Using a surrogate-assisted Bayesian framework to calibrate the runoff-generation scheme in the Energy Exascale Earth System Model (E3SM) v1

Donghui Xu1, Gautam Bisht1, Khachik Sargsyan2, Chang Liao1, and L. Ruby Leung1

1Atmospheric Sciences and Global Change Division, Pacific Northwest National Laboratory, Richland, WA, USA
2Chemistry, Combustion, and Materials Science Center, Sandia National Laboratories, Livermore, CA, USA

Correspondence: Donghui Xu (donghui.xu@pnnl.gov)

Received: 29 November 2021 – Discussion started: 16 December 2021
Revised: 10 June 2022 – Accepted: 14 June 2022 – Published: 1 July 2022

Abstract. Runoff is a critical component of the terrestrial water cycle, and Earth system models (ESMs) are essential tools to study its spatiotemporal variability. Runoff schemes in ESMs typically include many parameters so that model calibration is necessary to improve the accuracy of simulated runoff. However, runoff calibration at a global scale is challenging because of the high computational cost and the lack of reliable observational datasets. In this study, we calibrated 11 runoff relevant parameters in the Energy Exascale Earth System Model (E3SM) Land Model (ELM) using a surrogate-assisted Bayesian framework. First, the polynomial chaos expansion machinery with Bayesian compressed sensing is used to construct computationally inexpensive surrogate models for ELM-simulated runoff at 0.5° × 0.5° for 1991–2010. The error metric between the ELM simulations and the benchmark data is selected to construct the surrogates, which facilitates efficient calibration and avoids the more conventional, but challenging, construction of high-dimensional surrogates for the ELM simulated runoff. Second, the Sobol’ index sensitivity analysis is performed using the surrogate models to identify the most sensitive parameters, and our results show that, in most regions, ELM-simulated runoff is strongly sensitive to 3 of the 11 uncertain parameters. Third, a Bayesian method is used to infer the optimal values of the most sensitive parameters using an observation-based global runoff dataset as the benchmark. Our results show that model performance is significantly improved with the inferred parameter values. Although the parametric uncertainty of simulated runoff is reduced after the parameter inference, it remains comparable to the multi-model ensemble uncertainty represented by the global hydrological models in ISMIP2a. Additionally, the annual global runoff trend during the simulation period is not well constrained by the inferred parameter values, suggesting the importance of including parametric uncertainty in future runoff projections.

1 Introduction

Runoff is an essential source of freshwater, and its variability has profound socioeconomic impacts (Hall et al., 2014; Vörösmarty et al., 2000). Flooding in wet regions during peak streamflow is among the most impactful natural hazards of all weather-related events in terms of fatalities and material costs (Doocy et al., 2013). However, higher streamflow replenishes reservoirs that help provide water for agriculture and hydropower generation and transports nutrients to the floodplain. Drought is a form of hydrological extreme that can also result in immense damages to the ecosystem and agriculture (Mishra and Singh, 2010). It is associated with abnormally low runoff, especially in arid and semi-arid regions. Therefore, understanding the spatial and temporal patterns of runoff is crucial for flood control, water management, crop yield, ecosystem services, etc. The runoff variability has been impacted by human-induced land use and climate change (Milly et al., 2008; Fischer and Knutti, 2016; Bosmans et al., 2017; Dai, 2013; Xu et al., 2021a), and the changes are projected to be more significant towards the end of this century (Xu et al., 2021a).
The spatial and temporal patterns of runoff and its response to climate change for water security assessments and water management are commonly studied using Earth system models (ESMs; Milly et al., 2002; Hirabayashi et al., 2013; Schewe et al., 2014). Current generation ESMs have large uncertainty in the simulation of runoff and its changes under future scenarios. However, statistical methods have been applied recently to reduce uncertainty in model predictions (Yang et al., 2017; Gosling and Arnell, 2011; Lehner et al., 2019; Xu et al., 2021a). Uncertainties in ESM simulations of runoff stem from uncertain model inputs, model structural uncertainty, and parametric uncertainty (Sun et al., 2013; Giuntoli et al., 2018). Input uncertainties consist of uncertainties in atmospheric forcing and land surface cover data that can be reduced by improving observation quality as more data become available. Model structural uncertainty is due to knowledge gaps or simplifications of the physical processes of the Earth system. Specifically, the typical coarse resolution (~100 km) of ESMs cannot capture a few of the key physical factors that control runoff-generation processes such as terrain and soil variations. Downscaling methods have been developed to reduce model bias when projecting the changes in hydrological variables from the coarse resolution ESM simulation to a fine resolution (Tebaldi et al., 2005; Knutti et al., 2010; Xu et al., 2019). Recent development in the Energy Exascale Earth System Model (E3SM) has introduced a sub-grid topography-based downscaling of precipitation (Tesfa et al., 2020) to understand the role of topography in hydrological processes.

Over the past few decades, the land component of ESMs has continuously been improved by developing new representations of physical processes, such as implementing variable soil thickness (Brunke et al., 2016), solving the variably saturated flow in groundwater dynamics (Bisht et al., 2018), including land–river interactions (Decharme et al., 2019; Xu et al., 2021b), representing lateral subsurface flow (Swenson et al., 2019), and increasing spatial resolution (Haarsma et al., 2016). While these advances improve our understanding of the Earth system, they may not lead to reduced uncertainties in future projections (Knutti and Sedláček, 2012; Lehner et al., 2020). This is because parametric uncertainty may increase as new processes are included in the model. The uncertainty in ESM simulated runoff must be reduced before reliable conclusions can be drawn regarding ESM projections of future changes in the runoff characteristics.

The parametric uncertainties in simulated runoff can be reduced by model calibration (Gupta et al., 1998). Previous studies have shown that it is possible to constrain the uncertainty of runoff by calibrating the relevant model parameters at a regional scale (Ray et al., 2015; Sun et al., 2013; Sheng et al., 2017; Xie et al., 2007; Troy et al., 2008; Hou et al., 2012; Huang et al., 2013). Hou et al. (2012) and Huang et al. (2013) identified the most sensitive hydrologic parameters of the Community Land Model (CLM) for simulating runoff and surface energy fluxes at a few selected watersheds and flux tower sites in the USA. They found that reducing the dimensionality of uncertain parameters using sensitivity analysis speeds up the calibration processes (Huang et al., 2013). Consequently, Sun et al. (2013) successfully applied a Bayesian inversion approach to estimate the optimal parameters to improve the performance of runoff generation in CLM. Troy et al. (2008) proposed an efficient framework to calibrate the Variable Infiltration Capacity (VIC) model for the contiguous USA by interpolating the calibrated parameters from small gauged basins. While previous studies performed comprehensive model calibration of runoff at regional scales, it remains challenging to calibrate the land components of ESM at global scales due to (1) the lack of runoff observations and (2) the high computational cost of running a large ensemble of global land model simulations. For (1), it is common to validate land models with streamflow (i.e., flow rate accumulated from runoff within a drainage area) observation (Li et al., 2015; Krysanova et al., 2020; Beck et al., 2017; Zhang et al., 2016), as runoff is not directly measured. However, routing the runoff to simulate streamflow at a coarse resolution introduces additional uncertainties due to the representation of the stream network (Wu et al., 2011; Liao et al., 2022) and river channel geometry (Andreadis et al., 2013). A recent observation-based global runoff dataset (GRUN; Ghiggi et al., 2019a) provides a good benchmark for calibrating runoff-generation-related parameters without the need of coupling the land model with a river-routing model. For (2), tens of thousands of simulations are typically needed for parameter calibration when the parameter dimension is high, but it is not computationally feasible to run a large ensemble of ESM simulations at the global scale.

