Water Resources Research

RESEARCH ARTICLE
10.1029/2020WR028740

Key Points:
- A surrogate modeling based multi-objective optimization method using grid sampling is proposed for calibration of environmental models
- Multi-objective adaptive surrogate modeling-based optimization using grid sampling (MO-ASMOGS) uses only 10% or less of the total grid cells to obtain the same calibration results as MO-ASMO using 100% of the grids
- MO-ASMOGS is suitable for computationally intensive distributed environmental models such as continental or global-scale land surface models

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

Q. Duan,
qyduan@hhu.edu.cn

Citation:

Sun, R., Duan, Q., & Huo, X. (2021). Multi-objective adaptive surrogate modeling-based optimization for distributed environmental models based on grid sampling. *Water Resources Research*, *57*, e2020WR028740. https://doi.org/10.1029/2020WR028740

Received 2 SEP 2020
Accepted 26 OCT 2021

Author Contributions:

Conceptualization: Ruochen Sun, Qingyun Duan
Data curation: Xueli Huo
Formal analysis: Ruochen Sun
Funding acquisition: Ruochen Sun, Qingyun Duan
Methodology: Ruochen Sun, Xueli Huo
Software: Xueli Huo
Supervision: Qingyun Duan
Visualization: Ruochen Sun
Writing – original draft: Ruochen Sun
Writing – review & editing: Qingyun Duan

Multi-Objective Adaptive Surrogate Modeling-Based Optimization for Distributed Environmental Models Based on Grid Sampling

Ruochen Sun[1,2], Qingyun Duan[1,2], and Xueli Huo[3]

[1]State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing, China, [2]College of Hydrology and Water Resources, Hohai University, Nanjing, China, [3]School of Natural Resources and Environment, The University of Arizona, Tucson, AZ, USA

Abstract Parameter optimization is needed for reliable simulations and predictions of natural processes by environmental models. The surrogate modeling-based approach is an efficient way to reduce the number of model evaluations needed for optimization. However, building a surrogate of a distributed environmental model with many output variables over a large spatial domain is computationally intensive as it involves a large number of expensive model simulations on many spatial grid cells. In this study, a novel calibration method called the multi-objective adaptive surrogate modeling-based optimization using grid sampling (MO-ASMOGS) is introduced. This method constructs the response surface surrogate of the original model more efficiently by using both parameter and spatial grid sampling. The spatial grid sampling strategy utilizes the evolutionary elitism and adaptive sampling concepts, thus allowing the surrogate model to be built using a fraction of the total grid cells over a large region. We apply MO-ASMOGS to calibrating the Noah-MP model against two surface fluxes: the gross primary production (GPP) and the latent heat flux (LH), over two plant function types (PFTs) across the continental United States. The results demonstrate that the MO-ASMOGS method can significantly improve the GPP and LH simulations. The new method needs only a small portion of the total grid cells sampled for a given PFT to achieve comparable optimization results obtained by MO-ASMO using all grid cells. This method can be very valuable in improving model calibration of computationally intensive distributed environmental models.

1. Introduction

Computer-based environmental models (EMs) have been widely used to explore, simulate and predict the behaviors of the environmental systems and their complex reactions to changing conditions. Examples of environmental systems include hydrological, ecological, and climatic systems (Letcher & Jakeman, 2004). As researchers are seeking to improve their understanding of the real-world environmental systems, EMs are becoming progressively more complex. In order to represent the heterogeneity and the spatial variation of properties that control the environmental system processes, spatially distributed EMs, which represent various components of the earth system at large scale (Koch et al., 2018), have been proliferated and extensively applied in a wide variety of fields (Gan et al., 2019; Lin et al., 2018; Sun et al., 2018, 2021). These large, complex EMs usually involve many uncertain spatially varying parameters whose specification could significantly affect the model simulation capability (Lohmann et al., 2004; Xia et al., 2016). Those parameters usually cannot be directly measured and the assigned default values based on land surface characteristics (e.g., soil and vegetation types) are usually inappropriate (Gu et al., 2016; Huang et al., 2013; Rosero et al., 2010). To resolve this problem, many researchers resorted to model calibration to estimate model parameters.

Calibration of complex models has necessitated the use of automated, time-efficient optimization methods. There have been major advances in automatic calibration algorithms as a result of the development and breakthroughs in the field of optimization (Deb et al., 2002; Duan et al., 1992; Tolson & Shoemaker, 2007; Wang, 1991; Yapo et al., 1998). Automatic optimization methods have been widely used to calibrate hydrologic models. There also have been some attempts to use these methods to calibrate parameters of land surface models (LSMs) (Bastidas et al., 1999; Gupta et al., 1999; Xia et al., 2002), ecosystem models (Fox et al., 2009; Ricciuto et al., 2008), and numerical weather prediction models (Liu et al., 2005; Severijns & Hazeleger, 2005). With the continuous development, current EMs are incorporating an increasing number of natural processes and representing spatial heterogeneities of these processes (Letcher & Jakeman, 2004), and thus could simulate many variables simultaneously. Therefore, multi-objective optimization is more favorable to ensure that all major variables of interest are well simulated, and the use of additional variables can allow calibration to become more robust with multiple observations constraining model parameters (Bastidas et al., 2003). The multi-objective optimization problem is much more complex than the single-objective optimization, thus further increasing the difficulty of model calibration (Gupta et al., 1998). Conventional automatic optimization algorithms, especially multi-objective methods, require a huge amount of model evaluations to identify the optimal parameter sets (Gupta et al., 1999; Li et al., 2012; Vrugt et al., 2003). Meanwhile, it could take considerable CPU time to run a distributed EM over a large spatial domain for a multi-year simulation. This high computational demand has become one of the greatest challenges in practical applications of complex distributed EM calibration. Therefore, more advanced methods are needed to obtain good parameter estimates within an acceptable time.

To reduce the computational cost of model calibration, surrogate modeling-based optimization methods have been proven as an efficient way. The surrogate models, which are usually statistical models, provide efficient approximation to the actual numerical model outputs/objective functions (Lu et al., 2018). Surrogate modeling has been widely used in engineering fields like aerospace science, civil engineering, robotics, chemistry et al. (Forrester & Keane, 2009; Gorissen, 2010; Viana et al., 2014). Razavi et al. (2012) gave a comprehensive review of research efforts on surrogate modeling and the applications in the field of hydrology and water resources. In addition, the applications of surrogate modelling-based optimization methods to complex EMs are gaining more attentions (Duan et al., 2017; Fer et al., 2018; Gong et al., 2015; Huang et al., 2016; Lu et al., 2018; Müller et al., 2015; Ray et al., 2015; Wang et al., 2014). Among these studies, some tried to transform the multi-objective problem into a single-objective problem and used single-objective surrogate-based optimization methods (Duan et al., 2017; Gong et al., 2015; Lu et al., 2018). Since the essence of solving a multi-objective optimization problem is to find the Pareto optimal solutions, Gong et al. (2016) proposed the multi-objective adaptive surrogate modeling-based optimization (MO-ASMO) algorithm based on the nondominated sorting of multiple objectives to meet the special requirements of calibrating large, complex geophysical models. When applied to the Common Land Model (CoLM), the MO-ASMO method identified the Pareto optimal parameter sets with greater effectiveness and efficiency compared to the popular nondominated sorting genetic algorithm II (NSGA-II; Deb et al., 2002).

Most of the previous studies on parameter optimization of distributed EMs were limited to single sites where in-situ observations are available. It has been pointed out that parameters estimated at point scale are not easily transferable to new sites or to large-scale applications (Post et al., 2017). It means that the optimal parameter sets obtained for one site are possibly not the optimal values for other sites, although these sites are classified as the same type according to the plant functional type (PFT) and soil texture (McNeall et al., 2016; Rosolem et al., 2013). This explains why many distributed EMs using calibrated model parameters from individual sites still produce large spatial bias. Therefore, calibration of spatially distributed EMs at limited sites is not enough (Li, Duan, et al., 2018). Either because of the high spatial and temporal resolution or because of the massive scale (Sun et al., 2020), application of parameter optimization methods to spatially distributed EMs is not trivial due to the high computational demands. Surrogate modeling-based optimization methods can be adopted to save computational cost, but it is not enough for the optimization problems related to distributed EMs at large scales. This is because that the construction of an accurate surrogate model still requires many expensive model runs, especially in view of the fact that the spatially distributed EMs need to be run over the entire spatial domain with a huge number of grid cells over many years. Therefore, how to develop strategies for computationally efficient model calibration at continental-to-global scales is a vital issue needed to be addressed.

Currently, the relevant research is very limited. One strategy is to further reduce the number of model evaluations involved in the optimization algorithm. For example, some studies (Dagon et al., 2020; Lu & Ricciuto, 2019) used singular value decomposition to reduce the dimensionality of the model output, and then trained a neural network based surrogate model to emulate the model outputs. In this way, an accurate and effective surrogate system of model outputs at large spatial scale can be built based on only a few model runs. Another strategy is to reduce the computational time of a single model run based on a sparse grid idea. Troy et al. (2008) introduced a method to calibrate the Variable Infiltration Capacity (VIC; Liang et al., 1994) hydrologic model for a subset of the grid cells individually and then interpolated the parameters to the uncalibrated grid cells. However, the selection of grid cells to be calibrated was relatively arbitrary and how it affected the calibration results was not clear. Similarly, the idea of running the model on a subset of grid cells was also implemented by Huo et al. (2019) to perform the global sensitivity analysis (SA) of the Noah-MP (Niu et al., 2011) LSM. The results demonstrated that a relatively small grid sample size (5% of the total grid cells) was sufficient to identify the most important parameters of the spatially distributed Noah-MP model. Compared to SA, which is only needed to depict the general shape of the response surface of the original model, calibration is much more complex and needed to explore the specific details of the response surface. Therefore, how to sample grid cells from the large domain to facilitate the calibration process is worth studying.

In this study, we introduce a novel calibration method called multi-objective adaptive surrogate modeling-based optimization using grid sampling (MO-ASMOGS), to significantly reduce the computational burden in multi-objective parameter optimization of spatially distributed EMs. This method extends the advantage of the MO-ASMO method with a novel spatial grid sampling strategy. The essential idea of MO-ASMOGS is to construct and to adaptively update the response surface surrogate of the original model more efficiently by combining both parameter sampling and spatial grid sampling (and adaptive resampling). In other words, this novel calibration method only requires model evaluations on a small portion of the total grid cells to make sophisticated calibration of spatially distributed EMs computationally feasible. Different from MO-ASMO and other traditional surrogate modelling-based optimization methods which only rely on parameter sampling for modeling the response surface of the original simulation model, the MO-ASMOGS method introduces a novel spatial grid sampling strategy to better implement the new idea, which greatly facilitates the application of complex multi-objective calibration over large spatial domains and therefore makes the novel contribution of this study. Our proposed spatial grid sampling strategy involves two core concepts: evolutionary elitism and adaptive sampling (or multi-stage sampling). Implementation of the first concept allows for the more important (i.e., the more sensitive) grid cells to be selected with higher priorities. Implementation of the second concept allows the user to monitor the performance of the sampling-based optimization in an "online" adaptive manner. With the design of the spatial grid sampling strategy, the optimization performed on only a small portion of the total grid cells over a large region can improve all objective functions relative to those obtained with the default parameterization scheme and produce optimization results that are just as good as MO-ASMO that needs running the model on all of the grid cells. We apply this calibration method to the Noah-MP model to improve the model simulation capability of gross primary production (GPP) and latent heat flux (LH) for two PFTs across the continental United States (CONUS).

The rest part of the paper is organized as follows. Section 2 describes the detailed methodology. Section 3 briefly describes the Noah-MP model, the study region, various data and the experiment setup. Section 4 presents the detailed results, followed by discussion and conclusions in Section 5 and Section 6.

2. Methodology

2.1. The MO-ASMOGS Method

In this section, we present the multi-objective calibration method MO-ASMOGS, which combines the advantage of a surrogate modeling-based multi-objective optimization approach (MO-ASMO) with a novel spatial grid sampling scheme to calibrate spatially distributed EMs that use individual grid cells as primary calculation units. The essential idea of MO-ASMOGS is combining parameter sampling and spatial grid sampling to construct and adaptively update the surrogate model of the objective functions. Therefore, it only needs running the model on a representative subset of grid cells selected out of the total number of grid cells within the spatial domain, rather than on the entire grid cells. SA is commonly used as a prerequisite for parameter estimation to identify sensitive parameters. Similarly, our method first picks out the "most influential" grid cells, as the model outputs on some grid cells are not sensitive to the parameters to be optimized. Therefore, we run the model only on certain grid cells to construct surrogate models and conduct optimization. The overall goal of the MO-ASMOGS method is that the optimization carried out on the representative grid cells can improve all the objective functions relative to those obtained with the default parameterization scheme and produce comparable results to those obtained by performing optimization on all grid cells.

Figure 1a shows the flow chart of the MO-ASMOGS method, which consists of two phases: a spatial grid sampling phase (shown in blue) and a surrogate modeling-based optimization phase (shown in black). The former controls how the calibration will be conducted in the optimization phase by determining the parameters to be optimized, assigning reasonable ranges for model parameters, investigating the importance ranking of each parameter on each grid cell, selecting representative grid cells based on a novel spatial grid sampling strategy (Figure 1b). The second phase is to carry out surrogate modeling-based optimization using MO-ASMO. The input (parameter value)-output (objective function value) data needed to construct the surrogate model are obtained by running the original dynamic model with sampled parameter sets on the representative grid cells and computing the corresponding objective function values. The MO-ASMOGS method supports both one shot and adaptive (or multi-stage) modes. The one-shot mode means that the representative grid cells are only sampled once for later surrogate modeling-based optimization. If the user is not satisfied with the optimization results, more representative grid cells are added to increase the accuracy of the surrogate model. In the adaptive mode, additional grid cells are progressively added to update the surrogate model. The key steps of MO-ASMOGS are described as follows:

1. Problem definition: Specify model outputs of interest and determine the model parameters to be investigated. People can rely on their model expertise or refer to previous studies to determine the parameters.
2. Initial sampling: An initial sample set of spatial grid cells is generated using the proposed spatial grid sampling strategy. The Good Lattice Points (GLP) method with ranked Gram-Schmidt (RGS) decorrelation (Gong et al., 2016), is used to generate an initial set of parameter sample points.
3. Model evaluation: The EM is run with the initial set of parameter samples on the sampled grid cells.
4. Build the surrogate model: The Gaussian Processes Regression (GPR) is used to build surrogate models that emulate the response surface of the original EM to the change in parameter values. In this study, the surrogate model is an approximation of the objective functions, which are calculated from the raw model outputs on the sampled grid cells.
5. Multi-objective optimization: The multi-objective optimization algorithm NSGA-II is run on the surrogate models built in the previous step to obtain nondominated solutions.
6. Adaptive parameter sampling: These solutions are sorted in descending order of the crowding distance, and a portion of these sorted solutions with the largest crowding distances (represent the diversity of the nondominated solutions) are selected for evaluation using the EM model. Run the EM on the representative grid cells with the selected parameter sets. Append these new input-output data pairs to the original data pool. Steps 4–6 are repeated until the iteration limit is reached.
7. Adaptive grid sampling: If not satisfied with the optimization results from Step 6, add new grid cells based on the spatial grid sampling strategy and go back to step 3. Repeat steps 3–7 until the termination condition is met.

