




Article

Inverse Modeling of Hydrologic Parameters in CLM4 via Generalized Polynomial Chaos in the Bayesian Framework

Georgios Karagiannis ^{1,†} , Zhangshuan Hou ^{2,†} , Maoyi Huang ³ and Guang Lin ^{4,5,6,*} 

¹ Department of Mathematical Sciences, Durham University, Stockton Road, Durham DH1 3LE, UK; georgios.karagiannis@durham.ac.uk

² Energy and Environment Directorate, Pacific Northwest National Laboratory, Richland, WA 99352, USA; zhangshuan.hou@pnnl.gov

³ National Oceanic and Atmospheric Administration, Washington, DC 20230, USA; maoyi.huang@noaa.gov

⁴ Department of Mathematics, School of Mechanical Engineering, Purdue University, West Lafayette, IN 47907, USA

⁵ Department of Statistics (Courtesy), Purdue University, West Lafayette, IN 47907, USA

⁶ Department of Earth, Atmospheric, and Planetary Sciences (Courtesy), Purdue University, West Lafayette, IN 47907, USA

* Correspondence: guanglin@purdue.edu; Tel.: +1-765-494-1965

† These authors contributed equally to this work.

Abstract: In this work, generalized polynomial chaos (gPC) expansion for land surface model parameter estimation is evaluated. We perform inverse modeling and compute the posterior distribution of the critical hydrological parameters that are subject to great uncertainty in the Community Land Model (CLM) for a given value of the output LH. The unknown parameters include those that have been identified as the most influential factors on the simulations of surface and subsurface runoff, latent and sensible heat fluxes, and soil moisture in CLM4.0. We set up the inversion problem in the Bayesian framework in two steps: (i) building a surrogate model expressing the input–output mapping, and (ii) performing inverse modeling and computing the posterior distributions of the input parameters using observation data for a given value of the output LH. The development of the surrogate model is carried out with a Bayesian procedure based on the variable selection methods that use gPC expansions. Our approach accounts for bases selection uncertainty and quantifies the importance of the gPC terms, and, hence, all of the input parameters, via the associated posterior probabilities.

Keywords: uncertainty quantification; generalized polynomial chaos; Bayesian inversion; US-ARM; inverse modeling



Citation: Karagiannis, G.; Hou, Z.; Huang, M.; Lin, G. Inverse Modeling of Hydrologic Parameters in CLM4 via Generalized Polynomial Chaos in the Bayesian Framework.

Computation **2022**, *10*, 72. <https://doi.org/10.3390/computation10050072>

Academic Editor: Demos T. Tsahalidis

Received: 22 March 2022

Accepted: 29 April 2022

Published: 5 May 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Land surface models (LSMs) are used by scientists to quantitatively simulate the exchange of water and energy fluxes at the Earth–atmosphere interface. LSMs have progressed from simplistic schemes defining simply the surface boundary conditions for general circulation models (GCMs) to complicated models incorporating modules representing biogeochemical, hydrological, and energy cycles at the surface–atmosphere interface over the last few decades (Pitman [1]). The model parameters are frequently related with definite physical meaning and have an influence on the primary model outputs, such as water and energy fluxes. They are built on mathematical formulations of the laws of physics. The assumption for Inter-comparison of Land Surface Parameterization Schemes is that the parameters are quantifiable and transferable between sites with similar physical properties or site conditions (PILPS) (Bastidas et al. [2], Henderson-Sellers et al. [3,4]). However, default assignment of parameter values are actually inappropriate according to (Bastidas et al. [2], Rosero et al. [5]). In the meantime, given the high dimensionality of the parameter space and the complexity of the land surface system, more research is needed to determine

which parameters are more uncertain and the potential for using observations to constrain or calibrate the uncertain parameters in order to better capture uncertainty in the resulting land surface states (Hou et al. [6], Huang et al. [7]). The goal of this research is to quantify the uncertainty in a subset of parameters in a community LSM called the Community Land Model (CLM), which is the land component of the Community Earth System Model (CESM) (formerly known as the Community Climate System Model (CCSM) (Collins et al. [8], Gent et al. [9], Lawrence et al. [10])). It is possible that model structural uncertainty in LSMs is attributable to simplified assumptions or representations of actual processes or events. Many of these assumptions hold true only in certain situations. Furthermore, LSMs are prone to ambiguity in terms of input parameter values, owing to the fact that many input characteristics, such as those related to land cover and land use, as well as soil qualities, are difficult to assess at the scales at which they are utilized. A common technique in land surface modeling is to define a set of universally applicable default parameter values. Over the last two decades, the land surface modeling community has worked hard to resolve uncertainty in model parameters, data, and model structure. The purpose of this study is to reduce model parameter uncertainty using generalized polynomial expansion and Bayesian inversion.

In a high-dimensional parameter space, the computational cost of stochastic inversion is significant. Surrogate models can be utilized to solve this problem instead of numerical simulators. Ensemble simulations, which are required to build surrogate models, can be efficiently executed in a task-parallel manner on supercomputing facilities. Surrogate creation, on the other hand, is a lengthy procedure. Surrogates are not utilized very often in climate model or LSM calibration. Using observations of latent heat fluxes, the authors used different surrogates (e.g., polynomials and/or universal kriging) to calibrate hydrological parameters of CLM 4.0 in (Ray et al. [11], Huang et al. [12]).

For the model–data mismatch, two competing models were utilized to construct a composite of measurement error and (a crude estimate of) CLM structural error. In (Gong et al. [13]), the authors employed adaptive surrogate-based optimization to estimate the parameters of the Common Land Model employing six observables at once; 12 independent parameters were (deterministically) calibrated.

Using Bayesian compressive sensing (BCS) and polynomial chaos expansions, (Sargsyan et al. [14]) sought to create surrogates for five variables of interest from CLM4 with prognostic carbon and nitrogen modules turned on (i.e., CLM4-CN) (PCEs). They discovered that the input–output relationship in CLM4-CN could be made up of qualitatively diverse regimes (i.e., living or dead vegetation regimes associated with various locations in the parameter space), necessitating the use of clustering and classification to generate piecewise PCEs.

The applicability of employing gPC for CLM4 hydrological model calibration is investigated in this study. Based on [15], we provide a completely Bayesian technique that combines fully Bayesian statistics, variable selection, and generalized polynomial chaos surrogate models to handle the uncertainty quantification and model inversion problem in CLM4. The procedure produces a cheap mathematical/statistical approximation of the model output (latent heat flux) as a function of a set of model parameters. Bayesian inversion of the model parameters' given observations is performed by using the produced cheap gPC surrogate model instead of the expensive computer model output in the likelihood function, and then by performing Bayesian parametric inference facilitated by Markov chain Monte Carlo methods. The method allows for dimension reduction and the selection of the important model parameters that significantly influence the output parameter by computing inclusion posterior probabilities.

The layout of the paper is as follows: in Section 2, we describe the study site, input data, and the conducted numerical simulations, and present the parameters of interest that we calibrate; in Section 3, we present the inversion methodology using gPC in the Bayesian framework; in Section 4, we evaluate the inversion results; in Section 5, we draw our conclusions.

2. Dataset and Parameterization

The FLUXNET database (www.fluxdata.org, accessed on 20 March 2022) contains half-hourly observations of ecosystem CO₂, heat fluxes, and meteorological data of more than 250 sites worldwide and for a total of 960 site-years. The study site in this study is US-ARM (ARM Southern Great Plains site, Lamont, Oklahoma) (Fischer et al. [16], Torn [17]). It has a vegetation type of croplands, covered with temporary crops followed by harvest and a bare soil period (e.g., single and multiple cropping systems), a humid subtropical climate, and clay-type soil texture.

Observational data used in parameter estimation are observed latent heat fluxes and runoff measurements, which are processed and gap-filled to obtain daily and monthly averaged data. We consider 10 critical hydrological parameters that are likely to have dominant impacts on the simulation of surface and subsurface runoff, latent and sensible fluxes, and soil moisture, as suggested in existing literature (Hou et al. [6], Niu et al. [18,19], Oleson et al. [20,21]). The selected parameters are f_{max} , C_s , f_{over} , f_{drain} , $q_{drain,max}$ (denoted as Q_{dm} hereinafter), S_y , b , Y_s , K_s , and θ_s . Explanations of the 10 parameters and their prior information are shown in Table 1 in Hou et al. [6]. Prior distributions of the parameters are specified as independent truncated normal distributions with location, scale, and range parameters specified according to the values in Table 1 in Hou et al. [6]; in the cases that prior variance was not available, we considered uniform prior distributions. We recall that such values for the hyper-parameters are acceptable according to existing literature Oleson et al. [20,21]. It is important to highlight that, although truncated normals, these priors are essentially flat due to their large variances. Moreover, 256 samples were generated by using quasi Monte Carlo (QMC) sampling on the basis of the prior pdfs; QMC provides a good dispersion between sample points and can achieve a good uniformity in dimensionality at around 10 Caflisch [22], Wang and Sloan [23]. Numerical simulations corresponding to sampled parameter sets were conducted, which yield the data matrix of inputs (i.e., realizations of the 10 parameters) and outputs (i.e., latent heat fluxes), which enables the development of response surfaces or surrogates that can be used for sensitivity analysis, parameter ranking, and model calibration.

3. Bayesian Methodology

We describe a synergy of Bayesian methods aiming to quantify the importance of input CLM4 model parameters and calibrate these parameters against real measurements, as well as build a surrogate model describing the input–output relation in the CLM4 model, in the Bayesian framework.

3.1. Bayesian Inverse Problem Setup

Bayesian inverse methods allow for the uncertainty quantification of input parameters of a computer model from observations [24]. Bayesian inference is performed through the posterior distribution, which is derived according to the Bayes theorem. We need to specify two components: the likelihood function representing experimental information from the measurements, and the prior distribution representing the researcher's prior information about the uncertain parameters.

