

# Water Resources Research®

## RESEARCH ARTICLE

10.1029/2021WR029879

### Key Points:

- A combined framework by jointly using parameter optimization and data assimilation algorithms is proposed
- A new objective function is introduced in PO to make the optimization spatially effective
- The value of assimilation of satellite soil moisture data into the Community Land Model is investigated

### Correspondence to:

P. Abbaszadeh,  
pabbaszadeh@ua.edu

### Citation:

Zhang, C., Abbaszadeh, P., Xu, L., Moradkhani, H., Duan, Q., & Gong, W. (2021). A combined optimization-assimilation framework to enhance the predictive skill of Community Land Model. *Water Resources Research*, 57, e2021WR029879. <https://doi.org/10.1029/2021WR029879>

Received 26 FEB 2021  
Accepted 20 NOV 2021

## A Combined Optimization-Assimilation Framework to Enhance the Predictive Skill of Community Land Model

**Chong Zhang<sup>1,2</sup>, Peyman Abbaszadeh<sup>2,3</sup> , Lei Xu<sup>2,4</sup>, Hamid Moradkhani<sup>2,3</sup> , Qingyun Duan<sup>5,6</sup> , and Wei Gong<sup>7</sup> **

<sup>1</sup>College of Resource Environment and Tourism, Capital Normal University, Beijing, China, <sup>2</sup>Center for Complex Hydrological Systems Research, The University of Alabama, Tuscaloosa, AL, USA, <sup>3</sup>Department of Civil, Construction, and Environmental Engineering, The University of Alabama, Tuscaloosa, AL, USA, <sup>4</sup>State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China, <sup>5</sup>State Key Laboratory of Hydrology-Water Resources and Hydraulic Engineering, Hohai University, Nanjing, China, <sup>6</sup>College of Hydrology and Water Resources, Hohai University, Nanjing, China, <sup>7</sup>Institute of Land Surface System and Sustainable Development, Beijing Normal University, Beijing, China

**Abstract** Land Surface Model (LSM) predictions are affected by unknown parameters, uncertain meteorological forcings, inaccurate initial conditions, and incomplete process representation. While parameter optimization (PO) algorithms have been used to estimate the model parameter for best model performance, the use of data assimilation (DA) has been receiving increased prominence to characterize the uncertainties in such models. However, it is still an open question that how DA techniques can be used together with PO algorithms to enhance the performance of LSMs for different applications. In this study, a combined framework composed of the PO algorithm—Adaptive Surrogate Modeling Based Optimization (ASMO) and the DA algorithm—Evolutionary Particle Filter with Markov Chain Monte Carlo (EPFM), is used to improve the soil moisture (SM) estimates of Community Land Model (CLM) across the Tibetan Plateau in China. The SM data from the Soil Moisture Active Passive satellite are used as the benchmark data for model improvement. To demonstrate the usefulness and effectiveness of the proposed approach, the simulated SM values are validated against the SM observations collected from the in situ networks in the Tibetan Plateau, China. The findings revealed that the joint application of ASMO and EPFM algorithms results in more accurate and reliable SM estimates of CLM compared to when they are solely employed. This study suggests that the combined optimization-assimilation framework can be utilized for improving the predictive skill of the other large-scale complex LSMs while accounting for the uncertainties associated with both model parameters and state variables.

### 1. Introduction

Land surface model (LSM) is an integral component of the earth system modeling that simulates the states of terrestrial water, energy, carbon, and associated flux exchanges in land-atmosphere interactions (Clark et al., 2015; Lawrence et al., 2011, 2019). The LSM simulations are most often inaccurate and erroneous due to multiple sources of uncertainties. These include errors in atmospheric forcing data, model structural uncertainty due to inadequate conceptualization of underlying physics or non-uniqueness of model parameters, and initial condition uncertainty (Duan et al., 2019; Moradkhani et al., 2018). Failure to account for these sources of uncertainties may result in misleading LSM simulations (i.e., forecast hydrometeorological variables) under current and future climate conditions (Abbaszadeh et al., 2019; Liu et al., 2012; Wang et al., 2014). Given that many of the scientific and societal challenges to solve global and regional environmental issues depend on the accurate and reliable simulations of LSMs, it is essential to be able to characterize, quantify, and account for different sources of uncertainties involved in different layers of model simulations (Huang et al., 2016; Lievens et al., 2016; Vrugt et al., 2005).

LSMs contain several parameters that represent spatially and temporally heterogeneous properties of the real system. Parameters are not often easily measurable, rather they are estimated indirectly through either prior knowledge or field and laboratory tests with the consequent introduction of errors and uncertainties (Duan et al., 2019; Gupta et al., 2003). Thus, parameter optimization (PO) is needed to reduce their uncertainties through matching the model output(s) at a certain location(s) where the corresponding observation is available. There may be more than one set of parameters that are equally capable of reproducing similar (not necessarily identical) model outputs due to equifinality or non-uniqueness, which is also referred to as unidentifiability (Gong et al., 2015; Gou

et al., 2020; Wang et al., 2014). From another perspective, the model parameters can be optimized to compensate for structural errors, therefore it is presumed that explicitly acknowledging the structural error or uncertainty in a model is not necessary (Duan et al., 2019; Xu et al., 2017). Although this is true when estimating best-fitting parameters at a certain location where the observation is available, for accurate and realistic quantification of total uncertainty, an explicit representation of state uncertainty is required in LSM simulation (Liu et al., 2012; Moradkhani, Hsu, Gupta, & Sorooshian et al., 2005, Moradkhani et al., 2012). This has been the main topic of many studies (DeChant & Moradkhani, 2012; Liu & Gupta, 2007; Vrugt et al., 2005) where the benefit of “explicitly” accounting for initial condition uncertainty in improving hydrologic model predictions through data assimilation (DA) has been explored. DA is a well-known approach in the hydrometeorological community for characterizing the uncertainties associated with the state variables, model parameters, and input forcing data through probabilistically conditioning the states of the model on the observations (Abbaszadeh et al., 2018; Gavahi et al., 2020; Liu & Gupta, 2007). Over the past decades, both PO and DA algorithms have been successfully applied in many studies to improve the hydrologic model simulations (Abbaszadeh et al., 2020; Duan et al., 1994; Liu et al., 2012; Xu et al., 2020). However, for a typical LSM with high parameterization and strong non-linearity, addressing the uncertainties associated with the model predictions while properly representing the model state variables and parameters still is a great challenge, if not impossible (Gong et al., 2015, 2016b; Liu et al., 2012).