Note that the initial parameter sample points are only generated once. Every time the new grid cells are added, the original EM is only needed to be run on the new grid cells with the initial parameter samples to generate the training data.

Since typical EMs have default assignment of parameter values (reference point), we use the MO-ASMO with weight crowding distance (WMO-ASMO) to perform the surrogate modeling-based model optimization. An important feature of WMO-ASMO is the use of the weighted crowding distance when sorting the nondominated solutions. Therefore, this method can make the optimization focus on the region that can improve all of the objectives compared to what can be achieved with the default parameter set, which is suitable for the calibration of complex EMs. Please refer to Gong et al. (2016) for more details.

The MO-ASMOGS method is another extension to our developed ASMO methodology family. Although the MO-ASMO is one of the main components of the proposed MO-ASMOGS method, the new method has significant differences in the following aspects: (a) MO-ASMOGS is specifically designed for distributed EMs (e.g., LSMs) which are run on spatial grid cells and have gridded outputs. (b) Considering the spatial heterogeneity of the parameter sensitivity, some grid cells are more influential in reducing model errors

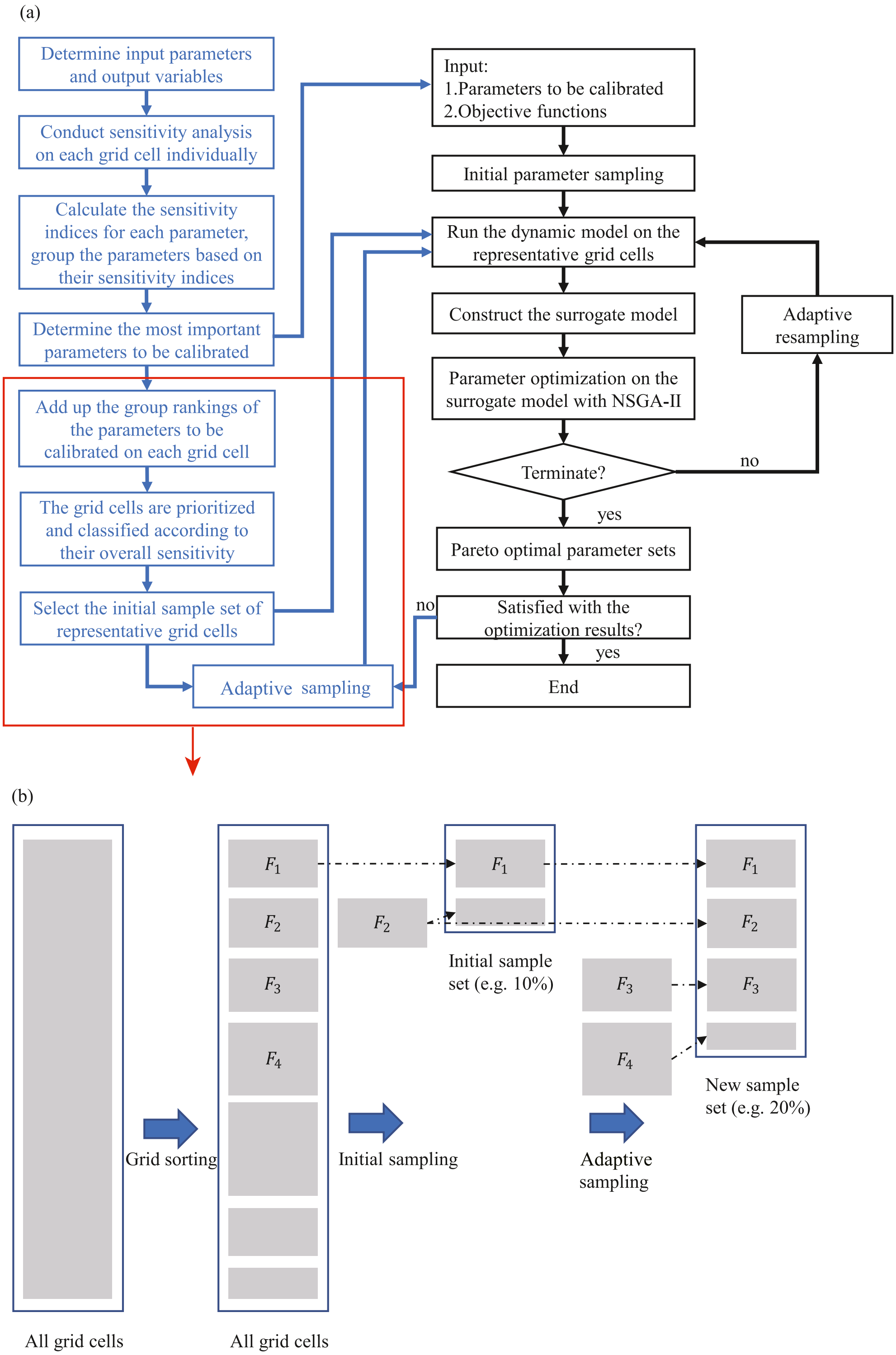

Figure 1. (a) Flowchart of the MO-ASMOGS method with the spatial grid sampling strategy shown in blue and the MO-ASMO method shown in black; (b) a schematic of the key steps in the spatial grid sampling strategy.

than others. Thus, the surrogate model in MO-ASMOGS is constructed and updated based not only on parameter sampling as in MO-ASMO, but also on spatial grid sampling. (c) The designed spatial grid sampling strategy is flexible to cater for the need of adding additional spatial grid samples to increase the accuracy of the surrogate model. Therefore, the core novelty of the MO-ASMOGS method is building and improving the response surface surrogate of the distributed EM more efficiently based on both parameter and spatial grid sampling (and adaptive resampling) through introducing a new spatial grid sampling strategy. This strategy is described in detail in Section 2.2.

2.2. Spatial Grid Sampling Strategy

Due to the land surface heterogeneity, the sensitivity of a parameter can vary with the grid cells. In other words, some regions are more sensitive and have greater potential for reducing model errors. Thus, we do not necessarily calibrate all grid cells to obtain an optimal set of parameters, and only need to run the dynamic model on some selected representative grid cells to obtain the input-output pairs for performing the surrogate modeling-based optimization. The overall goals of the optimization performed on the representative grid cells are (a) to improve the model performance in simulating all output variables of interest across the study region and (b) to produce comparable calibration results to those obtained by calibrating all grid cells. Because the objective function values, which are obtained by evaluating model outputs on the selected grid cells, directly affect the constructed response surface, the representative grid cells for running the dynamic model should be carefully selected to focus on the most "informative" regions. To meet the goals, the spatial grid sampling strategy have four criteria: (a) To reduce the computational cost as much as possible, the initial sample size of the representative grid cells cannot be too large. (b) To make sure the surrogate model constructed on the limited grid cells can better approximate the response surface of the original model, enough information should be collected from the "informative" region where the model outputs change drastically with parameter values. Thus, evolutionary elitism is needed to ensure that more informative grid cells can be selected with higher priority. (c) If there are many grid cells that have similar information but are scattered across the study region, the representative grid cells are uniformly sampled from them to maintain spatial diversity. (d) If the user is not satisfied with the optimization results and wants to enlarge the sample size and resume the optimization process with the updated/new sample, additional grid cells can be added progressively to update the surrogate model and proceed with the optimization. Consequently, we have designed the following procedures to implement the four criteria, and the schematic description of the last three steps in the spatial grid sampling strategy is shown in Figure 1b.

1. On each grid cell, calculate the sensitivity indices for each parameter through a prescribed SA method. The parameters are then divided into m groups based on the calculated sensitivity indices. Parameters categorized into the first group with ranking of 1 are labeled as "strongly influential," and the parameters in the last group with ranking of m are considered as "weakly influential." The number of groups is prespecified by the user, however, large number of groups is not recommended as this may translate into some of them being disproportionally emphasized or de-emphasized (Sun et al., 2020). The grouping of parameters can be done in a subjective and case-specific manner. There are many grouping strategies which can be used, and we adopt a new grouping strategy proposed by Sheikholeslami et al. (2019). Its biggest advantage is that it can group parameters into a certain number of groups with any size based on information gained from SA. Importantly, if the number of groups is not prespecified, the algorithm will efficiently determine an optimal number of groups.
   The SA and grouping procedure can be applied to all the tunable parameters to screen out the most sensitive parameter to be optimized, just like the way in Huo et al. (2019). However, if the most influential parameters have already been identified by other means, this procedure will only be applied to these parameters, which could also save computational resources.
2. Add up the group rankings of all focused parameters on each grid cell. Let S_g denotes the sum of group rankings on a grid cell. Sort all the grid cells in ascending order of S_g. Grid cells with the same S_g are further sorted in ascending order of aridity index and are grouped into the same set. Now, grid cells belonging to set F_1 have the smallest S_g, meaning that the model outputs on these grid cells are most sensitive to the parameter values. In another word, the grid cells belonging to set F_1 are the most influential ones in the optimization process and must be emphasized more than other grid cells in the sampling process. The sum of parameter group rankings is an indicator of sensitivity of objective function values

to the grid cells. If S_g is small, it means that the corresponding grid cell has high sensitivity and thus a large impact on objective function values. Based on this, we prioritize the grid cells and sample them for further implementation of model calibration.

3. Set an initial spatial grid sampling level, corresponding to a proportion of the total number of grid cells within the entire region. At the initial sampling level (such as 5% of the total number of grid points) with N grid cells, spatial grid samples are selected in order of S_g. If the size of F_1 is smaller than N, all members of the set are chosen. Then grid cells from the set F_2 are chosen next, followed by grid cells from the set F_3, and so on. This procedure is continued until no more sets can be accommodated. Say that the set F_l is the last set containing grid cells with the same S_g, and beyond this set no other set can be accommodated. In general, the number of grid cells in all sets from F_1 to F_l would be larger than N. To fill all sample slots of the sampling level, only a subset of grid cells from F_l can be selected to make the sample size reach N. The remainder grid cells are uniformly sampled from F_l to maintain spatial diversity. Note that for the one-shot mode of using MO-ASMOGS, the spatial grid sampling terminates at this step.
4. To implement the fourth criterion, a new set of grid cell samples with size M is generated and added to the existing samples. M is set to a proportion of the total number of grid cells (e.g., 5%). The additional sampling starts with F_l, and proceed with F_{l+1}, F_{l+2}, ..., as in the last step. When it finally comes to F_e to reach the sample size M, the remainder subset of the added sample set is generated using the progressive Latin hypercube sampling (PLHS, Sheikholeslami & Razavi, 2017) method. The PLHS method generates a series of small subsets so that the sample size can grow progressively during the analysis while the progressive union of subsets remains the proprieties of Latin hypercube. More detailed information can be referred to Sheikholeslami and Razavi (2017). In our spatial grid sampling strategy, the grid cells from the same set F can be divided into a series of subsets using PLHS for the adaptive or multi-stage sampling. This step can be repeated several times until the stopping criteria are met (iteration limit, total original model evaluation limit, total number of representative grid cells limit, etc.). The additional sampling can be done by simply selecting new grid cells from the left ones following (3). However, the spatial diversity of the grid cell samples can be ensured to the maximum extent through the PLHS approach when there are many grid cells with the same S_g.

In summary, two core concepts are involved in the new spatial grid sampling strategy: evolutionary elitism and adaptive (multi-stage) sampling. Implementation of the first idea allows more important or sensitive grid cells to be selected with higher priorities. By doing this, a relatively accurate surrogate model and further good optimization effectiveness can be obtained based on a small sample of grid cells. Implementation of the second idea allows the user to monitor the performance of the sampling-based optimization in an "online" adaptive manner. The surrogate model can be adjusted and improved by adding grid cells progressively, while maintaining the spatial diversity. Therefore, the MO-ASMOGS so designed can significantly reduce the computation burden of the calibration process and ensure satisfactory optimization results.

2.3. Objective Function Formulation

The objective function used in this study is formulated in the following way. For each of the target output of the Noah-MP (LH and GPP), we calculate the Root Mean Squared Error (RMSE) between the monthly simulated and observed values in the study period on each selected representative grid cell. We define the RMSE of the monthly GPP and LH on the j th grid point as follows:

$$\mathrm{RMSE}_v^j = \sqrt{\frac{\sum_{i=1}^{N}\left(S_v^{i,j} - O_v^{i,j}\right)^2}{N}} \tag{1}$$

where $S_v^{i,j}$ and $O_v^{i,j}$ are the simulated and observed monthly v (GPP or LH) in the th month on the j th grid. N is the number of total months.

Then we construct the cumulative distribution function (CDF) curve of these RMSE values (Figure S1) and calculate the area between the curve and the left vertical axis as the objective function (S). We can say that better overall model performance in the spatial domain is associated with smaller value of S. Let S_{LH} and S_{GPP} denote the objective function values with respect to LH and GPP, respectively. Therefore, the aim of

Table 1
Most Influential Parameters of the Noah-MP Model for the Two Plant Function Types (PFTs) (Huo et al., 2019)

Plant function type	GPP	LH
Grassland (G)	smcmax, vcmx25, rmf25	smcmax, bexp
Deciduous broadleaf forest (D)	vcmx25, rmf25	bexp

Note. GPP, gross primary production; LH, latent heat flux.

the multiobjective optimization in this study is to simultaneously minimize both of the objective function values.

To further compare model errors when using calibrated parameters with model errors when using the default parameters, we define the relative RMSE change as:

$$\Delta = \frac{\mathrm{RMSE}_c - \mathrm{RMSE}_d}{\mathrm{RMSE}_d} \times 100\% \quad (2)$$

where RMSE_c and RMSE_d are the RMSEs when the model is run with calibrated and default parameters, respectively. Negative value of Δ means that model error is reduced after calibration.

