boldsymbol{\beta}\|_1\right].$$

We cast this into a hierarchical Bayesian setting that removes the discontinuous nature of a ℓ_1 norm. We model β_m with a Gaussian prior with standard deviation s_m and, in turn, model all s_m with a Gamma prior

$$\pi\left(\beta_m|s_m^2\right) = \left(2\pi s_m^2\right)^{-\frac{1}{2}} \exp\left(-\frac{\beta_m^2}{2s_m^2}\right) \text{ and } \pi\left(s_m^2|\lambda^2\right) = \frac{\lambda^2}{2} \exp\left(-\frac{s_m^2\lambda^2}{2}\right).$$

Note that if we integrate out s_m^2, we recover the Laplace prior. This hierarchical formulation can be solved using a greedy algorithm commonly used in BCS and described in [4]. It returns non-zero values of β_m that can be estimated from the LS , revealing, in theory, the exact form of the polynomial i.e., the terms in (3.3) that are required to model $\mathbf{y}_c$. Further information on the use of BCS to develop surrogate models of CLM4 is in [46].

Cross-validation studies revealed that BCS could be somewhat imperfect i.e., if we start with a large M (e.g., $M = 10$) the non-zero β_m returned by BCS depend on the LS used. While some low-order terms are always chosen, a significant number of high-order terms were chosen quite often (we will provide an example of this uncertainty below). This uncertainty in the identity of high-order terms led us to use cross-validation to choose an appropriate M. Note that choosing an M^{th} order polynomial for surrogate modeling does not imply that we retain all the terms in the polynomial.

Using cross-validation to choose the polynomial order M: We divided the training set into K distinct LS/TS pairs, $K = 500$, to perform K-fold cross-validation. Polynomial models, with $M = 1 \ldots 5$ were fitted using the CLM4 runs in the LS to estimate β_m. The β_m were then used to predict $\log(LH)$ using $\mathbf{p}_i$ in the TS. Relative errors were calculated for both the LS and TS, for all K LS/TS pairs and then averaged to obtain the mean errors for a given order M, as described below.

Let the vector $\mathbf{y}_c^{(LS)}$ be the CLM4 output for an arbitrary month generated by the vector of parameter combinations $\boldsymbol{\xi}^{(LS)}$ contained in a learning set, LS. Let β_m be the model coefficients generated from LS. We define a relative error for the learning set, for a polynomial model of order M, as

$$E_M^{(LS)} = \frac{\|\mathbf{y}_c^{(LS)} - \sum_{m=1}^{M} \beta_m \boldsymbol{\Psi}_m\left(\boldsymbol{\xi}^{(LS)}\right)\|_2}{\|\mathbf{y}_c^{(LS)}\|_2}.$$

Since we construct K LS/TS pairs during K-fold cross-validation, it is more convenient to define the relative error for an individual test i.e.

$$E^{(LS)}_{M,l} = \frac{\|\mathbf{y}^{(LS)}_{c,l} - \sum_{m=1}^{M} \beta_{m,l} \mathbf{\Psi}_m \left(\boldsymbol{\xi}^{(LS)}_l\right) \|_2}{\|\mathbf{y}^{(LS)}_{c,l}\|_2},$$

the relative error for the l^{th} learning set, $l = 1 \ldots K$ and the mean relative error (for the learning set) over all K cross-validation tests

$$\overline{E^{(LS)}_M} = \frac{1}{K} \sum_{l=1}^{K} E^{(LS)}_{M,l}.$$

In an identical manner, we can develop the relative errors $E^{(TS)}_{M,l}$ and $\overline{E^{(TS)}_M}$ for the testing sets. Note that when computing the testing set errors, the model coefficients $\beta_{m,l}$ are always computed from the corresponding (i.e., l^{th}) learning set.

If the fitting is proper and no spurious terms are retained, then $\overline{E^{(LS)}_M} \approx \overline{E^{(TS)}_M}$, i.e., the fitted model is equally predictive for the LS and TS. In case of overfitting, the polynomial model will be more predictive for the LS. We will choose a value of M for developing surrogate models if

$$\eta = \frac{\overline{E^{(TS)}_M}}{\overline{E^{(LS)}_M}} \leq 1.05. \tag{3.8}$$

GP models : Fitting a polynomial model does not ensure that $\|y_c(\mathbf{p}) - y_1(\mathbf{p}; \mathbf{\Theta}_1)\|_2 / \|y_c(\mathbf{p})\|_2 < 0.1$. If $\Delta y(\mathbf{p}) = y_c(\mathbf{p}) - y_1(\mathbf{p}; \mathbf{\Theta}_1)$, where $\mathbf{p}$ are samples from the training set, is smoothly distributed in the (p_1, p_2, p_3) space, and the mean over the training set samples is zero, then the discrepancy can be modeled as multivariate Gaussian i.e., $\Delta y(\mathbf{p}) \sim \mathcal{N}(0, \Sigma)$. The key is to model Σ appropriately. Any covariance/correlation model can be used for Σ; in this study, we will use a variogram model involving distance (and a single range / lengthscale) in the (ξ_1, ξ_2, ξ_3) normalized parameter space. We will compute the empirical semi-variogram and fit various two-parameter variogram models – exponential, linear, Gaussian, spherical etc. – via maximum likelihood estimation. In general, the estimation will yield a magnitude (sill) and a lengthscale (range) of the variogram model, along with a goodness-of-fit metric. The form of the variogram model and its parameters (the sill and the range) constitute the parameter $\mathbf{\Theta}_2$.

3.2. Models for US-ARM. As a first step we examine polynomial fits to the LS data by BCS, for April, climatologically averaged over 2003-2006. In Fig. 1 we plot the distribution of $E^{(LS)}_{M,l}$ and $E^{(TS)}_{M,l}$ for $M = \{1, 2, 4\}$ generated via a 500-fold cross-validation test. The top, middle and bottom rows of plots are obtained for $M = 1, 2$ and 4. The distribution of errors from the LS (240 CLM4 runs), in the first column, is somewhat different from that of the TS errors (42 runs); however, for $M = 1$ and 2, the average of LS and TS errors are very similar. This is not the case for $M = 4$. We also plot the distribution of the number of terms retained in the polynomial by the BCS algorithm. For $M = 1$ and 2, there is little uncertainty; all the terms in the polynomial are retained. The same behavior, i.e., linear and quadratic models proving to be "well-behaved" was seen for other months too. This is not the case for the quartic model, where there is considerable uncertainty in the number of terms retained (it varies from 25 to 35), let alone the identity of the terms retained in polynomial. It is this uncertainty that led us to use cross-validation (CV) and (3.8) to choose the model order M.

In Fig. 2 we examine the order of polynomial model to use. These models are obtained by BCS-fitting of the model to LS data. The data is obtained from a 500-fold CV. On the left, we plot $\overline{E^{(LS)}_M}$ for

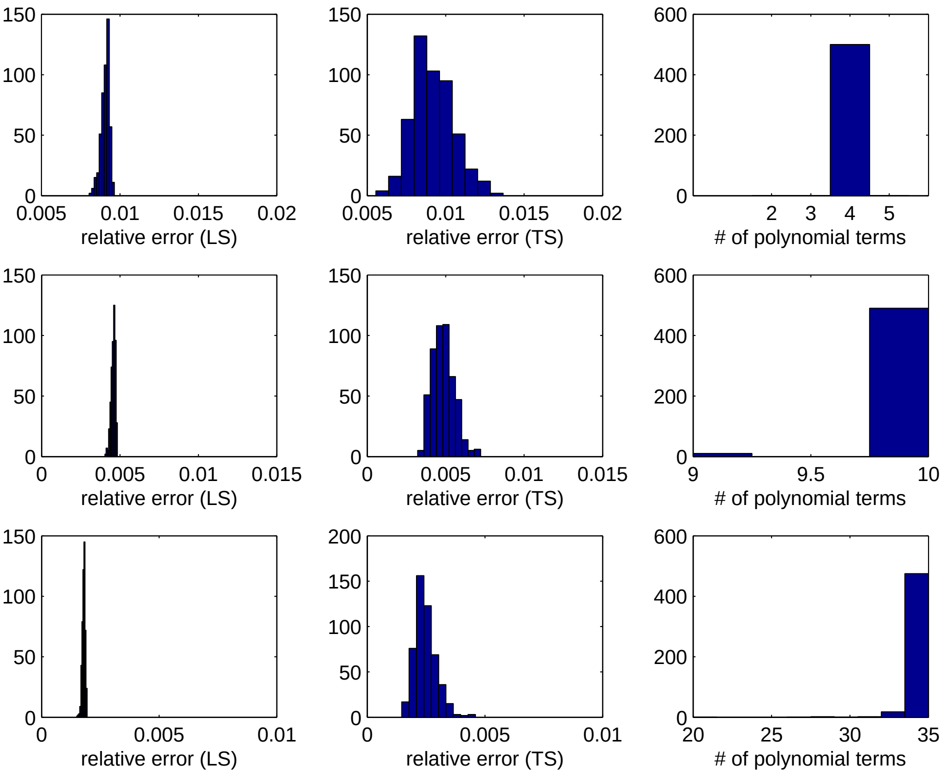

Figure 1. *Distribution of $E^{(LS)}_{M,l}$ and $E^{(TS)}_{M,l}$ for $M = \{1,2,4\}$ as calculated from a 500-fold cross-validation test. In the top row, we use $M = 1$. The corresponding values for M are 2 and 4 for the middle and bottom row of plots. In the first column, we plot the distribution of $E^{(LS)}_{M,l}$ from a LS of 240 CLM4 runs. In the second column, we plot the distribution of $E^{(TS)}_{M,l}$ from a TS of 42 runs. In the last column, we plot the distribution of the number of terms retained in the polynomial model by the shrinkage regression algorithm.*

all months using climatologically averaged CLM4 predictions over 2003-2006. We use $M = 1\ldots5$. On the right, we plot η for the same months. We see, on the left, that $\overline{E^{(LS)}_M}$ decreases as M increases i.e., model complexity improves predictive skill, even though shrinkage regression removes many of the polynomial terms. However, this improvement is largely due to overfitting, as is shown in the plot of η on the right. For cubic and higher-order models, $\overline{E^{(TS)}_M}$ is larger than $\overline{E^{(LS)}_M}$ and the improvement of predictive skill with model complexity is not seen. Since we wish to have models that are equally predictive everywhere, we see that quadratic models ($M = 2$) offer the best solution. Also, note that the relative errors are small, less than 2%. This allows us set $y_2(\mathbf{p};\mathbf{\Theta}_2) = 0$ in (3.2) i.e., skip any GP modeling for US-ARM, and yet meet the accuracy requirement for surrogate models ($\|y_c(\mathbf{p}) - y_1(\mathbf{p};\mathbf{\Theta}_1)\|_2/\|y_c(\mathbf{p})\|_2 < 0.1$).

We repeated the same process with the models created for each of the 48 months in 2003—2006 for US-ARM; the results are in the Supplementary Materials (Fig. S1) as well as in [41]. The same issues were observed – BCS proved to be inadequate and 500-fold cross-validation was required. $\eta < 1.05$ was obtained for linear and quadratic models, and quadratic surrogate models, which provided $< 4\%$ relative errors, were retained for further use.