Implementing PO on a LSM is most often a computationally intensive task (Gong et al., 2015, 2016b). Generally, using the traditional PO algorithms to calibrate a hydrologic model with more than 10 parameters requires  $10^5$ – $10^6$  (or even more) times of model runs (Li et al., 2013). LSMs typically contain 20–100 (or more) parameters, such that even with large computational resources and power, depending on the size of the study area, it can take several days (or weeks) for the model to run (Gan et al., 2014; Gong et al., 2015; Li et al., 2013). Sensitivity analysis is an efficient approach to make the optimization process computationally feasible. It allows identifying the most dominant parameters and reducing the dimension of the optimization problem (Gan et al., 2014; Hou et al., 2012; Saltelli et al., 2008). Another alternative is to use the surrogate modeling-based optimization that can alleviate the computational complexity by replacing the physical model with a statistical surrogate model, which describes the error response surface between the adjustable parameters and the performance measure (i.e., objective function) of the physical model (Huang et al., 2016; Razavi et al., 2012). The optimal solution achieved by the surrogate model can thus approximate the optimal solution of the physical model (Duan et al., 2017). Integrating sensitivity analysis and surrogate modeling-based optimization algorithms has been recognized as an efficient approach to implement PO on the LSMs (Gong et al., 2015, 2016b; Zhang et al., 2020).

The DA algorithms based on the Ensemble Kalman Filter and Particle Filter (PF) were designed to recursively estimate model state variables (and parameters of interest). In these algorithms, Monte Carlo sampling and sequential updating are applied to a set of prognostic and diagnostic state variables at each assimilation time step (DeChant & Moradkhani, 2012; Gavahi et al., 2020). The probability distributions of model states are recursively updated at each time step when a new observation becomes available. These approaches provide better state estimates (and parameters if DA is applied jointly) through which the modeling system evolves consistently over time and consequently results in improved predictions while accounting for uncertainties (Gupta et al., 2003; Lievens et al., 2016; Vrugt et al., 2005). Despite the widespread use of the Ensemble Kalman Filter and its variants in many studies, this approach is subject to some inherent limitations that result in suboptimal model performance. These include linear updating rules, Gaussian assumption of errors, and water balance violation. Given these problems, DA via particle filtering has received great attention in the hydrologic community and beyond (Moradkhani, Sorooshian, Gupta, & Houser, 2005, Moradkhani et al., 2018). However, particle filtering suffers from sample impoverishment and particle degeneracy problems. The literature indicates that these problems can be alleviated by using the resampling strategies, and also the Markov Chain Monte Carlo (MCMC) algorithm (Moradkhani et al., 2012), and/or through the utilization of metaheuristic techniques, such as the genetic algorithm (Abbaszadeh et al., 2018). This type of algorithms incorporating PF and auxiliary resampling procedures (e.g., PF-MCMC) is capable of quantifying a more complete representation of the posterior for a non-linear/non-Gaussian system such as the LSM (Abbaszadeh et al., 2020; Xu et al., 2020).

Although most of the DA algorithms offer a robust framework for updating state variables, their “optimality” relies on the unbiased state simulation, which is largely dependent on the accurate parameter estimation for the model (Liu et al., 2012; Liu & Gupta, 2007). Hence, there has been growing interest in the development of dual state-parameter DA techniques where both model state variables and parameters are recursively and jointly

updated throughout the assimilation period (Moradkhani, Hsu, Gupta, & Sorooshian et al., 2005; Moradkhani, Sorooshian, Gupta, & Houser, 2005). When dual state-parameter DA is applied on a LSM with strong non-linearity, it may render the updating process unstable and intractable as the parameters generally vary much slower than states in reality. Moreover, for the large study areas, it is required to have access to unaffordable computational resources. To mitigate these issues, some studies suggested employing a global PO algorithm in a loop external to the DA filter to adjust parameters (Vrugt et al., 2005). Such a combined optimization-assimilation framework merges the high efficiency of PO in minimizing the overall variance of the model outputs and the advantage of DA in explicitly dealing with the uncertainties of model structure and forcing data and initial conditions, thereby improving the predictive accuracy in an overall statistical sense. Therefore, here, we aim to investigate the usefulness and effectiveness of enhancing the simulation performance of LSM by jointly using the advanced PO and DA algorithms.

In this study, the Community Land Model (CLM) is used as the LSM to estimate the soil moisture (SM) across the Tibetan Plateau (TP) in China. We improve the predicted SM by assimilating the SM observation from the Soil Moisture Active Passive (SMAP) satellite into the CLM. The Adaptive Surrogate Modeling Based Optimization (ASMO) algorithm developed by Wang et al. (2014) is used to implement PO in a batch-processing mode. Subsequently, the Evolutionary PF and MCMC (EPFM) algorithm developed by Abbaszadeh et al. (2018) is used to executing DA on the CLM with and without PO treatment. The SM simulations improved by ASMO and EPFM are individually compared with the results of when ASMO and EPFM are used together (i.e., ASMO + EPFM). The in situ observed SM data collected from densely instrumented sites are used for independently validating the CLM-simulated SM improved by ASMO, EPFM, and their combination. This study investigates the extent to which the combined optimization-assimilation framework enhances the predictive skill of a high-dimensional LSM.