3. Case Study Configurations

3.1. Model and Study Area

The spatially distributed EM used for this study is the Noah-MP model. Noah-MP is a new generation LSM which models the states of terrestrial energy, water, carbon, and associated flux exchanges between the land surface and the atmosphere. It is an augmented version of Noah LSM (Chen & Dudhia, 2001; Chen et al., 1996; Ek et al., 2003) through the incorporation of the conceptual realism in biophysical and hydrological processes, and multiple parameterization options for key land-atmosphere interaction processes (Niu et al., 2011), such as the vegetation phenology, canopy stomatal resistance, surface exchange coefficient for heat, soil moisture factor for stomatal resistance, runoff and groundwater, frozen soil permeability, radiation transfer, and snowpack (Li, Chen, et al., 2018). Details about the model physical parameterizations of each physical process can be referred to Niu et al. (2011). Here, we use the same model version and the same physical options as in Huo et al. (2019), as suggested by Yang et al. (2011) and Ma et al. (2017).

Improvements in hydrological process representation and carbon cycle process representation are the main directions of LSM development. GPP and LH are arguably the most fundamental and key variables of all carbon cycle and hydrological quantities (Prentice et al., 2015). So, they are widely selected as the target variables for evaluating the model performance. Huo et al. (2019) conducted a parameter SA of the Noah-MP model for modeling GPP and LH using the newly developed grouping-based SA approach (Sheikholeslami et al., 2019). It identified the “strongly influential” parameters (Table 1) for the GPP and LH associated with two PFTs, grassland and deciduous broadleaf forest (DBF) across the CONUS (Figure 2). Since the Noah-MP is run with spatial resolution of 0.125°, there are 8,362 and 5,257 grid cells on which the dominant vegetation type is grassland and DBF, respectively. As a follow-up study of Huo et al. (2019), we want to optimize the most sensitive parameters to improve the simulated GPP and LH for each PFT based on the newly developed MO-ASMOGS method. In this study, we use both the adaptive mode and the one-shot mode of MO-ASMOGS to calibrate the model. In the adaptive mode, we begin with the initial spatial grid cell sampling level corresponding to 5% of all grid cells, then add additional grid cells to double the sample size. This process is repeated two times to get spatial grid sampling levels corresponding to 10% and 20%. In the one-shot mode, we directly select the representative grid cells at the grid sampling level of 40%. In this way, the computational time of these two modes for calibrating the model is nearly the same. To serve as a benchmark, we also conduct calibrations based on the MO-ASMO method by running the model on 100% of the grid cells. The detailed information of these parameters including names, physical meanings, default values, and value ranges are shown in Table 2. Note that the parameters smcmax and bexp actually refer to the spatially constant multipliers to be applied to the default parameter values which can be found in the Noah-MP look-up table based on soil texture classifications. Applying the spatially constant parameter

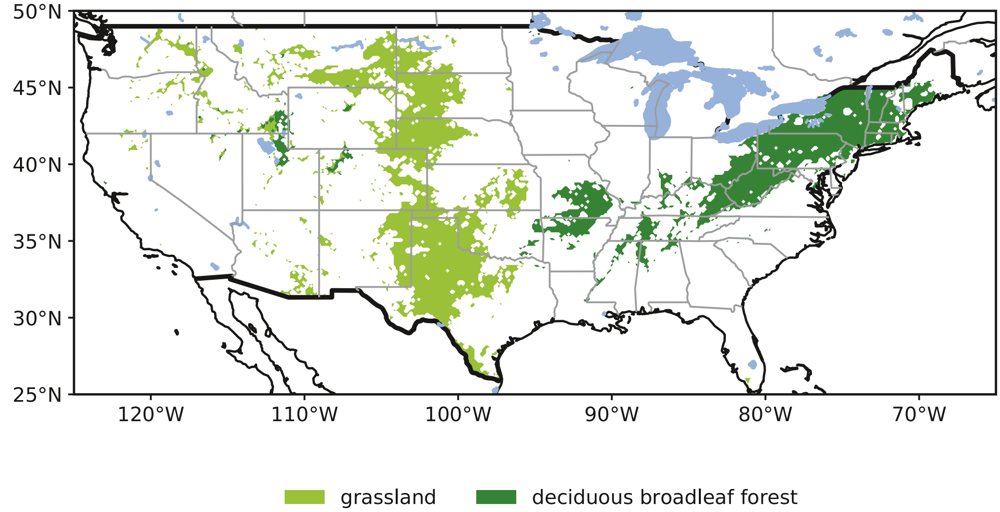

Figure 2. The spatial distribution of grassland and deciduous broadleaf forest (DBF) across the continental United States (CONUS).

multipliers to a priori parameter fields is a common form of spatial regularization for calibrating spatially distributed hydrological and land surface models (Mizukami et al., 2017). The vegetation parameters are PFT-specific, and are constant within the same PFT region.

3.2. Data and Setup

The associated datasets used as inputs of the Noah-MP model and the evaluation data for simulated GPP and LH are the same as in Huo et al. (2019) and Ma et al. (2017). The inputs of the Noah-MP model include meteorological forcing data and static geography data. The forcing data consisting of precipitation, air temperature, wind speed, surface pressure, specific humidity, downward solar radiation, and downward longwave radiation, are collected from Phase 2 of the North American Land Data Assimilation System (NLDAS-2) forcing datasets with spatial and temporal resolution of 0.125° and hourly. The static input data include the geographical location, soil category, vegetation category, green vegetation fraction of each grid. Here, we use the United States Geological Survey (USGS) 30-s global vegetation type and the State Soil Geographic Database (STATSGO) soil texture datasets to derive the vegetation and soil indices. Both the 30-s datasets are aggregated to 0.125° with the dominant soil and vegetation types to match the spatial resolution of the NLDAS-2 forcing. In addition, the National Environmental Satellite, Data, and Information Service (NESDIS) 0.144° monthly 5-year climatological green vegetation fraction data are used and remapped onto the model grids.

To evaluate the model simulated GPP and LH, we use the monthly GPP and LH data from the FLUXNET model tree ensemble (MTE) products produced by Max Planck Institute for Biogeochemistry as "observations." These monthly, 0.5° × 0.5° gridded datasets are available over the global continents for the period 1982–2011 (Jung et al., 2010, 2011), and have been widely adopted to evaluate LSMs-simulated (Anav et al., 2015; Gan et al., 2019; Ma et al., 2017; Xia et al., 2016) land-atmosphere water, carbon, and energy exchanges. The FLUXNET MTE products over the continental United States can be considered of high quality because most of the FLUXNET sites over the "data-rich" CONUS were incorporated (Jung et al., 2010). In this study, we use the nearest-neighbor interpolation to resample these evaluation data sets to 0.125° grids.

Considering the computational costs and the availability of the evaluation data, we select six representative years from 1982–2011 to run the Noah-MP model at a time step of 1 hr. The six selected years are the same as in Huo et al. (2019), which include two successive wet years (1982 and 1983), two successive dry years (1987 and 1988), and two successive medium years (2002 and 2003) according to annual precipitation and annual humidity index. The data of the first year in each 2-year period is used for "spin-up" and that of the rest is used for surrogate modeling and optimization. We aggregate the hourly outputs to monthly to match the temporal scale of the evaluation data.

In both the MO-ASMOGS and the MO-ASMO methods, the number of generations and population size are set to 100 for the embedded NSGA-II. We found that the nondominated solutions could be located in the nondominated region of the reference point more efficiently when the initial parameter sample size was 200 times of the number of parameters. Thus, the initial sample sizes are 800 and 600 for the optimization in the grassland and DBF systems, respectively. The total iteration number is set to 5, and the resampling percentage is set to 20%. Thus, the maximum numbers of Noah-MP evaluations are 900 and 700 for grassland and DBF, respectively. For the multi-objective optimizations conducted on the representative grid cells, the obtained approximate Pareto optimal parameter sets are then used to run the Noah-MP model on all grid cells for the specified PFT, thereby enabling further comprehensive verification.

In this study, we run the Noah-MP model in parallel with 24 processors (Intel Xeon 2.5 GHz CPU) and 128 GB memory. The computational time of one model run over the entire grassland region and the DBF region is 960s and 720s, respectively. Therefore, it will take about 10.2 days and 6.1 days to perform the calibration of all grid cells for the grassland and DBF regions, respectively. If running the model on only 10% of the total grid cells, the computational time can be reduced to 260s and 220s for the grassland and DBF regions respectively. Furthermore, the total computational time of the calibration using the MO-ASMOGS method is about 5.4 and 3.2 days for the grassland and DBF regions respectively.

4. Results

Through looking into the multi-objective optimization procedure (figures not shown), we found that the 20 approximate Pareto optimal points obtained in each iteration did not evolve consistently toward the right direction as the loop progresses. For example, the optimal pointes of loop 4 were worse than those of loop 2 points. Nonetheless, the overall evolution procedure was still steered toward the right direction to the nondominated region of the default point. After the iteration limit is reached, the final nondominated solutions can be determined. For those optimizations conducted using the MO-ASMOGS method with representative spatial grid cells, the Noah-MP model is then run in an ensemble mode with the obtained optimal parameter sets on the entire grid cells covered by the focused PFT. Finally, the simulated results and RMSE values corresponding to these optimal parameter sets are averaged on each grid cell.

Figure 3 shows the RMSE CDF curves of model simulations corresponding to the default parameterization scheme, the MO-ASMO case with all grid cells and the MO-ASMOGS cases of different grid sampling levels for the grassland and DBF, respectively. The top panel of the figure represents the results of LH, while the bottom panel represents the results of GPP. For the MO-ASMOGS cases, their RMSE CDF curves are obtained by running the Noah-MP model on all grid cells of each PFT covered region with the parameters optimized on representative grid cells. It can be seen that for both PFTs and both variables, the curves of the all-grid cases and the representative-grid cases lie above the curves of the default case, except the left tails of the 5% curves for grassland. The results indicate that the simulated monthly GPP and LH over the regions of interest are simultaneously improved by optimization, and the improvement is more significant in the DBF region. This is not surprising for the all-grid cases as the area to the left CDF curve is directly used as the objective function value to be minimized. Interestingly, optimization using the adaptive mode of MO-ASMOGS at the initial grid sampling level can lead to slightly better model performance than that obtained with the default parameter set.

Moreover, the CDF curves for the 10%, 20%, and 40% cases are almost identical to those of the all-grid cases for both variables and both PFT regions. In the grassland region, the CDF curves of the 5% cases lie under those of the all-grid cases. Even though the optimization results based on the initial sample of grid cells (5%) are not satisfactory, one additional sampling process (10%) can significantly improve the optimization results, which are comparable to those obtained using the parameters optimized by the MO-ASMO method in a traditional way (on all grid cells). In the DBF region, the CDF curve of the 5% case is slightly lower than that of the all-grid case for LH, while the reverse happens for GPP. These indicate that the optimization results based on the initial sample of grid cells are already good enough. Next, we will thoroughly compare

10.1029/2020WR028740

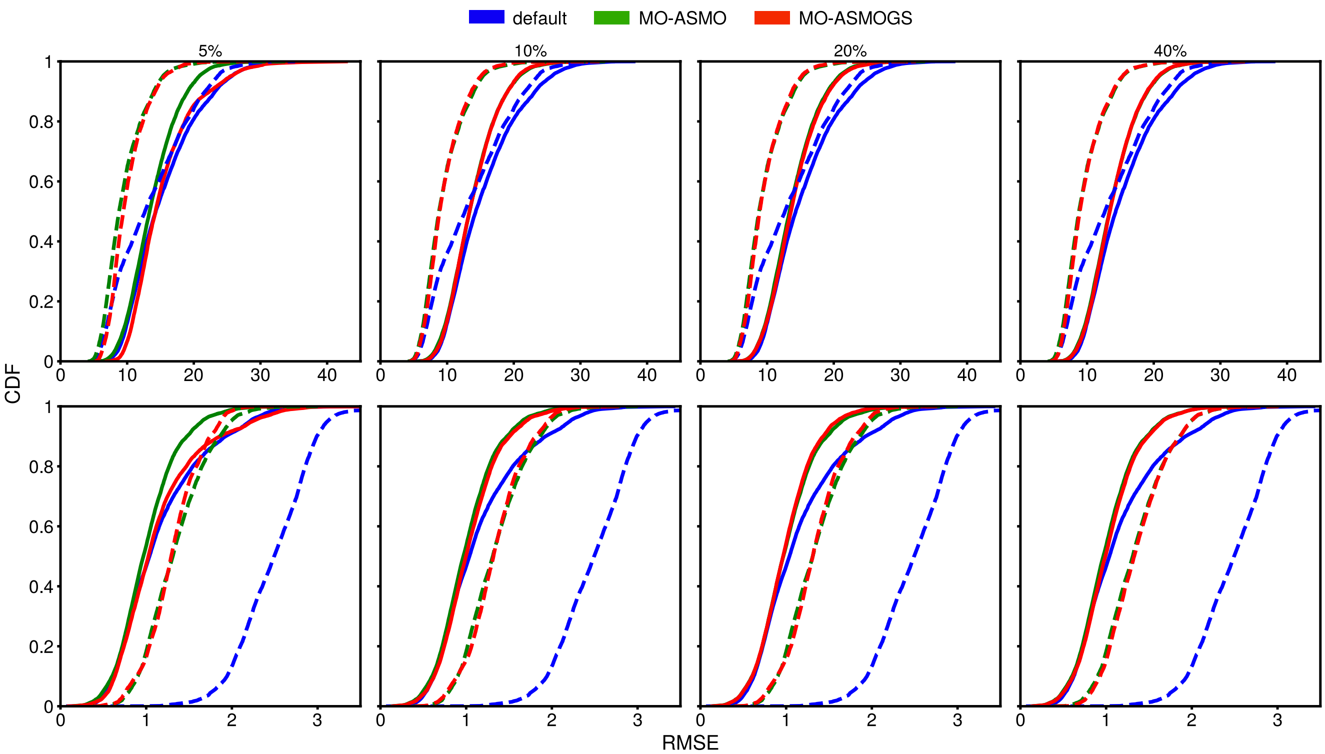

Figure 3. The cumulative distribution function (CDF) curves of root mean squared error (RMSE) values of simulated latent heat flux (LH) (top panel) and gross primary production (GPP) (bottom panel) over the grassland region (solid lines) and the deciduous broadleaf forest (DBF) region (dashed lines) using the default parameters (blue), the parameters optimized by the MO-ASMO method with all grid cells (green) and the parameters optimized by the MO-ASMOGS method with representative grid cells selected at different grid sampling levels (red).

the simulated LH and GPP by use of the default parameter set and parameters optimized in different calibration cases.