As a check of the sufficiency of the training data, we halved it and refitted the surrogate models. The smaller training set led to models that were less accurate, but nevertheless met the 10% surrogate model error threshold, indicating that there was no need for expanding the dataset.

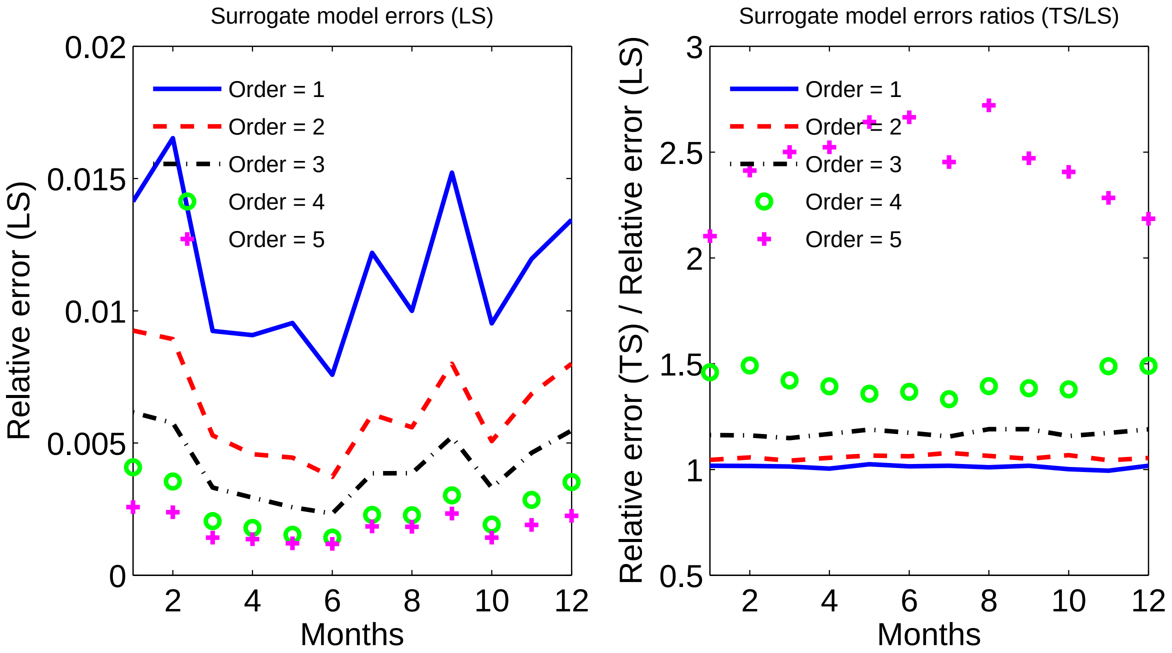

Figure 2. *Left: We plot $\overline{E_M^{(LS)}}$ for US-ARM, for all months using climatologically-averaged CLM4 predictions over 2003-2006. We use $M = 1\ldots5$. Right: We plot η for the same months. We see that, as expected, high-order polynomial models provide lower errors when fitted to LS. This is largely due to overfitting since $\eta \approx 1$ holds only for linear and quadratic models; in the rest of the models, higher predictive skill in the LS does not carry over to the TS.*

3.3. Models for US-MOz. We performed the same analysis, as described above, for US-MOz, but only using climatologically averaged LH predictions. The results were much the same. The shrinkage regression algorithm is imperfect and $\overline{E_M^{(LS)}}$ reduces with model complexity (when using the LS) but the same predictive skill of the surrogate models is not evident when tested using the TS. The plots of $\overline{E_M^{(LS)}}$ and $\overline{E_M^{(TS)}}$ for each month, for various M can be found in Fig. 3 (top). Again, quadratic models provide the best balance between minimizing $\overline{E_M^{(LS)}}$ while keeping $\eta \leq 1.05$. Note that $\overline{E_M^{(LS)}}$, $M = 2$, is between 15% and 20% (which does not meet our 10% surrogate model error threshold for acceptability) and hence we will augment the polynomial model in (3.2) with a GP approximation $y_2(\mathbf{p};\mathbf{\Theta}_2)$.

We construct GP models $y_2(\mathbf{p};\mathbf{\Theta}_2)$, for each month, using $\Delta y(\mathbf{p}) = y_c(\mathbf{p}) - y_1(\mathbf{p};\mathbf{\Theta}_1)$ computed from the LS data. In Fig. 3 (bottom left) we show the empirical semi-variogram for $\Delta y(\mathbf{p})$ in the normalized (p_1, p_2, p_3) space and its approximation using an exponential semi-variogram for the month of April. A better fit could not be obtained using other semi-variogram models such as spherical, linear etc. The resulting model, $y_2(\mathbf{p};\mathbf{\Theta}_2)$ in (3.2), is added to $y_1(\mathbf{p};\mathbf{\Theta}_1)$, and used to compute the relative error for the TS dataset. The relative errors are averaged over a 500-fold cross-validation test and plotted in Fig. 3 (bottom right) with a solid line. The errors without the GP augmentation are also plotted (dashed line). We see that including the GP surrogate halves the surrogate modeling error to bring it below the 10% relative error target that we have adopted for surrogate models.

We next attempted to construct surrogate models without climatological averaging the data i.e., using the 48-month time-series spanning 2004-2007. We found that we could construct only 40 (out of 48) such models that met the 10% relative error requirement. We conjecture that this may be due to meteorological anomalies or extremes. This difficulty was not seasonal in nature - after climatological averaging, surrogate models could be constructed for all the months. This also implies that for US-

MOz, we will only be able to calibrate CLM4 using climatologically averaged observations.

We checked whether the same level of accuracy could be obtained by halving the training data. We could not achieve the requisite 10% surrogate modeling error indicating that (1) the training data was barely sufficient for constructing acceptable (10% error) models and (2) a larger training dataset would lead to better surrogates.

4. Calibration.

In this section we use the surrogate models created in § 3 to calibrate 3 hydrological parameters of CLM4. Having established that quadratic polynomials and GPs with their covariance modeled using an exponential variogram suffice, we remake the surrogates using all the training data. We will use the surrogate models in an MCMC calibration effort to obtain PDFs of the parameters of interest. We address the following issues:

1. *Accuracy:* Does calibration improve predictive skill vis-à-vis the default CLM4 parameter setting?
2. *Impact of climatological averaging:* Does using the climatological mean of the observations have a significant impact on the parameter estimates?
3. *Impact of the structural error model:* The 48-month time-series model allows us to explore 2 structural error models of differing complexities. What are the ramifications of using a simple versus a complex structural error model?

4.1. Formulation.

Let $Y^{(obs)} = \{y_m^{(obs)}\}, m = 1 \dots N_m$ be the observed values of log-transformed latent heat surface fluxes, averaged over a month. We rewrite (3.2) for month m as

$$y_{c,m}(\mathbf{p}) = y_{s,m}(\mathbf{p}) + \delta_m = y_1(\mathbf{p}; \boldsymbol{\Theta}_{1,m}) + y_2(\mathbf{p}; \boldsymbol{\Theta}_{2,m}) + \delta_m,$$

where $y_{s,m}(\mathbf{p})$ is the surrogate model prediction for month m, for parameter setting $\mathbf{p}$. Note that $y_2(\mathbf{p}; \boldsymbol{\Theta}_{2,m})$ is zero for US-ARM. Let $Y_s(\mathbf{p}) = \{y_{s,m}(\mathbf{p})\}, m = 1 \dots N_m$. Since the surrogate model parameters were estimated from the training set, we will consider them known constants. We relate the observations to the model predictions as

$$Y^{(obs)} = Y_s(\mathbf{p}) + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} = \{\varepsilon_m\}, m = 1 \dots N_m, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Gamma). \tag{4.1}$$

Here, $\boldsymbol{\varepsilon}$ is a combination of structural and measurement errors. The errors in daily measurements of LH (*not* log(LH) as used in this paper) have been discussed in § 2.6; however, we average these measurements over a month and considerably reduce the stochastic component of the error. There is currently no systematic study of the error in these measurements. The parameter vector is $\mathbf{p} = \{p_k\} = \{F_{drai}, \log(Q_{dm}), S_y\}$ (for US-ARM) and $\{F_{drai}, \log(Q_{dm}), b\}$ for US-MOz. Per (2.3), the posterior distribution is given by

$$P\left(\mathbf{p}, \Gamma | Y^{(obs)}\right) \propto |\Gamma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left[Y^{(obs)} - Y_s(\mathbf{p})\right]^T \Gamma^{-1}\left[Y^{(obs)} - Y_s(\mathbf{p})\right]\right) \pi(\Gamma) \prod_{k=1}^{3} \pi(p_k), \tag{4.2}$$

where we have explicitly imposed independent priors on the elements of $\mathbf{p}$, as given by (3.1). We will consider two models for $\boldsymbol{\varepsilon}$:

1. *Uncorrelated errors:* We will assume that the monthly model-observation discrepancies ε_i are uncorrelated and can be modeled as $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. We will estimate σ^2 along with $\mathbf{p}$. We model $\Gamma = \text{diag}(\sigma^2)$. We will estimate the precision $\chi = \sigma^{-2}$ for convenience. The prior for χ is

$$\chi \sim \text{Gamma}\left(\frac{n_0}{2}, \frac{n_0 S_0^2}{2}\right)$$

where n_0 and S_0 are user-supplied values. The two parameters of the Gamma distribution are the shape and the rate (the reciprocal of the scale) respectively. Since the likelihood in (4.2) is Gaussian (with