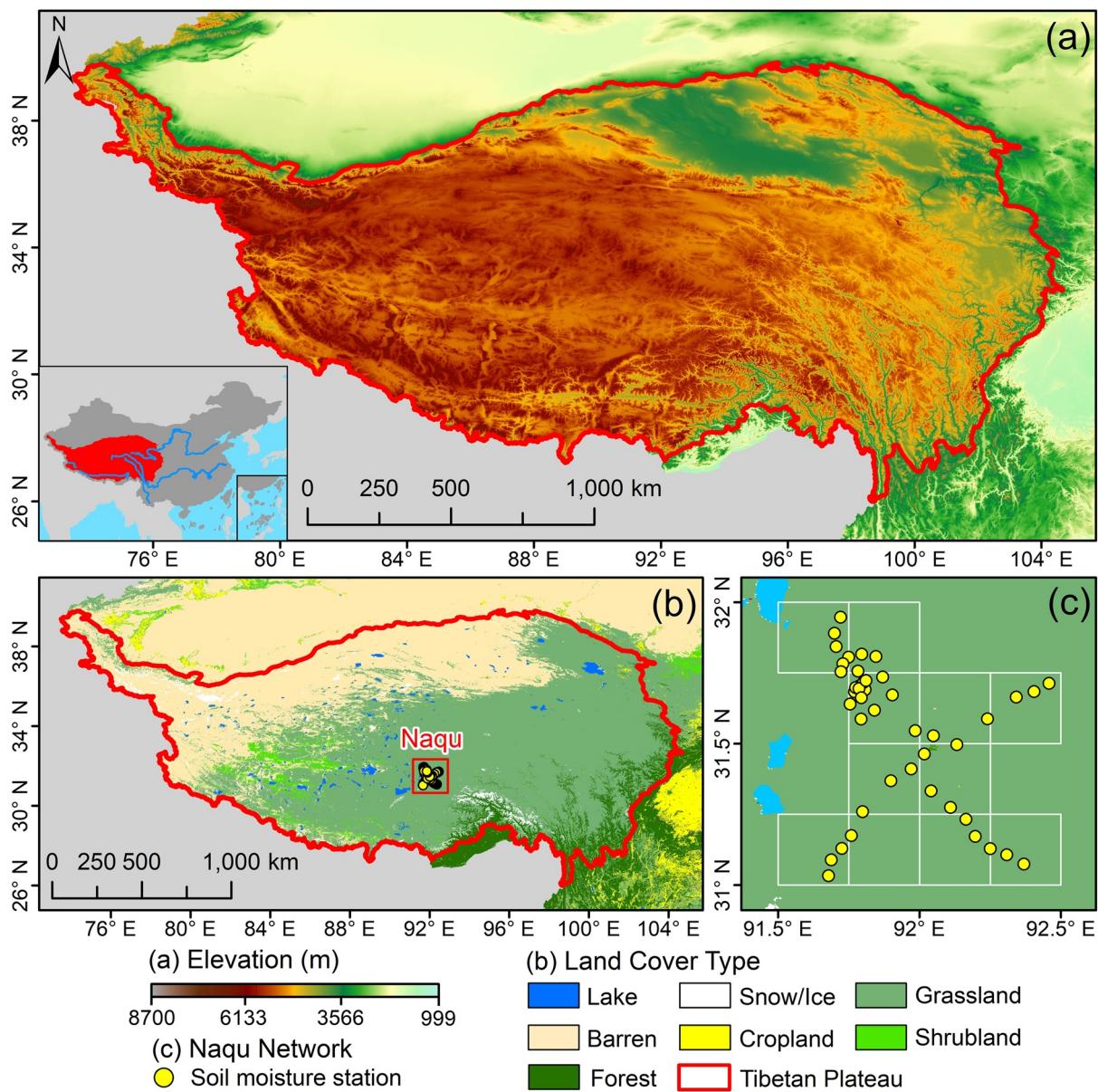
The rest of the paper is organized as follows: Section 2 describes the study area, the CLM structure, and the dataset used in this study. Section 3 presents the ASMO and EPFM algorithms for PO and DA, respectively, and explains the experimental design and all performance measures used to assess the effectiveness of the proposed algorithms. Section 4 elaborates the results including the parameter sensitivity analysis, the ASMO-based PO, the performance of different DA configurations (i.e., EPFM and ASMO + EPFM), and the independent validation using in situ observations. Section 5 discusses the strengths and weaknesses of the proposed optimization-assimilation framework and provides some suggestions to extend this work. Finally, Section 6 concludes the main findings of this study.

## 2. Study Area, Model, and Data Sets

### 2.1. Study Area

The Tibetan Plateau (TP, Figure 1) is the world's highest and vastest plateau, with the mean altitude exceeding 4,000 m above sea level and the total area reaching 2,500,000 km<sup>2</sup> (Chen et al., 2017; Yang et al., 2014). The topography in TP is characterized by a wide flat valley and mountainous terrain (Figure 1a). Most of the major rivers in Asia (e.g., Brahmaputra, Yellow, and Yangtze rivers) stem from TP, making it known as the "Asia's Water Tower." TP is also called the "Earth's Third Polar" due to its cold climatology and hold of the largest ice store following the two poles of the Earth (Qiu, 2008; Song et al., 2014; Yao et al., 2012). The major land cover type in TP is grassland and the second type is barren (Figure 1b). More than 1,500 lakes that covers an area of 38,000 km<sup>2</sup> are scattered on the TP (Song et al., 2014). TP has undergone major hydroclimatic shifts over the past half-century, primarily characterized by rapid warming, moistening, solar dimming, and wind stilling (Yang et al., 2011, 2014). About 82% of the plateau's glaciers and 10% of its permafrost have been degraded due to rapid warming (Qiu, 2008; Yao et al., 2012). The rate of degradation has proved to be continuing and even accelerating in recent years (Song et al., 2014). The increase of meltwater and precipitation further increases river discharge and expands lakes, which in turn increases the likelihood of flooding and threatening local living conditions (Yang et al., 2014; Yao et al., 2012).

SM estimation with high accuracy and fine spatiotemporal resolution is essential for understanding the complex feedbacks of land hydrological processes to the changing climate in TP (Yang et al., 2013). However, currently, the SM networks in TP are either densely instrumented at field scale or sparsely distributed at the local scale (Luo et al., 2020; Yang et al., 2013). Land surface modeling and remotely sensed observation are the two



**Figure 1.** Maps of (a) digital elevation, (b) land cover type, and (c) Naqu network soil moisture observation in the Tibetan Plateau, China.

promising ways to estimate SM at a larger spatial scale. Many studies have evaluated the accuracy of SM products from LSM simulation and satellite retrieval by using field-scale site observations in TP (Chen et al., 2017; Li et al., 2018; Yang et al., 2013). These comparisons are often made by using the simulated surface SM data of the upper ~5 cm soil depth due to the limited capability of satellite microwave sensors. The findings of these studies demonstrated that the LSMs tend to simulate a biased SM due to its inherent uncertainties, while satellite-based SM observations across the TP are more accurate probably due to its sparse vegetation cover and less human activities (Chen et al., 2017; Li et al., 2018). This motivated us to examine the value of the proposed optimization-assimilation framework that incorporates the satellite-based SM observations into the LSM to improve the accuracy and reliability of the SM estimates in TP.