Figures 4 and 5 show the spatial distribution of the RMSE of simulated GPP and LH in both grassland and DBF regions across the CONUS. The subpanels show the RMSE on each grid cell using the default parameters and optimized parameters obtained by calibrating the Noah-MP model on all grid cells, and on representative grid cells selected at four grid sampling levels (5%, 10%, 20%, and 40%). The simulated GPP using the default parameters shows large errors in the DBF region while the large errors of simulated LH mainly concentrate in the grassland region. For both variables, there are noticeable reductions in RMSE values of all-grid cases and cases of different grid sampling levels. In Figure 4, there is some obvious deterioration in RMSE of the 5% case in the central America covered by the grassland, while the errors of GPP of the 10%, 20%, 40% cases are greatly reduced in most parts of the study region. In Figure 5, it is also the case that the simulated LH using the optimized parameter values are significantly improved compared to the default case across much of the region, despite a similar deterioration of the 5% case in the Nebraska and Kansas states compared to other grid sampling level cases.

To further quantitatively investigate the changes in model performance using the optimized parameters compared to that using the default parameters, Figures 6 and 7 present the distributions of the relative RMSE change Δ (Equation 2) for GPP and LH, respectively. The subpanels present the relative RMSE changes when using the optimized parameters obtained by calibrating the Noah-MP model on all grid cells, and on representative grid cells selected at four grid sampling levels. Figure 6 shows that using the optimized parameter values significantly reduces the RMSE of simulated GPP for nearly the entire DBF region. However, the RMSE values increase by over 60% in the north part of the grassland region for the all-grid, 10%, 20%, and 40% cases. The large relative changes of RMSE are partly due to the small RMSE values of the default case. Relative changes in small values can appear to be more significant than they are. This is because a small absolute change in the value can result in a large percentage change. As for the 5% case of

Table 2
The Detailed Information of the Parameters to be Calibrated

Parameters	Meaning	Unit	Range	Default values (*G*/*D*)
smcmax	porosity, saturated soil moisture	--	[0.83, 2.0]	1
bexp	Clapp and Hornberger "b" parameter	--	[0.1, 2.5]	1
vcmx25	maximum rate of carboxylation at 25°C	umol $CO_2/(m^2.s)$	[0, 200]	40.0/60.0
rmf25	leaf maintenance respiration at 25°C	umol $CO_2/(m^2.s)$	[0, 20]	1.8/3.0

grassland, the deterioration in model performance is observed on more grid cells but with milder degree compared to other optimization cases. For LH shown in Figure 7, the spatial patterns of relative RMSE changes for the all-grid, 10%, 20%, and 40% cases are very similar, with reduced RMSE on most grid cells but increased RMSE in the northeastern DBF region and the northern grassland region. In comparison, for the 5% case, there are more grid cells with increased model errors in the grassland region, and the deterioration in model performance is less sever in the northeastern DBF region.

Figures 8 and 9 show the histograms of the relative RMSE change Δ estimated over the entire region for GPP and LH, respectively. In addition, Table 3 shows the percentages of grid cells on which the RMSE is reduced after calibration in both PFT covered regions. It is clear that the distributions of Δ and the number

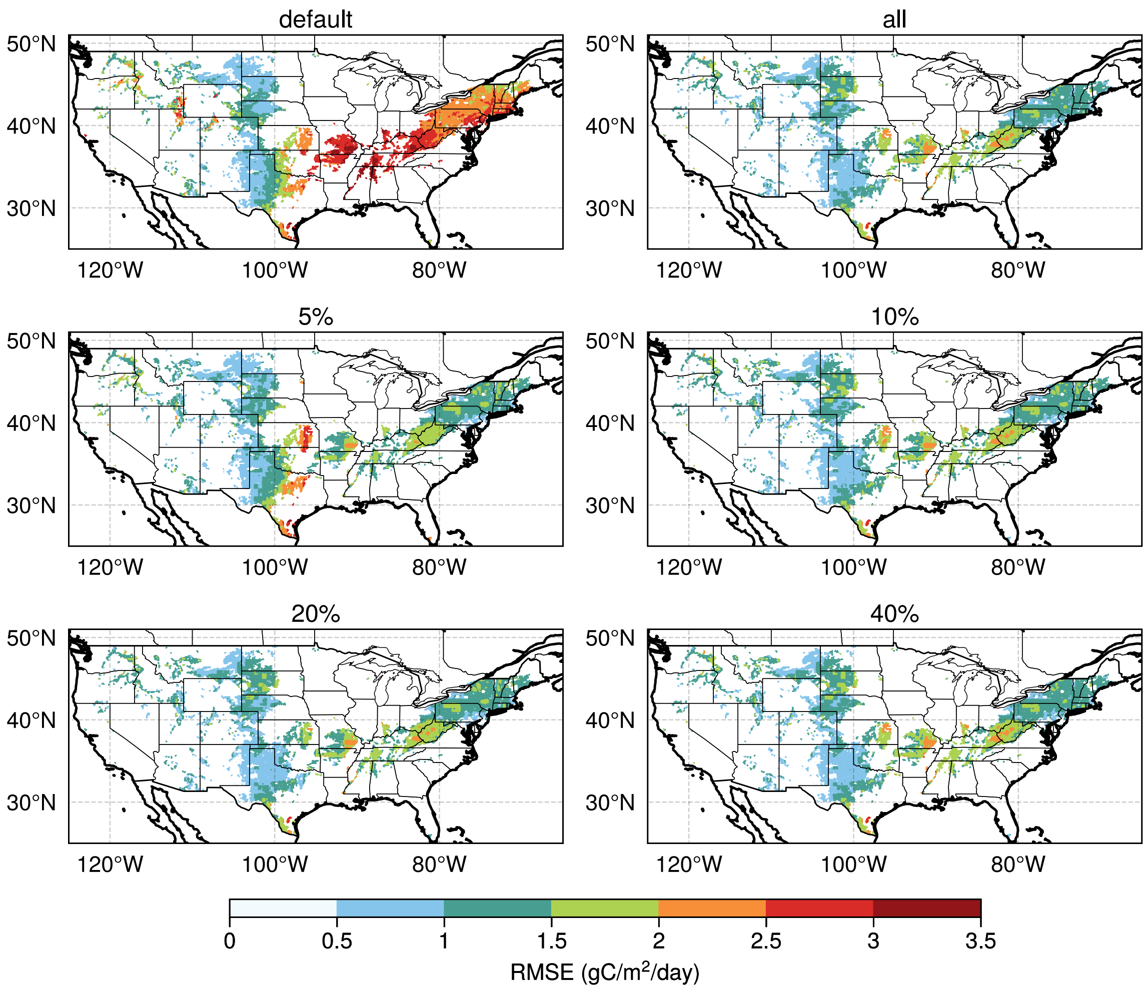

Figure 4. The spatial distribution of root mean squared error (RMSE) values of simulated gross primary production (GPP) using the default parameters, the parameters optimized by the MO-ASMO method with all grid cells and the parameters optimized by the MO-ASMOGS method with representative grid cells selected at different grid sampling levels.

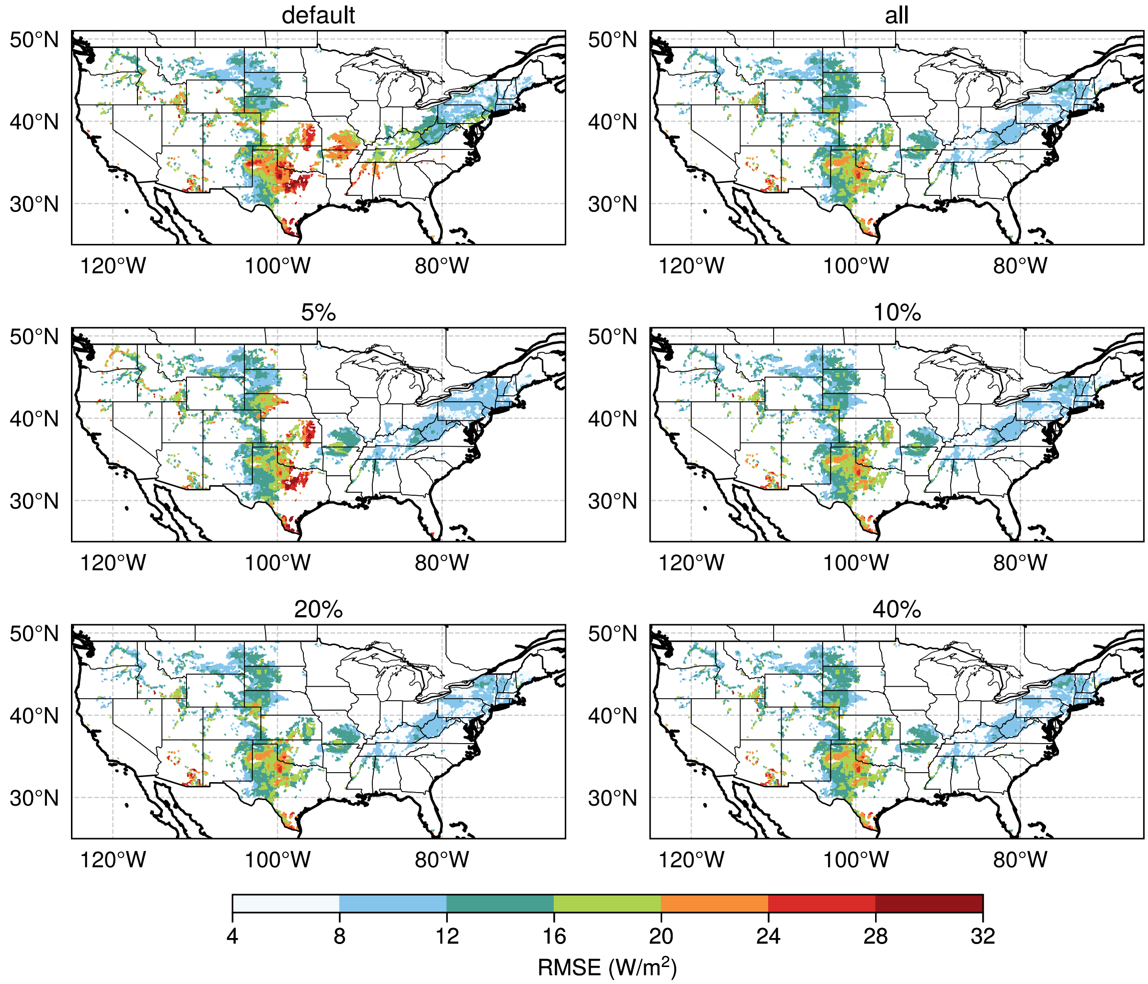

Figure 5. Same as in Figure 4, but for latent heat flux (LH).

of RMSE-reduced grid cells for the all-grid, 10%, 20% and 40% cases are very similar. Overall, using the optimized parameters of these calibration cases reduces the monthly model errors of GPP on about 75% of the total points, while the number is about 70% for LH. In comparison, the effect of the optimization process for the 5% case is much worse in the grassland region. The simulated GPP results on more than half of the total grid cells become worse after calibration, indicating that using the MO-ASMOGS method at the initial grid sampling level is not enough to improve the overall model performance. The improvement of model performance is much more significant for the simulations in the DBF region, especially for the GPP simulation. This, from a side, indicates that the assigned default values of vcmx25 and rmf25 for the DBF are inappropriate.

Furthermore, the spatial distributions of the RMSE differences between the MO-ASMO calibration case and each of the four MO-ASMOGS calibration cases are also presented in Figures S2 and S3. When compared to the all-grid case, the absolute changes in RMSE values are very small for the 10%, 20% and 40% cases. The model performance of the 20% case slightly outperforms that of the all-grid case in terms of GPP simulation, while the model performance of the 10% case slightly outperforms that of the all-grid case in terms of LH simulation. In contrast, there are clear disparities of model errors between the 5% case and the all-grid case.

Figure 10 shows the approximate Pareto optimal parameters obtained in each calibration case along with the default parameters, which provides further proof of the former results. All of the parameter values are normalized to (0, 1). It is evident that the optimal parameters obtained in the all-grid, 10%, 20% and 40% cases are close to each other, but are remarkably different from the default ones. In addition, all the optimal