The computational cost of model calibration can be significantly reduced by using an Uncertainty Quantification (UQ) framework that develops surrogate models of complex physical models. UQ frameworks include the following steps: (1) construction of a surrogate model that can mimic the behavior of a physical model, (2) identification of sensitive parameters to reduce the dimensionality of uncertain parameters, and (3) use of the parameter inference process to constrain the parametric uncertainty by comparing surrogate model prediction against observation. The surrogate modeling approach has received wide attention in hydrological applications (Razavi et al., 2012; Ivanov et al., 2021; Wang et al., 2014) to calibrate large-scale land models in terms of different hydrological processes (Gong et al., 2015; Lu et al., 2018; Müller et al., 2015; Huang et al., 2016; Ray et al., 2015; Sargsyan et al., 2014; Ricciuto et al., 2018). Multiple methods falling into the class of surrogate models include Gaussian process models, artificial neural networks, support vector machines, and polynomial chaos expansions (PCEs). In this study, we rely on PCEs as convenient machinery for uncertain input parameter representation and surrogate construction. The PCE surrogate captures the complex, nonlinear behavior of the physical model through a learned polynomial expansion. This method also provides convenient
global sensitivity analysis (Dwelle et al., 2019). Furthermore, we employ Bayesian compressive sensing (BCS) to arrive at sparse PCEs that include only polynomial terms relevant to the model, thus facilitating PCE surrogate construction in the presence of a large number of uncertain inputs and a relatively small number of model simulations (Sargsyan et al., 2014). Once the surrogate model is constructed, it replaces the expensive physical model in simulation-intensive studies such as global sensitivity analysis and parameter inference.

The objective of this work is to use the UQ framework to improve the performance of runoff generation at a monthly scale and quantify the associated parametric uncertainty in the E3SM Land Model version 1 (ELM v1; E3SM; Golaz et al., 2019). This study is organized in the following manner. We briefly describe the runoff-generation process in ELM v1, the UQ framework, and the data used in this study in Sects. 2, 3, and 4, respectively. In Sect. 5, we first present the validation of the surrogate models, sensitivity of simulated runoff to the uncertain parameters, dimensional reduction of uncertain parameters, and estimation of optimal parameters. Then we evaluate the performance of ELM-simulated runoff with the optimal parameters, the runoff sensitivity to precipitation, and the changes due to the use of optimal parameters on ELM-simulated water- and energy-related variables against various benchmarks using the International Land Model Benchmarking (ILAMB) package (Collier et al., 2018). Last, we present the simulated runoff uncertainty associated with parameters and their impacts on runoff trends at global scale. Section 6 discusses the limitations of this work, followed by the conclusions in Sect. 7.

2 E3SM Land Model

2.1 Runoff-generation scheme in ELM v1

The ELM v1 (hereafter, v1 is omitted) was developed based on the Community Land Model 4.5 (CLM4.5; Oleson et al., 2013) to understand the water availability and water cycle extremes (Leung et al., 2020). The new physical processes added in ELM to better represent the terrestrial water cycle include a variably saturated flow model (Bisht et al., 2013) to understand the water availability and water cycle extremes. The runoff generation in ELM is based on the simple TOPMODEL-based runoff parameterization (SimTop; Niu et al., 2005) in which the total runoff \(R_{total}\) consists of the following three components: surface runoff \(R_{over}\) (e.g., saturation excess runoff), surface water runoff \(R_{h2osfc}\) (e.g., surface water drainage from depressions/wetlands), and subsurface runoff \(R_{drai}\), as seen in the following:

\[
R_{total} = R_{over} + R_{h2osfc} + R_{drai}.
\]

A fraction of the flux of water reaching the soil surface \(q_{liq}\) generates surface runoff, and the fraction is determined by the saturation fraction \(f_{sat}\) of the grid cell as follows:

\[
R_{over} = f_{sat} q_{liq}
\]

\[
f_{sat} = f_{max} \exp(-0.5 f_{over} z_{water})
\]

where \(f_{max}\) represents the maximum saturation fraction for a given grid cell that is calculated with high-resolution compound topographic indices, \(f_{over}\) is a decay factor, and \(z_{water}\) is the water table depth.

ELM includes surface water storage to represent inland/wetland surface water dynamics (Ekici et al., 2019). When the surface water storage is fully filled, surface water runoff is generated as follows:

\[
R_{h2osfc} = k_{h2osfc} f_{connected} (W_{sfc} - W_{c}) \frac{1}{\Delta t},
\]

where \(k_{h2osfc}\) represents the linear storage coefficient, \(f_{connected}\) is the interconnected fraction of the inundated areas, \(W_{sfc}\) is the mass of surface water, \(W_{c}\) is the mass of surface water when the storage is full, and \(\Delta t\) is the model time step. \(W_{sfc}\) is formulated as follows:

\[
W_{sfc} = \frac{d}{2} \left[ 1 + \text{erf} \left( \frac{d}{\sigma_{micro} \sqrt{2}} \right) \right] + \frac{\sigma_{micro} e^{-d^2/2 \sigma_{micro}^2}}{\sqrt{2 \pi}}
\]

where erf represents the error function, \(d\) is the height of the surface water relative to the cell averaged elevation, and \(\sigma_{micro}\) is the standard deviation of the microtopographic distribution that characterizes subgrid elevation variation. Given the surface water height from the previous equation, the surface water fraction \(f_{h2osfc}\) of a cell is estimated with the following:

\[
f_{h2osfc} = \frac{1}{2} \left[ 1 + \text{erf} \left( \frac{d}{\sigma_{micro} \sqrt{2}} \right) \right].
\]

The inundation areas are assumed to be randomly distributed within the grid cell, and the interconnected fraction of the inundated areas can be estimated based on percolation theory, as follows:

\[
f_{connected} = \begin{cases} (f_{h2osfc} - f_{c})^\mu \text{ if } f_{h2osfc} > f_{c} \\ 0 \text{ if } f_{h2osfc} \leq f_{c} \end{cases}
\]

where \(f_{c}\) is the threshold below which the inundated areas are not connected, and \(\mu\) is a scaling exponent. The default parameter values in ELM of \(f_{c}\) and \(\mu\) are 0.4 and 0.14 for all the global cells, respectively.

The subsurface runoff is parameterized as an exponential function of water table depth and includes an ice impedance factor \(Q_{eq}\) to account for the reduction in the bottom drainage when ice is present in the soil (Swenson et al.,...
2012), as follows:

\[ R_{\text{drai}} = \Theta_{\text{ice}} q_{\text{drai, max}} \exp (-f_{\text{drai}} z_y) \]  
\[ \Theta_{\text{ice}} = 10^{-\Omega_{\text{max}}} \]  

where \( q_{\text{drai, max}} \) is the maximum drainage rate, \( f_{\text{drai}} \) is the
decay factor, \( \frac{\Theta_{\text{ice}}}{\Omega_{\text{max}}} \) represents the ice-filled fraction of the pore
space for the soil under the water table, and \( \Omega \) is an adjustable
parameter.

We follow the work of Huang et al. (2013) in selecting un-
certain parameters and their corresponding ranges (Table 1). There are three additional parameters included in this study
for surface water storage drainage and impacts of ice to sub-
surface runoff and soil water dynamics, which represent new
features in ELM compared to CLM4.0 used in Huang et
al. (2013). All the parameter prior distributions are assumed
to be a uniform distribution.

2.2 Model configuration

We ran ELM globally at a spatial resolution of 0.5° × 0.5°,
driven by the Global Soil Wetness Project forcing dataset
(GSWP3v1) from 1991 to 2010, featuring 3 h, 0.5° × 0.5°
global atmosphere forcing. GSWP3v1 has been dynamically
downscaled and bias corrected based on the reanalysis data of
Compo et al. (2011). The default configuration of ELM was
used with a 30 min time step. With the default configuration,
the hydrologic representations of ELM are the same as those
in CLM4.5, as new model features such as the variably satu-
rated flow model and subgrid topography are not included.
Except the uncertain parameters listed in Table 1, the default
values of all other ELM parameters were used in this study.