## 2.2. Community Land Model

The Community Land Model (CLM) version 4 is a LSM that has an advanced hydrology scheme, representing several aspects of water and energy fluxes, as well as state variables in the land surface process (Lawrence

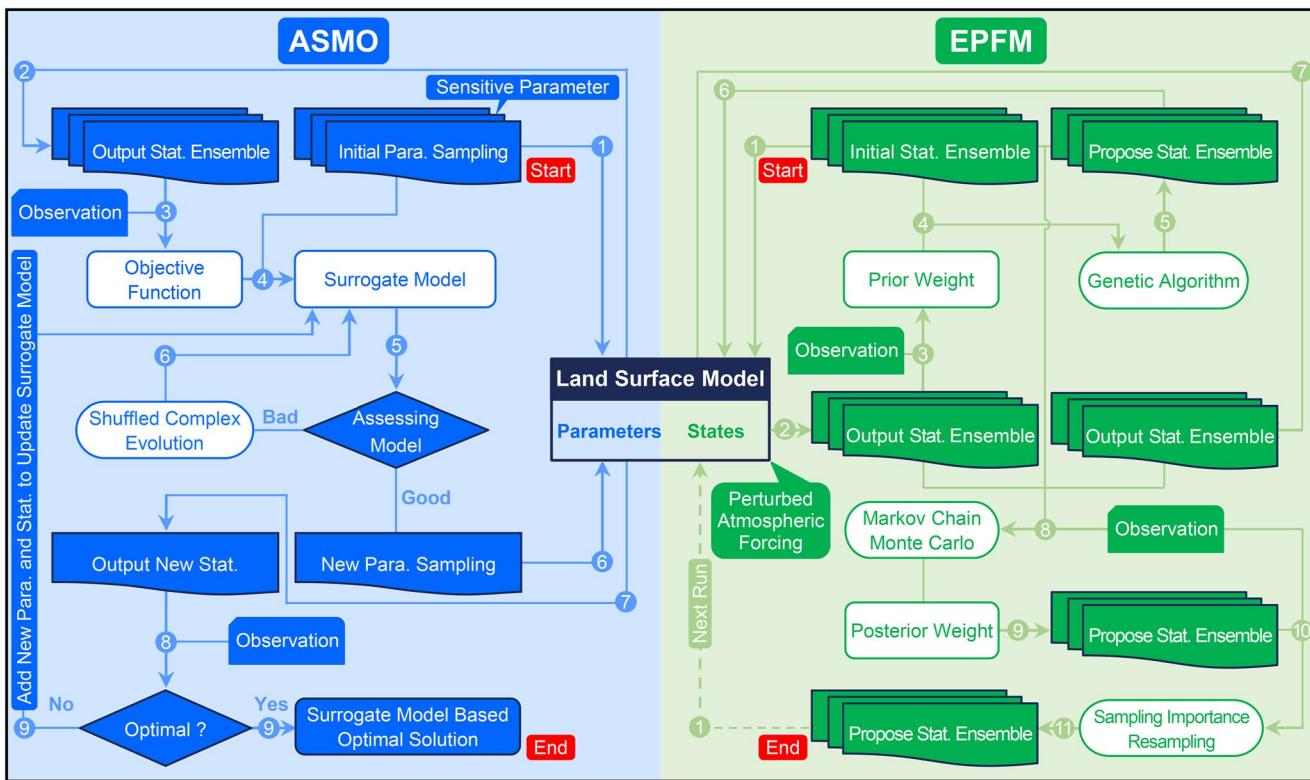
et al., 2011; Oleson et al., 2010). Structurally, subgrid-scale heterogeneity induced by varying land units (e.g., lake, wetland, glacier, and vegetated area) is considered in model design through subdividing one grid cell into several fractions that, respectively, represent a single plant functional type. These fractions within a grid cell have no direct interaction mutually in model simulation (Lawrence et al., 2011; Oleson et al., 2010). By default, the soil column in a total depth of ~35.2 m is divided into 10 layers with the thickness of each layer increasing from top to bottom. The soil hydraulic and thermal parameters in the model are derived by using the pedotransfer functions of sand and clay contents, as well as the organic properties of soil (Clapp & Hornberger, 1978). Beneath the soil column, an aquifer with a 5 m depth is parameterized as the low boundary condition to simulate the unconfined groundwater storage changes (Niu et al., 2007). The modified Richard's equation is adopted to describe one-dimensional water flow and energy transport in vertical soil layers, in which the SM simulation is governed by infiltration, surface and subsurface discharge, gradient diffusion, vegetation absorption, and interactions with groundwater (Oleson et al., 2010).

### 2.3. Meteorological Forcing Data Set

The China meteorological forcing dataset, which is freely available at the data sharing repository of the National Tibetan Plateau Data Center (He et al., 2020; Yang & He, 2018), is used to drive the CLM simulation in TP. This dataset contains all seven near-surface meteorological forcing data, including precipitation rate, surface pressure, specific humidity, 2-m air temperature, 10-m wind speed, downward shortwave radiation, and longwave radiation. They are created by fusing satellite-based products, reanalysis datasets, and in situ gage records. The dataset is available from January 1979 to the present and is now being updated continuously (He et al., 2020). The temporal and spatial resolutions of this dataset are 3 hr and 0.1° grids, respectively. We resampled it to 0.25° grids by nearest interpolation in order to be consistent with the spatial resolution of satellite-based SM observation (see Section 2.4). He et al. (2020) compared all seven variables in the China meteorological forcing dataset and the Global Land Data Assimilation System against the in situ observations of meteorological sites in China. Their findings demonstrated that the China meteorological forcing dataset has superior quality rather than the Global Land Data Assimilation System in China. They reported that the correlation coefficient (mean bias error) of precipitation rates between China meteorological forcing dataset and site observations is 0.47 (~0.02 mm/day) while the correlation coefficient between Global Land Data Assimilation System and site observations is 0.23 (~0.27 mm/day). Such improvement is expected as more in situ measurements were incorporated into the generation of the China meteorological forcing dataset. The China meteorological forcing dataset is available at <https://data.tpdc.ac.cn/en/data/8028b944-daaa-4511-8769-965612652c49/>.