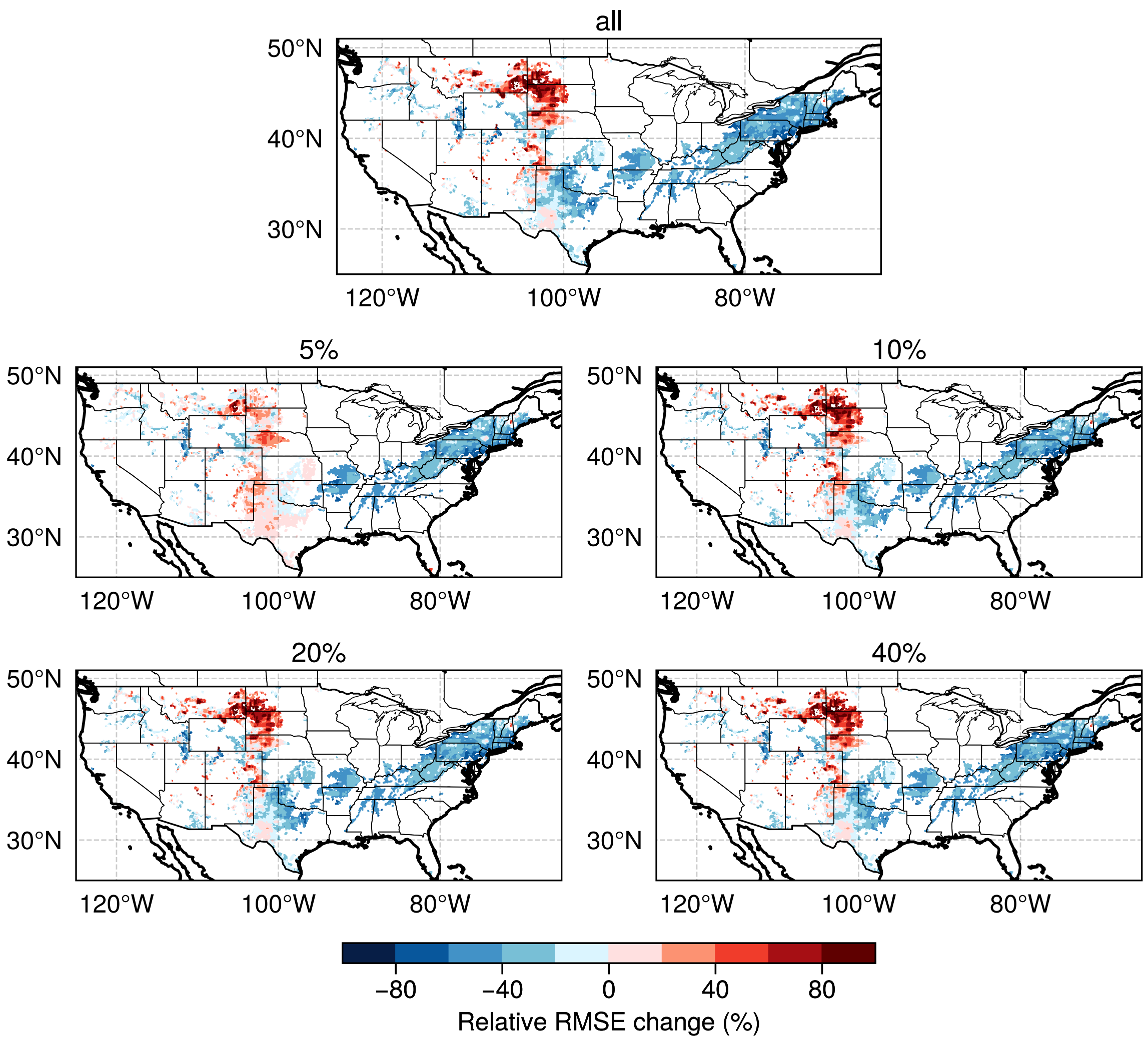

Figure 6. The spatial distribution of relative root mean squared error (RMSE) change Δ of simulated gross primary production (GPP) when using the parameters optimized by the MO-ASMO method with all grid cells and the parameters optimized by the MO-ASMOGS method with representative grid cells selected at different grid sampling levels.

parameters obtained in the 5% case are different from the others for the grassland, while the optimized vcmx25 and rmf25 are similar to the ones obtained in other calibration cases for the DBF. Using the two modes of MO-ASMOGS leads to nearly the same calibration results, indicating the robustness of our proposed spatial grid sampling strategy. Therefore, the users can use the method in a flexible way to better meet their requirements.

Finally, we give some insight into the spatial distribution of the representative spatial grid cells selected by the MO-ASMOGS method (Figure S4). Because the parameters needed to be calibrated are all related to vegetation and soil, we also present the bar charts to show the numbers of representative grid cells clustered in each soil type over each PFT region. It is shown that at the initial spatial grid sampling level (5%), the representative grid cells are distributed across each PFT region, indicating that even the most strongly influential grid cells exhibit spatial heterogeneity. Further investigation shows that the distribution of the number of representative grid cells in each soil type is very similar to that of the total number of grid cells in each soil type. This means that there are strongly influential grid cells in each soil type dominated region, and one soil type dominated region with more grid cells is more likely to contain strongly influential grid cells. The spatial heterogeneity of the importance of grid cells with the same soil type is mainly due to the subgrid variability of the soil attributes (e.g., soil texture, stone fragment distribution). In addition, the climate condition such as precipitation could also impact the sensitivity of a parameter on different grid cells

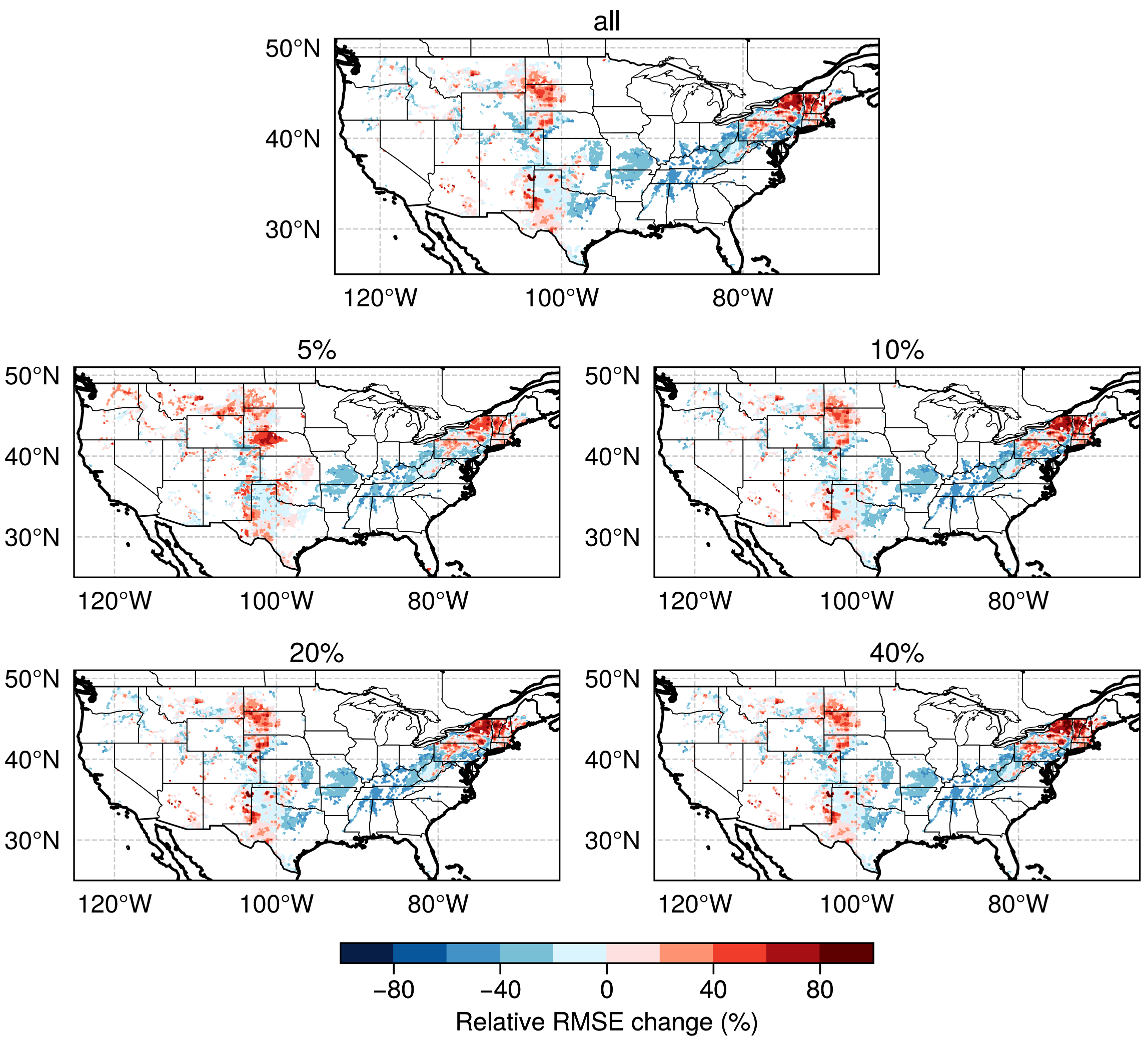

Figure 7. Same as in Figure 6, but for latent heat flux (LH).

with the same soil type. Therefore, there is no such a clear spatial pattern of important grid cells that we can simply select representative grid cells from the specific regions. Using the spatial sampling strategy in the MO-ASMOGS method, the variability of the importance of each grid cell in the region covered with the same soil type and vegetation type is specifically considered. With the introduction of evolutionary elitism, the grid cells which are more sensitive to reduce the RMSE are selected with higher priority. For example, there are much more grid cells of silt loam than those of sand, however, the representative grid cells selected from sand covered region are more than those selected from silt loam covered region, indicating the sand region is slightly more sensitive than the silt loam region. With the additional sampling, the overall shapes of the frequency distribution remain the same, demonstrating that the spatial diversity of representative grid cells is maintained.

5. Discussion

Different from other multi-objective surrogate modeling-based optimization methods, the MO-ASMOGS method is specifically designed for calibration of distributed EMs, especially when running the model on a large scale (e.g., continental or global scale). Generally, spatially distributed EMs use individual grid cells as primary calculation units and produce distributed outputs. Taking advantage of these features, we developed MO-ASMOGS, which combines the MO-ASMO method with a novel spatial grid sampling strategy. This method only requires model evaluations on a small portion of the total grid cells to improve all the

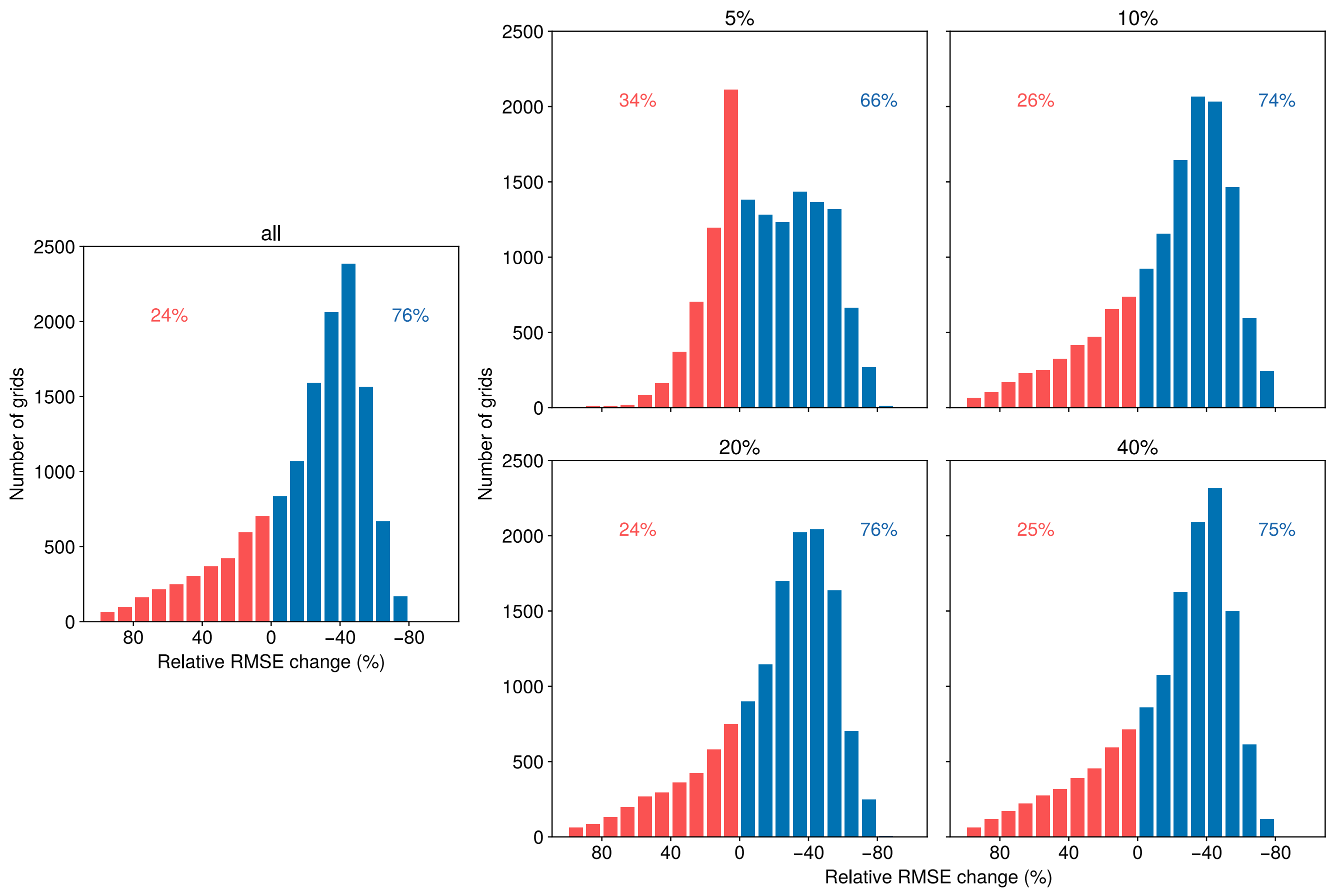

Figure 8. The histograms of relative root mean squared error (RMSE) change Δ of simulated gross primary production (GPP) over the entire region for different calibration cases. Blue means the RMSE of simulated GPP using the default parameters is reduced after calibration, while red means the opposite.

objective functions relative to those obtained with the default parameter values and to produce optimization results as good as those by calibrating all grid cells. To guarantee the calibration efficiency and effectiveness, the proposed spatial grid sampling strategy introduced evolutionary elitism and adaptive grid sampling to make sure the selected grid cells are highly informative by considering the contrasting parameter sensitivities over the spatial domain. Since the method is specifically designed for distributed EMs, the model being distributed is the reason that our method works. Because each grid cell is modeled independently, we can easily do the adaptive sampling without influencing the simulation results of the sampled grid cells. It should be noted that our method may not work well when the simulations between grid cells have strong connections, or the targeted model output is aggregated from model simulations of all grid cells. For example, we cannot calibrate a distributed hydrologic model against the observed streamflow at basin outlet by only running the model on a few grid cells within the basin. In addition, it may be unnecessary to use the MO-ASMOGS method when the model spatial resolution is so low that the total number of grid cells is not large. For example, the grid resolution of the Noah-MP model is 0.125° in this study, so there are 8,362 grid cells for grassland over the CONUS. If the grid resolution is 0.5°, there would be about 520 grid cells. Hence the computational cost is affordable, and it may not have enough room for implementing the adaptive grid sampling. Although there are some limitations for applying the proposed method, we believe it has great potential in calibration of large-scale distributed EMs, as many EM communities have great expectations on moving toward hyper-resolution models (Archfield et al., 2015; Wood et al., 2011).