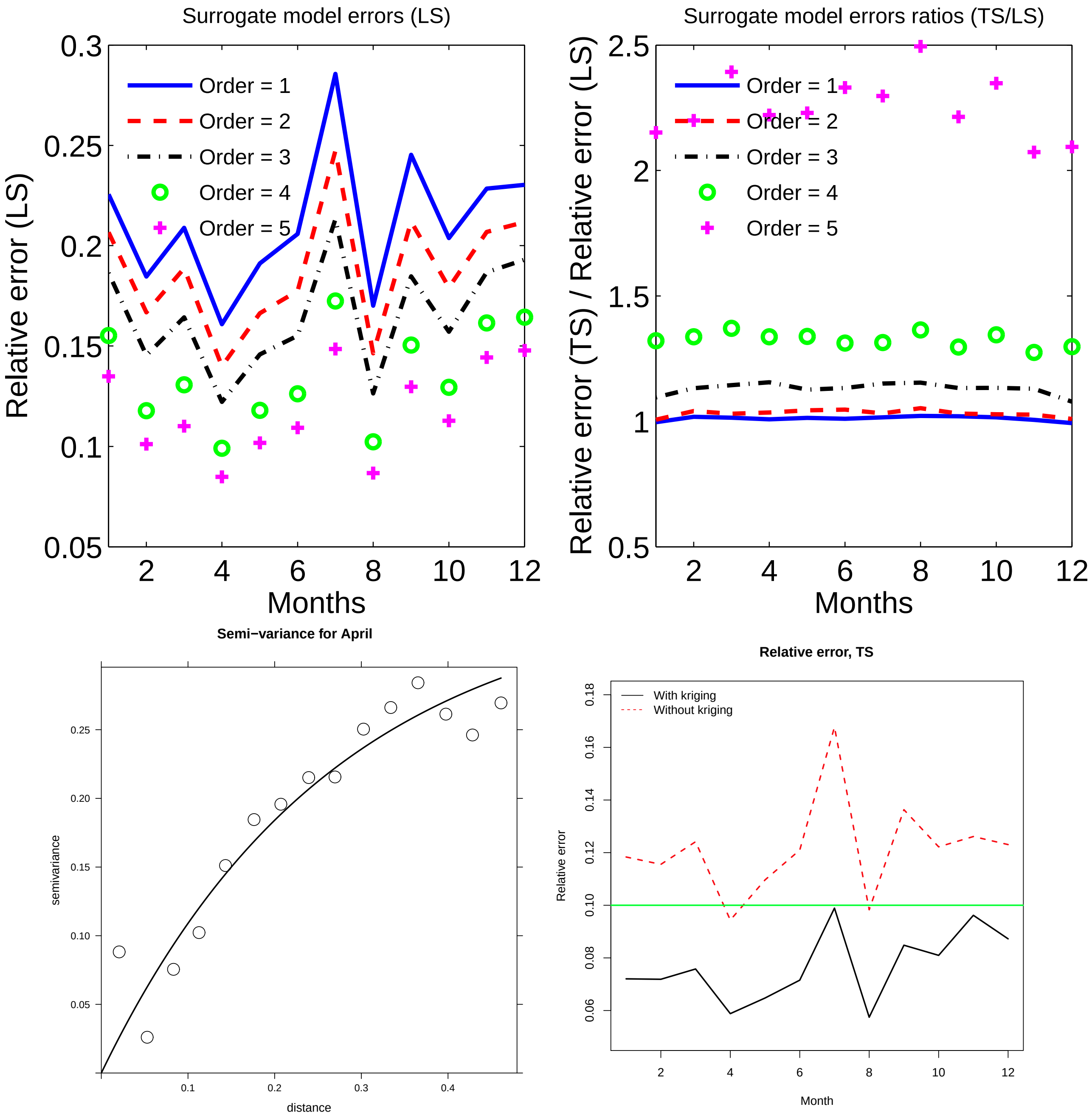

Figure 3. *Top left: We plot $\overline{E_M^{(LS)}}$ for US-MOz, for all months using climatologically-averaged CLM4 predictions over 2004-2007. We use $M = 1 \ldots 5$. Top right: We plot η for the same months. We see that, as expected, high-order polynomial models provide lower errors when fitted to LS. This is largely due to overfitting since $\eta \approx 1$ holds only for linear and quadratic models; in the rest of the models, higher predictive skill in the LS does not carry over to the TS. Note that none of the polynomial models provide surrogate modeling errors less than 10% for all months. Bottom left: Empirical semi-variogram for the discrepancy $y_c(\mathbf{p}) - y_1(\mathbf{p}; \mathbf{\Theta}_1)$ in the $\xi_1 - \xi_2 - \xi_3$ space (in symbols) and its approximation using an exponential variogram. Results are for $\log(LH)$ in April, for US-MOz, climatologically-averaged over 2004-2007. Bottom right: The relative error obtained using a quadratic polynomial model and a GP model is plotted (solid line) for all 12 month, for US-MOz, using climatologically averaged CLM4 predictions for 2004-2007. The error obtained without the GP surrogate is plotted with a dashed line. The horizontal line is the 10% accuracy threshold for surrogate models. These errors were computed using only the TS data from a 500-fold CV test.*

a known mean, conditional on $\mathbf{p}$), the inverse-Gamma distribution for σ^2 is a conjugate prior, and allows us to sample χ using a Gibbs sampler [13, 41]. This circumvents issues regarding mixing and efficiency of sampling. We use $n_0 = 0.1$ and $S_0^2 = 0.01$. The prior is essentially flat for $\chi > 3$.

2. *Temporally correlated errors:* We will model $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \Gamma)$. We assume a stationary distribution and model Γ using a two-parameter variogram. The variogram model will be chosen by fitting to the defect $\boldsymbol{\gamma} = \{Y^{(obs)} - Y_s(\mathbf{p}_{opt})\}$, where $\mathbf{p}_{opt}$ is obtained via a deterministic optimization method. The variogram model's parameters, sill (σ^2) and range (τ), are calibrated along with $\mathbf{p}$.

The inverse problem in (4.2) was solved using a combination of a Gibbs sampler (for χ) and the adaptive MH sampler DRAM [20] (for the parameters without conjugate priors). Convergence of the chain was monitored using the Raftery-Lewis (RL) statistic [39]. The RL statistic ensures that the sampler has collected sufficient samples to estimate (in our case) the median value of each parameter within a tight tolerance. It does so by recursively downsampling the chain (e.g., retain every alternate sample in the stream of samples collected by the MCMC method) till the chain resembles a first-order Markov process. It then checks whether there are sufficient samples in the downsampled (or thinned) chain to approximate the stationary solution of the Markov process within the specified tolerance. The code was written in R [38] and we used the DRAM implementation in FME [49], which contains the MH-Gibbs combination discussed above.

Posterior predictive test (PPT) and error metrics: MCMC solution yields the posterior distribution $P(\mathbf{p}, \sigma^2 | Y^{(obs)})$ (or $P(\mathbf{p}, \sigma^2, \tau | Y^{(obs)})$, if using temporally correlated errors) which is checked using posterior predictive tests (PPT). We choose N_s samples from the posterior distribution and generate a set of predictions $Y_l^{(ppt)} = \{y_{l,m}^{ppt}\} = \{y_{s,m}(\mathbf{p}_l) + \boldsymbol{\varepsilon}_l\}, l = 1 \ldots N_s, m = 1 \ldots N_m$, where $\boldsymbol{\varepsilon}_l \sim \mathcal{N}(0, \Gamma_l), \Gamma_l = \mathrm{diag}(\sigma_l^2)$ or $\Gamma_l = \Gamma(\sigma_l^2, \tau_l)$. Thus for each observation $y_m^{(obs)}$, we obtain N_s predictions $y_{l,m}^{ppt}, l = 1 \ldots N_s$. The quality of these predictions is gauged using the mean absolute error (MAE), continuous rank probability score (CRPS) and the verification rank histogram (VRH). CRPS and MAE are integrated measures of the error in the ensemble predictions vis-à-vis observations. The VRH is a metric that is used to probe the calibration further. The details of these metrics are in [16, 15], but they are summarized below.

MAE: The *MAE* is calculated as

$$MAE = \frac{1}{N_m N_s} \sum_{l=1}^{N_s} \sum_{m=1}^{N_m} |y_m^{(obs)} - y_{l,m}^{ppt}|$$

CRPS; The CRPS is calculated as a mean over N_m $CRPS_m$, the CRPS for month m. For a given month m, we use N_s predictions $\{y_{l,m}^{ppt}\}, l \ldots N_s$ to compute the cumulative distribution function (CDF) $F_m(y)$. We use it in the computation of $CRPS_m$ as:

$$CRPS_m = \int_{\infty}^{\infty} \left(F_m(y) - \mathbb{H}(y - y_m^{(obs)}) \right) dy.$$

$\mathbb{H}(z)$ is the Heaviside function.

VRH: For each month m, we sort the predictions and the observations to find the rank of the observation. The N_m ranks are binned and used to create a histogram. In a perfect calibration, the ranks of the observed values should resemble draws from a uniform distribution. If the observations' ranks are clustered at the lower or upper end, the calibration in under-dispersive i.e., model predictions are not sufficiently sensitive to the model parameters. If the observations' ranks are clustered in the middle of the distribution, the calibration is over-dispersive. In either case, a change in CLM4 or the structural error model is indicated.

4.2. Calibration using US-ARM data.

The observational dataset for US-ARM consists of $N_m = 48$ months of log(LH) readings (2003-2006). As a first step towards calibration, we use the surrogate models to perform a deterministic calibration using a box-constrained optimization method (L-BFGS-B, [6]) to obtain $\mathbf{p}_{opt} = \{F_{drai}, \log(Q_{dm}), b\} = \{0.97, \log(10^{-2}), 0.1\}$. Note that the "optimal" values for two of the parameters are at the edge of the prior distribution. In Fig. 4, (top left) we plot 48 months of observations of log(LH), and the predictions using surrogate models generated using $\mathbf{p}_{opt}$ and $\mathbf{p}_{def}$, the default values of $\{F_{drai}, \log(Q_{dm}), b\} = \{2.5, \log(5.5 \times 10^{-3}), 9.76\}$. We see that $\mathbf{p}_{opt}$ provides far better predictions than $\mathbf{p}_{def}$, which are largely over-predictions. Further, we clearly see that the model-data discrepancy is correlated in time. We assume that the temporally correlated discrepancies are stationary and model Γ using a variogram. In Fig. 4 (top right), we plot the empirical semi-variogram and a fit with a spherical variogram model,

$$\rho(t) = \sigma^2 \left[\left(\frac{3t}{2\tau} - \frac{t^3}{2\tau^3} \right) \mathbb{H}(\tau - t) + \mathbb{H}(t - \tau) \right]$$

obtaining $\sigma^2_{opt} = 0.1515$ and $\tau_{opt} = 7.32$ months. Here t is time measured in months. Note that when τ is small i.e., uncorrelated errors, the variogram model reduces to an i.i.d. Gaussian model for the errors. Fits with exponential, linear, etc. variogram models were inferior. Note that the empirical semi-variogram shows a decline (an improvement in correlation) beyond 8 months. This is because CLM4 is seen to consistently underpredict LH during the winter months, leading to correlated winter errors with an approximately 12-month period. This, in turn, leads to a downturn in the semi-variance. The spherical variogram excludes the periodicity in errors. Since the variation in latent heat fluxes is seasonal, errors with a 12-month periodicity can be expected and the correlation timescale of the structural error should be less than a year. The structural error would repeat every season. In Fig. 4 (bottom row) we plot the auto-correlation function (ACF) and the partial auto-correlation function (PACF) of the error (difference between observations and predictions generated using $\mathbf{p}_{opt}$). Clearly, the errors are correlated; a Durbin-Watson test rejects the hypothesis of i.i.d Gaussian errors comprehensively (p-value of 6×10^{-11}).