3 Uncertainty Quantification framework

A detailed derivation of the PCE-based Uncertainty Quan-
tification framework and BCS method used in this work
is presented in Sargsyan et al. (2014) and Debusschere et
al. (2016). In this study, we used the Uncertainty Quantifi-
cation Toolkit (UQTk; Debusschere et al., 2004, 2016) that
includes implementations of PCE construction with BCS and
subsequent global sensitivity analysis. Only a brief descrip-
tion of the construction of the PCE-based surrogate for the
ELM simulations is summarized below.

3.1 Polynomial chaos expansion

Let \( M \) denote a physical model (e.g., ELM) with uncertain
parameters \( \mathbf{X} \), where \( \mathbf{X} = [X_1, X_2, \ldots, X_D] \), and \( D \) repre-
sents the total number of uncertain parameters. In this study,
the uncertain parameters \( \mathbf{X} \) are listed in Table 1 and \( D = 11 \).
A scalar quantity of interest (QoI), \( \hat{y} \) (e.g., runoff at a spe-
cified time from a specified location), obtained using a sample
of random parameters, \( \mathbf{x} \), can be expressed as a polynomial
expansion, as follows:

\[ \hat{y} = M (\mathbf{x}) = \sum_{\alpha} c_{\alpha} \Psi_{\alpha} (\mathbf{x}) \]  

where \( \Psi_{\alpha} \) is a polynomial, and \( c_{\alpha} \) is the corresponding coef-
ficient. In practice, \( \mathbf{x} \) is scaled to \([-1, 1] \) from the original un-
certainty input range. The polynomial expansion in Eq. (10)
is written, with respect to multivariate orthogonal polynomi-
als, as follows:

\[ \Psi_{\alpha} (\mathbf{x}) = \prod_{i=1}^{D} \Psi_{\alpha_i} (x_i) \]  

where \( \Psi_{\alpha_i} (x_i) \) is a univariate polynomial, whose form is asso-
ciated with the prior distribution of uncertain input vari-
ables \( \mathbf{X} \) (e.g., Legendre polynomials are used when the input
variable follows a uniform distribution), and \( \alpha_i \) is a member
of the multi-index \( \alpha = [\alpha_1, \alpha_2, \ldots, \alpha_D] \), which repre-
sents the degrees of the univariate polynomial terms. Read-
ers should refer to Dwelle et al. (2019) for details about
the selection of polynomial terms and an illustration of how the
multi-index is used to construct a PCE-based surrogate. In
practice, Eq. (11) is approximated with a truncated PCE
by only selecting terms with a total degree of polynomi-
als smaller than a certain value \( p \) (Xiu and Karniadakis,
2002; Lin and Karniadakis, 2009). This leads to a finite set
\( A_p = (\alpha : \sum_{i=1}^{D} |\alpha_i| \leq p) \) for the multi-index \( \alpha \) to take the follow-

\[ M (\mathbf{x}) \approx M_{\text{PC}} (\mathbf{x}) = \sum_{\alpha \in A_p} c_{\alpha} \Psi_{\alpha} (\mathbf{x}) = \sum_{j=0}^{P} c_{j} \Psi_{j} (\mathbf{x}) \]  

where \( j \) represents the counter-index of any possible multi-
index \( \alpha \) in \( A_p \) in a predefined order (see details in Ap-
pendix B of Dwelle et al., 2019). The coefficients \( c_j \) for
the \( P + 1 \) polynomial bases are computed using training sim-
ulations of \( M (\mathbf{x}) \) (e.g., ELM) to construct the truncated PCE
approximation in Eq. (12). The number of the polynomial
basis is determined by both the input dimension \( D \) and the
total degree for truncation \( p \) (Xiu and Karniadakis, 2002) is
as follows:

\[ P + 1 = \frac{(D + p)!}{D!p!} \]  

The value \( P \) increases rapidly as the number of uncertain
input variables increases. For example, 11 uncertain para-
ters (e.g., \( D = 11 \)) with a truncated PCE order of \( p = 4 \) lead
to 1365 coefficients to solve in Eq. (12). It is computa-
tionally prohibitive to run 1365 global ELM simulations, so we
adopted the BCS method of Sargsyan et al. (2014) that re-
quires a much smaller number of ELM simulations to con-
struct a PCE-based surrogate. The BCS method computes
only a sparse set of \( c_j \) to construct the surrogate of a form
given by Eq. (12) because not all \( \Psi_j (\mathbf{x}) \) are relevant for the
given QoI (Sargsyan et al., 2014).
Table 1. Uncertain parameters’ information. Note: DEM is for digital elevation model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Definition</th>
<th>Default value</th>
<th>Priors</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_{\text{max}}$</td>
<td>Maximum saturated fraction for a grid cell (-)</td>
<td>Derived from a high-resolution DEM.</td>
<td>$U(0.01, 0.907)$</td>
</tr>
<tr>
<td>$f_{\text{over}}$</td>
<td>Decay factor for surface runoff $(\text{m}^{-1})$</td>
<td>0.5</td>
<td>$U(0.1, 5)$</td>
</tr>
<tr>
<td>$f_{\text{dr}}$</td>
<td>Decay factor for subsurface runoff $(\text{m}^{-1})$</td>
<td>2.5</td>
<td>$U(0.1, 5)$</td>
</tr>
<tr>
<td>$q_{\text{dr}, \text{max}}$</td>
<td>Maximum subsurface drainage rate $(\text{kg m}^{-2} \text{s}^{-1})$</td>
<td>$5.5 \times 10^{-3}$</td>
<td>$U(1 \times 10^{-6}, 1 \times 10^{-1})$</td>
</tr>
<tr>
<td>$b$</td>
<td>Clapp–Hornberger exponent (-)</td>
<td>Determined by plugging the soil type into the equations of means from Table 5 of Cosby et al. (1984).</td>
<td>Uniform distributions with ±50% of the means as the lower and upper bounds.</td>
</tr>
<tr>
<td>$\psi_s$</td>
<td>Saturated soil matrix potential (mm)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$K_s$</td>
<td>Hydraulic conductivity $(\text{mm s}^{-1})$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\theta_s$</td>
<td>Porosity (-)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$f_c$</td>
<td>Surface water fraction threshold for outflow (-)</td>
<td>0.4</td>
<td>$U(0.1, 0.7)$</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Scaling exponent for estimating connected surface water fraction (-)</td>
<td>0.14</td>
<td>$U(0.04, 0.24)$</td>
</tr>
<tr>
<td>$\Omega$</td>
<td>Adjustable parameter for ice impedance factor (-)</td>
<td>6</td>
<td>$U(0.6, 60)$</td>
</tr>
</tbody>
</table>

### 3.2 Global sensitivity analysis

In this study, we performed variance-based, global sensitivity analysis using Sobol’ indices (Sobol’, 2001). For a PCE-based surrogate model, the main Sobol’ index, $S_i$, for the uncertain parameter $X_i$ can be estimated as follows:

$$S_i = \frac{\sum_{j \in \Pi_i} c_j^2 ||\Psi_j||^2}{\sum_{j=0}^{\Pi} c_j^2 ||\Psi_j||^2},$$  \hspace{1cm} (14)

where $\Pi_i$ denotes all the indices of polynomial basis terms in Eq. (12) that only involve parameter $X_i$, and $||\Psi_j||$ is the norm of the polynomial $\Psi_j(x)$. The main Sobol’ index $S_i$ can be interpreted as the fraction of variance in the output that is associated with the uncertainty model parameter $X_i$ only when other parameters are fixed at constant values. Similarly, one can estimate the Sobol’ index for any pair of parameters $X_i$ and $X_j$ to represent parameter interaction sensitivity with the coefficients $c_j$ (Sargsyan et al., 2014).