In the MO-ASMOGS method, the novel spatial grid sampling strategy is the most important part and makes the method different from MO-ASMO and other multi-objective surrogate modeling-based optimization methods. With the ideas of introducing evolutionary elitism and adaptive sampling in the spatial grid sampling strategy, the response surface surrogate of the distributed EM when applied to a large region can be

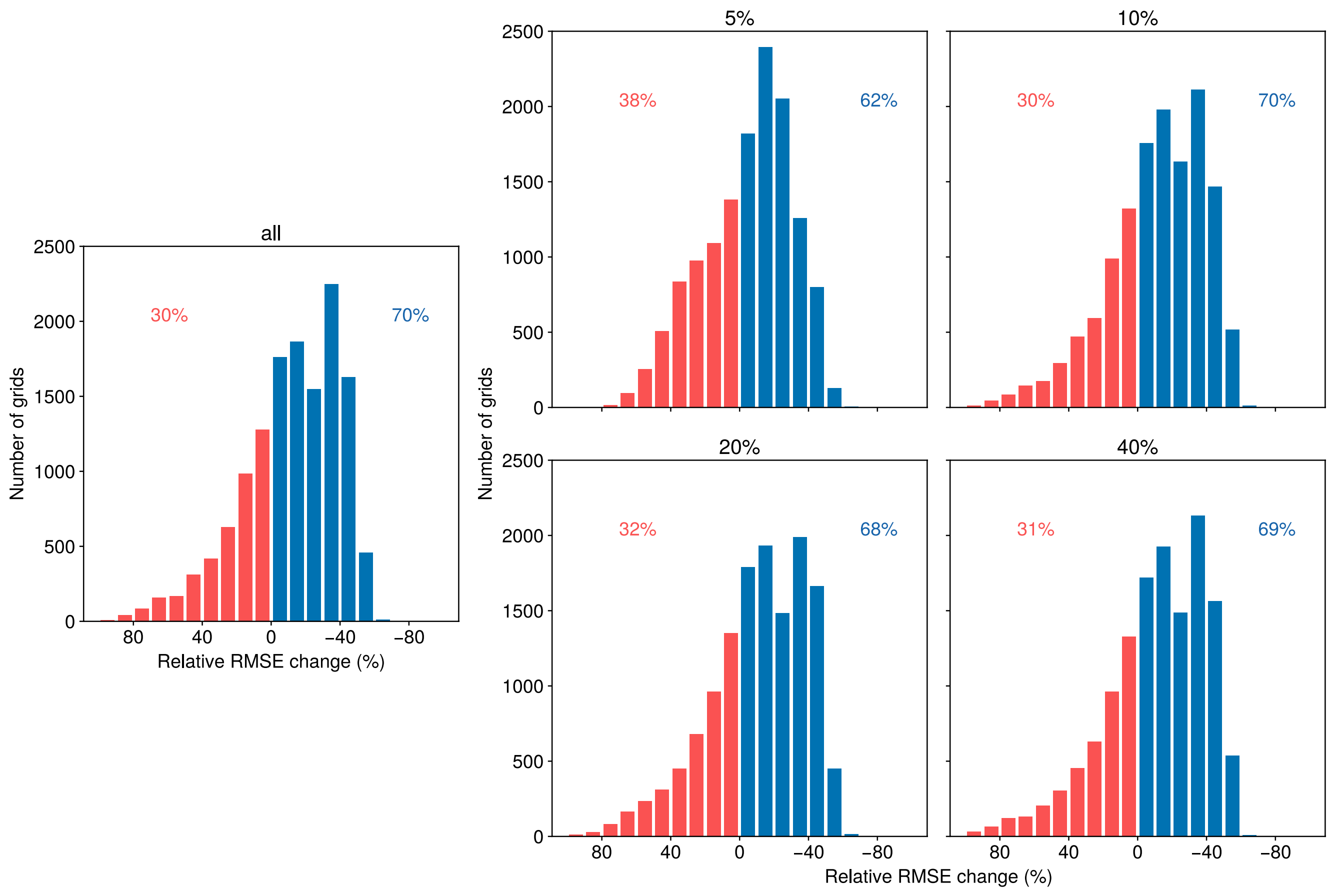

Figure 9. Same as in Figure 8, but for latent heat flux (LH).

constructed and adjusted more efficiently based on both parameter and spatial grid sampling. Therefore, the MO-ASMOGS method can significantly reduce the computation burden of the calibration process and ensure satisfactory optimization effectiveness. Note that the PLHS sampling method is used in our spatial grid sampling strategy to better maintain the spatial diversity during the adaptive or multi-stage sampling process. Ensuring the sample diversity of grid cells is important as it provides better chances to uniformly explore the unevaluated region. However, the PLHS method is only used as an auxiliary measure. In other words, the MO-ASMOGS method is not a simple application of using MO-ASMO on some sampled grid cells based on the PLHS method. To save huge computational time of calibrating the distributed EMs over a large region with many grid cells, our idea is introducing the grid sampling to the response surface surrogate modeling process. However, the time saving should not be at the cost of deteriorating the optimization effectiveness. How to select grid cells is thus critical to construct a reliable surrogate model and obtain satisfactory optimization results. To prove the effect of our spatial grid sampling strategy, we also conducted another experiment for comparison over the grassland region. In this experiment, we simply applied the MO-ASMO method on sampled grid cells with sample size of 5%, 10% and 20% using the PLHS method. This setup is similar to that of the adaptive mode of using the MO-ASMOGS but does not consider the importance of each grid cell and evolutionary elitism for selecting the grid cells. The CDF curves of RMSE values of simulated LH and GPP (Figure S5) show that although the model performance using the optimized parameters obtained by the new calibration is better than that using the default parameters, it cannot compare with the model performance based on calibration of all

Table 3
The Fraction of the Grids on Which the Root Mean Squared Error (RMSE) Values are Reduced After Optimization for Each Target Variable Within Each Plant Function Type (PFT) Region

	GPP		LH	
Grid levels	Grassland	DBF	Grassland	DBF
all	61%	99%	64%	80%
5%	45%	99%	53%	77%
10%	59%	99%	64%	78%
20%	62%	99%	62%	79%
40%	60%	99%	63%	78%

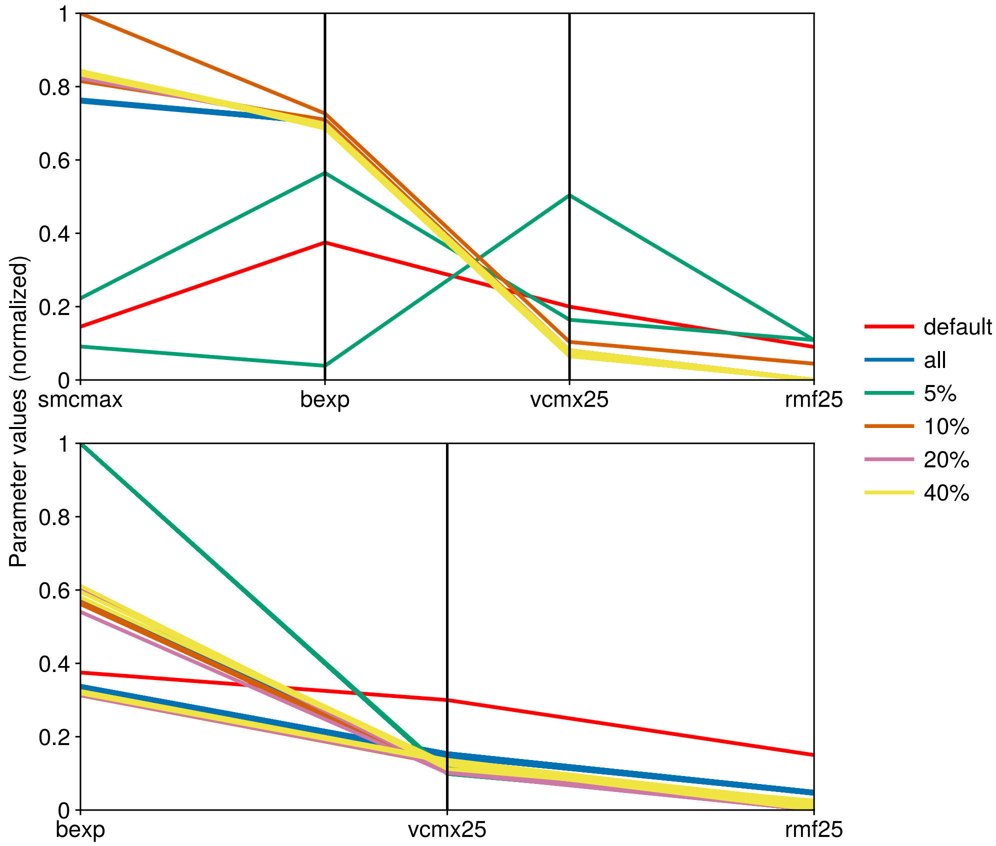

Figure 10. The default parameter values and the approximate Pareto optimal parameter values obtained by the MO-ASMO method with all grid cells and by the MO-ASMOGS method with representative grid cells selected at different grid sampling levels for grassland (top) and deciduous broadleaf forest (DBF) (bottom).

grid cells. Further assessment results (not shown) also demonstrate the optimization effectiveness in the supplementary experiment is inferior to that of the all-grid case. When using the MO-ASMOGS method, even though the optimization results based on the initial sample of grid cells (5%) are not satisfactory, one additional sampling process can significantly improve the optimization results. Therefore, this comparison demonstrates that a direct application of simply combining the MO-ASMO method and the PLHS-based spatial grid sampling does not meet our requirement, thus highlighting the important impact of implementing the novel spatial grid sampling strategy on surrogate modeling and further optimization.

In this study, we tested the proposed method for two PFTs over the CONUS. In both experiments, we started at the initial sampling level of 5% of the total grid cells. For grassland, the optimization results obtained at the initial sampling level are not good. However, the optimization results of the 5% case for DBF are nearly as good as those by calibrating all grids. The optimization results obtained at the initial sampling level are already good enough and further adding more grid cells does not improve the results much, meaning that there may still be room to enhance the computational efficiency by further reducing the initial sample size. Therefore, we conducted another calibration experiment for DBF. In this experiment, we began with a smaller initial sample size of 0.5% of the total grid cells (26 grid cells) and used more iterations of adaptive sampling with added sample size of 0.5% of the total grid cells (26 grid cells) in each iteration. The optimization based on the initial grid samples can produce significantly better model performance than that obtained with the default parameter set, indicating the default values are unreasonable. However, the optimization results based on the initial sample of grid cells are not comparable to those obtained using the parameters optimized by the MO-ASMO method on all grid cells (not shown). This illustrates that using the small initial sample size cannot generate a good approximation of the response surface. Figure S6 shows that the optimization results of the 1% case (after one iteration of adaptive grid sampling) are still not stable, the simulated LH is inferior to that obtained with the parameters optimized using all grid cells (Figure S7) while the performance of simulated GPP is much better. After two iterations of adaptive grid sampling, the method converges and the optimization results by calibrating 1.5% of the total grid cells are nearly the same as those by calibrating all grid cells, or even slightly better. To demonstrate the robustness of the optimization results of this case with small initial sample size, we repeated the experiment five times. The mean and standard deviation of the objective function values over five replicates are shown in Table S1. The mean objective function values of the 1.5%, 2% cases are very close to those of the all-grid case, and the standard deviations of the two cases are also very small. These results indicate the above conclusion about the optimization results of MO-ASMOGS is robust. For the DBF calibration case, the calibrated parameters are fewer than those of the grassland case. In addition, the distribution of soil texture types across the DBF region is less heterogenous than that across the grassland region (Figure S4), suggesting that fewer grid cells are needed to represent the whole region situation for the DBF. Therefore, we can know that our proposed method would give a better result for less complex calibration problems. We can also know that how to properly define the grid sampling size mainly depends on the calibration problem. Many studies (e.g., Wang et al., 2014) have discussed how the initial sample size of parameter points affects the optimization results of surrogate modeling-based optimization algorithms, and mentioned that initial sample size can be neither too small nor too large. Our proposed MO-ASMOGS method uses both parameter sampling and spatial grid sampling to construct and adaptively update the response surface surrogate. Hence, too few initial grid samples could lead to very poor approximation of the real model response over a large region, then the surrogate model is easy to get stuck at local optima and the convergence speed would be slow. In addition, because the computational time is not proportional to the number of grid cells modeled, the computational time of the calibration with a small size of initial grid samples but many adaptive sampling iterations may be even larger than the calibration with a relatively larger size of initial grid samples.

The robustness of the calibration results can be proved by our spatial grid sampling strategy and the experiment design of this study. Since the grid cells are sorted and selected based on the sum of parameter importance rankings on each grid, evolutionary elitism is ensured. At each grid sampling level, the most influential spatial grids could always be selected out. In another word, if each grid cell has a different S_g value, the top n% of the sorted grid cells would be chosen at the n% grid sampling level, and there would be no spatial grid sampling variability. In practice, when there are many grid cells with the same S_g, the grid cells belonging to the same set are further sampled using the PLHS approach to ensure the spatial diversity. By using the spatial grid sampling strategy, the grid cells chosen at lower levels are always contained in the sample set chosen at higher grid sampling levels. In the adaptive mode, we began with an initial spatial grid sampling level corresponding to 5% of all grid cells, then added additional grid cells to double the sample size. This process was repeated two times to get spatial grid sampling levels of 10% and 20%. By doing this, we can test the robustness of calibration results obtained at the 10% spatial grid sampling level. For example, there are 8,362 grids in the grassland region, the 10% spatial grid sampling level contains all grid cells belonging to the first three sets ($F_1 \sim F_3$) and another 40% of grid cells belonging to the fourth set (F_4). The 20% spatial grid sampling level contains all grid cells belonging to the first four sets ($F_1 \sim F_4$). It is already known that the calibration results barely changed after 60% of grid cells belonging to F_4 were added, indicating the calibration results are barely affected by the grid sampling variability.

This study has shown the superiority of the optimized parameter values over the default parameter values in improving the simulation of GPP and LH simultaneously. However, the improvement of model performance is not perfect. As shown in the results, use of the approximate Pareto optimal parameter sets did not reduce the model errors on all grid cells within a PFT region consistently, instead it even increased the model errors on some grid points, especially for the grassland. One possible reason is that the potential heterogeneity of the parameters within a PFT was ignored (Prihodko et al., 2008). It has been found that values of optimized model parameters can vary among different sites covered by the same PFT (Raoult et al., 2016; Rosolem et al., 2013; Wang et al., 2007). For example, there could be various soil types within one PFT region, and a uniform assignment of the multiplier to all the soil types which was adopted in this study may be inappropriate. The spatially constant multiplier smcmax could also make the porosity of some soil types get unrealistic in the grassland region, although this would not affect the effectiveness of our proposed method. Therefore, our future work includes using the pedo-transfer function and optimizing the transfer function coefficients to alter the spatial patterns of soil parameterization (Demirel et al., 2018; Mizukami et al., 2017; Samaniego et al., 2017). Another possible reason is the intrinsic model structural errors. Similarly, McNeall et al. (2016) found that the structural error in the climate model (FAMOUS) was responsible for the inconsistence between the regions of parameter space where FAMOUS best simulated the forest fraction in the Amazon forests and other forests. It was pointed out that the structural error was possibly caused by a missing process in the vegetation model. We will look into the Noah-MP model to quantify the impact of model structural error.