Next we use the dataset to estimate $\mathbf{p}$ with a temporally-correlated structural error model. We use the spherical variogram above to model Γ, and estimate $\{F_{drai}, \log(Q_{dm}), b, \sigma^2, \tau\}$. The priors are $\sigma^2 \sim Exp(\sigma^2_{opt})$ and $\tau \sim Exp(\tau_{opt})$. Note that the exponential priors are informative, and we will need to check their impact on the parameter estimates. In Fig. 5, we plot the priors (symbols), the marginalized posterior distributions for $\{F_{drai}, \log(Q_{dm}), b, \sigma^2, \tau\}$, along with their default values (or σ^2_{opt} or τ_{opt}). There is considerable uncertainty in the parameter estimates; the marginalized PDFs are not narrow. For $\log(Q_{dm})$, the default value and the peak of the posterior PDF agree. For F_{drai}, there is considerable disagreement between the peak of the PDF and default parameter value. The calibrated value of the Clapp-Hornberger exponent b bears little resemblance to the default CLM4 value. The exponential priors adopted for σ^2 and τ accomplish two functions - they use the "optimal values" from the L-BFGS-B fit, while expressing a prior belief that MCMC calibration could calibrate them to smaller values. Small values of σ^2 define $\mathbf{p}$ that are more predictive. A small τ indicates that the structural error is uncorrelated in time. The PDFs in Fig. 5 show that the PDF of σ^2 peaks to the left of σ^2_{opt}. The MCMC calibration provides realizations of $\mathbf{p}$ that have smaller disagreements with observations. The PDF for τ peaks to the left of τ_{opt}, but is far from zero. The calibration indicates that errors are correlated, though the correlation timescale is less than the 7.72 months obtained by L-BFGS-B fit. Thus the spherical variogram does not reduce to i.i.d. Gaussian errors. 10^5 MCMC steps (and model invocations) were required to obtain converged posterior distributions.

The ACF and PACF plotted in Fig. 4 (bottom row) show that in the vicinity of $\mathbf{p}_{opt}$, the correlated error model is strongly preferred. However, the PDFs in Fig. 5 are quite wide, indicating that the 5 parameters may be too many to be resolved from a 48-month time-series. Consequently, we repeat the calibration after modeling the structural error as uncorrelated i.i.d. Gaussians. This calibration has one less parameter to estimate (no τ). The prior on σ^2 was the conjugate inverse Gamma distribution,

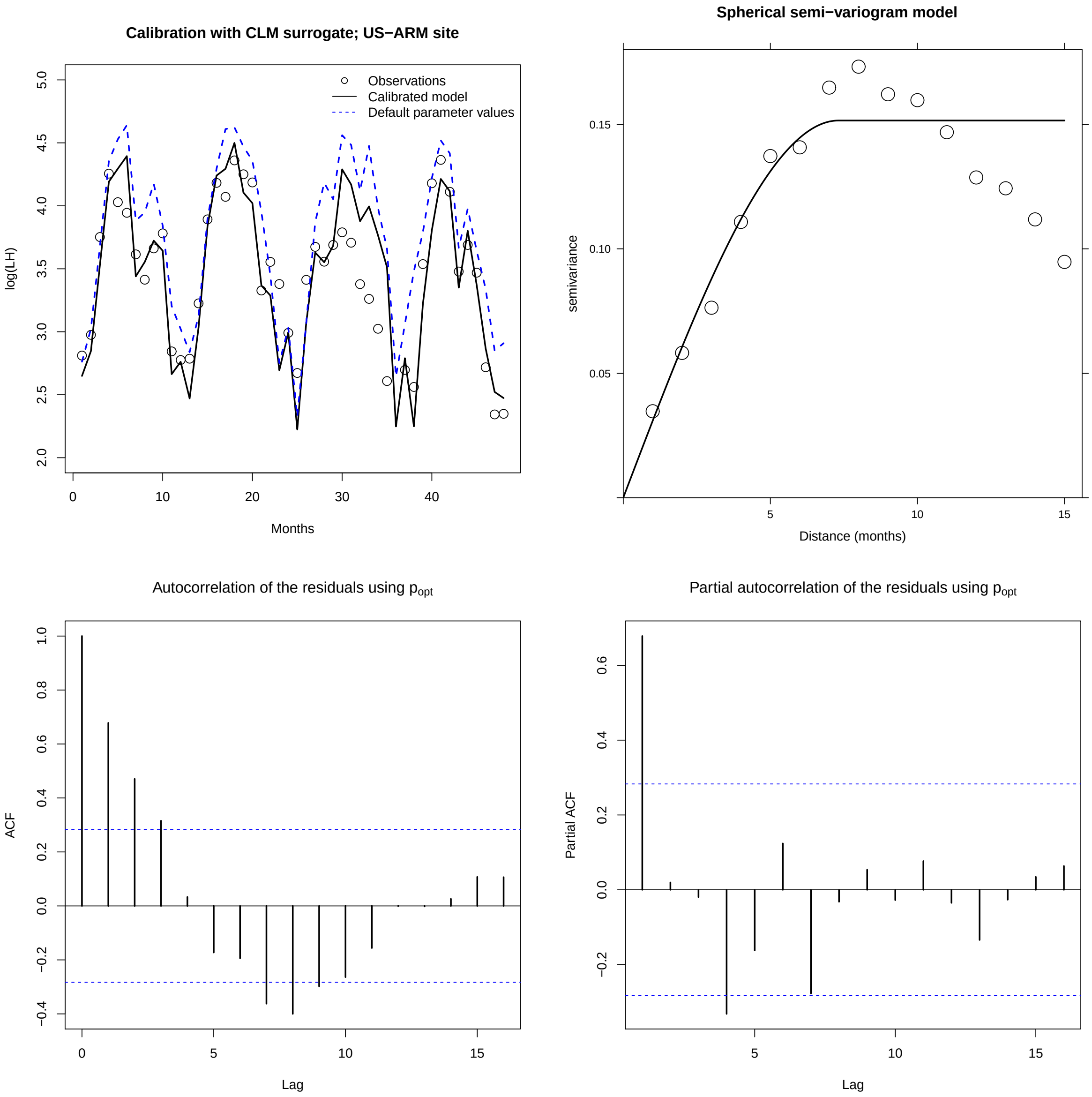

Figure 4. *Top left: Plots of log(LH) as observed at US-ARM over 2003-2006 (plotted with symbols). We plot the CLM4 predictions (using surrogates) generated with $\mathbf{p}_{opt}$. The predictions with default values of $\mathbf{p}$, $\mathbf{p}_{def}$, are plotted with a dashed line. Top right: We plot the empirical semi-variogram calculated from the defects $\boldsymbol{\gamma}$ and a spherical variogram fit to the data. Bottom row: We plot the auto-correlation function (ACF, left) and partial ACF (PACF, right) of the discrepancy between observations and $\mathbf{p}_{opt}$ predictions (black line top-left subfigure). The dashed lines indicate the bounds outside of which the null hypothesis of no auto-correlation is rejected with a significance level of 5%. The existence of auto-correlated errors is quite clear.*

as discussed earlier. The marginalized posterior distributions are plotted in Fig. 5 using dashed lines. We see that the peaks of the PDFs of F_{drai} and $\log(Q_{dm})$ are approximately at the same location as the PDFs obtained using the temporally correlated structural error model; however, the PDFs obtained using the uncorrelated structural error model are sharper. The PDF for b, the Clapp-Hornberger exponent, shows that the default value is far too large. The PDF for σ^2 is narrower for the uncorrelated structural error model and peaks to the left i.e., calibration may be slightly more predictive than the one performed with temporally correlated errors. Comparing with $\mathbf{p}_{opt}$, we find that the deterministic

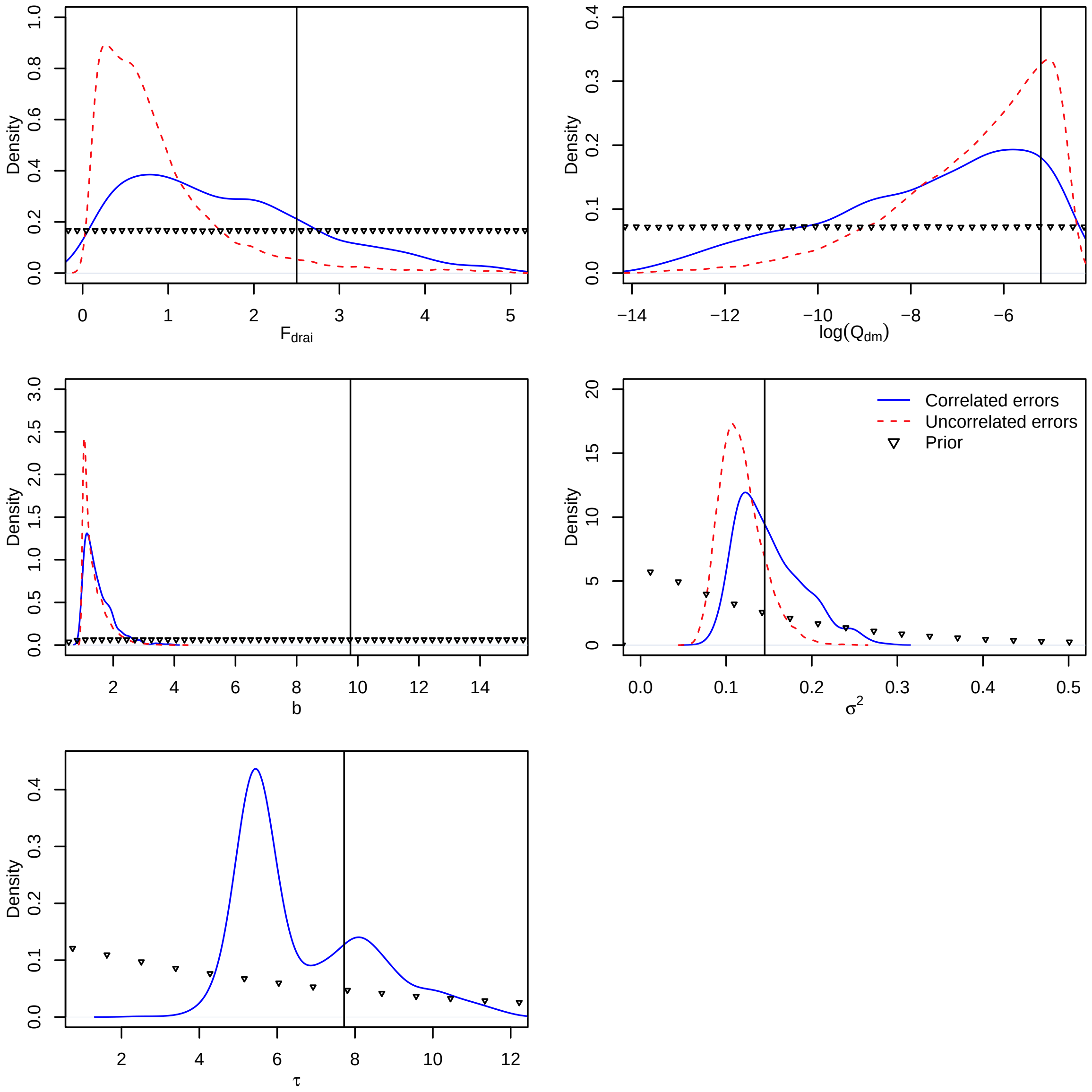

Figure 5. *Marginalized posterior distributions for* $\{F_{drai}, \log(Q_{dm}), b, \sigma^2, \tau\}$*, after calibrating to US-ARM data. The vertical line is the default value or* σ^2_{opt} *or* τ_{opt}*. The symbols denote the prior distribution. The solid line denotes calibration using a temporally correlated structural error model while the dashed line is obtained when we assume the structural error is uncorrelated and can be modeled as i.i.d. Gaussian.*

calibration converges to the peak of the PDF for F_{drai} (at $F_{drai} = 0.97$). It reached the boundaries for the other two parameters.