### 3.3 Parameter inference

Parameter inference is used to determine a set of model parameters that reduces the error between observation and model prediction. The model inverse problem can be solved with the Bayes’ theorem, as follows:

$$p(X|y) = \frac{L(y|x)p(X)}{p(y)},$$  \hspace{1cm} (15)

where $p(X|y)$ is the posterior distribution of parameter $X$ given observation $y$, $L(y|x)$ is the likelihood function, $p(X)$ represents the prior distribution of $X$, and $p(y)$ is merely a normalizing constant for the purposes of parameter calibration. The discrepancy between the model and observations, $\epsilon = y - M(X)$, should be included in the likelihood function. It is common to assume the error term (e.g., $\epsilon$) follows a Gaussian distribution with a vanishing mean:

$$\epsilon_i \sim N(0, \sigma^2), \hspace{0.5cm} i = 1, 2, \ldots, N,$$  \hspace{1cm} (16)

where $N$ is the number of observations used to infer the parameters (e.g., time series of monthly runoff), and the
standard deviation, $\sigma$, can be inferred from the data (see Sect. 3.4). Then, the likelihood function can be written as follows:

$$L(y|X) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i - M_i(X))^2}{2\sigma^2}\right].$$

(17)

The logarithm of Eq. (17) leads to the least squares objective function that is used for deterministic parameter estimation in practice (Sargsyan et al., 2015), as follows:

$$\log L(y|X) = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - M_i(X))^2 - \frac{N}{2} \log(2\pi\sigma^2).$$

(18)

The posterior distribution in Eq. (18) is difficult to compute in practice; hence, we estimate it through samples obtained by the Markov Chain Monte Carlo (MCMC) method. Specifically, 1000 iterations are used as the burn-in period in this study, and the sampling of the posterior distribution is saved every 10 iterations. We run MCMC for 10,000 steps, resulting in 1800 samples to construct the posterior distribution. We have employed adaptive MCMC method of Haario et al. (2001), in which the parameter space is searched according to proposal steps with a covariance that is updated spontaneously.

### 3.4 Quantity of interest

In this study, the physical model $M$ and the QoI $\hat{y}$ correspond to ELM and runoff, respectively. The development of a surrogate model for the simulated runoff for each grid cell for each month of a 20-year simulation would require 240 (= 12 months $\times$ 20 years) PCE-based surrogates. Although developing a PCE-based surrogate is not expensive, it is computationally expensive to train 240 PCEs for each of the 70,302 grid cells in the global domain. The parameter inference process for 240 PCEs for each grid cell will further increase the computational cost. We reduce the number of QoIs by training the surrogate model for the root mean square error (RMSE) between the simulated runoff and observations instead of training the surrogate model to predict monthly runoff. The RMSE is given as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (R^\text{sim}_i - R^\text{obs}_i)^2},$$

(19)

where $R^\text{sim}_i$ and $R^\text{obs}_i$ represent the grid-level simulated total runoff and observed total runoff, respectively, for $i$th month in the simulated period, and $N$ represents the number of simulation months. Consequently, only one surrogate model is needed for each grid cell to quantify the performance of ELM in capturing the monthly runoff variation for a given uncertain parameter set. The selection of RMSE as QoI in constructing surrogate models significantly reduces the computational burden of the surrogates’ construction and parameter inference. We performed ELM simulations using 200 parameter sets that were randomly sampled from the range specified in Table 1. A set of 175 ELM simulations were used for training the surrogate models, and the other 25 simulations were used for validating their performances. The performance of the PCE-based surrogate model can be affected by the truncated order (Dwelle et al., 2019). For each grid cell, we train the surrogate with $p = 1, 2, \ldots, 7$ separately and picked the order that minimizes the relative norm 2 error (RE) of validation simulations as follows:

$$\text{RE} = \frac{||\text{RMSE}^\text{PC}_\text{val} - \text{RMSE}^\text{M}_\text{val}||_2}{||\text{RMSE}^\text{M}_\text{val}||_2},$$

(20)

where $\text{RMSE}^\text{PC}_\text{val}$ and $\text{RMSE}^\text{M}_\text{val}$ represent the PCE-simulated and ELM-simulated vector of RMSE of the 25 validation simulations, respectively. Then, the trained surrogate models, $\text{RMSE}^\text{PC}$, can be plugged into the likelihood function of Eq. (18) seamlessly, as follows:

$$\log L(y|X) = -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (0 - \text{RMSE}^\text{PC}_i)^2 - \frac{N}{2} \log(2\pi\sigma^2).$$

(21)

3.5 Calibration procedure

In summary, the following procedures were implemented to determine the optimal parameter values and their joint probability distribution:

1. Run ELM with 200 parameter sets randomly sampled with the range specified in Table 1.
2. Construct PCE-based surrogate models to mimic the RMSE between the ELM and GRUN runoff dataset with 175 simulations and validate the performance of the surrogate models with the other 25 simulations.
3. Implement sensitivity analysis with the surrogate models to reduce parameter dimensionality for calibration by ignoring the parameters with negligible Sobol’ index (e.g., less than 0.05).
4. Estimate the Bayesian posterior of the most sensitive parameters for each grid through MCMC process with the runoff dataset of Ghiggi et al. (2019a).

It has been shown that small surrogate error can result in significant deviation of the inferred parameter (Laloy and Jacques, 2019). To further search the optimal parameters and construct the runoff posterior uncertainty, we ran ELM simulations with additional 100 samples from the posteriors of the three most sensitive parameters for all global grid cells, and default values were used for less sensitive parameters.

The parameters with the minimum RMSE between simulations and reference runoff data from the 100 ELM simulations were determined as the optimal parameter value for each grid cell.

4 Data

4.1 Observation-based runoff data

The 0.5° × 0.5° observation-based global runoff dataset (GRUN) dataset of Ghiggi et al. (2019a) was used in this study as the observation within the calibration framework for parameter inference. The GRUN dataset was generated from a trained random forest (RF) model (Breiman, 2001) that used precipitation and near-surface temperature to predict monthly runoff. The training runoff data were derived from the Global Streamflow Indices and Metadata Archive (GSIM; Gudmundsson et al., 2018; Do et al., 2018), and only the gauges with contributing area comparable to cell area of 0.5° × 0.5° were used. GSWP3 atmospheric forcing was used for training and reconstruction of the monthly global runoff.

4.2 Model benchmarks

The ILAMB package (Collier et al., 2018) was used to evaluate the simulated water and energy cycles from the calibrated ELM against various benchmarks. Specifically, a gridded energy flux data (FLUXCOM; Jung et al., 2019) that was generated by machine learning with flux tower measurements was used to evaluate latent and sensible heat fluxes, the Global Land Evaporation Amsterdam Model version 3 (GLEAMv3; Martens et al., 2017) product was used to evaluate global evapotranspiration (ET), and Gravity Recovery And Climate Experiment (GRACE; Kim et al., 2009) data were used to evaluate a terrestrial water storage anomaly (TWSA). Details about ILAMB can be found at https://www.ilamb.org (last access: 23 June 2021).

The Inter-Sectoral Impact Model Intercomparison Project (ISIMIP) archived simulations from multiple global hydrological models and land surface model forced by the same atmosphere forcings (Warszawski et al., 2014). We used 13 available models from the second-phase water sector (ISIMIP2a; Gosling et al., 2019) to provide a benchmark for the uncertainty of annual runoff magnitude and trend. Only the models in ISIMIP2a that were driven by the GSWP3 forcing without accounting for human activity impacts were selected here to be consistent with ELM’s configuration.