We consider that the output value observed $u^f \in \mathbb{R}^{d_u}$ is associated to some unknown input ζ via a forward model (e.g., CLM4) $u(\cdot) : \mathbb{R}^{d_\zeta} \rightarrow \mathbb{R}$, and possibly contaminated by some additive observational noise ε^f (residuals); namely $u^f = u(\zeta) + \varepsilon^f$. A reasonable assumption is to model ε^f as a normally distributed variable with mean zero and variance σ^2 . This is justified by central limit theorem arguments, and the fact that, in the noise term, there are several insignificant random measurement errors accumulated from observations and modeling errors due to model parameterization. Thus, the likelihood is:

$$\mathcal{L}(u^f|\zeta) = N_{d_u}(u^f|u(\zeta), \text{diag}(\sigma_1^2, \dots, \sigma_{d_u}^2)).$$

To account for the uncertainty about the input parameters we assigned prior distribution $\zeta \sim \pi_\zeta(\cdot) = \prod_{i=1}^{d_\zeta} \pi_{\zeta_i}(\cdot)$, where $\pi_{\zeta_i}(\zeta_i)$ are considered to be shifted beta distributions

Beta($a_{\xi_i}, b_{\xi_i}, \xi_{i,\min}, \xi_{i,\max}$), with known $\xi_i \in [\xi_{i,\min}, \xi_{i,\max}]$ and $a_{\xi_i} > 0, b_{\xi_i} > 0$. Here, $\xi_{i,\min}$, and $\xi_{i,\max}$ can be specified because a reasonable range of the model parameter is often a priori known. In addition, a_{ξ_i} and b_{ξ_i} can be specified by using the method of moments or the maximum entropy method [25] because, often, reasonable values for the mean and variance of the model parameter are a priori known. A computationally convenient choice for the priors of the variance $\{\sigma_j^2\}$ is the inverse gamma prior distribution $IG(a_{\sigma^2}, b_{\sigma^2})$ with parameters $a_{\sigma^2} > 0$, and $b_{\sigma^2} > 0$, as it is a semi-conjugate prior to the likelihood. By considering the likelihood variance σ^2 as random unknown parameters and treating the problem in the Bayesian framework, we let the data decide a proper value for σ^2 through the posterior distribution. The prior hyper-parameters are considered to be fixed values and are pre-defined by the researcher.

The posterior distribution of ξ, σ^2 given the observations u^f is

$$\pi_{\xi, \sigma^2}(\xi, \sigma^2 | u^f) = \frac{\mathcal{L}(u^f | \xi, \sigma^2) \pi_{\xi}(\xi) \pi_{\sigma^2}(\sigma^2)}{\int \mathcal{L}(u^f | \xi, \sigma^2) \pi_{\xi}(\xi) \pi_{\sigma^2}(\sigma^2) d(\xi, \sigma^2)} \tag{1}$$

by using the Bayes' theorem, and marginal distribution $\pi_{\xi}(\xi | u^f)$ quantifies the uncertainty about the model parameters ξ .

The posterior distribution density (1) is usually intractable. If $u(\cdot)$ was known or was cheap to compute, inference on ξ could be obtained using standard Markov chain Monte Carlo (MCMC) methodology [26]. In Algorithm 1, we present a simple MCMC sampler that updates ξ and σ^2 iteratively in two stages. The output sample $\{(\xi^{(t)}, \sigma^{2,(t)})\}_{t=1}^T$ of Algorithm 1 can be used to perform inference on ξ and σ^2 , e.g., expectation of any function $h(\cdot)$ of ξ can be approximated as

$$E(h(\xi)) \approx \frac{1}{T} \sum_{t=1}^T h(\xi^{(t)})$$

for large T .

In the present application of the CLM4 model, $u(\cdot)$ is prohibitively expensive to compute iteratively because the forward model (CLM4) is too expensive to run. This prevents us from directly using this method, and, particularly, Algorithm 1. To overcome this issue, we built a cheap but accurate proxy (called surrogate model) for $u(\cdot)$, and plugged it into (1). The construction of such a surrogate model in the Bayesian framework is discussed in Section 3.2.

Algorithm 1 Updates of a single MCMC swap.

Update ξ :

- Simulate ξ from a Metropolis–Hastings transition probability that targets $\pi_{\xi | \sigma^2}(\xi | u^f, \sigma^2)$, where $\pi_{\xi | \sigma^2}(\xi | u^f, \sigma^2) \propto N(u^f | u(\xi), \sigma^2) N(\xi | \mu_{\xi}, \sigma_{\xi}^2)$.

Update σ^2 :

- Draw σ^2 from $IG\left(\frac{1}{2} + a_{\sigma^2}, \frac{1}{2} |u^f - u(\xi)|_2^2 + b_{\sigma^2}\right)$.
-

3.2. Surrogate Model Specification

We describe a fully Bayesian procedure for building a surrogate model to be used as a cheap but accurate proxy of $u(\cdot)$ in (1) based on gPC expansions and MCMC methods. The highlight is that, apart from evaluating a surrogate model, the procedure is able to quantify the importance of each PC basis via inclusion posterior probabilities, which allows for the selection of the important input parameters: Fmax, Cs, Fover, Fdrai, Qdm, Sy, B, Psis, Ks, thetas.

3.2.1. Generalized Polynomial Chaos Expansion

We consider the output parameter $u(\xi)$ as a function of the vector of random input variables $\xi \in \Xi \subseteq \mathbb{R}^{d_\xi}$ that follow distribution $f(\cdot)$.

The output parameter $u(\xi)$ can be modeled as an expansion of PC bases $\{\psi_\alpha(\cdot)\}$ and PC coefficients $\{c_\alpha\}$:

$$u(\xi) = \sum_{\alpha \in \mathbb{N}_0^{d_\xi}} c_\alpha \psi_\alpha(\xi), \tag{2}$$

for $\xi \sim f(\cdot)$ [27]. We denote multi-indices $\alpha := (\alpha_1, \dots, \alpha_{d_\xi})$ of size d_ξ that are defined on a set of non-negative integers $\mathbb{N}_0^{d_\xi} := \{(\alpha_1, \dots, \alpha_{d_\xi}) : \alpha_j \in \mathbb{N} \cup \{0\}\}$. The family of polynomial bases $\{\psi_\alpha(\cdot); \alpha \in \mathbb{N}_0^{d_\xi}\}$ contains multidimensional orthogonal polynomial bases with respect to the probability measure $f(\cdot)$ of ξ . Each multidimensional PC basis $\psi_\alpha(\cdot)$ results as a tensor product of univariate orthogonal polynomial bases $\psi_{\alpha_j}(\cdot)$ of degree $\alpha_j \in \mathbb{N}_0^1$, namely:

$$\psi_\alpha(\xi) = \prod_{j=1}^{d_\xi} \psi_{\alpha_j}(\xi_j), \alpha_j \in \mathbb{N}_0^1, \tag{3}$$

where $E_f(\psi_{\alpha_j}(\xi) \psi_{\alpha_{j'}}(\xi)) = Z_j \delta_0(j - j')$, for $j, j' = 1, \dots, d_\xi$ and $Z_j = E_f(\psi_{\alpha_j}^2(\xi))$.

Often, the PC bases $\{\psi_\alpha(\cdot)\}$ are chosen so that they are orthogonal with respect to the distribution $f(\cdot)$. A quite general family of polynomials is the Askey family [28], whose members are associated with standard distributions. In this work, we focus on the use of Jacobi polynomial bases [28], which can be defined recursively as:

$$\begin{aligned} \psi_0(z^{(\xi)}) &= 1; \\ \psi_1(z^{(\xi)}) &= \frac{1}{2}[a - b + (a + b - 2)z^{(\xi)}]; \\ \psi_j(z^{(\xi)}) &= \frac{(2j + a + b - 1)[(2j + a + b)(2j + a + b - 2)z^{(\xi)} + a^2 + b^2]}{2j(j + a + b)(2j + a + b - 2)} \psi_{j-1}(z^{(\xi)}) \\ &\quad - \frac{2(j + a + b - 1)(j + b - 1)(2j + a + b)}{2j(j + a + b)(2j + a + b - 2)} \psi_{j-2}(z^{(\xi)}), \\ &\quad j = 2, \dots, p_\xi - 1, \end{aligned}$$

where $z^{(\xi)}$ is a linear transformation $z^{(\xi)} : [\xi_{\min}, \xi_{\max}] \rightarrow [-1, 1]$, and ξ_{\min}, ξ_{\max} are the minimum and maximum of ξ .

In practice, for computational reasons, we truncate (2) by considering a finite sub-set of PC bases based on the *total truncated rule*. Hence, the expansion form becomes:

$$u_{p_\xi}(\xi) := \sum_{\alpha \in A} c_\alpha \psi_\alpha(\xi), \tag{4}$$

which uses, for only a finite set of multi-indices, A , such that $A = \{\alpha \in \mathbb{N}^{d_\xi} : \sum_{i=1}^{d_\xi} \alpha_i \leq p_\xi\}$ with cardinality $m_\xi := \frac{(p_\xi + d_\xi)!}{p_\xi! d_\xi!}$. Other truncation rules can be adopted [14,29].

The PC coefficients $\{c_\alpha; \alpha \in \mathbb{N}_0^{d_\xi}\}$ are equal to $c_\alpha = E_f(u(\xi) \psi_\alpha(\xi)) / Z_\alpha$, where $Z_\alpha = E_f((\psi_\alpha(\xi))^2)$, for $\alpha \in \mathbb{N}_0^{d_\xi}$ [27]; however, they are intractable. Moreover, the number of unknown PC coefficients is of order $d_\xi^{p_\xi}$ and grows rapidly with the dimension d_ξ and PC degree p_ξ . This causes computational problems such as over-fitting [15,30], especially when the dimensionality of ξ is high. If we consider a smaller PC degree or carelessly omit PC bases, it may lead to a significant increase in bias and hence produce inaccurate surrogate models. Hence, there is a particular interest in keeping only the inputs or bases that significantly affect the output in the gPC. The Bayesian procedure in Section 3.2.2 effectively addresses these matters.