### 2.4. SMAP Soil Moisture

The SMAP satellite was launched in January 2015 by the National Aeronautics and Space Administration to map global SM and detect whether soils are frozen or thawed (Entekhabi et al., 2010). It carries two L band microwave sensors, of which one is passive radiometer (centered at 1.41 GHz) and the other is active radar (centered at 1.26 GHz). Unfortunately, due to the malfunction of SMAP's radar in July 2015, the SMAP's radiometer has been the only operating sensor on the satellite retrieving the SM at 36 km spatial resolution across the globe. The SMAP Level-3 Radiometer Global Daily SM dataset version 6 (O'Neill et al., 2019) is produced by adopting the Single Channel Algorithm-V polarization, which uses the vertically polarized brightness temperature data for SM estimation given the knowledge of soil texture. There have been many studies that successfully used the older versions of this dataset to study the SM in TP (Chen et al., 2017; Li et al., 2018). The enhanced version of this product is used as a benchmark dataset for implementing PO and DA in this study. We transformed the projected coordinate system of this dataset from its original global cylindrical 36 km Equal-Area Scalable Earth Grid version 2.0 to the geographic coordinate system of World Geodetic System 1984, and then resampled it to 0.25° grids by nearest interpolation. The systematic bias of SM estimates between SMAP observation and CLM simulation is removed by the empirical matching of the cumulative distribution function. The dataset of SMAP-observed SM can be retrieved at <https://nsidc.org/data/SPL3SMP/versions/6>.



**Figure 2.** The algorithmic flowchart of the Adaptive Surrogate Modeling based Optimization (ASMO) and the Evolutionary Particle Filter and MCMC (EPFM) used to conduct PO and data assimilation (DA) in this study. Note that in this combined framework, the ASMO algorithm is used in a batch-processing mode, and the EPFM DA algorithm is implemented afterward.

### 2.5. In Situ Observed Soil Moisture

To independently validate the simulated SM of CLM, we collected the observed SM data for the time interval of half-hour throughout years 2016 and 2017 from the 44 sites in the Naqu network (Figure 1c). Each site has four sensors installed at the soil depths of 5, 10, 20, and 40 cm, respectively. The shallow surface SM at 5 cm soil depth (comparable to the monitoring depth of SMAP satellite) is used for the validation. The half-hourly SM data are aggregated to a daily scale in order to be consistent with the temporal resolution of CLM outputs. As shown in Figure 1c, the Naqu network is installed within a spatial extent of  $1.0^{\circ}$  with 44 sites nested in eleven  $0.25^{\circ}$  grids if referring to the spatial grids of CLM simulation. It is important to note that the direct comparison of grid size SM simulation with the corresponding in situ observation is not reasonable as each site measures the SM at point scale and does not represent the SM within the model simulation scale. To address this problem, the in situ observed SM data from all sites are spatially averaged and compared with the average of CLM-simulated SM values across eleven  $0.25^{\circ}$  grid cells (Figure 1c). Four percentile ranges (i.e., 10%, 68%, 90%, and 95% confident intervals) of all site-observed SM data are calculated to represent the extent of spatial variability in SM, and explore whether the simulated SM improved by the optimization-assimilation framework becomes more consistent with the average of site-observed SM (see Section 4.4).

## 3. Methodology and Experiment Design

This section presents the PO algorithm—Adaptive Surrogate Modeling Based Optimization (ASMO), the DA algorithm—Evolutionary PF with MCMC (EPFM), designing four experiments using the ASMO and EPFM algorithms, and the performance measures used to assess the usefulness and effectiveness of the proposed algorithms. Figure 2 illustrates the structure of the ASMO and EPFM algorithms. Here we briefly describe the main components and implementation process of ASMO and EPFM algorithms in the following two Sections 3.1

and 3.2. For more information related to ASMO and EPFM, we refer the readers to Wang et al. (2014) and Abbaszadeh et al. (2018), respectively.

### 3.1. Adaptive Surrogate Modeling Based Optimization (ASMO)

The ASMO is a PO algorithm that can effectively reduce the number of model runs required for optimizing parameter values in a large complex physical model (Wang et al., 2014). The main idea of ASMO is to replace the physical model with an effective surrogate model to approximate the optimal solution of the physical model within an optimization framework. An iterative process designed in ASMO uses the knowledge acquired from previous iterations to guide future optimal solution acquisition. The ASMO algorithm consists of three main components including the parameter sensitivity analysis for identifying important parameters, the surrogate-based modeling for emulating physical model behavior, and the adaptive sampling strategy for searching optimal solutions. We describe the ASMO algorithm in detail below.

Suppose that a physical model can be formulated as follows:

$$Y = f(\mathbf{u}, \theta) \quad (1)$$

where  $\mathbf{u}$  and  $\theta$  denote the forcing datasets and the model parameters, respectively.  $Y$  is the output variables of the model  $f(\cdot)$ . For the problem of single-objective optimization, the output variable of interest is a member of  $Y$ , expressed as  $y$ . The parameters to be optimized are a subset of  $\theta$ , expressed as  $\theta$ .  $\theta$  is also known as the sensitive parameters of  $y$  that can be identified by performing a sensitivity analysis experiment in the ASMO algorithm (Figure 2).

Each of the sensitive parameter in  $\theta$  is sampled  $n$  times in its parameter space to obtain the set of sample points  $\theta_i = \{\theta_1, \theta_2, \dots, \theta_n\}$ . Strictly speaking, the initial parameter sampling requires prior knowledge of parameters' probability distribution and some space-filling criteria. However, if such prior knowledge is not available, the sampling can be performed following the uniform probability distribution within the adjustable range of each parameter. A sample size of more than 10 times the number of sensitive parameters is better to describe the surrogate model (Jones et al., 1998). These sample points are then put into the physical model instead of the default parameter values, to obtain the model outputs. The corresponding objective function values (i.e.,  $\varphi_i = \{\varphi_1, \varphi_2, \dots, \varphi_n\}$ ) are calculated based on the model outputs and associated observation data (Figure 2). Afterward, a statistical surrogate model is built to describe the error response surface between initial parameter samples ( $\theta$ ) and objective function values ( $\varphi$ ):

$$\varphi = f^*(\theta) \quad (2)$$

Wang et al. (2014) and Gong et al. (2015) compared five commonly used surrogate models and found that the Gaussian processes regression (GPR) model (Rasmussen & Williams, 2006) outperformed others in the optimization of hydrologic models. Therefore, GPR is used to build the surrogate model in this study. A GPR model can be formulated by mean function  $m(\theta)$  and covariance function  $cov(\theta, \theta')$ :

$$m(\theta) = E[f^*(\theta)] \quad (3)$$

$$cov(\theta, \theta') = E[(f^*(\theta) - m(\theta)) \times (f^*(\theta') - m(\theta'))] \quad (4)$$