4.3 Evaluation metrics

There were two metrics used to evaluate ELM’s performance of simulating runoff at a monthly scale with calibrated parameters, including the Nash–Sutcliffe efficiency (NSE; Nash and Sutcliffe, 1970) and the Kling–Gupta efficiency (KGE; Gupta et al., 2009) which are computed as follows:

\[
\text{NSE} = 1 - \frac{\sum_{i=1}^{N} (R_i^{\text{sim}} - R_i^{\text{obs}})^2}{\sum_{i=1}^{N} (R_i^{\text{obs}} - \mu^{\text{obs}})^2},
\]

\[
\text{KGE} = 1 - \sqrt{(\rho - 1)^2 + \frac{\sigma^{\text{sim}}}{\sigma^{\text{obs}}} - 1 + \left(\frac{\mu^{\text{sim}}}{\mu^{\text{obs}}} - 1\right)^2},
\]

where \(R_i^{\text{sim}}\) and \(R_i^{\text{obs}}\) represent cell-level simulated total runoff and observed total runoff, respectively, for the \(i\)th month, \(\mu^{\text{obs}}\) is the corresponding averaged observation, \(\rho\) is the correlation coefficient between simulation and observation, \(\sigma^{\text{sim}}\) is the standard deviation in simulations, \(\sigma^{\text{obs}}\) is the standard deviation in observations, and \(\mu^{\text{sim}}\) is the simulation mean. Both NSE and KGE vary from \(-\infty\) to 1, and a perfect model performance is indicated by NSE = 1 and KGE = 1. NSE < 0 and KGE < −0.41 mean the simulations are worse estimates than the mean of observations, indicating a bad model performance (Knoben et al., 2019).

The sensitivity of runoff to precipitation is a critical aspect for runoff simulation evaluation, considering changes in precipitation will continue in the future (Trenberth, 2011). Therefore, we evaluated the sensitivity of runoff to precipitation anomalies with the calibrated parameters. The sensitivity was quantified by the slope of linear regression (\(\beta\)) between runoff anomalies (\(\Delta R\)) and precipitation anomalies (\(\Delta P\)) as follows:

\[
\Delta R = \beta \Delta P + \epsilon.
\]

The interception \(\epsilon \approx 0\), implies the mean runoff is related to mean precipitation.

We also evaluated the impacts of parameters on the runoff trend. Specifically, the magnitude of runoff trend was calculated with Sen’s slope (Sen, 1968), which is nonparametric and not sensitive to the outliers. Then, Mann–Kendall test was used to determine if the trend is significant or not at confidence level \(\alpha = 0.05\).

5 Results

5.1 Refinement of \(f_{\text{drai}}\) for arid regions

The proposed prior for \(f_{\text{drai}}\) is not suitable for all the climate regions, such as the simulations with the full range of
For example, the simulated runoff from an example grid cell with $f_{\text{drai}} < 0.4$ shows higher magnitudes and lower variabilities compared to simulations with $f_{\text{drai}} \geq 0.4$ (Fig. 1a). Lower $f_{\text{drai}}$ can lead to unrealistically high subsurface runoff according to the exponential function of baseflow drainage (Eq. 8) for the arid regions, where the precipitation is not enough to maintain the water table at a reasonable level. Such simulations with $f_{\text{drai}} < 0.4$ result in high nonlinearity in the simulated runoff, and hence, the PCE-based surrogate model cannot capture the model behaviors (Fig. 1b). The performance of surrogate models is improved by constraining the lower bound of $f_{\text{drai}}$ to 0.4 (Fig. 1c). Therefore, $f_{\text{drai}}$ is refined as [0.4, 5] for areas that are identified as arid climate in the Köppen climate classification (Fig. S1 in the Supplement), and [0.1, 5] is used in other regions.

5.2 Validation of surrogate models

The PCE-based surrogate models can mimic the variations in RMSE between ELM-simulated runoff and the GRUN runoff with the truncated order determined in Fig. S2. Specifically, the surrogate models exhibit good performance for the validation simulations with RE < 0.1 for 70% of the global domain (Fig. 2a). The global averaged RE of surrogate models for the validation simulations is around 0.1, with the largest error over the arid regions (Fig. 2b). While 41% of the arid region shows an acceptable performance in the surrogate models when narrowing the range of $f_{\text{drai}}$ with RE less than 0.15, the RE of other arid areas remain high (Fig. 2a). Additional simulations were performed to investigate if the lower performance of surrogate models for arid regions is due to insufficient number of training simulations. We randomly selected 20 grid cells from the arid region and ran 2000 ELM simulations with random samples from the parameter priors, as summarized in Table 1. The RE of surrogate models for the 20 grid cells remained large (e.g., RE > 0.2), even as the number of training simulations were increased (Fig. S3). Thus, the lower performance of surrogate models over the arid regions is not dependent on the number of training simulations.

Most surrogate models with large RE are in extremely dry arid regions; for example, RE > 0.2 are mainly from grid cells with annual runoff < 0.05 mm d$^{-1}$ (Fig. 3). The RE of surrogate models tends to decrease for areas with relatively higher annual runoff that are still from arid regions (annual runoff < 0.5 mm d$^{-1}$ in Fig. 3). However, the runoff uncer-
Figure 2. Relative norm 2 error of the surrogate models for the validation simulations. Panel (a) shows the spatial distribution of the errors, and panel (b) shows the average errors for the grid cells in each climate defined by Köppen climate classification.

Figure 3. Plot of relative norm 2 error (RE) of surrogate models for the validation simulations vs. averaged annual runoff magnitude with all the grid cells from the arid region.

5.3 Global sensitivity analysis

The most significant ELM parameters identified for runoff generation are $f_{\text{over}}$, $f_{\text{drain}}$, $\psi_s$, $f_c$, and $\Omega$, based on the spatial distribution of the main Sobol’ indices (Fig. 4), while the other six parameters have negligible contributions to the runoff variations (Fig. S4). In equatorial regions, $f_{\text{drain}}$ and $f_{\text{over}}$ are equally sensitive and account for 39% and 36% of the average runoff variations, respectively, as indicated by the size of circles in Fig. 5a, while $\psi_s$ is the secondary sensitive parameter. For the arid regions, $f_{\text{drain}}$ is the most sensitive parameter, and $f_{\text{over}}$, $f_c$, $K_s$, and $\psi_s$ are secondary sensitive parameters with a similar value for the main Sobol’ indices (Fig. 5b). Although other parameters show negligible main Sobol’ indices for arid regions, they have shown sensitivities when interacting with each other, as denoted by the thickness of the lines between each pair of parameters in Fig. 5b. The complex joint sensitivity results in high nonlinearity in the runoff variations, representing a possible reason for the poor performance of PCE for arid regions. The most significant uncertain parameters for the warm temperate region are the same as those for the equatorial region (Fig. 5c). Snow and polar climates have similar sensitivity pattern, with $f_c$ and $\psi_s$ being the two most important uncertain parameters (Fig. 5d and e). In colder regions, the contribution of surface water storage drainage, which is controlled by $f_c$, is large to the total runoff because of prominent surface water areas.

tainty in extremely dry areas will have negligible impact on the global water cycle. Surrogate model with RE > 0.15 is considered as not sufficiently accurate, and such grid cells are excluded in the sensitivity analysis presented next.
(Pekel et al., 2016). The hydraulic conductivity and groundwater drainage when ice is present in the soil is controlled by $\Omega$, which has a significant impact on runoff-generation process when the soil is partially or fully frozen. The surface water storage and ice impedance factor, which were not included in the version of the model used in previous study (Hou et al., 2012; Huang et al., 2013), are found to be the most sensitive parameters in cold regions. Besides the arid region, other regions show smaller sensitivities to parameter interactions.