3.2.2. Bayesian Training Procedure

We describe a stochastic and automatic Bayesian procedure that accurately evaluates the PC coefficients and the gPC surrogate model, while it allows for the selection of the the significant PC bases, and hence input model parameters. This procedure is able to trade off efficiently between the bias (caused by omitting bases) and the over-fitting, which are both required. Furthermore, it can select the significant PC bases and estimate the PC coefficients simultaneously, while providing credible intervals.

We assume that there is a training dataset $\mathcal{D} = \{(u_j, \xi_j)\}_{j=1}^{n_\xi}$ available, where n_ξ is the size of the dataset, ξ_j denotes the random input value, and $u_j := u(\xi_j)$ denotes the output value corresponding to the j -th input value ξ_j . Then, we model

$$u_j = u_{p_\xi}(\xi_j) + \epsilon_j, \text{ for } j = 1, \dots, n_\xi, \tag{5}$$

where $\epsilon_j \in \mathbb{R}$ is the error term, associated with the j -th observation. Equation (5) can be written in matrix form:

$$u = Xc + \epsilon, \tag{6}$$

where $u := (u_j; j = 1 : n_\xi)^\top$, $\epsilon := (\epsilon_j; j = 1 : n_\xi)^\top$, $X_a := (\psi_a(\xi_j); j = 1 : n_\xi)$, and $X := (X_a; a \in A)$ is an $n_\xi \times m_\xi$ dimensional matrix of basis functions.

The likelihood $\mathcal{L}(u|c, \sigma^2) := \mathcal{L}(\{u_j\}|\{\xi_j\}, c, \sigma^2)$ is:

$$\begin{aligned} \mathcal{L}(u|c, \sigma^2) &= \prod_{j=1}^{n_\xi} \mathbf{N}(u_j | \psi(\xi_j)^\top c, \sigma^2); \\ &= \mathbf{N}(u | X^\top c, \mathbb{I}_m \sigma^2), \end{aligned} \tag{7}$$

where $\mathbf{N}(\cdot | \mu, \sigma^2)$ denotes the Gaussian probability density function with mean μ and variance σ^2 . The likelihood function is considered as a proximity measure of the truncated gPC expansion to the observations. We specify a hierarchical prior model $\pi(c, \gamma, \sigma^2, \lambda, \rho)$:

$$\begin{aligned} c_a | \gamma_a, \sigma^2, \lambda &\sim \gamma_a \mathbf{N}(c_a | 0, \sigma^2 / \lambda) + (1 - \gamma_a) \delta_0(c_a), \quad a \in A; \\ \gamma_a &\sim \text{Bernoulli}(\rho), \quad a \in A; \\ \sigma^2 | a_\sigma, b_\sigma &\sim \text{IG}(a_\sigma, b_\sigma); \\ \lambda | a_\lambda, b_\lambda &\sim \text{G}(a_\lambda, b_\lambda); \\ \rho | a_\rho, b_\rho &\sim \text{Beta}(a_\rho, b_\rho), \end{aligned} \tag{8}$$

where $a_\lambda, b_\lambda, a_\rho,$ and b_ρ are fixed prior hyper parameters, and predetermined by the researcher. In the Bayesian framework, inference on the uncertain model parameters can be performed based on the posterior distribution

$$\pi(c, \gamma, \sigma^2, \lambda, \rho | \mathcal{D}) = \frac{\mathcal{L}(u|c, \sigma^2) \pi(c, \gamma, \sigma^2, \lambda, \rho)}{\int \mathcal{L}(u|c, \sigma^2) \pi(c, \gamma, \sigma^2, \lambda, \rho) d(c, \gamma, \sigma^2, \lambda, \rho)}. \tag{9}$$

Particular interest lies in the computation of the inclusion probabilities $\pi(\gamma_a | \mathcal{D})$ that refer to the marginal posterior probability that the a -th basis is important, and the $\pi(c_a | \mathcal{D}, \gamma_a)$ that refers to the posterior density of the a -th PC coefficient.

To fit the Bayesian model, we resort to Markov chain Monte Carlo (MCMC) methods because the posterior distribution (9) is intractable and cannot be sampled directly. The conjugate prior model (8) allows for the design of a Gibbs sampler [31,32] whose updates involve sampling from the full conditional distributions of the parameter updated. In Algorithm 2, we represent a pseudo-code of one sweep of the Gibbs samples, along with the associated full conditional distributions. We highlight that the procedure is fully automatic because there is no need to tune the algorithmic parameters involved in Algorithm 2. The

notation c_{-a} (and X_{-a}) refers to the vector c (and matrix X) excluding the a -th element (and column). Moreover, we denote $m_{\xi,\gamma} = \sum_{a \in A} \gamma_a$.

Algorithm 2 Updates of the Gibbs swap.

Update $\{(\gamma_a, c_a)\}$: For $a \in A$,

1. Compute μ_a , and s_a^2 , where

$$\begin{aligned} \mu_a &= \left(X_a^\top X_a + \lambda \right)^{-1} X_a^\top (u - X_{-a} c_{-a}), \\ s_a^2 &= \sigma^2 (X_a^\top X_a + \lambda)^{-1} \end{aligned}$$

2. Update γ_a : draw γ_a from Bernoulli($P_a^{(\gamma)}$), where

$$P_a^{(\gamma)} = \left[1 + \frac{1 - \rho}{\rho} \sqrt{\frac{2\pi\sigma^2}{\lambda}} \mathbf{N}(0 | \mu_a, s_a^2) \right]^{-1},$$

3. Update c_a : draw c_a from $\pi(c_a | u, X, \gamma, c_{-a}, \rho, \sigma^2, \lambda)$, where

$$\pi(c_a | u, X, \gamma, c_{-a}, \rho, \sigma^2, \lambda) = \begin{cases} \delta_0(c_a) & , \text{ if } \gamma_a = 0 \\ \mathbf{N}(c_a | \mu_a, s_a^2) & , \text{ if } \gamma_a = 1 \end{cases}$$

Update σ^2 :

- Sample σ^2 from $\text{IG}\left(\frac{n_\xi}{2} + \frac{m_{\xi,\gamma}}{2} + a_\sigma, b_\sigma + \frac{n_\xi}{2} |u - Xc|_2 + \frac{1}{2} \lambda |c|_2^2\right)$.

Update λ : ,

- Sample λ from $\text{G}\left(\frac{m_{\xi,\gamma}}{2} + a_\lambda, \frac{1}{2\sigma^2} |c_a|_2^2 + b_\lambda\right)$.

Update ρ :

- Sample ρ from $\text{Beta}(m_{\xi,\gamma} + a_\rho, m_\xi - m_{\xi,\gamma} + b_\rho)$.
-

In order to evaluate the surrogate model, as well as quantify the importance of the input model parameters, we consider two Bayesian procedures. One is the *Bayesian model averaging* [33] and the other is the *median probability model*.

Bayesian Model Averaging

Bayesian model averaging is most suitable when we are interested in the predictive ability of the surrogate model.

Let $\left\{ \left(\gamma^{(t)}, c^{(t)}, \sigma^{2,(t)}, \lambda^{(t)}, \rho^{(t)} \right) \right\}_{t=1}^T$ be a Gibbs sample from Algorithm 2. Estimates and associated standard errors of $\{c_a\}$ can be computed as $\hat{c}_a = \frac{1}{T} \sum_{t=1}^T c_a^{(t)}$ and $\text{s.e.}(\hat{c}_a) = s_a^c \sqrt{\frac{q_a^c}{T}}$, where s_a^c is the sample standard deviation and q_a^c is the sample integrated autocorrelation time of $\{c_a^{(t)}; t = 1 : T\}$ for $a \in A$ [26,33]. Estimates for μ , v , and $u(\xi)$ can be computed by Monte Carlo integration using the ergodic average of quantities in Gibbs sample; for instance:

$$\begin{aligned}
 \hat{u}(\xi) &= \frac{1}{T} \sum_{t=1}^T \left(\sum_{a \in A} c_a^{(t)} \psi_a(\xi) \right) = \sum_{a \in A} \hat{c}_a \psi_a(\xi); & (10) \\
 \hat{P}_a &= \frac{1}{T} \sum_{t=1}^T \mathbb{1}(\gamma_a); \\
 \hat{\mu} &= \frac{1}{T} \sum_{t=1}^T c_0^{(t)} = \hat{c}_0, & \hat{\vartheta} &= \frac{1}{T} \sum_{t=1}^T \sum_{\alpha \in A - \{0\}} (c_\alpha^{(t)})^2 Z_\alpha; \\
 \hat{\sigma}^2 &= \frac{1}{T} \sum_{t=1}^T \sigma^{2,(t)}, & \hat{\lambda} &= \frac{1}{T} \sum_{t=1}^T \lambda^{(t)}; \\
 \hat{c}_a &= \frac{1}{T} \sum_{t=1}^T c_a^{(t)}, & \hat{\rho} &= \frac{1}{T} \sum_{t=1}^T \rho^{(t)}.
 \end{aligned}$$

Median Probability Model Based Evaluation

A parsimonious (or sparse) surrogate model involving only significant basis functions and important input model parameters can be obtained by examining the estimated inclusion probabilities $\pi(\gamma_a|\mathcal{D})$. A suitable probabilistic basis selection mechanism is the median probability model (MPM) [34].

Let $\left\{ \left(\gamma^{(t)}, c^{(t)}, \sigma^{2,(t)}, \lambda^{(t)}, \rho^{(t)} \right) \right\}_{t=1}^T$ be a Gibbs sample from Algorithm 2. The posterior distribution that the PC basis is significant can be estimated as $\hat{P}_a = \hat{\pi}(\gamma_a|\mathcal{D}) = \frac{1}{T} \sum_{t=1}^T \gamma_a^{(t)}$. We call $\{P_a\}$ the marginal inclusion posterior probabilities. The inclusion parameters are estimated as $\hat{\gamma}_a^{(MPM)} = \mathbb{1}(\hat{P}_a \in (0.5, 1))$, for $a \in A$. MPM consists of PC coefficients $\{c_a\}$, and PC bases $\{\psi_a(\cdot)\}$, whose marginal inclusion probabilities are such that $\hat{P}_a > 0.5$ (and hence $\hat{\gamma}_a^{(MPM)} = 1$). Hereafter, we will refer to them as significant PC coefficients and PC bases.