We next perform PPTs for both the calibrations and plot their results in Fig. 6. We use $N_s = 200$ runs in our posterior predictive tests. Above, we plot the median predictions from PPTs generated using both the calibrations. The error-bars denote the inter-quartile range (IQR). Observations and predictions using $\mathbf{p}_{opt}$ are also plotted. There is little doubt that calibration draws predictions closer to observations; $\mathbf{p}_{def}$ causes over-predictions. Further, the IQR captures all the observations except in the latter half of 2005 (months 30-36), when all observations are systematically lower than the predictions. The observations tend to be near the upper end of the IQR. There is little to choose between the PPTs generated using the competing structural error models. Lower left, we plot the VRH for the two calibrations. An ideal calibration would have yielded a uniform distribution; clearly, we are far from being so. The low observations during months 30-36 (which lead to low ranks for the observations) are clearly seen in the peak at the lower end. Otherwise, the observation ranks are clustered in the range 100-150, for both the calibrations. In Table 1, we tabulate the CRPS and MAE for the two calibrations; they are the same. Lower right, we plot an individual realization of predictions generated by the two calibrations. The observations and the mean prediction are plotted for reference. The prediction generated using correlated errors, which varies smoothly around the mean is plotted with crosses. The prediction with unfilled circles varies in an uncorrelated fashion around the mean. We see that these variations, due to differing structural error models, are insignificant compared to the seasonal variations and are hardly distinguishable. This can be seen from Fig. 5 - σ^2 is around 0.1, whereas log(LH) varies between 2.5—4.5 during a year. This also provides an estimate of the relative magnitudes of the structural error vis-à-vis predictions.

Given the small differences in both the posterior distributions of the parameters and the predictive skill of the models when the two structural error models are used, the simpler structural error model based on uncorrelated errors is preferable. However, the use of the temporally correlated model does reveal the timescales of the structural error (around 5.5 months). This, in turn, can help identify and improve parameterizations of physical processes that may be contributing to them and potentially result in reduced model structural uncertainty.

Finally, we explore the impact of climatological averaging. This reduces the time-series from 48 months to 12; we model the structural error as uncorrelated to reduce the dimensionality of the calibration problem. The deterministic calibration revealed $\mathbf{p}_{opt} = \{0.1, \log(5.9 \times 10^{-4}), 1.0\}$, which shows that the optimization has reached the edge of the prior distribution for 2 out of 3 parameters. The deterministic optimization was seen to be sensitive to the starting guess and we report the best of 10 runs, starting from different guesses. In Fig. 7, top and middle rows, we plot the marginalized posterior PDFs with solid lines; with dashed lines, we plot the calibration obtained without climatological averaging and with uncorrelated structural errors. We see modest changes in the calibrations for F_{drai} and $\log(Q_{dm})$. Further, we see that, like the calibration studies above, the peaks of the PDF do not agree with the default values of the parameters. The calibrations for b are similar and very different from the default value. We also see that σ^2 is far smaller when the observations are climatologically averaged, as it reduces the impact of outliers e.g., the low log(LH) observations during months 30-36. Further, the peak of the PDF corresponds to the value obtained via deterministic calibration.

In Fig. 7 (lower left) we plot the results from the PPT, along with the prediction using $\mathbf{p}_{def}$ and $N_s = 200$. Clearly, the default CLM parameters over-predict log(LH) and the calibration largely rectifies this shortcoming. The IQR of the predictions (the error bars) captures the observations. Lower right, we plot the VRH from the calibration. Clearly, the calibration is not ideal, but since the histogram reflects just 12 ranks, it is difficult to draw conclusions regarding the finer aspects of the calibration. In Table 1, we mention the MAE and CRPS for the calibration. These error metrics are almost half of those achieved with the non-averaged data. The MCMC method required 50,000 model invocations to reach a converged 4-dimensional posterior distribution (125,000 for the 5-dimensional one), when tested using the Raftery-Lewis method.

4.3. Calibration with US-MOz data.

We next estimate $\{F_{drai}, \log(Q_{dm}), S_y\}$ using data from US-MOz to check the variation of these parameters with sites. We could not construct accurate surrogates for US-MOz without climatological averaging, and consequently, we will perform calibration only with climatologically averaged data. The data (latent heat surface fluxes) spans 2004-2007, climatologically averaged monthly and log-transformed. Note that the surrogate models for US-MOz consist of a quadratic and a GP component. The model-observation mismatch is modeled as uncorrelated-in-

Table 1

CRPS and MAE for the four calibrations performed for US-ARM and US-MOz. The units of CRPS and MAE are the same as those of observations.

Calibration test case	MAE	CRPS
US-ARM, 48-months of data, correlated errors	0.37	0.18
US-ARM, 48-months of data, uncorrelated errors	0.37	0.18
US-ARM, climatologically-averaged data, uncorrelated errors	0.203	0.096
US-MOz, climatologically-averaged data, uncorrelated errors	0.205	0.098

time. The MCMC chain was run 50,000 steps to convergence. $\mathbf{p}_{opt} = \{2.639, \log(4.43 \times 10^{-3}), 0.2\}$, but the optimization method was seen to converge to multiple (local) minima depending upon the starting guess; the figures provided here correspond to the best of 10 runs. Note that the second parameter is not far from its default value (see § 3).

In Fig. 8 we plot the marginalized PDFs for the CLM4 parameters being calibrated, along with the prior. F_{drai} and $\log(Q_{dm})$ how strong disagreement with the default CLM values, though S_y peaks close to it. The PDF for F_{drai} and $\log(Q_{dm})$ are bimodal, which also explains the inaccuracy in $\mathbf{p}_{opt}$. The deterministic method correctly captured the peak in the S_y PDF, but converged to the smaller peaks (in fact, locations in the PDF with zero slope) in the PDFs for $\log(Q_{dm})$ and F_{drai}. MCMC, being a global optimization method, has the practical benefit of being resilient to many of the complexities of the optimization surface and locates the peak of the PDF which our 10 attempts with a deterministic optimization method failed to capture.

The three parameters show complex interdependence. There is a negative correlation between F_{drai} and S_y, with high values of F_{drai} compensating for lower S_y and a weak positive correlation between $\log(Q_{dm})$ and S_y. The plots of the samples that reveal these correlations are provided in the Supplementary Materials (Fig. S2) as well as in [41]. In Fig. 8, bottom left, we plot the PPT runs using $N_s = 200$. We see minor improvement over the default parameters. Thus the net contribution of the calibration are not new values of $\{F_{drai}, \log(Q_{dm}), S_y\}$ but rather the variability/uncertainty in their values that can be supported by the LH observations. Bottom right, we plot the VRH, which is inconclusive due to the small number of ranks being histogrammed. The MAE and CRPS values are in Table 1, and the PPT for US-MOz is seen to have errors similar to US-ARM.

Finally, we check if the calibration performed with surrogates improves the predictive skill of CLM4 (not the surrogates). We repeat the PPTs performed with surrogates for US-ARM and US-MOz using CLM4. Due to the cost of the simulation, only 32 CLM4 runs (instead of 200 for the surrogates) were used. In Fig. 9, we compare the PPTs performed using surrogates and CLM4. The error bars plot the median and IQR for surrogates; the $+$ symbol and dashed lines are the corresponding CLM4 plots. We see that the predictions using CLM4 are very close to those obtained using surrogates. Thus the improvement in the predictive skill of the CLM4 surrogates carry over to the original model itself.

4.4. Discussions. The four calibrations discussed above have led to parameter estimates that are clearly more predictive than CLM4's default settings. Further, they have demonstrated the importance of using MCMC for the calibration. Deterministic methods, in our case L-BFGS-B, showed a significant sensitivity to the starting guess and frequently fell into local minima that we later isolated in the PDFs of the parameters (for US-MOz) in Fig. 8. Further, the posterior distribution of the parameters bears no resemblance to a Gaussian and methods such as Ensemble Kalman Filters (which assume Gaussian distributions) should not be used to estimate them. Finally, the PDFs for the parameters are quite wide and parameter estimates are uncertain. The width of the PDFs could be due to the fact that the surrogates (and by implication, CLM4) are not sufficiently responsive to our three calibration parameters. This suspicion is bolstered by the VRH in Fig. 6 which shows ranks clustered at the top end, indicating an under-dispersive posterior prediction. The under-dispersed nature could be a reflection of model shortcomings or because we have varied only 3 parameters in this study. While these parameters are the most sensitive individually, their interaction with other parameters (which are currently held constant) need not have an insignificant effect on LH prediction.

The estimates could perhaps be improved i.e., the PDFs made narrower, by using a second observation stream. However, the previous calibration effort [53] identified that runoff, when used in conjunction with latent heat fluxes, was not very informative on the parameters of interest and our experiments with sensible heat fluxes (not presented here) removed it as a contender.

We found that climatological averaging had a modest impact on the PDFs of the estimated parameters. Note that the climatologically averaged dataset is quarter the size of the original one. The muted impact of such a drastic decrease in the observational dataset size seems to imply that the original observations were dominated by seasonal variability i.e., they could be approximated as minor variations about a repeated annual profile (the climatological mean). The smooth observational time-series obtained after climatological averaging also led to smaller structural error estimates and tighter posterior predictions (see CRPS and MAE in Table 1).

One of the main aims in this study was to model and estimate the structural error and explore the impact of the model on parameter estimation and prediction accuracy. We examined an uncorrelated-in-time and a temporally-correlated structural error model. Their impact on the parameter PDFs was modest and the effect on posterior predictions, smaller still. The latter was due to seasonal variation in LH, which dwarfed the structural error magnitude. From a purely predictive point of view, the simpler uncorrelated-in-time structural error model is preferable. However, the temporally-correlated error model identified the correlation timescale of the error, which in turn can be used to identify (models of) physical processes which may be responsible for it.

Different priors were used for the two structural error models. The uncorrelated-in-time structural error model used a non-informative conjugate prior; the other used informative exponential priors. Yet the estimates for the structural error magnitude from the two competing models are not too dissimilar and both are unequivocally better than the estimate obtained using L-BFGS-B. This implies that (1) L-BFGS-B failed to find the global optimal for the parameters and (2) the impact of the exponential priors was rather muted.

The use of surrogates proved to be a mixed blessing. It allowed us to develop converged PDFs of the parameters without recourse to approximations (except the surrogates themselves) and examine the impact of surrogate error models and climatological averaging. These would have been very time-consuming had we used CLM4 natively as in [53]. Yet the structural error that we estimate is that of the surrogate and not of CLM4. While that does not impact the correlation timescale of the structural error, its magnitude, σ^2, should be considered an approximation to CLM4's structural error.

Our calibration can, in principle, be compared with [53] in two ways: by comparing the posterior distributions of the parameters and by comparing the predictive skills of the two models. The parameters' posterior distributions do not agree. While our PDFs for US-ARM are unimodal, those in [53] are multimodal. In case of US-MOz, our PDFs are multimodal, as are the ones in [53], but the modes are quite different. Further, the study in [53] developed 4 separate posterior distributions, for four different values of “reference acceptance probability”, which has no counterpart in our conventional MCMC method. It is unclear which distribution one should compare to. Comparing the predictive skill of the calibrated model is far more difficult due to the difference in the dimensionality of the inverse problem (10 parameters in [53] to our 3). A 10-dimensional calibration will result in a larger predictive variability compared to our three-dimensional one; this larger variability can be captured by metrics such as CRPS. However [53] compared their calibration to observations using an ensemble mean prediction and its Root Mean Square Error (RMSE). Without variability information, ensemble predictions from two posterior distributions of differing dimensionalities cannot be directly compared.