That is to say, the GPR model can be expressed as follows:

$$f^*(\theta) \sim GP(m(\theta), cov(\theta, \theta')) \quad (5)$$

The GPR model needs to undergo a training procedure by using  $\theta$  and  $\varphi$ . Suppose  $\theta^*$  is the predicting inputs of the parameter set, namely, the samples that we want to evaluate using the GPR model; and  $\varphi^*$  is the predicting outputs, the joint distribution of inputs and outputs for training and predicting procedures follows the joint Gaussian distribution:

$$\begin{bmatrix} \varphi \\ \varphi^* \end{bmatrix} \sim N \left( 0, \begin{bmatrix} cov(\theta, \theta) + \sigma_n^2 I & cov(\theta, \theta^*) \\ cov(\theta^*, \theta) & cov(\theta^*, \theta^*) \end{bmatrix} \right) \quad (6)$$

where  $\sigma_n^2$  and  $I$  are the noise term and characteristic length, respectively.  $cov(\theta^*, \theta)$  denotes the covariance matrix of the predicting parameter inputs ( $\theta^*$ ) and the initial parameter samples ( $\theta$ ). The rest (e.g.,  $cov(\theta, \theta)$ ,  $cov(\theta, \theta^*)$ , and so on) can be understood in the same manner. The covariance matrix is computed via the following covariance function:

$$cov_{v=5/2}(r) = \left( 1 + \frac{\sqrt{5}r}{I} + \frac{5r^2}{3I^2} \right) \exp\left(-\frac{\sqrt{5}r}{I}\right) \quad (7)$$

where  $r = |\theta_1^* - \theta_2^*|$  is the Euclidian distance between each sample point in the relevant pair of matrixes. Note that the covariance function is the Matérn kernel with parameter  $v = 5/2$ . It is commonly used to define the statistical covariance between measurements made at two points. GPR has high flexibility to represent different interpolation models through using various types of covariance functions. The values of several hyper-parameters in the model building can control the behavior of GPR. For example, in Equation 6, if the noise term  $\sigma_n^2$  is equal to 0, GPR will be used as a multidimensional interpolation approach; if the characteristic length  $I$  is large, GPR will be smooth and insensitive to the changes of model outputs. To obtain the appropriate hyper-parameter values, thus, the shuffled complex evolution algorithm (Duan et al., 1992) is used to maximize the marginal likelihood function:

$$\log[p(\varphi|\theta)] = -\frac{1}{2}\varphi^T [cov(\theta, \theta) + \sigma_n^2 I]^{-1} \varphi - \frac{1}{2}\log|cov(\theta, \theta) + \sigma_n^2 I| - \frac{n}{2}\log 2\pi \quad (8)$$

Doing so is to find the optima of an error response surface described by GPR in the multiparameter space. The minimum of the response surface corresponds to a new sample point, which is considered as the optimal parameter sample ( $\theta_s$ ) of the surrogate model.  $\theta_s$  is always generated using an adaptive sampling strategy, which refers to the process of making full use of the information gained from existing sample points to favor new samples in the promising parameter space (Wang et al., 2014). Afterward,  $\theta_s$  is substituted into the physical model to simulate the model output of interest, the corresponding objective function ( $\varphi_s$ ) is calculated accordingly (Figure 2). If  $\varphi_s$  meets the convergence criterion (this is a check to determine whether the new parameter sample ( $\theta_s$ ) is optimal), the optimization terminates; otherwise,  $\theta_s$  and  $\omega_s$  are put respectively into  $\theta$  and  $\omega$  to update the surrogate model, and the next loop of surrogate modeling-based optimization is performed iteratively until the convergence criteria meet (Figure 2).

In this study, the convergence criterion of the ASMO algorithm is defined as the objective function value of the surrogate-based parameter sample remaining unchanged within 20 times of iteration, or the number of iterations reaching the maximum number of model runs (in this study the maximum is 80, excluding initial samples). The initial samples of sensitive parameters are drawn by using the Good Lattice Points approach (Korobov, 1959). Two sensitivity analysis approaches, Sum-Of-Trees (SOT; Breiman et al., 1984) and Multivariable Adaptive Regression Spline (MARS; Friedman, 1991), are used together to ensure the robustness of sensitive parameter identification (Gan et al., 2014; Saltelli et al., 2008). The selection of parameter sampling and sensitivity analysis approaches is at the discretion of the user for the ASMO algorithm, but the aforementioned ones are recommended because their applicability and reliability have been validated in our previous studies for optimizing hydrologic models (Gong et al., 2016a; Gou et al., 2020; Wang et al., 2014), LSMs (Gong et al., 2015; Zhang et al., 2020), and a weather forecast model (Di et al., 2019, 2020; Duan et al., 2017).

The selected CLM parameters along with their notations, definitions, default values, and adjustable ranges are summarized in Table 1. For simplicity, 15 parameters are numbered from P01 to P15 (Table 1). P01–P04 are the hydrologic parameters that directly influence the movement and storage of water in the soil column and aquifer. For example, P02 ( $f_{drat}$ ) is the decay factor controlling subsurface flow and P04 ( $S_y$ ) is the average specific yield dominating the storage capacity of the aquifer. P05–P15 are the functional parameters in the pedotransfer functions that represent the soil hydraulic and thermal properties in the model (Oleson et al., 2010). It is expected that changes in their values can indirectly influence SM by altering soil properties. For example, the mineral soil porosity is calculated by  $\theta_{min} = 0.489 - 0.00126(\%sand)$ , where 0.489 and 0.00126 represent the intercept- and slope-related parameters (recognized as P05 and P06, respectively). A reduction of P05 value corresponds to a decrease of soil porosity that could in turn reduce the saturated SM (Göhler et al., 2013).