In our case study, we optimized 4 and 3 strongly influential parameters of the Noah-MP model for simulating GPP and LH over two PFT regions across CONUS. The case study is a follow-up study of Huo et al. (2019), in which we have performed the parameter sensitivity analysis of the Noah-MP model for simulating GPP and LH. In addition, due to the lack of gridded ground observations with high quality, we did not test the MO-ASMOGS method for other model outputs. The FLUXNET MTE products used in this study also limit the calibration to monthly scale. However, we believe that application of the MO-ASMOGS method is not limited by the model temporal resolution and model outputs selected as the calibration subject. Instead, the proposed method is flexible and is applicable to other types of distributed EMs particularly for large-scale applications. In the future, we will explore multiple avenues to improve our approach. First, we will apply the proposed calibration method to other large complex EMs with different model outputs. We will also try to use various satellite remote sensing data with higher spatiotemporal resolution to constrain model parameters based on our method. Second, as many data are becoming available and accessible, they can provide more prior information for model parameter estimation. For example, some studies (Beck et al., 2020; Rakovec et al., 2016) have used transfer functions to link model parameters to landscape and even climate characteristics. As described in Section 2.2, the land cover, soil texture and climatic characteristics have been considered when designing the spatial grid sampling strategy, however, some information is only used implicitly. We will explore how to fully use the prior information of soil texture, land cover and climate maps to improve our method. Third, we will test the scalability of the spatial grid sampling strategy, surrogate modeling and multi-objective optimization methods on other complex models with more adjustable parameters. For example, when the number of parameters to be optimized is large, the representative grid cell screening based on the sum of parameter importance rankings with regard to each target variable may be less effective. The Pareto rank-based multiple-criteria implementation of the sensitivity analysis approach proposed by Rosolem et al. (2012) could be adopted as an alternative way to identify the most influential parameters and to further choose representative grids. Fourth, we also anticipate that the method we develop could be adapted for including other advanced methods like the formal Bayesian approach (Huang et al., 2016; Sun et al., 2017) and the dimensionality reduction techniques in machine learning (Xi, Li, Wang, 2020; Xi, Yuan, et al., 2020), which can facilitate the data-model integration for complex EMs.

6. Conclusions

In this study, we proposed a novel multi-objective calibration method called MO-ASMOGS, which is aimed to significantly reduce the computational costs associated with calibration of spatially distributed EMs, especially for large-scale applications. The MO-ASMOGS method introduces a novel spatial grid sampling strategy and combines it with an advanced surrogate modeling-based multi-objective optimization approach MO-ASMO. With the newly introduced spatial grid sampling strategy, which makes the novel contribution of this study, the MO-ASMOGS differs from its predecessor and other surrogate modeling-based multi-objective optimization methods in the following aspects. (a) MO-ASMOGS is specifically designed for distributed EMs (e.g., LSMs) while other methods theoretically can be applied to any multi-objective optimization problems. (b) The surrogate model in MO-ASMOGS is constructed and updated based not only on parameter sampling as in MO-ASMO and other methods, but also on spatial grid sampling. (c) To reduce the computational time as much as possible, the initial sample set of spatial grid cells is not generated uniformly like the initial parameter sampling does in MO-ASMO and other methods. Instead, evolutionary elitism is adopted in our spatial grid sampling strategy to ensure that more important grid cells are selected with higher priorities. If the user is still not satisfied with the optimization results, this strategy is flexible to allow the user to add additional spatial grid samples progressively in an adaptive manner. Therefore, through both running the model on representative grid cells and reducing the number of model runs, a considerable amount of computational cost can be saved.

We applied the new calibration method to the Noah-MP model, which is the follow-up study of Huo et al. (2019), to explore the possibility of being able to improve the model performance of simulating two key variables of water and carbon cycles. To test the MO-ASMOGS method, we used both the adaptive mode (5%, 10%, 20%) and the one-shot mode (40%) to optimize the most important parameters to GPP and LH for two PFTs (grassland and DBF) across the CONUS. The computational time of these two modes for calibrating the model is near the same. The calibration results at different spatial grid sampling levels were compared with those obtained by calibration of all grid cells. As demonstrated by the results, parameter optimization can significantly improve the simulated GPP and LH compared to those simulated using the default parameter set. Use of the optimized parameters reduced the model errors on about 75% of the total grid cells for the simulated GPP and 70% of the total grid cells for the simulated LH. This indicates the inappropriate assignment of the default parameter values. The 10%, 20%, and 40% cases produced similar optimization results to those of the all-grid case. In addition, even the 5% case produced results were comparable to those of the all-grid case for the DBF system. An additional experiment was conducted to illustrate the benefits of using the spatial sampling strategy. For the DBF case, we began with a smaller initial sample size at 0.5% of the total grid cells with increment of 0.5% of the total grid cells in adaptive sampling iterations. The results showed that MO-ASMOGS used as little as 1.5% of the total grids to achieve comparable results with MO-ASMO using 100% of the grids, demonstrating that the method would give a better result for less complex calibration problems. Our proposed calibration method showed similar effectiveness but was much more efficient than the traditional calibration strategy based on evaluating cell-to-cell performance of the distributed EMs on all grid cells. Therefore, the computational burden associated with the calibration of large complex distributed EMs can be significantly reduced. The proposed MO-ASMOGS method can be applied to various distributed EMs with different model outputs.

Data Availability Statement

The NLDAS-2 forcing data set is available from the website https://disc.gsfc.nasa.gov/datasets?keywords=NLDAS, and the static geography data are publicly available from the Noah-MP official website https://ral.ucar.edu/solutions/products/noah-multiparameterization-land-surface-model-noah-mp-lsm. The monthly GPP and LH data from the FLUXNET model tree ensemble (MTE) products are available at https://www.bgc-jena.mpg.de/geodb/projects/Home.php. All the data generated in this work for the figures and tables are available through Mendeley Data (DOI: 10.17632/r7xhzvtr7y.1).

Acknowledgments
This study was jointly supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA2006040104), the National Natural Science Foundation of China (42101046, 51979004), the China Postdoctoral Science Foundation (2019M661714), and the Fundamental Research Funds for the Central Universities (B200202031, 2019B04614). The numerical calculations in this paper have been done on the computing facilities in the High Performance Computing Center (HPCC) of Nanjing University. The authors would like to thank Dr. Wei Gong at Beijing Normal University for providing the source code of MO-ASMO.

References

Anav, A., Friedlingstein, P., Beer, C., Ciais, P., Harper, A., Jones, C., et al. (2015). Spatiotemporal patterns of terrestrial gross primary production: A review. *Reviews of Geophysics*, *53*(3), 785–818. https://doi.org/10.1002/2015RG000483

Archfield, S. A., Clark, M., Arheimer, B., Hay, L. E., McMillan, H., Kiang, J. E., et al. (2015). Accelerating advances in continental domain hydrologic modeling. *Water Resources Research*, *51*(12), 10078–10091. https://doi.org/10.1002/2015WR017498

Bastidas, L. A., Gupta, H. V., Hsu, K. L., & Sorooshian, S. (2003). Parameter, structure, and model performance evaluation for land-surface schemes. In Q. Duan, H. V. Gupta, S. Sorooshian, A. N. Rousseau, & R. Turcotte (Eds.), *Calibration of watershed models, water science and application* (Vol. *6*, pp. 229–237): American Geophysical Union.

Bastidas, L. A., Gupta, H. V., Sorooshian, S., Shuttleworth, W. J., & Yang, Z. L. (1999). Sensitivity analysis of a land surface scheme using multicriteria methods. *Journal of Geophysical Research: Atmospheres*, *104*(D16), 19481–19490. https://doi.org/10.1029/1999JD900155

Beck, H. E., Pan, M., Lin, P., Seibert, J., van Dijk, A. I., & Wood, E. F. (2020). Global fully distributed parameter regionalization based on observed streamflow from 4,229 headwater catchments. *Journal of Geophysical Research: Atmospheres*, *125*(17), e2019JD031485. https://doi.org/10.1029/2019JD031485

Chen, F., & Dudhia, J. (2001). Coupling an advanced land surface–hydrology model with the Penn State–NCAR MM5 modeling system. Part I: Model implementation and sensitivity. *Monthly Weather Review*, *129*(4), 569–585. https://doi.org/10.1175/1520-0493

Chen, F., Mitchell, K., Schaake, J., Xue, Y., Pan, H. L., Koren, V., et al. (1996). Modeling of land surface evaporation by four schemes and comparison with FIFE observations. *Journal of Geophysical Research: Atmospheres*, *101*(D3), 7251–7268. https://doi.org/10.1175/1520-0493(2001)129<0569:caalsh>2.0.co;2

Dagon, K., Sanderson, B. M., Fisher, R. A., & Lawrence, D. M. (2020). A machine learning approach to emulation and biophysical parameter estimation with the Community Land Model, version 5. *Advances in Statistical Climatology, Meteorology and Oceanography*, *6*(2), 223–244. https://doi.org/10.5194/ascmo-6-223-2020

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, *6*(2), 182–197. https://doi.org/10.1109/4235.996017

Demirel, M. C., Mai, J., Mendiguren, G., Koch, J., Samaniego, L., & Stisen, S. (2018). Combining satellite data and appropriate objective functions for improved spatial pattern performance of a distributed hydrologic model. *Hydrology and Earth System Sciences*, *22*(2), 1299–1315. https://doi.org/10.5194/hess-22-1299-2018

Duan, Q., Di, Z., Quan, J., Wang, C., Gong, W., Gan, Y., et al. (2017). Automatic model calibration: A new way to improve numerical weather forecasting. *Bulletin of the American Meteorological Society*, *98*(5), 959–970. https://doi.org/10.1175/BAMS-D-15-00104.1

Duan, Q., Sorooshian, S., & Gupta, V. (1992). Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resources Research*, *28*(4), 1015–1031. https://doi.org/10.1029/91WR02985

Ek, M. B., Mitchell, K. E., Lin, Y., Rogers, E., Grunmann, P., Koren, V., et al. (2003). Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *Journal of Geophysical Research: Atmospheres*, *108*(D22). https://doi.org/10.1029/2002JD003296

Fer, I., Kelly, R., Moorcroft, P. R., Richardson, A. D., Cowdery, E. M., & Dietze, M. C. (2018). Linking big models to big data: Efficient ecosystem model calibration through Bayesian model emulation. *Biogeosciences*, *15*(19), 5801–5830. https://doi.org/10.5194/bg-15-5801-2018

Forrester, A. I., & Keane, A. J. (2009). Recent advances in surrogate-based optimization. *Progress in Aerospace Sciences*, *45*(1–3), 50–79. https://doi.org/10.1016/j.paerosci.2008.11.001

Fox, A., Williams, M., Richardson, A. D., Cameron, D., Gove, J. H., Quaife, T., et al. (2009). The REFLEX project: Comparing different algorithms and implementations for the inversion of a terrestrial ecosystem model against eddy covariance data. *Agricultural and Forest Meteorology*, *149*(10), 1597–1615. https://doi.org/10.1016/j.agrformet.2009.05.002

Gan, Y., Liang, X. Z., Duan, Q., Chen, F., Li, J., & Zhang, Y. (2019). Assessment and reduction of the physical parameterization uncertainty for Noah-MP Land Surface Model. *Water Resources Research*, *55*(7), 5518–5538. https://doi.org/10.1029/2019WR024814

Gong, W., Duan, Q., Li, J., Wang, C., Di, Z., Dai, Y., et al. (2015). Multi-objective parameter optimization of common land model using adaptive surrogate modeling. *Hydrology and Earth System Sciences*, *19*(5), 2409–2425. https://doi.org/10.5194/hess-19-2409-2015

Gong, W., Duan, Q., Li, J., Wang, C., Di, Z., Ye, A., et al. (2016). Multiobjective adaptive surrogate modeling-based optimization for parameter estimation of large, complex geophysical models. *Water Resources Research*, *52*(3), 1984–2008. https://doi.org/10.5194/hess-19-2409-2015

Gorissen, D. (2010). *Grid-enabled adaptive surrogate modeling for computer aided engineering, PhD thesis*. Ghent University.

Gu, L., Pallardy, S. G., Yang, B., Hosman, K. P., Mao, J., Ricciuto, D., et al. (2016). Testing a land model in ecosystem functional space via a comparison of observed and modeled ecosystem flux responses to precipitation regimes and associated stresses in a Central US forest. *Journal of Geophysical Research: Biogeosciences*, *121*(7), 1884–1902. https://doi.org/10.1002/2015JG003302

Gupta, H. V., Bastidas, L. A., Sorooshian, S., Shuttleworth, W. J., & Yang, Z. L. (1999). Parameter estimation of a land surface scheme using multicriteria methods. *Journal of Geophysical Research: Atmospheres*, *104*(D16), 19491–19503. https://doi.org/10.1029/1999JD900154

Gupta, H. V., Sorooshian, S., & Yapo, P. O. (1998). Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resources Research*, *34*(4), 751–763. https://doi.org/10.1029/97WR03495

Huang, M., Hou, Z., Leung, L. R., Ke, Y., Liu, Y., Fang, Z., & Sun, Y. (2013). Uncertainty analysis of runoff simulations and parameter identifiability in the Community Land Model: Evidence from MOPEX basins. *Journal of Hydrometeorology*, *14*(6), 1754–1772. https://doi.org/10.1175/JHM-D-12-0138.1

Huang, M., Ray, J., Hou, Z., Ren, H., Liu, Y., & Swiler, L. (2016). On the applicability of surrogate-based Markov chain Monte Carlo-Bayesian inversion to the Community Land Model: Case studies at flux tower sites. *Journal of Geophysical Research: Atmospheres*, *121*(13), 7548–7563. https://doi.org/10.1002/2015JD024339

Huo, X., Gupta, H., Niu, G. Y., Gong, W., & Duan, Q. (2019). Parameter sensitivity analysis for computationally intensive spatially distributed dynamical environmental systems models. *Journal of Advances in Modeling Earth Systems*, *11*(9), 2896–2909. https://doi.org/10.1029/2018MS001573

Jung, M., Reichstein, M., Ciais, P., Seneviratne, S. I., Sheffield, J., Goulden, M. L., et al. (2010). Recent decline in the global land evapotranspiration trend due to limited moisture supply. *Nature*, *467*(7318), 951–954. https://doi.org/10.1038/nature09396

Jung, M., Reichstein, M., Margolis, H. A., Cescatti, A., Richardson, A. D., Arain, M. A., et al. (2011). Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. *Journal of Geophysical Research: Biogeosciences*, *116*(G3). https://doi.org/10.1029/2010JG001566

Koch, J., Demirel, M. C., & Stisen, S. (2018). The SPAtial EFficiency metric (SPAEF): Multiple-component evaluation of spatial patterns for optimization of hydrological models. *Geoscientific Model Development*, *11*(5), 1873–1886. https://doi.org/10.5194/gmd-11-1873-2018

Letcher, R. A., & Jakeman, A. J. (2004). *Types of environmental models*: Eolss Publishers.