### 5.4 Parameter dimensionality reduction

The ELM simulated runoff is significantly sensitive to three or fewer parameters with Sobol’ index $>0.05$ for 81.3% of the total grid cells (Fig. S5). Therefore, we sampled only the three most sensitive parameters in each grid cell in the Markov chain Monte Carlo (MCMC) process to perform parameter inference, as mentioned in Sect. 3.3. The posteriors of the three calibrated parameters ($f_{\text{drai}}$, $f_{\text{over}}$, $\psi_s$) at an example grid cell ($56.75^\circ$ W, $11.25^\circ$ S) are much more constrained than the priors after the MCMC simulation with the surrogate model (Fig. 6a, b, and c). The third parameter, $\psi_s$, has a relatively wider posterior than the first two parameters because its sensitivity is much smaller (e.g., Sobol’ index $=0.08$). The Gelman–Rubin R statistic of Gelman and Rubin (1992) computed with five MCMC chains (after the burn-in period) is 1.002, 1.004, and 1.003 for $f_{\text{drai}}$, $f_{\text{over}}$, and $\psi_s$, respectively, suggesting that our MCMC simulation has converged (see the convergence curve in Fig. S6). ELM simulations with a large number of samples from parameter priors are needed to identify the optimal parameter that minimizes RMSE; for example, 10,000 surrogate simulations are used to find the parameters that yield $\text{RMSE} = 1$ (Fig. 6d). In contrast, due to the reduced parameter dimensionality and narrowed range, much fewer samples (e.g., 100) are needed to find the better parameter values (e.g., corresponding to $\text{RMSE} < 1$) when they are sampled from the parameter posteriors (Fig. 6d). The spatial distribution of the parameter values at 5% and 95% of the posteriors is shown in Figs. S7 and S8, respectively.

### 5.5 Optimal parameter values

The procedure described in Sect. 3.5 is used to find the optimal parameter values for the three most sensitive parameters for each grid cell. For the grid cells with RE $>0.15$ for surrogate models, the optimal parameter value is determined from the training and validation simulations (e.g., 200 simulations with random parameter values from priors) that yield minimum RMSE. The optimal parameter values show clear regional patterns (Fig. 7). Specifically, the optimal $f_{\text{over}}$ tends to be lower than the default value for the equatorial and partial snow areas (Fig. 7a). The optimal $f_{\text{over}}$ is found to be higher than the default value for the arid areas, while it is around the default value on average for the warm temperate areas (Fig. 7a). For the same water table depth, lower $f_{\text{over}}$ leads to higher saturation fraction (Eq. 3), that in turn leads to larger surface runoff (Eq. 2). The calibrated $f_{\text{drai}}$ is lower than the default value for both equatorial and arid regions (Fig. 7b). The optimal $f_{\text{drai}}$ for warm temperate areas show different patterns, with higher values over eastern USA and Europe but lower values over southeastern China. The generation of subsurface runoff depends on $f_{\text{drai}}$ (Eq. 8), with lower $f_{\text{drai}}$ leading to larger subsurface runoff. $\psi_s$ affects the runoff generation through its impact on soil water movement, such as the soil water flux being larger at saturation with higher $\psi_s$. As shown in Fig. 7c, higher $\psi_s$ are needed to minimize the RMSE for all regions that show sensitivity to this parameter, except for some grid cells from polar area. Over the high latitudes of Northern Hemisphere, higher $f_c$ and lower $\Omega$ are found in the optimal parameters (Fig. 7d, e). The surface water storage can store more water at higher $f_c$ by reducing surface water runoff (Eqs. 4, 7), thus leading to a lower and delay peak runoff than the default values. Furthermore, the lower $\Omega$ values have fewer impacts of ice on hydraulic conductivity (Eq. 7.89 in Olson et al., 2016) and drainage (Eq. 9), leading to higher runoff for the winter seasons.

### 5.6 Evaluation of ELM with the optimal parameters

The ELM-simulated runoff with the optimal parameter values shows improved skills of capturing the spatiotemporal variation in monthly runoff at a global scale with higher NSE and KGE compared to the simulation with default parameter values (Fig. 8). Specifically, the median of NSE and KGE from all global grid cells increases from $-0.88$ and $-0.05$ to 0.06 and 0.31, respectively. Over the western USA coast, southeast and midwest of USA, western Europe, and equatorial areas, the performance of the calibrated ELM is better with NSE $>0.5$ and KGE $>0.7$. While the performance of other areas (e.g., western USA, Sahara and Arabian desert, central and eastern Asia, and partial high latitude regions) is improved compared to simulations with the default parameter values, the NSE and KGE still have negative values. The higher model errors in those regions cannot be resolved by calibration as (1) the simulation resolution is too coarse to resolve the topographic impacts (Chegwidden et al., 2020), (2) the snow-melting processes are not calibrated in this study, and the onset of snowmelt in ELM is poorly represented (Toure et al., 2018), and (3) the hydrology of arid areas is not well understood (Pilgrim et al., 1988). Except for the calibration period, ELM with the optimal parameters also shows an improved performance in runoff simulation for another period (2011–2013; Fig. S9).

Compared to the reference runoff (Fig. 9a), the ELM simulation with default parameter values tends to overestimate the sensitivity of runoff to precipitation ($\beta$ in Eq. 24) for the equatorial and arid regions but underestimates $\beta$ in the warm temperate regions, such as the eastern USA, China, and east-
Figure 4. Spatial distribution of the main Sobol’ index for the sensitivity parameters.

Table 2. ILAMB benchmark scores for latent heat flux, sensible heat flux, evapotranspiration, and terrestrial water storage anomaly with default and optimized parameters in ELM. A description of each score metric can be found at http://redwood.ess.uci.edu/CMIP6_bnchmrk1_9_8/ (last access: 23 June 2021).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Data source</th>
<th>Parameter</th>
<th>Bias score</th>
<th>RMSE score</th>
<th>Seasonal cycle score</th>
<th>Spatial distribution score</th>
<th>Overall score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latent heat flux</td>
<td>FLUXCOM</td>
<td>Default</td>
<td>0.740</td>
<td>0.680</td>
<td>0.910</td>
<td>0.993</td>
<td>0.800</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Optimal</td>
<td>0.730</td>
<td>0.677</td>
<td>0.909</td>
<td>0.992</td>
<td>0.797</td>
</tr>
<tr>
<td>Sensible heat flux</td>
<td>FLUXCOM</td>
<td>Default</td>
<td>0.682</td>
<td>0.643</td>
<td>0.932</td>
<td>0.940</td>
<td>0.768</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Optimal</td>
<td>0.680</td>
<td>0.636</td>
<td>0.932</td>
<td>0.933</td>
<td>0.763</td>
</tr>
<tr>
<td>ET</td>
<td>GLEAM3.3</td>
<td>Default</td>
<td>0.714</td>
<td>0.675</td>
<td>0.870</td>
<td>0.971</td>
<td>0.781</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Optimal</td>
<td>0.705</td>
<td>0.672</td>
<td>0.873</td>
<td>0.967</td>
<td>0.778</td>
</tr>
<tr>
<td>TWSA</td>
<td>GRACE</td>
<td>Default</td>
<td>0.901</td>
<td>0.554</td>
<td>0.818</td>
<td>0.003</td>
<td>0.566</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Optimal</td>
<td>0.900</td>
<td>0.545</td>
<td>0.817</td>
<td>0.004</td>
<td>0.562</td>
</tr>
</tbody>
</table>
Figure 5. Averaged main Sobol’ index and joint Sobol’ index for different climates defined by the Köppen climate classification. Only the cells with the relative norm 2 errors of PCE-based surrogate models for validating simulations of less than 0.15 are used for estimating the averaged sensitivity for each climate region. The size of the circles and thickness of the lines are proportional to main Sobol’ index and joint Sobol’ index, respectively. The legend in the right bottom subplot shows the Sobol’ index for the corresponding size of circle and thickness of line.

ern coast of Australia (Fig. 9b). The simulation with optimal parameter values is able to more accurately estimate $\beta$ than the simulation with default parameter values with improved spatial correlation coefficient from 0.22 to 0.56 and lower RMSE from 1.22 to 0.65 (Fig. 9c). However, some significant discrepancy of $\beta$ still exists in the simulation with optimal parameter values (e.g., eastern China), implying that the sensitivity is not well constrained in ELM for certain regions even after model calibration.