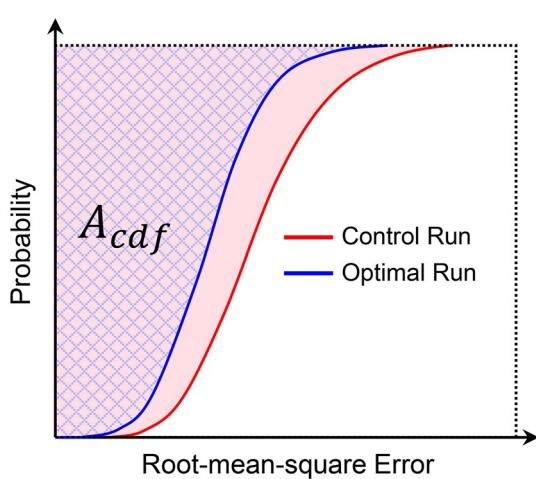
To stabilize the optimization effects spatially, a new objective function called the area index of the cumulative distribution function ( $A_{cdf}$ ) regarding root-mean-square error (RMSE) values is applied in this study (Huo, 2019).

**Table 1**  
Selected Parameters and Their Feasible Ranges

No.	Parameter	Description <sup>a</sup>	Default value	Unit	Adjustable range
P01	$f_{\text{over}}$	Decay factor of surface flow (Equation 7.62)	0.5	$\text{m}^{-1}$	[0.1, 5.0]
P02	$f_{\text{drai}}$	Decay factor of subsurface flow (Equation 7.157)	2.5	$\text{m}^{-1}$	[0.1, 5.0]
P03	$Q_{\text{dm}}$	Maximum subsurface drainage (Equation 7.159)	0.0055	$\text{kg m}^{-2} \text{s}^{-1}$	[0.000001, 0.1]
P04	$S_y$	Average specific yield for aquifer (Equation 7.165)	0.2	—	[0.16, 0.24]
P05	poro_a	Parameter 1 (Slope) of mineral soil porosity in PTF (Equation 7.82)	0.00126	—	[0.001134, 0.001386]
P06	poro_b	Parameter 2 (Intercept) of mineral soil porosity in PTF (Equation 7.82)	0.489	—	[0.4401, 0.5379]
P07	bsw_a	Parameter 1 (Slope) of exponent B in PTF (Equation 7.84)	0.159	—	[0.1431, 0.1749]
P08	bsw_b	Parameter 2 (Intercept) of exponent B in PTF (Equation 7.84)	2.91	—	[2.619, 3.201]
P09	suc_a	Parameter 1 (Slope) of saturated matric potential in PTF (Equation 7.87)	0.0131	—	[0.01179, 0.01441]
P10	suc_b	Parameter 2 (Intercept) of saturated matric potential in PTF (Equation 7.87)	1.88	—	[1.692, 2.068]
P11	xk1	Parameter 1 of saturated hydraulic conductivity in PTF (Equation 7.90)	0.0070556	—	[0.00635004, 0.00776116]
P12	xk2	Parameter 2 of saturated hydraulic conductivity in PTF (Equation 7.90)	0.884	—	[0.7956, 0.9724]
P13	xk3	Parameter 3 of saturated hydraulic conductivity in PTF (Equation 7.90)	0.0153	—	[0.01377, 0.01683]
P14	tk1	Parameter 1 of saturated thermal conductivity in PTF (Equation 6.66)	8.8	—	[7.92, 9.68]
P15	tk2	Parameter 2 of saturated thermal conductivity in PTF (Equation 6.66)	2.92	—	[2.628, 3.212]

<sup>a</sup>The equation numbers in the description refer to those in Oleson et al. (2010). Parameters present in one equation are described by using the serial numbers (e.g., parameter 1 and parameter 2). PTF is the abbreviation of pedotransfer function. For convenience, 15 parameters are numbered from P01 to P15.

As the schematic diagram of  $A_{\text{cdf}}$  is shown in Figure 3, first, the RMSE value between simulation and observation in each grid cell of the study area is calculated. The cumulative distribution curve of RMSE values in all grid cells is then fitted to determine the area between the curve and the vertical  $y$ -axis. If the RMSE values of some grid cells become smaller due to ASMO optimization, the corresponding cumulative distribution curve of RMSE values could bulge toward the left, indicating the increase in the number of grid cells with smaller RMSE values. Hence the decrease of  $A_{\text{cdf}}$  value can reasonably represent the overall optimization at a spatial scale after parameter tuning (Huo, 2019).



**Figure 3.** Schematic diagram of the objective function in ASMO algorithm.

### 3.2. Evolutionary Particle Filter With Markov Chain Monte Carlo (EPFM)

The EPFM is a DA algorithm that uses Sampling Importance Resampling (SIR), MCMC, and genetic algorithm (GA) collectively to refine the prior distribution of the state variables of interest (i.e., particles) and generate a more accurate and informative posterior distribution in particle filtering (Abbaszadeh et al., 2018). The GA-based crossover and mutation operators are used to simultaneously mitigate the degeneracy of particles by intensifying the particle diversity and generating a more reliable posterior distribution without a need to increase the ensemble size. This GA-MCMC framework is the core part of the EPFM algorithm. Here we describe the EPFM algorithm in detail below.

Suppose that the dynamic evolution of a nonlinear physical model can be formulated as the following two differential equations:

$$x_t = f(x_{t-1}, u_t, \theta) + \omega_t \quad (9)$$

$$y_t = h(x_t) + \theta_t \quad (10)$$

where  $x_t$  and  $x_{t-1}$  are the vectors of state variables at time step  $t$  and  $t - 1$ , respectively.  $u_t$  and  $\theta$  are the vectors of uncertain forcing data and model parameters, respectively.  $\omega_t$  is the model error and  $y_t$  is a vector of observation data.  $\vartheta_t$  is the measurement error. In most cases,  $\omega_t$  and  $\vartheta_t$  are assumed to be independent of each other and can be expressed as the white noises following a Gaussian distribution.