There could be a number of causes of differences in the two calibrations. In [53], the authors calibrated 10 parameters to our 3; we have kept the remaining 7 fixed at their defaults. In addition, the calibration in [53] used CLM4 directly and does not incur errors due to surrogate modeling; given that such errors are around 4%, this is probably a minor contributor to the difference. Also, the convergence criterion used in [53] is based on the mean statistics of the posterior samples during the burn-in period, not convergence statistics on their PDFs. Reconciling the differences between these two calibrations is left for future work.

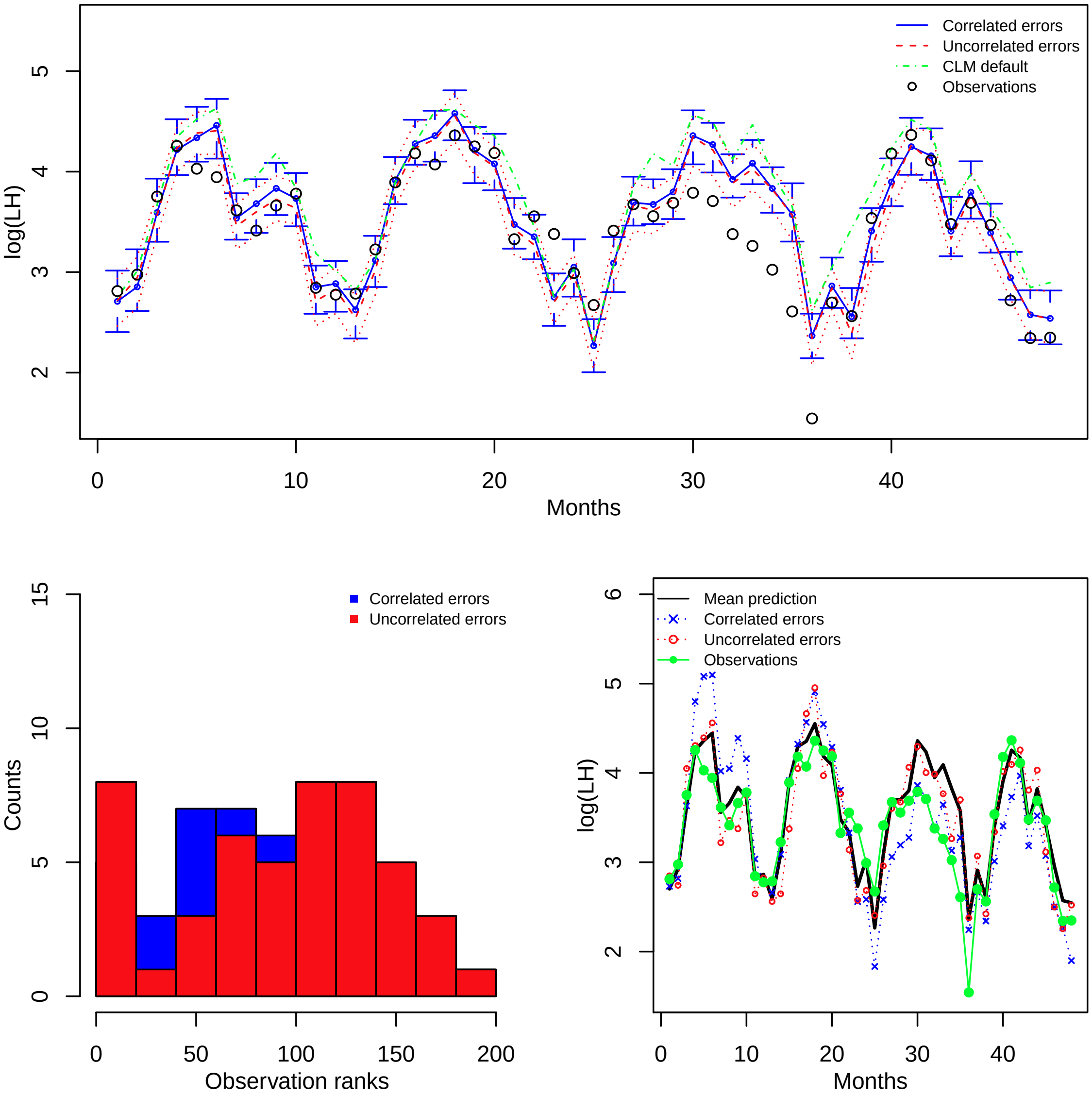

Figure 6. *Top: Results from the PPT performed using posterior distributions generated using both the correlated and uncorrelated models for the structural error, for US-ARM. The PPT tests were performed with 200 samples. The solid line is the median prediction, from the correlated-errors calibration; the dashed line is the corresponding prediction from the uncorrelated-error calibration. The error bars denote the inter-quartile range (IQR). The observations of log(LH) are plotted with symbols. The prediction with* $\mathbf{p}_{def}$ *is plotted with a dotted line. Lower left: VRH for both the calibrations, using blue for correlated-errors calibration and red for the other. Lower right: Comparison of two realizations of predictions vis-à-vis the observations (solid circles). We plot the average prediction from the PPT, generated using correlated structural errors, with a solid line. One realization of these predictions is plotted with crosses; it shows the smooth variation in time that the observations show. The plot with unfilled circles shows a prediction generated using the uncorrelated structural error model. Compared to the seasonal variation in log(LH), the variation in predictions due to the two different structural error models is not very noticeable.*

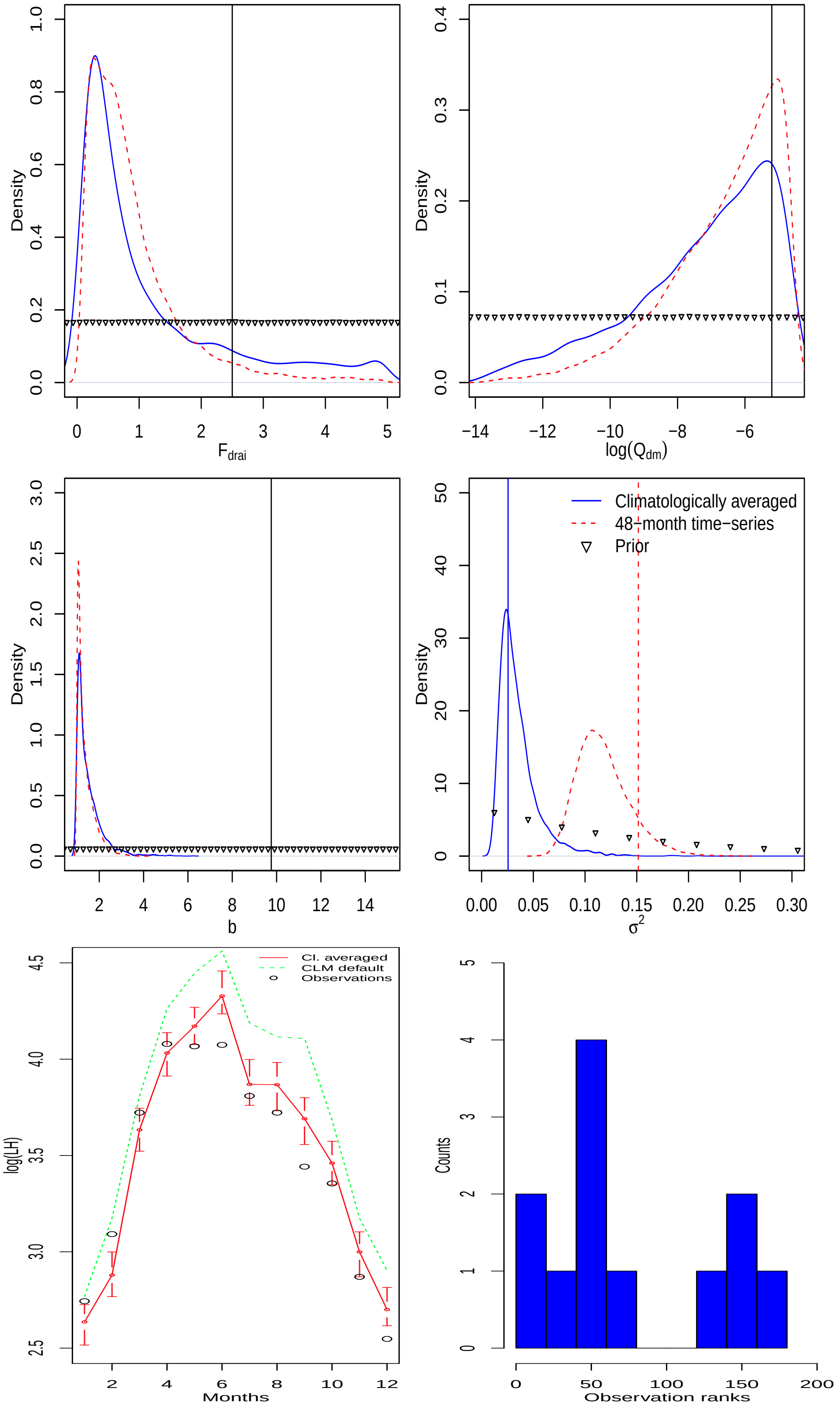

Figure 7. *Top and middle rows: Posterior distributions of* $\{F_{drai}, \log(Q_{dm}), b, \sigma^2\}$ *generated using climatologically-averaged log(LH) observations at US-ARM, plotted using solid lines. The dashed lines are the PDFs generated without climatological averaging (i.e., with a 48 month time-series) and using uncorrelated structural errors. The default parameter values are plotted as vertical lines. Bottom row, left: Predictions from the PPT, with 200 samples from the posterior distribution developed using climatologically averaged data. The error bars are the IQR and largely capture the observations. The prediction with* $\mathbf{p}_{def}$*, plotted with dashes, is clearly an over-prediction. Right: VRH for the calibration.*