After the selection of the significant PC bases, inference about the uncertain parameters can be made by re-running Algorithm 2 for fixed γ equal to $\hat{\gamma}^{(MPM)}$ in order to obtain a larger Gibbs sample. The new Gibbs sample can be used to perform inference and estimation. For instance, estimates for μ , v , and $u(\xi)$ can be computed by Monte Carlo integration and using the equations of the estimators in (10). Note that a number of the coefficients $\{c_a^{(t)}\}$ will be constantly zero for all $t = 1, \dots, T$. The reason is because, unlike in the BMA approach, here, we consider only a single subset of PC bases, and hence the inclusion parameter is a fixed parameter equal to $\gamma^{(t)} = \gamma^{(MPM)}$, for $t = 1, \dots, T$.

The MPM approach allows the selection of the important input model parameters that significantly affect the output. If an input parameter ζ_j is not represented by any significant PC basis in the gPC expansion, it would be reasonable to consider that input parameter ζ_j does not significantly affect the output model and, hence, is omitted from the analysis.

4. Analysis of the US-ARM Data-Set

Here, we consider the US-ARM data set. The main interest lies in computing the posterior distributions of the 10 (random) input parameters of CLM4 $\xi = (\text{Fmax}, \text{Cs}, \text{Fover}, \text{Fdrai}, \text{Qdm}, \text{Sy}, \text{B}, \text{Psis}, \text{Ks}, \theta_s)$, given an observed latent heat flux (LH) measurement $u^{(f)}$ as shown in Table 1.

We apply the above methodology, which involves two stages: (i) building a surrogate model to replace the accurate but expensive forward model CLM4 according to the methodology in Section 3.2, and (ii) conducting inversion of the 10 inputs according to the theory in Section 3.1.

Table 1. Observed value $u^{(f)}$ for the US-ARM data-set.

Months	Measurement			
January-April	15.542	22.017	41.365	59.095
May-July	58.377	58.813	45.107	41.362
August-December	31.250	28.645	17.635	12.778

Surrogate Model Building Step

For each month, we built a surrogate model that maps the input of CLM4 ξ to the output LH u . For this type of dataset, Hou et al. [6] and Sun et al. [35] observed that the dependency between the output parameters LH corresponding to different months is weak and, hence, can be neglected. Therefore, here, we built surrogate models for each month independently. An advantage of assuming independence is that it leads to a simpler parameterization for the statistical model, which is easier to treat and interpret.

To build the gPC expansion, we followed the procedure in Section 3.2. For the design of the gPC expansion, we considered PC bases from the Jacobi polynomial family, and the total truncated rule with polynomial degree 4. The bases in the gPC expansion had the following arrangement: first, the main effects as Fmax, Cs, Fover, Fdrai, Qdm, Sy, B, Psis, Ks, θ , and, then, the interactions by increasing order, while interactions of the same order were arranged so that their interacting bases correspond to main effects in the aforesaid order. This rule identifies the order of the columns of the design matrix X , as well as that of the elements of the vector C in (6). The parameters of the Jacobi PC bases were set according to the prior information of the input of CLM4 in ([6], Table 1) by matching the moments of the corresponding shifted beta distribution, to which, they are orthogonal. We considered the prior model (8) with hyper-parameters $a_\lambda = 10^{-3}$, $b_\lambda = 10^{-3}$, $a_\sigma = 10^{-3}$, $b_\sigma = 10^{-3}$, $a_\rho = 1$, and $b_\rho = 1$. This choice of hyper-parameters leads to weakly informative priors. This is a reasonable choice because there is lack of prior information about the parameterization of the surrogate model. For training the gPC expansion, we used the training data set US-ARM. We ran Algorithm 2 for $2 \cdot 10^5$ iterations, where the first 10^5 were discarded as burn-in.

In Figures 1 and 2, we present the marginal posterior probabilities $\{\Pr(\gamma_j|\mathcal{D})\}$ computed by the ergodic average of the occurrences of the corresponding PC bases in the Gibbs sample. We observe that, during the period May-August, the marginal inclusion probabilities are higher than those of the rest months. This indicates that the input parameters of CLM4 may have a larger impact on the output LH during those months. From the hydrology point of view, this is expected because LH is higher on average and has a larger variability during these months, and the effects of hydrological parameters are expected to be more pronounced in the summer months.

We can infer that the input parameters Fmax, Cs, Fdrai, Qdm, and B are significant according to the MPM rule. This is because these input parameters are represented by significant PC basis functions; namely, the corresponding marginal inclusion probabilities in Figures 1 and 2 are greater than 0.5. The order of the indices in Figures 1 and 2 is the same as that mentioned above in the surrogate model building step section; namely, first, the main effects are followed by their interactions in the order of Fmax, Cs, Fover, Fdrai, Qdm, Sy, B, Psis, Ks, θ . The first 11 indexes correspond to the main effects, indices 12–66 correspond to the second-order interactions, indices 67–286 correspond to third-order interactions, and the rest of the indices correspond to fourth-order interactions. From the hydrology perspective, this is reasonable, because these parameters are major factors controlling the drainage and runoff generation, which, in turn, impact heat fluxes. The results are also consistent with the previous work in (Hou et al. [6]).

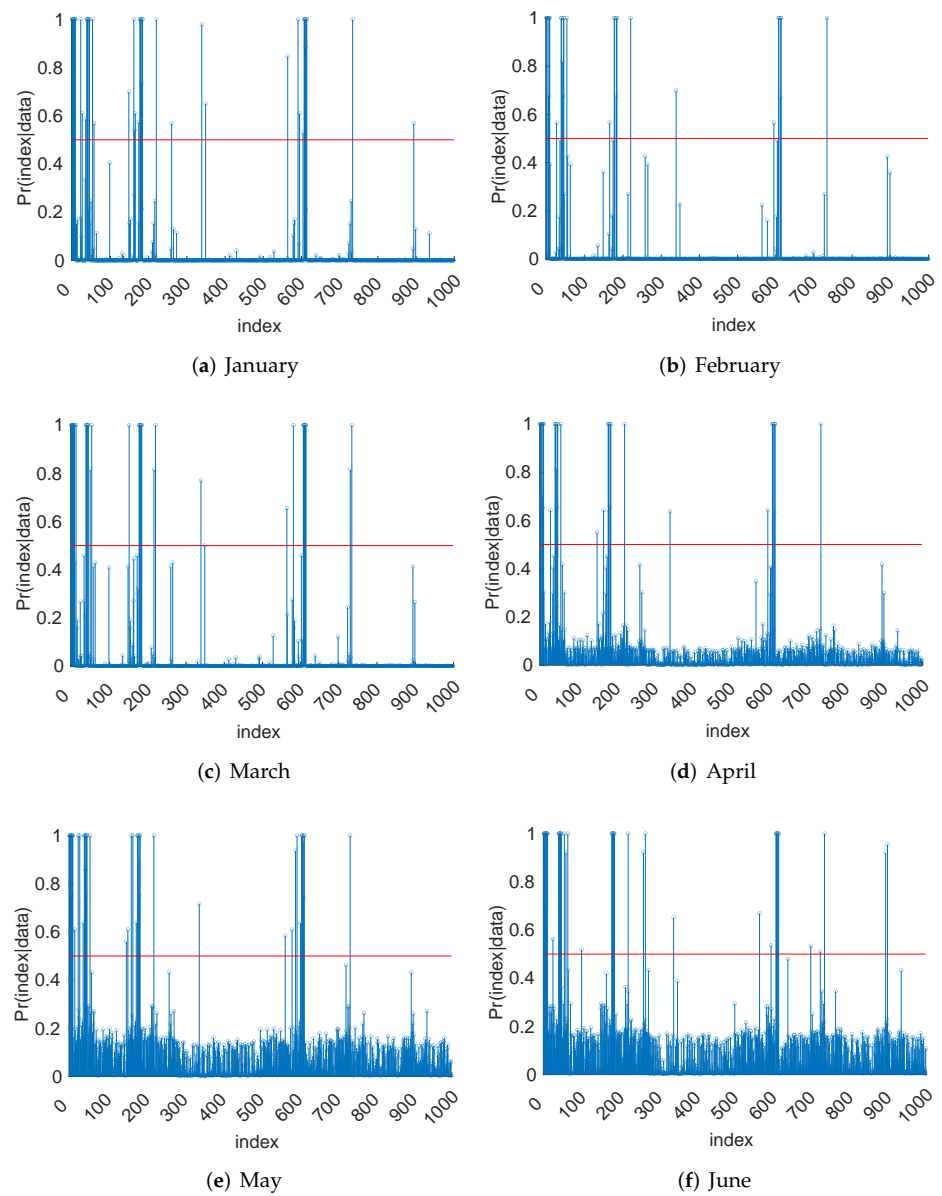


Figure 1. Plots of the posterior marginal inclusion probabilities: January–June.

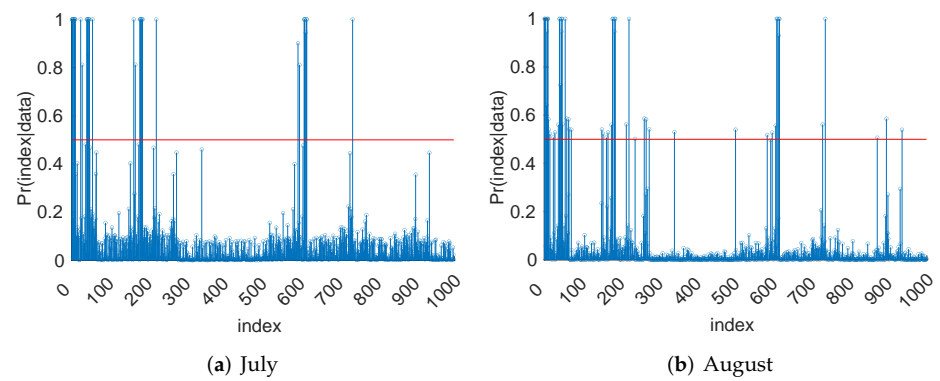


Figure 2. Cont.