According to the evaluation with the ILAMB package, ELM shows a similar performance in simulating other variables (e.g., latent heat flux, sensible heat flux, ET, and TWSA) with optimal parameter values compared to the use of the default parameter values (Table 2). However, both the default and optimal simulations fail to capture the spatial variation in TWSA with a spatial distribution score of less than 0.05. This is because the coarse resolution (e.g., several hundred kilometers) of GRACE product (Seyoum et al., 2019) cannot resolve the spatial variability in TWSA for our model resolution.

5.7 Parametric uncertainty

The parameter priors listed in Table 1 result in significant uncertainties in the total runoff, with the global average annual runoff for 1991–2010 varying from 30 999–76 496 km$^3$ yr$^{-1}$ (Fig. 10a). After parameter inference, the uncertainty of the runoff constructed using simulations with parameter posteriors is constrained to 35 389–49 741 km$^3$ yr$^{-1}$. The constrained annual runoff uncertainty captures the reference runoff (38 443 km$^3$ yr$^{-1}$) and is consistent with previous global runoff studies (Schellekens et al., 2017; Rodell et al., 2015; Clark et al., 2015; Haddeland et al., 2011). The simulation with the optimal parameter values yields an averaged global annual runoff of 42 156 km$^3$ yr$^{-1}$, overestimating the reference runoff by 9.6%. The overestimation is mainly from the Amazon, Asia, and eastern Europe (Fig. 10b), and Ghiggi et al. (2019a) reported a similar spatial bias pattern between global hydrological model simulations in ISIMP2a. The simulation with the default parameter values shows smaller biases in terms of the annual runoff magnitude as compared to the reference runoff data, with an overestimation of 5.3% on average. However, the smaller biases of annual runoff with the default parameters are because of the canceling out of the monthly errors to some extent. For example, the default parameters tend to overestimate the runoff during the wet periods but underestimate the runoff during the dry periods in the Amazon basin (Fig. S10a). While the default simulation shows higher RMSE and lower NSE at a monthly scale, it yields smaller biases at an annual scale than the optimal parameter (Fig. S10b). Therefore, the calibrated simulation shows better performance in capturing the spatiotemporal variability (higher NSE and KGE in Fig. 8b and e), but it does not lead to a reduced bias at an annual scale. We further acknowledge that 200 simulations with 11 random parameters may not be sufficient to capture the full variations in simulated runoff.

The runoff uncertainties associated with parameters are constrained significantly with the parameter posteriors at basin scale as well (Fig. 11). Noticeably, the posterior uncertainty of annual runoff is larger over the equatorial regions (e.g., Paraná, the Amazon, Godâvári River, and Congo River) than other regions. The simulation with optimal parameter values yields larger overestimation of total runoff compared to the simulation using the default parameter values for the selected basins, except for the Mississippi River, Godâvári River, and Loire basin (Table S1 in the Supplement). The reason for the overestimations is that the optimal parameters are determined by maximizing NSE at a monthly scale, which cannot ensure that the annual runoff is appropriately con-
stratified. There exist significant discrepancies between simulations and GRUN for basins located at high latitudes (e.g., Mackenzie, Volga, Ob, Yenisei, and Lena rivers) even when the posterior uncertainties are considered (Fig. 11), highlighting the importance of snow-melting processes in snow-dominated regions. However, the large difference between ELM and the reference runoff in Yangtze River basin may be caused by the bias of the reference runoff since a previous study reported annual discharge to be around 900 km$^3$ yr$^{-1}$ (Yang et al., 2015).

Despite being constrained by the parameter inference process, the parametric uncertainty of ELM-simulated annual runoff is considerable. Specifically, the posterior uncertainty of global runoff simulated by ELM is comparable to that of the multimodel ensemble constructed with the 13 global hydrological models from ISIMIP2a (Fig. 12a). The parametric uncertainty affects not only the magnitude of global runoff but also the trend for the simulation period during which a rapid increase in temperature has occurred (Fig. S11a). The Sen’s slope (Sen, 1968) for the reference runoff data is found to be 54.7 km$^3$ yr$^{-1}$, but this increasing trend is not significant according to the Mann–Kendall test (Fig. 12b). Other studies also reported no significant changes in the global runoff with observed streamflow data (Alkama et al., 2013; Dai et al., 2009; Milliman et al., 2008; Alkama et al., 2011). However, the default and calibrated ELM simulations yielded the Sen’s slope to be 188.9 and 133.8 km$^3$ yr$^{-1}$, respectively. Although the Sen’s slope is reduced with the optimal parameters, the increasing trend remains significant. Likewise, all the other global hydrological models of ISIMIP2a exhibit significant increasing trend in the annual runoff, with the Sen’s slope varying from 93 to 272 km$^3$ yr$^{-1}$ (Fig. 12b). Considering that the GRUN dataset and all model simulations are forced by the same atmosphere forcing (i.e., GSWP3), the differences in the global runoff trends can be attributed to the model structural/parametric uncertainty. We note that there exists a significant trend in GSWP3 precipitation at a global scale, with an increase of 246.1 km$^3$ yr$^{-1}$ during the simulation period (Fig. S11b). But it remains unclear how the runoff responds to the increase in precipitation at a global scale because the concurrent increased temperature (Fig. S11a) leads to more ET, which can potentially balance the increased precipitation to some extent. The inconsistency of the global runoff trend between the model simulations and observation-based data can be caused by uncertainties of different sources. For example, the accuracy of GRUN is limited by the coverage of the streamflow gauges, as over half of the global areas are ungauged (Alkama et al., 2013). The
model parametric uncertainty is significant, as ELM simulations with parameters posteriors show a wide range of annual runoff trend, from no trend to a significant increasing trend (Fig. 12b). This highlights the necessity of including parametric uncertainty in future runoff projections, since runoff trend is not well constrained even if the model performance in the control period is improved.

6 Limitations

We note there can be other better choices of priors for the parameter whose range covers several orders of magnitude. For example, sampling $q_{\text{drai,max}}$ on a uniform distribution (e.g., $U[10^{-6} 10^{-1}]$) results in fewer prior samples with values less than $10^{-2}$. A log-transformed uniform distribution can be a good alternative to guarantee enough samples over each range of the desired values. Using a log-transformed uniform distribution for $q_{\text{drai,max}}$ prior does not impact our results significantly because the simulated runoff is not sensitive to $q_{\text{drai,max}}$ (Fig. S4), and more samples over smaller values of $q_{\text{drai,max}}$ will not lead to more variation in runoff. However, careful selection of prior distributions can be important for sensitive parameters in a future application of surrogate-assisted calibration framework.
Figure 8. Evaluation of simulated monthly runoff for 1991–2010 at grid level with default and optimal parameters. Panels (a) and (b) show the NSE metrics between the GRUN runoff and simulated runoff with default and optimal parameters, respectively. Panel (c) shows the comparison of the probability density function (PDF) of NSE metrics from all the global grid cells. Panels (d), (e), and (f) illustrate the evolution with the KGE metric.