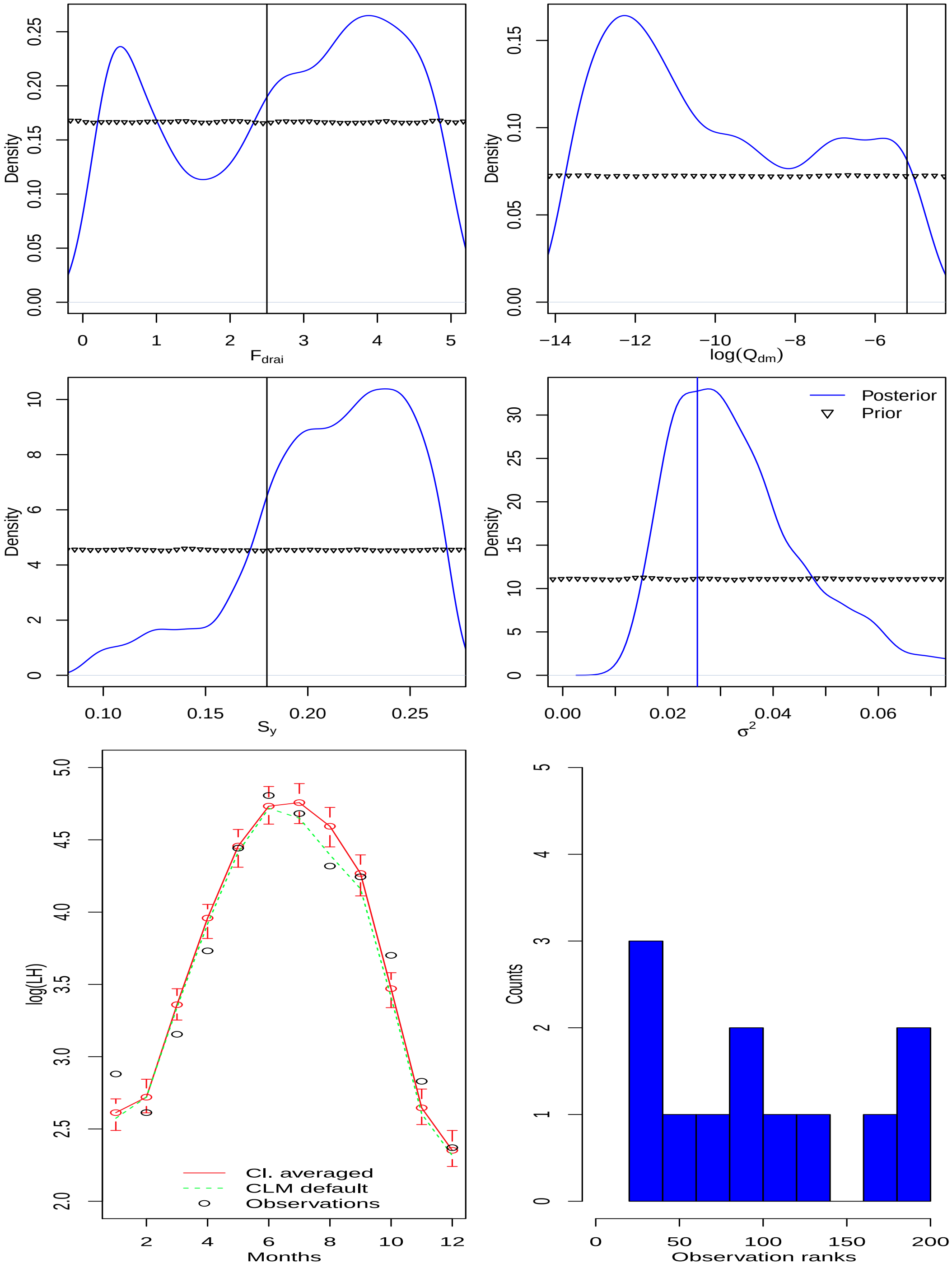

Figure 8. *Top and middle rows: Posterior distributions for* $\{F_{drai}, \log(Q_{dm}), S_y, \sigma^2\}$ *for US-MOz, using climatologically averaged observations. The priors are plotted with symbols and the default values are vertical lines. The vertical line for* σ^2 *is the value obtained using deterministic calibration. Bottom left: PPT results from the Bayesian calibration using US-MOz data. Bottom right: The VRH for the calibration.*

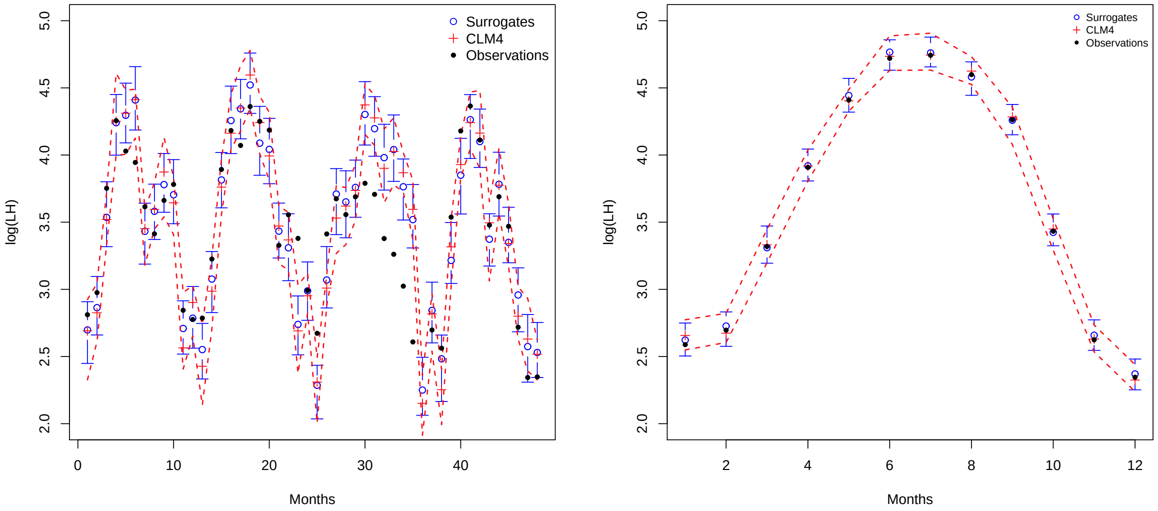

Figure 9. *Left: Plot of the PPT (200 runs) conducted with surrogates for US-ARM. The median prediction is plotted with a ○, and the error bars denote the IQR. The + symbol is the median prediction from a 32-run PPT performed with CLM4; the dashed lines denote the IQR. Observations are plotted with ●. We see that the two PPTs are very similar and the improvement in predictive skill of the surrogates holds true for the original model too. Right: The same comparison, with the same outcome, for US-MOz.*

5. Conclusions. We have investigated the Bayesian calibration of three hydrological parameters of CLM4 using observations of monthly-averaged latent heat fluxes, collected over a 4-year period. The choice of these parameters was deliberate. They are the most important hydrological parameters that control the seasonality of heat fluxes at the two sites studied here (as revealed by the sensitivity studies in [22]). Accurate prediction of the seasonality of heat fluxes is a fundamental requirement of any climate model. The seasonal nature of the processes involved allowed us to filter out fast time-scale (e.g., daily) variations (“noise” from the viewpoint of seasonal variations) and focus on the accuracy of the calibration under two competing error models. Such a Bayesian calibration and error estimation study has not been done before for CLM4. Finally, it allowed us to present an example of how complex models, such as CLM4, could be subjected to rigorous statistical calibration and uncertainty quantification.

In this study, we computed the posterior distribution of the parameters using surrogate models of CLM4 and MCMC. The surrogate models were constructed using polynomial trend functions and GP modeling. The Bayesian inverse problem posed to estimate the parameters incorporated two alternative representations of the structural error (or the model—data discrepancy). We investigated their impact on the parameter estimates and the predictive skill, after calibration. We also explored the impact of using the climatological mean of the observations for the calibration. We demonstrated our method on data from two sites, US-ARM and US-MOz, each with three unknown parameters.

We developed an approach to construct surrogate models for CLM4. In particular, we investigated a shrinkage regression method, Bayesian Compressive Sensing (BCS), to fit a polynomial model to a training set of CLM4 runs. BCS was augmented with cross-validation to construct a robust procedure for devising polynomial surrogates for computationally expensive models. The method is general, and can be used elsewhere.

We found that Bayesian calibration led to posterior distributions of parameters that improved the predictive skill of CLM4. The marginal PDFs of the parameters were quite wide i.e., there is a considerable amount of uncertainty in the parameter estimates. The choice of the structural error model impacts the parameters' PDFs modestly and its effect on the posterior predictions is marginal. However, the more sophisticated model allowed us to estimate the time-scale of the structural error, which can help identify and improve models of the physical processes that contribute to the error.

In the single case of US-ARM where we could check the effect of climatological averaging, its impact was rather muted on the estimated parameters. We conjecture that this may be because seasonal variability is the dominant signal in LH observations. Since this is largely preserved during climatological averaging, the PDFs of the estimated parameters did not change much.

The parameter estimates that we developed for the two sites do not agree between themselves, nor do they agree with the default values used in CLM4. This is not entirely surprising since the parameters depend on the hydrologic regimes associated with local soil properties, topological and geologic conditions. These vary significantly in North America. The default CLM4 parameter values were developed to better constrain simulated hydrologic budgets at continental and global scales. Consequently, they are "globally averaged" constants in some sense, and are not expected to be equally predictive locally. Consequently if CLM4 is to be used at individual sites such as flux towers or watersheds, recalibration is recommended. As observational data from a single site is likely to contain measurement errors (which will then propagate into parameters estimated from them), we would advocate an estimation procedure that quantifies uncertainty, e.g., the Bayesian one that we have developed. Finally, it is unknown whether the parameter estimates developed for a site can be re-used for predictive CLM4 runs at similar sites. The transferability of parameter estimates across sites under similar hydrologic regimes is now being investigated under a follow-up study [42].

Our calibration yielded PDFs which are at variance with those developed in a previous calibration study. The two investigations are similar, but not identical, with respect to observations, the calibration parameters and the numerical method. We have speculated about the causes of this discrepancy, but identifying the causes is beyond the scope of this study. We will investigate it in the future.

REFERENCES

[1] *NACP Site: Tower Meteorology, Flux Observations with Uncertainty, and Ancillary Data*. http://daac.ornl.gov/NACP/guides/NACP_Site_Tower_Met_and_Flux_v2.html.

[2] J. D. Annan, J. C. Hargreaves, N. R. Edwards, and R. Marsh, *Parameter estimation in an intermediate complexity Earth system model using an ensemble Kalman filter*, Ocean modeling, 8 (2005), pp. 135–154.

[3] M. Aubinet, T. Vesala, and D. Papale (Eds.), *Eddy Covariance: A Practical Guide to Measurement and Data Analysis*, Springer Atmospheric Sciences, Springer Verlag, 2012.

[4] S. D. Babacan, R. Molina, and A. K. Katsaggelos, *Bayesian compressive sensing using Laplace priors*, IEEE Transactions on Signal Processing, 19 (2010).

[5] S. Brooks and A. Gelman, *General methods for monitoring convergence of iterative simulations*, Journal of Computational and Graphical Statistics, 7 (1998), pp. 434–445.

[6] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, *A limited memory algorithm for bound-constrained optimization*, SIAM Journal on Scientific Computing, 16 (1995), pp. 1190–1208.

[7] R. Craiu, J. Rosental, and C. Yang, *Learn from thy neighbor: Parallel-chain regional adaptive MCMC*, Journal of the Americal Statistical Association, (2009), pp. 1454–1466.

[8] W. N. Edeling, P. Cinnella, R. P. Dwight, and H. Bijl, *Bayesian estimates of parameter variability in the k-ε turbulence model*, Journal of Computational Physics, 258 (2013), pp. 73–94.

[9] M. Emory, R. Pecnik, and G. Iaccarino, *Modeling structural uncertainties in Reynolds-Averaged computations of shock/boundary layer interactions*, in 49th AIAA Aerospace Sciences Meeting, 2011.

[10] J. W. Hurrell et. al, *The Community Earth System model: A framework for collaborative research*, Bulletin of the American Meteorological Society, 94 (2013), pp. 1339–1360.