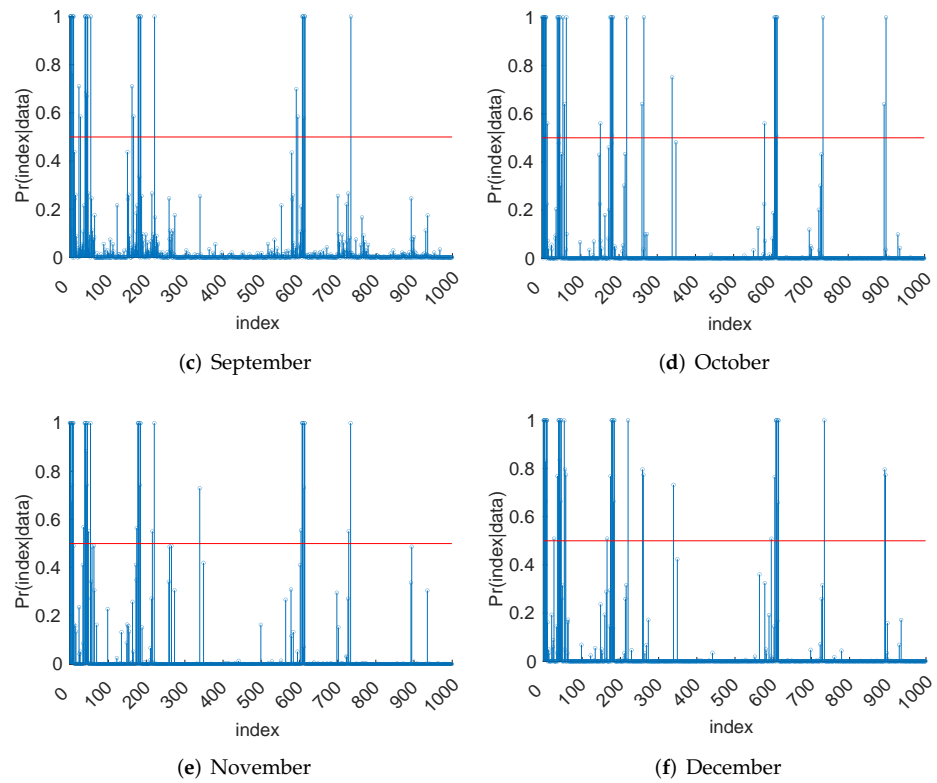


Figure 2. Plots of significant posterior marginal inclusion probabilities: July–December.

In Figures 3 and 4, we use box-plots to represent the posterior density estimates of the PC coefficients generated by the Gibbs sampler. In these figures, we consider only the coefficients corresponding to bases characterized as significant according to the MPM rule; namely, those with $\Pr(\gamma_j|D) < 0.5$. We observe that significant coefficients do not include zero in the range where the main posterior mass is concentrated. This shows that the method is consistent. Moreover, we observe that the significant PC coefficients that correspond to the period May–August have, in general, larger absolute values. This indicates that the variance of the output LH during those months is larger compared to that of the rest months, which is as expected. Finally, we can possibly infer that different input parameters ζ may have a different effect of LH at different times, since we can observe different significant coefficients, or bases, in different months.

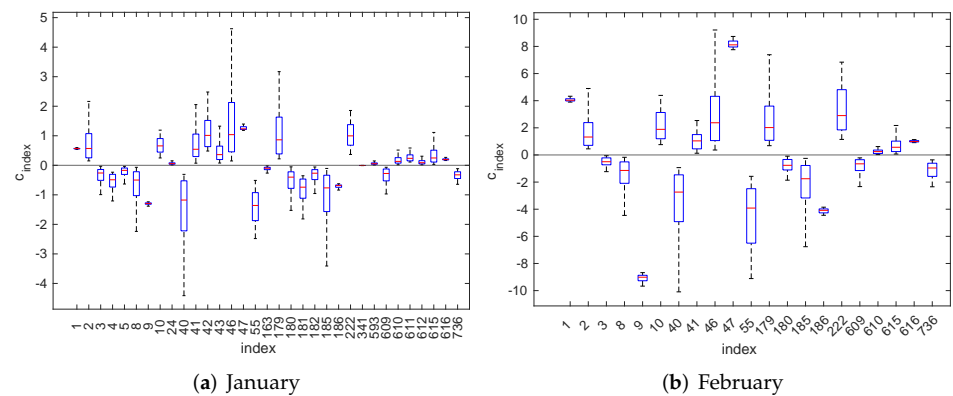


Figure 3. Cont.

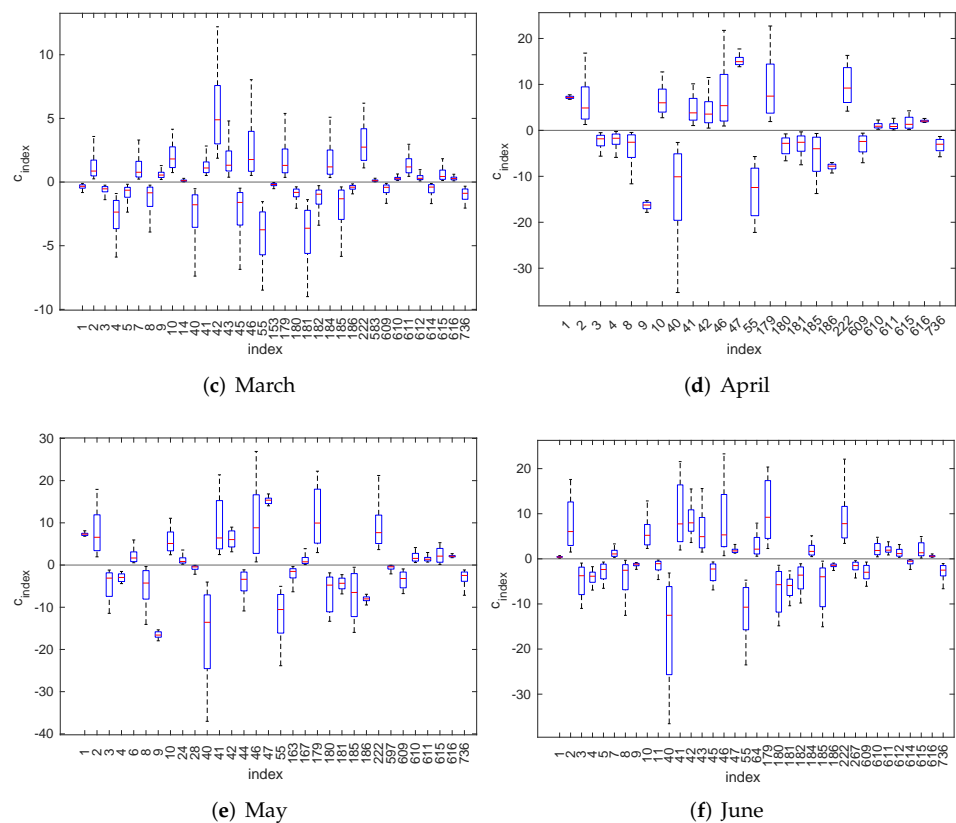


Figure 3. Box-plots of the posterior PC coefficients: January–June.

The gPC expansion of LH as a function of the input parameters ζ can be evaluated according to the estimator in (10), which is the ergodic average of the Gibbs sample.

4.1. Inversion Step

We calibrate the 10 CLM parameters $\zeta = (F_{max}, C_s, F_{over}, F_{drai}, Q_{dm}, S_y, B, P_{sis}, K_s, \theta)$ against the measurement of the parameter LH $u^{(f)}$ in Table 1, as in Section 3.1. We considered beta priors on the CLM parameters whose hyper-parameters are specified by using the method of moments and based on the prior information in Hou et al. [6].

Calibration was performed by running the MCMC sampler (Algorithm 1) and evaluating the posterior distributions according to the procedure in Section 3.1. Even though, in the previous step, we detected that the input model parameters F_{max} , C_s , F_{drai} , Q_{dm} , and B are the significant ones, we also considered them for calibration to obtain information about them as well. In order to make the MCMC sampler tractable, we replaced the forward model CLM4 $u(\cdot)$, in Algorithm 1, with the estimated gPC expansion that serves as a surrogate model. We ran the MCMC sampler (Algorithm 1) for $2 \cdot 10^4$ iterations and discarded the first 10^4 as burn-in.

In Figures 5 and 6, we present the estimated posterior densities of the input parameters of CLM4, as generated by the MCMC sampler. The blue bars correspond to the histogram estimate, whereas the red line corresponds to the kernel density estimate. These posterior distributions allow us to find a reasonable range of input values that correspond to the given value of output u^f . The associated box-plots of the marginal posteriors of the model parameters at each individual figure indicate the range of the main posterior density. We can see that we have successfully managed to shorten the ranges of the possible values for most of the input model parameters. For instance, we observe that the main density on the marginal posterior distribution of the density of Q_{dm} is around the area $[-4, -2.5]$ in \log_{10} scale.