Li, J., Chen, F., Zhang, G., Barlage, M., Gan, Y., Xin, Y., & Wang, C. (2018). Impacts of land cover and soil texture uncertainty on land model simulations over the central Tibetan Plateau. *Journal of Advances in Modeling Earth Systems*, *10*(9), 2121–2146. https://doi.org/10.1029/2018MS001377

Li, J., Duan, Q., Wang, Y. P., Gong, W., Gan, Y., & Wang, C. (2018). Parameter optimization for carbon and water fluxes in two global land surface models based on surrogate modelling. *International Journal of Climatology*, *38*, e1016–e1031. https://doi.org/10.1002/joc.5428

Li, M., Yang, D., Chen, J., & Hubbard, S. S. (2012). Calibration of a distributed flood forecasting model with input uncertainty using a Bayesian framework. *Water Resources Research*, *48*(8). https://doi.org/10.1029/2010WR010062

Liang, X., Lettenmaier, D. P., Wood, E. F., & Burges, S. J. (1994). A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research: Atmospheres*, *99*(D7), 14415–14428. https://doi.org/10.1029/94JD00483

Lin, P., Rajib, M. A., Yang, Z. L., Somos-Valenzuela, M., Merwade, V., Maidment, D. R., et al. (2018). Spatiotemporal evaluation of simulated evapotranspiration and streamflow over Texas using the WRF-Hydro-RAPID modeling framework. *JAWRA Journal of the American Water Resources Association*, *54*(1), 40–54. https://doi.org/10.1111/1752-1688.12585

Liu, Y., Gupta, H. V., Sorooshian, S., Bastidas, L. A., & Shuttleworth, W. J. (2005). Constraining land surface and atmospheric parameters of a locally coupled model using observational data. *Journal of Hydrometeorology*, *6*(2), 156–172. https://doi.org/10.1175/JHM407.1

Lohmann, D., Mitchell, K. E., Houser, P. R., Wood, E. F., Schaake, J. C., Robock, A., et al. (2004). Streamflow and water balance intercomparisons of four land surface models in the North American Land Data Assimilation System project. *Journal of Geophysical Research: Atmospheres*, *109*(D7). https://doi.org/10.1029/2003JD003517

Lu, D., & Ricciuto, D. (2019). Efficient surrogate modeling methods for large-scale Earth system models based on machine-learning techniques. *Geoscientific Model Development*, *12*(5), 1791. https://doi.org/10.5194/gmd-12-1791-2019

Lu, D., Ricciuto, D., Stoyanov, M., & Gu, L. (2018). Calibration of the E3SM Land Model Using Surrogate-Based Global Optimization. *Journal of Advances in Modeling Earth Systems*, *10*(6), 1337–1356. https://doi.org/10.1002/2017MS001134

Ma, N., Niu, G. Y., Xia, Y., Cai, X., Zhang, Y., Ma, Y., & Fang, Y. (2017). A systematic evaluation of Noah-MP in simulating land-atmosphere energy, water, and carbon exchanges over the continental United States. *Journal of Geophysical Research: Atmospheres*, *122*(22), 12–245. https://doi.org/10.1002/2017JD027597

McNeall, D., Williams, J., Booth, B., Betts, R., Challenor, P., Wiltshire, A., & Sexton, D. (2016). The impact of structural error on parameter constraint in a climate model. *Earth System Dynamics*, *7*(4), 917–935. https://doi.org/10.5194/esd-7-917-2016

Mizukami, N., Clark, M. P., Newman, A. J., Wood, A. W., Gutmann, E. D., Nijssen, B., et al. (2017). Towards seamless large-domain parameter estimation for hydrologic models. *Water Resources Research*, *53*(9), 8020–8040. https://doi.org/10.1002/2017WR020401

Müller, J., Paudel, R., Shoemaker, C. A., Woodbury, J., Wang, Y., & Mahowald, N. (2015). CH4 parameter estimation in CLM4.5bgc using surrogate global optimization. *Geoscientific Model Development*, *8*(1), 3285–3310. https://doi.org/10.5194/gmd-8-3285-2015

Niu, G. Y., Yang, Z. L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., et al. (2011). The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *Journal of Geophysical Research: Atmospheres*, *116*(D12). https://doi.org/10.1029/2010JD015139

Post, H., Vrugt, J. A., Fox, A., Vereecken, H., & Hendricks Franssen, H. J. (2017). Estimation of Community Land Model parameters for an improved assessment of net carbon fluxes at European sites. *Journal of Geophysical Research: Biogeosciences*, *122*(3), 661–689. https://doi.org/10.1002/2015JG003297

Prentice, I. C., Liang, X., Medlyn, B. E., & Wang, Y. P. (2015). Reliable, robust and realistic: The three R's of next-generation land-surface modelling. *Atmospheric Chemistry and Physics*, *15*(10), 5987–6005. https://doi.org/10.5194/acp-15-5987-2015

Prihodko, L., Denning, A. S., Hanan, N. P., Baker, I., & Davis, K. (2008). Sensitivity, uncertainty and time dependence of parameters in a complex land surface model. *Agricultural and Forest Meteorology*, *148*(2), 268–287. https://doi.org/10.1016/j.agrformet.2007.08.006

Rakovec, O., Kumar, R., Mai, J., Cuntz, M., Thober, S., Zink, M., et al. (2016). Multiscale and multivariate evaluation of water fluxes and states over European river basins. *Journal of Hydrometeorology*, *17*(1), 287–307. https://doi.org/10.1175/JHM-D-15-0054.1

Raoult, N. M., Jupp, T. E., Cox, P. M., & Luke, C. M. (2016). Land-surface parameter optimization using data assimilation techniques: The adJULES system V1.0. *Geoscientific Model Development*, *9*, 2833–2852. https://doi.org/10.5194/gmd-9-2833-2016

Ray, J., Hou, Z., Huang, M., Sargsyan, K., & Swiler, L. (2015). Bayesian calibration of the Community Land Model using surrogates. *SIAM/ASA Journal on Uncertainty Quantification*, *3*(1), 199–233. https://doi.org/10.1137/140957998

Razavi, S., Tolson, B. A., & Burn, D. H. (2012). Review of surrogate modeling in water resources. *Water Resources Research*, *48*(7). https://doi.org/10.1029/2011WR011527

Ricciuto, D. M., Davis, K. J., & Keller, K. (2008). A Bayesian calibration of a simple carbon cycle model: The role of observations in estimating and reducing uncertainty. *Global Biogeochemical Cycles*, *22*(2). https://doi.org/10.1029/2006GB002908

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

Rosolem, R., Gupta, H. V., Shuttleworth, W. J., de Gonçalves, L. G. G., & Zeng, X. (2013). Towards a comprehensive approach to parameter estimation in land surface parameterization schemes. *Hydrological Processes*, *27*(14), 2075–2097. https://doi.org/10.1002/hyp.9362

Rosolem, R., Gupta, H. V., Shuttleworth, W. J., Zeng, X., & De Gonçalves, L. G. G. (2012). A fully multiple-criteria implementation of the Sobol′ method for parameter sensitivity analysis. *Journal of Geophysical Research: Atmospheres*, *117*(D7). https://doi.org/10.1029/2011JD016355

Samaniego, L., Kumar, R., Thober, S., Rakovec, O., Zink, M., Wanders, N., et al. (2017). Toward seamless hydrologic predictions across spatial scales. *Hydrology and Earth System Sciences*, *21*(9), 4323–4346. https://doi.org/10.5194/hess-21-4323-2017

Severijns, C. A., & Hazeleger, W. (2005). Optimizing parameters in an atmospheric general circulation model. *Journal of Climate*, *18*(17), 3527–3535. https://doi.org/10.1175/JCLI3430.1

Sheikholeslami, R., & Razavi, S. (2017). Progressive Latin hypercube sampling: An efficient approach for robust sampling-based analysis of environmental models. *Environmental Modelling & Software*, *93*, 109–126. https://doi.org/10.1016/j.envsoft.2017.03.010

Sheikholeslami, R., Razavi, S., Gupta, H. V., Becker, W., & Haghnegahdar, A. (2019). Global sensitivity analysis for high-dimensional problems: How to objectively group factors and measure robustness and convergence while reducing computational cost. *Environmental Modelling & Software*, *111*, 282–299. https://doi.org/10.1016/j.envsoft.2018.09.002

Sun, R., Duan, Q., & Wang, J. (2021). Understanding the spatial patterns of evapotranspiration estimates from land surface models over China. *Journal of Hydrology*, *595*, 126021. https://doi.org/10.1016/j.jhydrol.2021.126021

Sun, R., Hernández, F., Liang, X., & Yuan, H. (2020). A calibration framework for high-resolution hydrological models using a multiresolution and heterogeneous strategy. *Water Resources Research*, *56*(8), e2019WR026541. https://doi.org/10.1029/2019WR026541

Sun, R., Yuan, H., & Liu, X. (2017). Effect of heteroscedasticity treatment in residual error models on model calibration and prediction uncertainty estimation. *Journal of Hydrology*, *554*, 680–692. https://doi.org/10.1016/j.jhydrol.2017.09.041

Sun, R., Yuan, H., & Yang, Y. (2018). Using multiple satellite-gauge merged precipitation products ensemble for hydrologic uncertainty analysis over the Huaihe River basin. *Journal of Hydrology*, *566*, 406–420. https://doi.org/10.1016/j.jhydrol.2018.09.024

Tolson, B. A., & Shoemaker, C. A. (2007). Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, *43*(1). https://doi.org/10.1029/2005WR004723

Troy, T. J., Wood, E. F., & Sheffield, J. (2008). An efficient calibration method for continental-scale land surface modeling. *Water Resources Research*, *44*(9). https://doi.org/10.1029/2007WR006513

Viana, F. A., Simpson, T. W., Balabanov, V., & Toropov, V. (2014). Metamodeling in multidisciplinary design optimization: How far have we really come? *AIAA Journal*, *52*(4), 670–690. https://doi.org/10.2514/1.J052375

Vrugt, J. A., Gupta, H. V., Bastidas, L. A., Bouten, W., & Sorooshian, S. (2003). Effective and efficient algorithm for multiobjective optimization of hydrologic models. *Water Resources Research*, *39*(8). https://doi.org/10.1029/2002WR001746

Wang, C., Duan, Q., Gong, W., Ye, A., Di, Z., & Miao, C. (2014). An evaluation of adaptive surrogate modeling based optimization with two benchmark problems. *Environmental Modelling & Software*, *60*, 167–179. https://doi.org/10.1016/j.envsoft.2014.05.026

Wang, Q. J. (1991). The genetic algorithm and its application to calibrating conceptual rainfall-runoff models. *Water Resources Research*, *27*(9), 2467–2471. https://doi.org/10.1029/91WR01305

Wang, Y. P., Baldocchi, D., Leuning, R. A. Y., Falge, E. V. A., & Vesala, T. (2007). Estimating parameters in a land-surface model by applying nonlinear inversion to eddy covariance flux measurements from eight Fluxnet sites. *Global Change Biology*, *13*(3), 652–670. https://doi.org/10.1111/j.1365-2486.2006.01225.x

Wood, E. F., Roundy, J. K., Troy, T. J., Van Beek, L. P. H., Bierkens, M. F., Blyth, E., et al. (2011). Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water. *Water Resources Research*, *47*(5). https://doi.org/10.1029/2010WR010090

Xi, J., Li, A., & Wang, M. (2020). HetRCNA: A novel method to identify recurrent copy number alternations from heterogeneous tumor samples based on matrix decomposition framework. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *17*(2), 422–434. https://doi.org/10.1109/TCBB.2018.2846599

Xi, J., Yuan, X., Wang, M., Li, A., Li, X., & Huang, Q. (2020). Inferring subgroup-specific driver genes from heterogeneous cancer samples via subspace learning with subgroup indication. *Bioinformatics*, *36*(6), 1855–1863. https://doi.org/10.1109/tcbb.2018.2846599

10.1029/2020WR028740

Xia, Y., Cosgrove, B. A., Mitchell, K. E., Peters-Lidard, C. D., Ek, M. B., Kumar, S., et al. (2016). Basin-scale assessment of the land surface energy budget in the National Centers for Environmental Prediction operational and research NLDAS-2 systems. *Journal of Geophysical Research: Atmospheres*, *121*(1), 196–220. https://doi.org/10.1002/2015JD023889

Xia, Y., Pitman, A. J., Gupta, H. V., Leplastrier, M., Henderson-Sellers, A., & Bastidas, L. A. (2002). Calibrating a land surface model of varying complexity using multicriteria methods and the Cabauw dataset. *Journal of Hydrometeorology*, *3*(2), 181–194. https://doi.org/10.1175/1525-7541(2002)003<0181:CALSMO>2.0.CO;2

Yang, Z. L., Niu, G. Y., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., et al. (2011). The community Noah land surface model with multiparameterization options (Noah-MP): 2. Evaluation over global river basins. *Journal of Geophysical Research: Atmospheres*, *116*(D12). https://doi.org/10.1175/1525-7541(2002)003<0181:calsmo>2.0.co;2

Yapo, P. O., Gupta, H. V., & Sorooshian, S. (1998). Multi-objective global optimization for hydrologic models. *Journal of Hydrology*, *204*(1–4), 83–97. https://doi.org/10.1016/S0022-1694(97)00107-8