[11] Y. Q. Luo et al, *A framework for benchmarking land models*, Biogeosciences, 9 (2012), pp. 3857–3874.

[12] G. Evensen, *Data assimilation : The ensemble Kalman filter*, Springer, 2007.

[13] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian data analysis*, Chapman & Hall/ CRC, 2004, ch. Model checking and improvement.

[14] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice*, Chapman and Hall, 1996.

[15] Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery, *Probabilistic forecasts, calibration and sharpness*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69 (2007), pp. 243–268.

[16] Tilmann Gneiting and Adrian E Raftery, *Strictly proper scoring rules, prediction, and estimation*, Journal of the American Statistical Association, 102 (2007), pp. 359–378.

[17] M. Gohler, J. Mai, and M. Cuntz, *Use of eigendecomposition in a parameter sensitivity analysis of the Community Land Model*, Journal of Geophysical Research: Biogeosciences, 118 (2013), pp. 904–921.

[18] L. Gu, W. J. Massman, R. Leuning, S. G. Pallardy, T. Meyers, P. J. Hanson, J. S. Riggs, K. P. Hosman, and B. Yang, *The fundamental equation of eddy covariance and its application in flux measurements*, Agricultural and Forest Meteorology, 152 (2012), pp. 135–148.

[19] L. Gu, T. Meyers, S. G. Pallardy, P. J. Hanson, B. Yang, M. Heuer, K. P. Hosman, J. S. Riggs, D. Sluss, and S. D. Wullschleger, *Direct and indirect effects of atmospheric conditions and soil moisture on surface energy partitioning revealed by a prolonged drought at a temperate forest site*, Journal of Geophysical Research, 111 (2006). D16102.

[20] Heikki Haario, Marko Laine, Antoinietta Mira, and Eero Saksman, *DRAM-Efficient adaptive MCMC*, Statistics and Computing, 16 (2006), pp. 339–354.

[21] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*, Springer, 2009.

[22] Z. Hou, M. Huang, L. R. Leung, G. Lin, and D. M. Ricciuto, *Sensitivity of surface flux simulations to hydrologic parameters based on an uncertainty quantification framework applied to the Community Land Model*, Journal of Geophysical Research, 117 (2012). D15108.

[23] M. Huang, Z. Hou, L. R. Leung, Y. Ke, Y. Liu, Z. Fang, and Y. Sun, *Uncertainty analysis of runoff simulations and parameter identifiability in the Community Land Model - Evidence from MOPEX basins*, Journal of Hydrometeorology, (2013).

[24] L. Ingber, *Very fast simulated annealing*, Mathematical and Computer Modeling, 12 (1989), pp. 967–973.

[25] C. Jackson, M. K. Sen, and P. L. Stoffa, *An efficient stochastic Bayesian approach to optimal parameter and uncertainty estimation for climate model predictions*, Journal of Climate, 17 (2004), pp. 2828–2841.

[26] H. Järvinen, P. Räisänen, M. Laine, J. Tamminen, A. Lin, E. Oja, A. Solonen, and H. Haario, *Estimation of ECHAM5 climate model closure parameters with adaptive MCMC*, Atmospheric Chemistry and Physics, 10 (2010), pp. 9993–10002.

[27] B. J.Cosby, G. M. Hornberger, R. B. Clapp, and T. R. Ginn, *A statistical exploration of the relationships of soil moisture characteristics to the physical properties of soils*, Water Resources Research, 20 (1984), pp. 682–690.

[28] J.-C. Jouhaud, P. Sagaut, B. Enaux, and J. Laurenceau, *Sensitivity analysis and multiobjective optimization for LES numerical parameters*, Journal of Fluid Engineering, 130 (2008), p. 021401.

[29] M. C. Kennedy and A. O'hagan, *Bayesian calibration of computer models (with discussion)*, Journal of the Royal Statistical Society B, 63 (2001), pp. 425–464.

[30] J. Laurenceau and P. Sagaut, *Building efficient response surfaces of aerodynamic functions with kriging and cokriging*, AIAA Journal, 46 (2008), pp. 498–507.

[31] David M. Lawrence, Keith W. Oleson, Mark G. Flanner, Peter E. Thornton, Sean C. Swenson Peter J. Lawrence, Xubin Zeng, Zong-Liang Yang, Samuel Lewis, Koichi Sakaguchi, Gordon B. Bonan, and Andrew G Slater, *Parameterization improvements and functional and structural advances in version 4 of the community land model*, Journal of Advances in Modeling Earth Systems, 3 (2011).

[32] H. Lei, M. Huang, L. R. Leung, D. Yang, X. Shi, J. Mao, D. J. Hayes, C. R. Schwalm, Y. Wei, and S. Liu, *Sensitivity of global terrestrial gross primary production to hydrologic states simulated by the community land model using two runoff parameterizations*, Journal of Advances in Modeling Earth Systems, 6 (2014).

[33] G. Leng, M. Huang, Q. Tang, H. Gao, and L. R. Leung, *Modeling the effects of groundwater-fed irrigation on terrestrial hydrology over the conterminous united states*, Journal of Hydrometeorology, 15 (2014), pp. 957–972.

[34] G.-Y. Niu, Z.-L. Yang, R. E. Dickinson, and L.E. Gulden, *A simple TOPMODEL-based runoff parameterization (SIMTOP) for use in global climate models*, Journal of Geophysical Research, 111 (2005), p. D211106.

[35] G.-Y. Niu, Z.-L. Yang, R. E. Dickinson, L.E. Gulden, and H. Su, *Development of a simple groundwater model for use in climate models and evaluation with Gravity Recovery and Climate Experiment data*, Journal of Geophysical Research, 112 (2007), p. D07103.

[36] K. W. Oleson, D. M. Lawrence, G B. Bonan, M. G. Flanner, E. Kluzek, P. J. Lawrence, S. Levis, S. C. Swenson, and P. E. Thornton, *Technical description of version 4.0 of the Community Land Model (CLM)*, 2010.

[37] V. R. N. Pauwels, N. E. C. Verhoest, G. J. M. De Lannoy, V. Guissard, C. Lacau, and P. Defoumy, *Optimization of a coupled hydrology-crop growth model through the assimilation of observed soil moisture and leaf area index values using an ensemble Kalman filter*, Water Resources Research, 43 (2007). W04421.

[38] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0.

[39] A. Raftery and Steven M. Lewis, *Implementing MCMC*, in Markov Chain Monte Carlo in Practice, W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, eds., Chapman and Hall, 1996, pp. 115–130.

[40] C. E. Rasmussen and C. K. I Williams, *Gaussian process for machine learning*, MIT Press, 2006.

[41] J. Ray, Z. Hou, M. Huang, and L. Swiler, *Bayesian calibration of the community land model using surrogates*, SAND Report SAND2014-0867, Sandia National Laboratories, Livermore, CA 94551-0969, February 2013. Unclassified and unlimited release.

[42] H. Ren, Z. Hou, M. Huang, Y. Sun, T. Tesfa, and L. R. Leung, *Hydrological parameter sensitivity and transferability across 431 MOPEX basins and a new basin classification system*, Journal of Hydrology, (2014). Under review.

[43] W. J. Riley, S. C. Biraud, M. S. Torn, M. L. Fischer, D. P. Billesbach, and J. A. Berry, *Regional CO2 and latent heat surface fluxes in the Southern Great Plains: Measurements, modeling and scaling*, Journal of Geophysical Research – Biogeosciences, 114 (2009). G04009.

[44] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, *Design and analysis of computer experiments*, Statistical Science, 4 (1989), pp. 409–435.

[45] T. Santer, B. Williams, and W. Notz, *The design and analysis of computer experiments*, Springer, New York, NY, 2003.

[46] K. Sargsyan, C. Safta, H. N. Najm, B. J. Debusschere, D. Ricciuto, and P. Thornton, *Dimensionality reduction for complex models via Bayesian compressive sensing*, International Journal for Uncertainty Quantification, (2014). In press.

[47] T. W. Simpson, V. Toropov, V. Balabanov, and F. A. C. Viana, *Design and analysis of computer experiments in multidisciplinary optimization: A review of how far we have come or not*, in Proceedings of the 12th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference, no. AIAA Paper 2008-5802, Victoria, British Columbia, Canada, 2008.

[48] W. C. Skamarock and J. B. Klemp, *A time-split nonhydrostatic atmospheric model for weather research and forecasting applications*, Journal of Computational Physics, 227 (2008), pp. :3465–3485.

[49] K. Soetaert and T. Petzoldt, *Inverse modeling, sensitivity and Monte Carlo in R using package FME*, Journal of Statistical Software, 33 (2010), pp. 1–28.

[50] A. Solonen, P. Ollinaho, M. Laine, H. Haario, J. Tamminen, and H. Järvinen, *Efficient MCMC for climate model parameter estimation: Parallel adaptive chains and early rejection*, Bayesian Analysis, 7 (2012), pp. 715–736.

[51] C. B. Storlie and J. C. Helton, *Multiple predictor smoothing methods for sensitivity analysis: Description of techniques*, Reliability Engineering and System Safety, 94 (2008), pp. 28–54.

[52] C. B. Storlie, L. P. Swiler, J. C. Helton, and C. J. Sallaberry, *Implementation and evaluation of nonparametric regression procedures for sensitivity analysis of computationally demanding models*, Reliability Engineering and System Safety, 94 (2009), pp. 1735–1763.

[53] Y. Sun, Z. Hou, M. Huang, F. Tian, and L. Ruby Leung, *Inverse modeling of hydrologic parameters using surface flux and runoff observations in the Community Land Model*, Hydrology and Earth System Sciences, 17 (2013), pp. 4995–5011.

[54] A. E. Suyker and S. B. Verma, *Evapotranspiration of irrigated and rainfed maize-soybean cropping systems*, Agricultural and Forest Meteorology, 149 (2009), pp. 43–452.

[55] L. Tomassini, P. Reichert, R. Knutti, T. F. Stocker, and M. E. Borsuk, *Robust Bayesian uncertainty analysis of climate system properties using Markov chain Monte Carlo methods*, Journal of Climate, 20 (2007), pp. 1239–1254.

[56] W. N. Venables and B. D. Ripley, *Modern Applied Statistics in S*, Springer-Verlag, new York, NY, 2002.

[57] B. Yang, Y. Qian, G. Lin, L. R. Leung, P. J. Rasch, G. J. Zhang, S. A. McFarlane, C. Zhao, Y. Zhang, H. Wang, M. Wang, and X. Liu, *Uncertainty quantification and parameter tuning in the CAM5 Zhang-Mcfarlane convection scheme and impact of improved convection on the global circulation and climate*, Journal of Geophysical Research: Atmospheres, 118 (2013), pp. 395–415.

[58] B. Yang, Y. Qian, G. Lin, R. Leung, and Y. Zhang, *Some issues in uncertainty quantification and parameter tuning: A case study of convective parameterization in the WRF regional climate model*, Atmospheric Chemistry and Physics, 12 (2012), pp. 2409–2427.

[59] X. Zeng, B. A. Drewniak, and E. M. Constantinescu, *Calibration of the crop model in the Community Land Model*, Geosciences Model Development Discussions, 6 (2013), pp. 379–398.