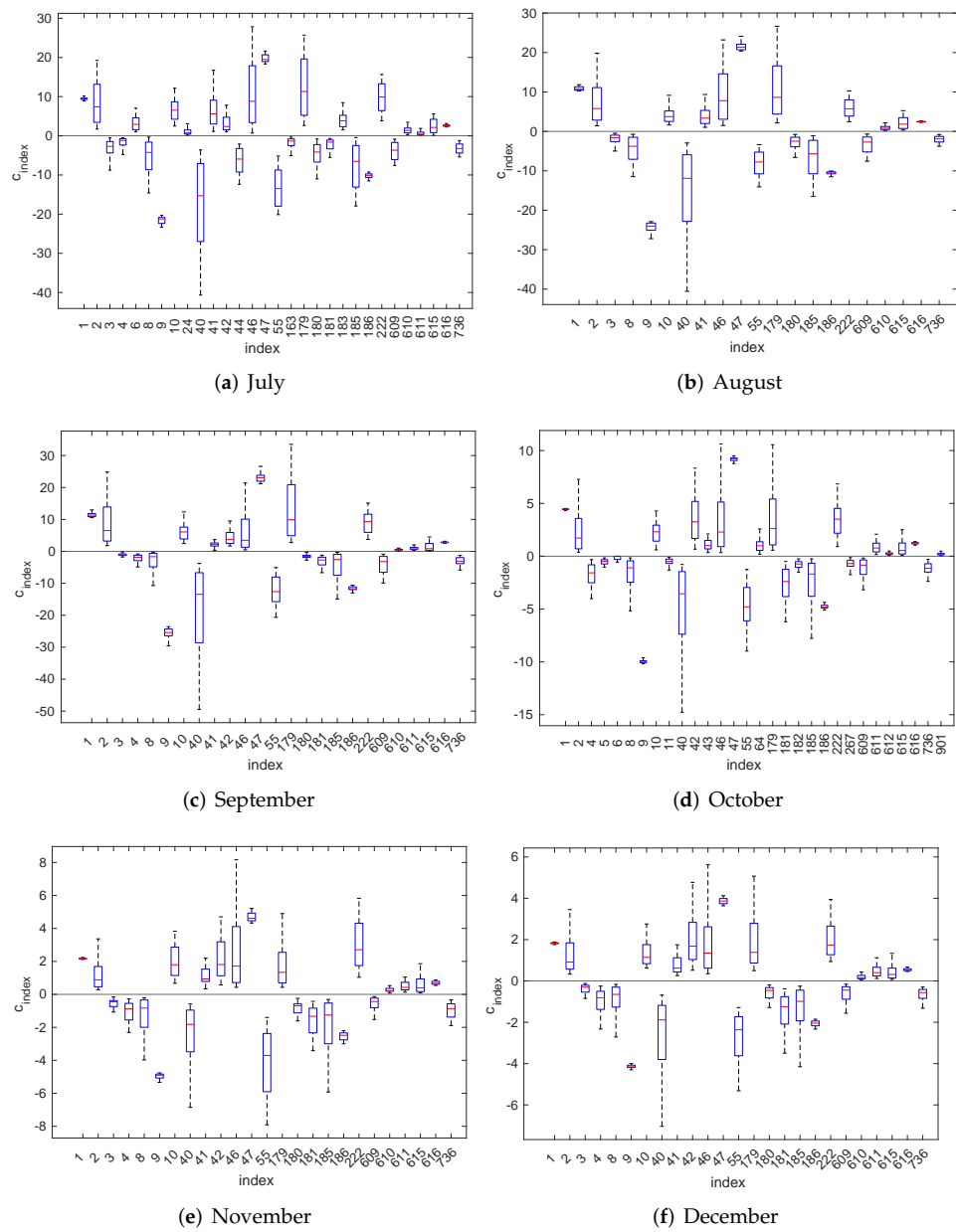


Figure 4. Box-plots of significant posterior PC coefficients: July–December.

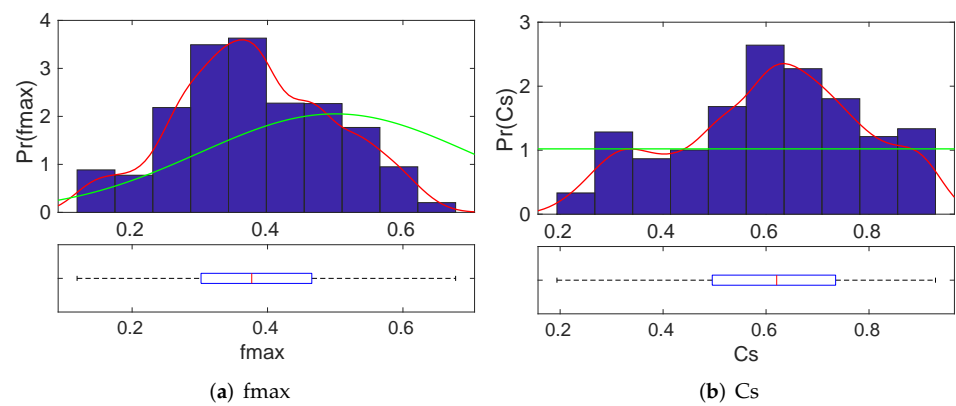


Figure 5. Cont.

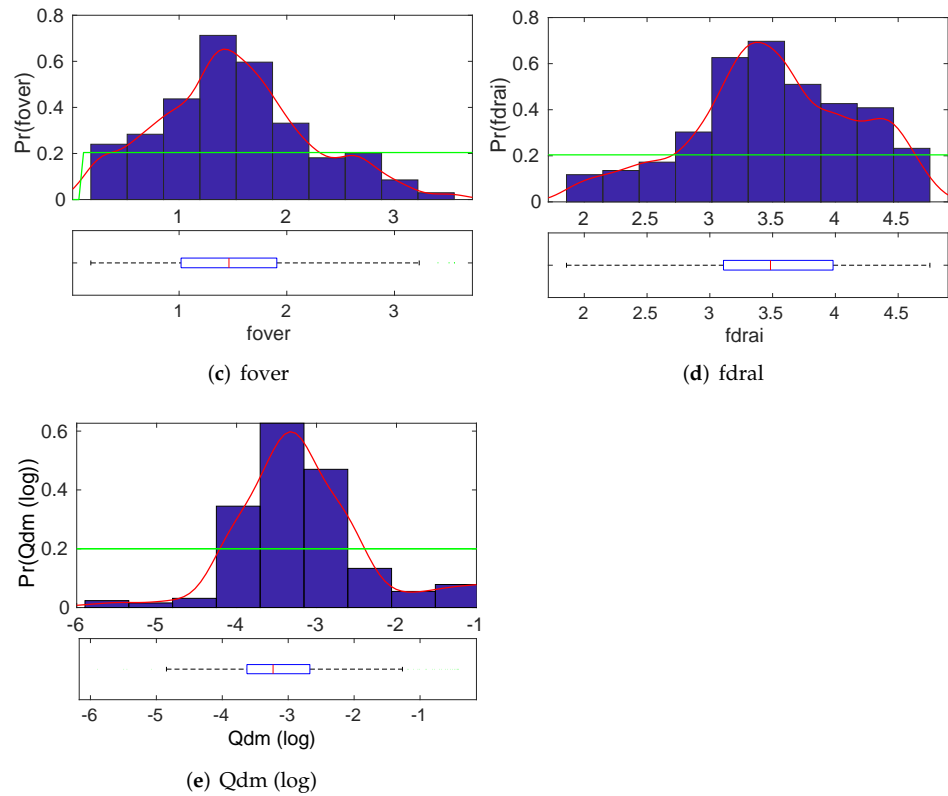


Figure 5. A posteriori distributions of the input parameters of CLM4 for a given output u^f . The red line indicates the posterior density, whereas the green line indicates the prior density of the parameter. (BMA evaluation).

4.2. Model Validation

We validated the effectiveness of the Bayesian inversion procedure. We evaluated the predictive distribution of the output parameter LH by using the derived gPC surrogate model and the MCMC sample of the input parameters generated by Algorithm 1. In Figures 7 and 8, we present the resultant predictive distributions of the output LH for the 12 months. The blue bars are the histogram estimate, the red line is the kernel density estimate of the predictive distribution, and the green arrow represents the observed output value of LH. The plots show that, for each month, the observed output value u^f for the LH lies below the modes of each of the marginal predictive distributions. This implies that the proposed methodology is valid and that the surrogate model derived from the method is able to produce accurate predictions.

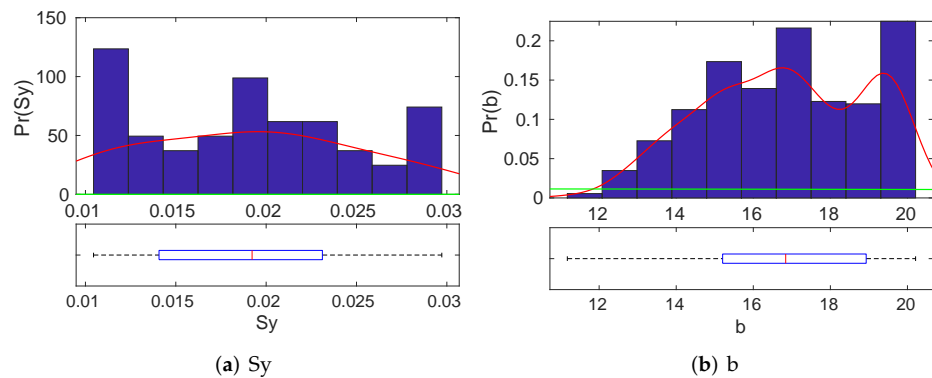


Figure 6. Cont.