By using RMSE instead of the simulated runoff as the QoI, only one PCE-based surrogate model is constructed for each grid cell to represent the ELM performance of simulating a monthly runoff time series. Although selecting RMSE as QoI significantly reduces the computational burden of the surrogates’ construction and parameter inference, the corresponding surrogates cannot be used to estimate posterior uncertainty of physical model outputs. For example, we still need to run ELM simulations after the parameter inference to construct the runoff posterior uncertainty (Sect. 3.5). Additionally, the objective of this study is to minimize RMSE at a monthly scale; hence, an improved model performance at an annual scale is not guaranteed. Including both monthly and annual performance metrics in objective function may balance the performance at different temporal scales. However, only one objective is accepted in the Uncertainty Quantification framework used in this study.

We further acknowledge the poor performance of PCE-based surrogate model in capturing the ELM-simulated runoff over extremely arid regions (Figs. 2 and 3). This can be attributed to the limitation of polynomial-based surrogate models in capturing highly non-smooth or strongly nonlinear relationships. Machine learning algorithms (Dagon et al., 2020) and deep neural networks (Tsai et al., 2021) are alternative techniques for surrogate modeling, which are better at capturing non-smooth or nonlinear functions, but future research is needed to investigate the capability.

The calibrated parameters have a significant impact on baseflow index, which is the ratio between subsurface runoff and total runoff. For example, the baseflow of Amazon basin with default and optimal parameters are 0.53 and 0.70, respectively (Fig. S12). Mortatti et al. (1997) reported the baseflow index of the Amazon basin to be 0.70 with the isotopic tracer method, which is consistent with our simulation with optimal parameter values. However, accurate separation of surface runoff and subsurface runoff over other regions is not guaranteed, though the total runoff has been calibrated to match with the reference runoff dataset. The global baseflow index dataset of Beck et al. (2013) that derived from observed streamflow provides us the benchmark for evaluating the baseflow index simulated in ELM. Constraining the baseflow index during the ELM validation and calibration study will be investigated in the future.

We further note that uncertainty in the reference runoff data of GRUN used in the parameter inference is inevitable. While Ghiggi et al. (2019a) found that GRUN outperformed...
Figure 9. Sensitivity of runoff to precipitation ($\beta$) estimated from the (a) GRUN runoff dataset, (b) ELM simulation with default parameter, and (c) ELM simulation with optimal parameter. The inserts show the scatterplots with density for cell-to-cell comparison of $\beta$ between GRUN and ELM simulations.

other global hydrological models and multimodel ensemble, lower accuracy over mountainous regions due to the coarse resolution has been reported. Additionally, the irrigation and water management impacts on streamflow was included for some regions during the training process of GRUN (Ghiggi et al., 2019a), but irrigation and water management are not active in the ELM configuration used in this study. This inconsistency may explain the significant overestimation of ELM simulated runoff compared to GRUN for certain regions, for example, the Yangtze River basin (Fig. 10).

Another limitation of this study is that the snow-melting processes were not calibrated. A poor representation of the snow-melting process can result in the poor skill of runoff generation in snow-dominant regions, where snowmelt is an important contribution to runoff (Jenicek and Ledvinka, 2020). This could explain the low performance (i.e., negative NSE) of calibrated ELM over the Northern Hemisphere’s high latitudes and mountainous regions. However, including parameterizations of snow processes such as snow albedo, solar absorption, and snow aging (Lawrence et al., 2011) can introduce more uncertain parameters, which will make calibration more challenging (Huang et al., 2013). In the future, a dedicated calibration on the snow-melting process is needed to improve the runoff generation in snow-dominated regions.

7 Conclusion

In this study, we applied an UQ framework to calibrate the runoff-generation-relevant parameters in the ELM v1 using...
an observation-based runoff dataset as the benchmark. The parameters with higher sensitivity are identified through the sensitivity analysis with the PCE-based surrogate models. While different sensitivity patterns are found for different regions, 81.3% of the global cells show significant sensitivities to 3 or fewer parameters of the 11 selected parameters. The results of our sensitivity analysis are consistent with those of previous studies over the North American continent (Huang et al., 2013; Sun et al., 2013), with runoff showing the largest sensitivity to the subsurface runoff parameter. The Bayesian posterior distribution of the highly sensitive parameters at each grid cell is estimated with MCMC simulations, using the surrogate model to construct the likelihood function. Additional ELM simulations with parameter samples from the posterior run to estimate the optimal parameter values and construct the parametric uncertainty for the simulated runoff. While the optimal parameter values improve the model performance of runoff significantly, the parametric uncertainty is comparable to the uncertainty in a multimodel ensemble in ISIMIP2a, which is appreciable. Furthermore, the parameters are found to impact the annual global runoff trend for our simulation period. Specifically, the simulations with parameter posteriors show a wide range of the annual runoff trends at a global scale, from no trend to a significant increasing trend. In summary, parameter calibration is necessary to improve model performance, and parametric uncertainties should be considered for comprehensive analysis of runoff and its projections.

**Code and data availability.** The current version of ELM is available from E3SM project (https://github.com/E3SM-Project/E3SM/releases/tag/v1.1.0, last access: 24 June 2022). The UQTk code and documentation are available from https://www.sandia.gov/uqtoolkit/ (last access: 24 June 2022). The exact version of ELM, exact version of UQTk source code, and scripts to produce the plots in this study are archived on Zenodo (https://doi.org/10.5281/zenodo.5815500,
Figure 11. Annual runoff at basin scale from the default ELM simulation, optimal ELM simulation, and GRUN runoff dataset for the simulation period (1991–2010). The red and blue shaded areas represent the uncertainties constructed from the simulations with parameter sampled on priors and posteriors, respectively.

Figure 12. (a) Annual global runoff from 13 global hydrological models participated in ISIMIP2a and ELM simulated runoff uncertainty constructed using simulations with parameter posteriors. (b) Sen’s slope for the global annual runoff for the GRUN runoff dataset and simulations. The violin plots (Hintze and Nelson, 1998) are generated with Sen’s slope of ELM simulations with parameter posteriors. The white point is the median value, and the gray line represents range of the 25%–75% percentile. The MATLAB function of Bechtold (2016) was used to create the violin plot. The cross signs are the Sen’s slopes estimated from the ISIMIP2a model simulations.
Author contributions. DX and GB designed the study. DX ran the simulations, performed the analysis, visualized the results, and prepared the first draft of the paper. GB mentored DX through this study. KS helped with the uncertainty quantification methodology. CL and LRL investigated the results. All authors contributed to the discussion and review of the results and writing the paper.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher’s note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements. This research was conducted at the Pacific Northwest National Laboratory, operated by Battelle for the U.S. Department of Energy (contract no. DE-AC05-76RL01830). Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy’s National Nuclear Security Administration (contract no. DE-NA-0003525). We thank the modelling groups listed in Fig. 12, and the Inter-Sectoral Impact Model Intercomparison Project (ISIMP) for making the ISIMP2a multi-model dataset available. We also thank two anonymous reviewers for their constructive comments that improved the manuscript.

Financial support. This work was supported by the Earth System Model Development program area of the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, as part of the multi-program, collaborative integrated Coastal Modeling (ICoM) project (grant no. KP1703110/75415). Chang Liao was supported through Next-Generation Ecosystem Experiments-Tropics, funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, at Pacific Northwest National Laboratory (grant no. 830403000/71073).

Review statement. This paper was edited by Dan Lu and reviewed by two anonymous referees.

References