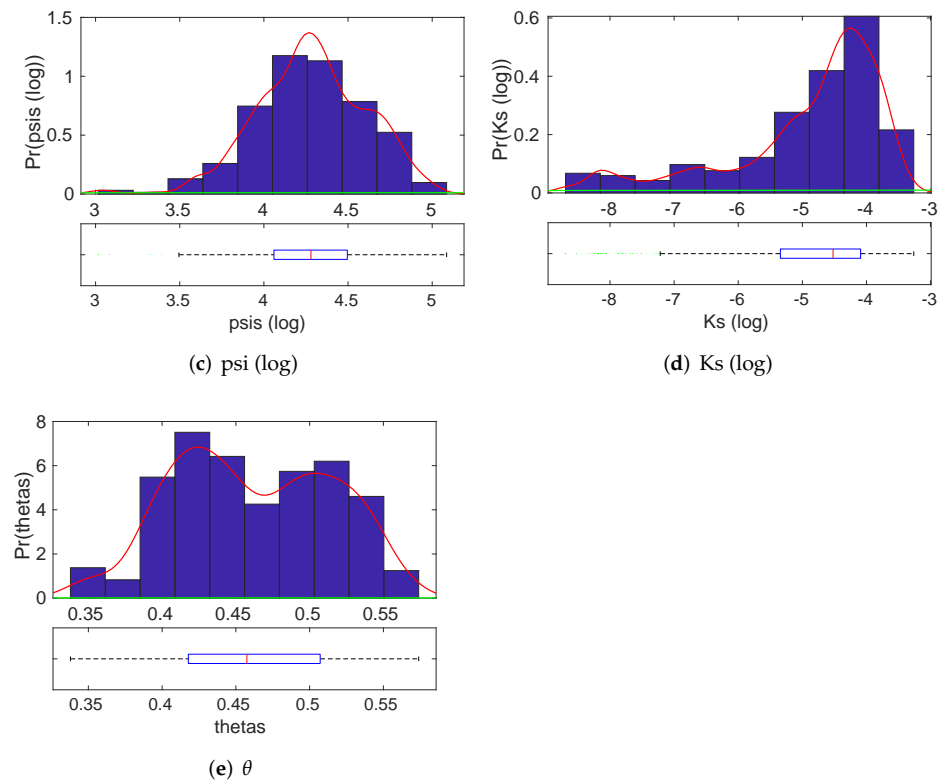


Figure 6. A posteriori distributions of the input parameters of CLM4 for a given output u^f . The red line indicates the posterior density, whereas the green line indicates the prior density of the parameter. (BMA evaluation).

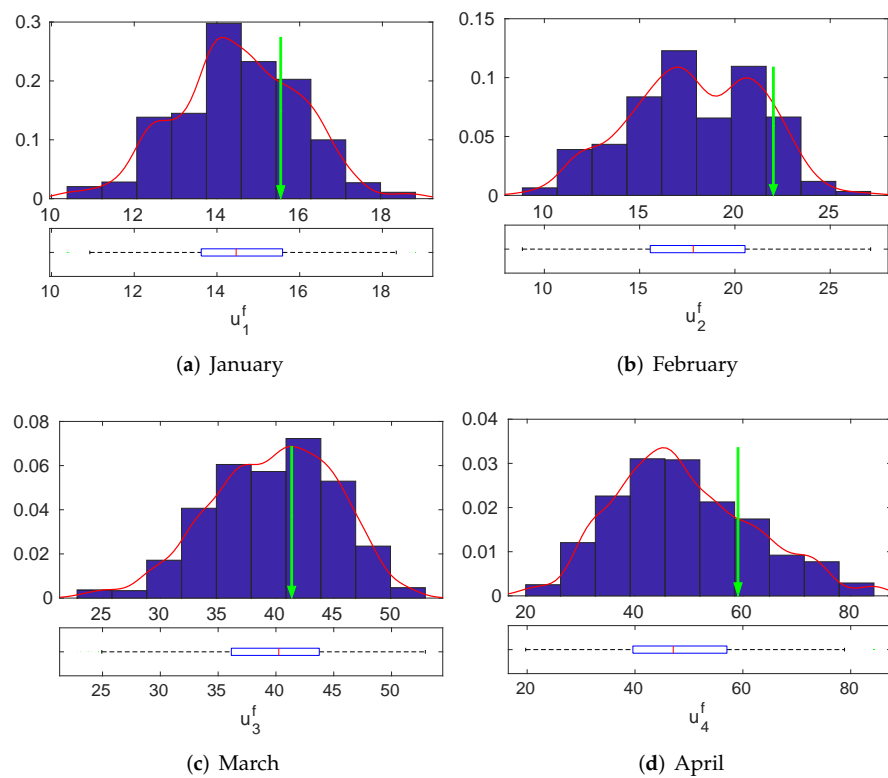


Figure 7. Cont.

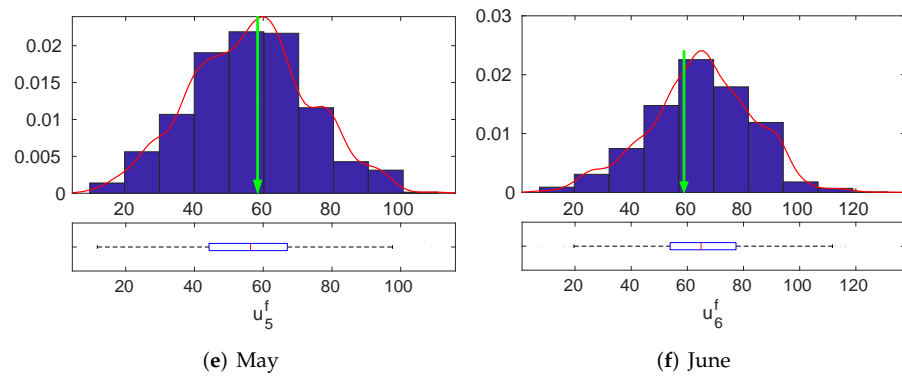


Figure 7. Distributions of the output LH associated to input parameters of CLM4 drawn by the a posteriori distributions: January–June.

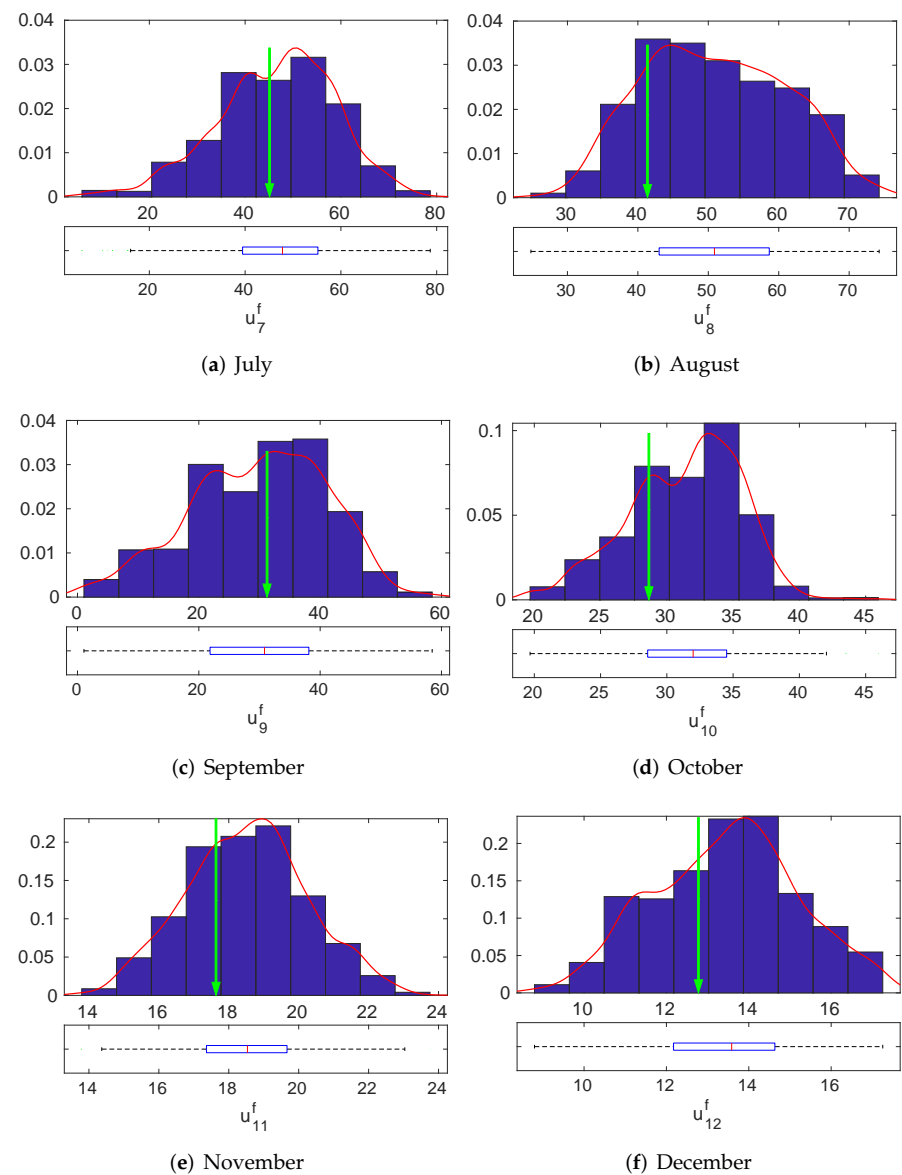


Figure 8. Distributions of the output LH associated to input parameters of CLM4 drawn by the a posteriori distributions: July–December.

5. Conclusions

We focused on evaluating the uncertainties associated with the hydrologic parameters in CLM4 in the Bayesian framework. We presented a Bayesian methodology for the uncertainty quantification (UQ) framework that couples a generalized polynomial chaos model with the Bayesian variable selection methods. We presented a fully Bayesian methodology that involves two steps: building a surrogate step, which is the construction of a surrogate model to express the input–output mapping; and the inverse modeling step, which is the evaluation of the posterior distribution of the input parameters for a given value of the output LH.

For the construction of the surrogate model, we propose a Bayesian procedure, based on variable selection methods, that uses gPC expansions and accounts for bases selection uncertainty. The advantage of this approach is that it can quantify the significance of the gPC terms, and, hence, the importance of the input parameters, in a probabilistic manner. The input posterior distributions were evaluated according to Bayesian inverse modeling. Our numerical experiments suggested that the method presented is suitable for performing the inverse modeling of hydrologic parameters in CLM4, and able to effectively describe the uncertainty related to these parameters.

Our future work involves the comparison of the proposed method against other methods based on neural networks, and generalized linear models, on which, we are currently working.

Author Contributions: G.K. and G.L. derived the models and performed the computations. Z.H. and M.H. worked on the interpretation of the results from the analysis of the application. All authors have read and agreed to the published version of the manuscript.

Funding: The National Science Foundation (DMS-1555072, DMS-1736364, DMS-2053746, and DMS-2134209), and Brookhaven National Laboratory Subcontract 382247, and U.S. Department of Energy (DOE) Office of Science Advanced Scientific Computing Research program DE-SC0021142.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Supplementary data related to this article can be found at <https://github.com/georgios-stats/Inverse-modeling-of-hydrologic-parameters-in-CLM4-via-generalized-polynomial-chaos-in-the-Bayesian-f>, accessed on 20 March 2022.

Acknowledgments: Guang Lin would like to thank the support from the National Science Foundation (DMS-1555072, DMS-1736364, DMS-2053746, and DMS-2134209), and Brookhaven National Laboratory Subcontract 382247, and U.S. Department of Energy (DOE) Office of Science Advanced Scientific Computing Research program DE-SC0021142.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pitman, A. The evolution of, and revolution in, land surface schemes designed for climate models. *Int. J. Climatol. J. R. Meteorol. Soc.* **2003**, *23*, 479–510. [CrossRef]
2. Bastidas, L.; Gupta, H.V.; Sorooshian, S.; Shuttleworth, W.J.; Yang, Z.L. Sensitivity analysis of a land surface scheme using multicriteria methods. *J. Geophys. Res. Atmos.* **1999**, *104*, 19481–19490. [CrossRef]
3. Henderson-Sellers, A.; Pitman, A.; Love, P.; Irannejad, P.; Chen, T. The project for intercomparison of land surface parameterization schemes (PILPS): Phases 2 and 3. *Bull. Am. Meteorol. Soc.* **1995**, *76*, 489–504. [CrossRef]
4. Henderson-Sellers, A.; Chen, T.; Nakken, M. *Predicting Global Change at the Land-Surface: The Project for Intercomparison of Land-Surface Parameterization Schemes (PILPS) (Phase 4)*; Technical Report; American Meteorological Society: Boston, MA, USA, 1996.
5. Rosero, E.; Yang, Z.L.; Wagener, T.; Gulden, L.E.; Yatheendradas, S.; Niu, G.Y. Quantifying parameter sensitivity, interaction, and transferability in hydrologically enhanced versions of the Noah land surface model over transition zones during the warm season. *J. Geophys. Res. Atmos.* **2010**, *115*, D03106. [CrossRef]
6. Hou, Z.; Huang, M.; Leung, L.R.; Lin, G.; Ricciuto, D.M. Sensitivity of surface flux simulations to hydrologic parameters based on an uncertainty quantification framework applied to the Community Land Model. *J. Geophys. Res. Atmos.* **2012**, *117*, D15. [CrossRef]

7. Huang, M.; Hou, Z.; Leung, L.R.; Ke, Y.; Liu, Y.; Fang, Z.; Sun, Y. Uncertainty analysis of runoff simulations and parameter identifiability in the Community Land Model: Evidence from MOPEX basins. *J. Hydrometeorol.* **2013**, *14*, 1754–1772. [[CrossRef](#)]
8. Collins, W.D.; Rasch, P.J.; Boville, B.A.; Hack, J.J.; McCaa, J.R.; Williamson, D.L.; Briegleb, B.P.; Bitz, C.M.; Lin, S.J.; Zhang, M. The formulation and atmospheric simulation of the Community Atmosphere Model version 3 (CAM3). *J. Clim.* **2006**, *19*, 2144–2161. [[CrossRef](#)]
9. Gent, P.R.; Yeager, S.G.; Neale, R.B.; Levis, S.; Bailey, D.A. Improvements in a half degree atmosphere/land version of the CCSM. *Clim. Dyn.* **2010**, *34*, 819–833. [[CrossRef](#)]
10. Lawrence, D.M.; Oleson, K.W.; Flanner, M.G.; Thornton, P.E.; Swenson, S.C.; Lawrence, P.J.; Zeng, X.; Yang, Z.L.; Levis, S.; Sakaguchi, K.; et al. Parameterization improvements and functional and structural advances in version 4 of the Community Land Model. *J. Adv. Model. Earth Syst.* **2011**, *3*. [[CrossRef](#)]
11. Ray, J.; Hou, Z.; Huang, M.; Sargsyan, K.; Swiler, L. Bayesian calibration of the Community Land Model using surrogates. *SIAM/ASA J. Uncertain. Quantif.* **2015**, *3*, 199–233. [[CrossRef](#)]
12. Huang, M.; Ray, J.; Hou, Z.; Ren, H.; Liu, Y.; Swiler, L. On the applicability of surrogate-based Markov chain Monte Carlo-Bayesian inversion to the Community Land Model: Case studies at flux tower sites. *J. Geophys. Res. Atmos.* **2016**, *121*, 7548–7563. [[CrossRef](#)]
13. Gong, W.; Duan, Q.; Li, J.; Wang, C.; Di, Z.; Dai, Y.; Ye, A.; Miao, C. Multi-objective parameter optimization of common land model using adaptive surrogate modeling. *Hydrol. Earth Syst. Sci.* **2015**, *19*, 2409–2425. [[CrossRef](#)]
14. Sargsyan, K.; Safta, C.; Najm, H.N.; Debusschere, B.J.; Ricciuto, D.; Thornton, P. Dimensionality reduction for complex models via Bayesian compressive sensing. *Int. J. Uncertain. Quantif.* **2014**, *4*, 63–93. [[CrossRef](#)]
15. Karagiannis, G.; Lin, G. Selection of polynomial chaos bases via Bayesian model uncertainty methods with applications to sparse approximation of PDEs with stochastic inputs. *J. Comput. Phys.* **2014**, *259*, 114–134. [[CrossRef](#)]
16. Fischer, M.L.; Billesbach, D.P.; Berry, J.A.; Riley, W.J.; Torn, M.S. Spatiotemporal variations in growing season exchanges of CO₂, H₂O, and sensible heat in agricultural fields of the Southern Great Plains. *Earth Interact.* **2007**, *11*, 1–21. [[CrossRef](#)]
17. Torn, M. *AmeriFlux US-ARM ARM Southern Great Plains Site-Lamont*; Technical Report, AmeriFlux; Lawrence Berkeley National Laboratory: Berkeley, CA, USA, 2016.
18. Niu, G.Y.; Yang, Z.L.; Dickinson, R.E.; Gulden, L.E. A simple TOPMODEL-based runoff parameterization (SIMTOP) for use in global climate models. *J. Geophys. Res. Atmos.* **2005**, *110*, D21. [[CrossRef](#)]
19. Niu, G.Y.; Yang, Z.L.; Dickinson, R.E.; Gulden, L.E.; Su, H. Development of a simple groundwater model for use in climate models and evaluation with Gravity Recovery and Climate Experiment data. *J. Geophys. Res. Atmos.* **2007**, *112*, D7. [[CrossRef](#)]
20. Oleson, K.; Niu, G.Y.; Yang, Z.L.; Lawrence, D.; Thornton, P.; Lawrence, P.; Stöckli, R.; Dickinson, R.; Bonan, G.; Levis, S.; et al. Improvements to the Community Land Model and their impact on the hydrological cycle. *J. Geophys. Res. Biogeosci.* **2008**, *113*, G1. [[CrossRef](#)]
21. Oleson, K.W.; Lawrence, D.M.; Gordon, B.; Flanner, M.G.; Kluzek, E.; Peter, J.; Levis, S.; Swenson, S.C.; Thornton, E.; Feddema, J.; et al. Technical Description of Version 4.0 of the Community Land Model (CLM). 2010. Available online: https://www.researchgate.net/publication/277114326_Technical_Description_of_version_40_of_the_Community_Land_Model_CLM (accessed on 20 March 2022).
22. Cafilisch, R.E. Monte carlo and quasi-monte carlo methods. *Acta Numer.* **1998**, *7*, 1–49. [[CrossRef](#)]
23. Wang, X.; Sloan, I.H. Low discrepancy sequences in high dimensions: How well are their projections distributed? *J. Comput. Appl. Math.* **2008**, *213*, 366–386. [[CrossRef](#)]
24. Marzouk, Y.; Xiu, D. A Stochastic Collocation Approach to Bayesian Inference in Inverse Problems. *Commun. Comput. Phys.* **2009**, *6*, 826–847. [[CrossRef](#)]
25. Berger, J.O. *Statistical Decision Theory and Bayesian Analysis*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.
26. Robert, C.P.; Casella, G. *Monte Carlo Statistical Methods*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2004.
27. Xiu, D. *Numerical Methods for Stochastic Computations: A Spectral Method Approach*; Princeton University Press: Princeton, NJ, USA, 2010.
28. Xiu, D.; Karniadakis, G.E. The Wiener–Askey Polynomial Chaos for Stochastic Differential Equations. *SIAM J. Sci. Comput.* **2002**, *24*, 619–644. [[CrossRef](#)]
29. Blatman, G.; Sudret, B. Adaptive sparse polynomial chaos expansion based on least angle regression. *J. Comput. Phys.* **2011**, *230*, 2345–2367. [[CrossRef](#)]
30. Doostan, A.; Owhadi, H. A non-adapted sparse approximation of PDEs with stochastic inputs. *J. Comput. Phys.* **2011**, *230*, 3015–3034. [[CrossRef](#)]
31. Hans, C. Model uncertainty and variable selection in Bayesian LASSO regression. *Stat. Comput.* **2010**, *20*, 221–229. [[CrossRef](#)]
32. Geman, S.; Geman, D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **1984**, 721–741. [[CrossRef](#)]
33. Hoeting, J.A.; Madigan, D.; Raftery, A.E.; Volinsky, C.T. Bayesian model averaging: A tutorial. *Stat. Sci.* **1999**, *14*, 382–401.
34. Barbieri, M.M.; Berger, J.O. Optimal Predictive Model Selection. *Ann. Stat.* **2004**, *32*, 870–897. [[CrossRef](#)]
35. Sun, Y.; Hou, Z.; Huang, M.; Tian, F.; Ruby Leung, L. Inverse modeling of hydrologic parameters using surface flux and runoff observations in the Community Land Model. *Hydrol. Earth Syst. Sci.* **2013**, *17*, 4995–5011. [[CrossRef](#)]