Following Bayes' Law, the posterior distribution of the state variables at time step  $t$  can be formulated as follows:

$$p(x_t|y_{1:t}) = p(x_t|y_{1:t-1}, y_t) = \frac{p(y_t|x_t)p(x_t|y_{1:t-1})}{\int p(y_t|x_t)p(x_t|y_{1:t-1})dx_t} \quad (11)$$

$$p(x_t|y_{1:t-1}) = \int p(x_t, x_{t-1}|y_{1:t-1})dx_{t-1} = \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})dx_{t-1} \quad (12)$$

where  $p(y_t|x_t)$  is the likelihood for time step  $t$ ,  $p(x_t|y_{1:t-1})$  is the prior distribution, and  $p(y_t|y_{1:t-1})$  is the normalization factor. Generally, the marginal likelihood function  $p(y_{1:t})$  and its relevant normalization factor  $p(y_t|y_{1:t-1})$  can be expressed as follows:

$$p(y_{1:t}) = p(y_1) \prod p(y_t|y_{1:t-1}) \quad (13)$$

$$p(y_t|y_{1:t-1}) = \int p(y_t, x_t|y_{1:t-1})dx_t = \int p(y_t|x_t)p(x_t|y_{1:t-1})dx_t \quad (14)$$

Since it is difficult to solve Equation 11 analytically for nonlinear physical models due to the propagation of non-Gaussian distribution, we rely on an approximate solution, in which the posterior distribution of state variables  $p(x_t|y_{1:t})$  is estimated using a set of random replicates with associated weights:

$$p(x_t|y_{1:t}) \approx \sum_{i=1}^{N_{ens}} w_t^{i+} \delta(x_t - x_t^i) \quad (15)$$

where  $w_t^{i+}$ ,  $\delta(\cdot)$ , and  $N_{ens}$  are the posterior weight of  $i$  th particle, the Dirac delta function, and the total number of particles (or ensemble size), respectively. The posterior weight is calculated as:

$$w_t^{i+} = \frac{w_t^{i-} \cdot p(y_t|x_t^i)}{\sum_{i=1}^{N_{ens}} w_t^{i+} \cdot p(y_t|x_t^i)} \quad (16)$$

Note that  $w_t^{i-}$  is the prior particle weights and  $w_t^{i+}$  is the posterior particle weight.  $p(y_t|x_t^i)$  can be computed from the normalized likelihood  $L(y_t|x_t^i)$ . For simplicity, the likelihood is assumed to follow the Gaussian distribution:

$$p(y_t|x_t^i) = \frac{L(y_t|x_t^i)}{\sum_{i=1}^{N_{ens}} L(y_t|x_t^i)} \quad (17)$$

$$L(y_t|x_t^i) = \frac{1}{\sqrt{(2\pi R_t)}} \cdot \exp \left[ -\frac{1}{2R_t} (y_t - y_t^i)^2 \right] \quad (18)$$

To obtain approximate samples from  $p(x_t|y_{1:t})$ , a sampling operation is essential. Numerous studies have proposed a variety of Sequential Importance Sampling algorithms based on the selection of different importance functions (Arulampalam et al., 2002; Doucet et al., 2001). In the processing of sampling, particles with lower weights are more likely to be eliminated while particles with higher weights have a higher chance to be retained. However, a common problem with this sampling is that it inevitably leads to a loss of diversity among particles due to the degeneracy of random measures. Moradkhani, Hsu, Gupta, & Sorooshian et al. (2005) proposed the SIR algorithm to sample the particles with a probability greater than the uniform probability. We refer the readers to Moradkhani, Hsu, Gupta, & Sorooshian et al. (2005) to learn more about the principle of SIR algorithm.

Abbaszadeh et al. (2018) proposed a GA-MCMC framework to further mitigate particle degeneracy and improve particle filtering. In this approach, GA-MCMC is performed before the resampling step. The GA operators, crossover and mutation, are used to shuffle the particles and generate new particles that result in a more informative prior distribution. See Abbaszadeh et al. (2018) for more information on how the crossover and mutation operators are

used within the particle filtering. In summary, each particle in the GA algorithm is treated as a chromosome. Each chromosome is made up of several genes, just as each particle is made up of several state variables. GA is used to select and generate good offspring from parents according to Darwin's evolution theory. This is equivalent to selecting good particles and states. Here, the roulette wheel selection is chosen to select good chromosomes from parents by reference to the fitness value. The weights of particles are considered as the fitness value  $f_t^i$ :

$$f_t^i = w_t^{i+} \quad (19)$$

Note that the particles are sorted in descending order of their fitness values before performing roulette wheel selection. The roulette wheel selection is designed to correlate the fitness values with the probability of selection. Therefore, the probability of  $i$  th chromosome (or particle) being selected is:

$$P_t^i = \frac{f_t^i}{\sum f_t^i} \quad (20)$$

Given that the selection of a chromosome is analogous to randomly choosing a point on the wheel in which the size of each sector is proportional to the fitness value ( $f_t^i$ ) of each chromosome, the chromosomes with larger weights are more likely to be selected as new offspring chromosomes while those with smaller weights are less likely to be selected. This approach allows us to know how to select the parent particles. The next step is the crossover operation to generate new offspring from parents. In the GA-MCMC framework, the arithmetic crossover operation is suggested, that is, a pair of new particles are produced via the combination of parent particles. The formulas for the crossover operation are as follows:

$$x_{t-1}^{i'} = \xi \cdot x_{t-1}^i + (1 - \xi) \cdot x_{t-1}^j \quad (21)$$

$$x_{t-1}^{j'} = (1 - \xi) \cdot x_{t-1}^i + \xi \cdot x_{t-1}^j \quad (22)$$

where  $x_{t-1}^i$  and  $x_{t-1}^j$  are the parent particles,  $x_{t-1}^{i'}$  and  $x_{t-1}^{j'}$  are the pair of new offspring particles.  $\xi$  is a uniform random number with varying values between 0 and 1. The change in the value of  $\xi$  directly determines the transmission of information from the parent particle to the offspring particle. It is noteworthy that if  $\xi = 1$  the crossover will not occur, otherwise, all information in  $x_{t-1}^j$  ( $x_{t-1}^i$ ) will be transferred into  $x_{t-1}^{i'}$  ( $x_{t-1}^{j'}$ ) if  $\xi = 0$ . To maintain the stability of particle diversity in the selection process, a crossover probability ( $\rho_c$ ) needs to be assigned to determine how many particles are involved in the crossover operation.

Moreover, a mutation strategy is suggested to further enhance the particle diversity. This operation allows us to alter each chromosome (particle) by changing its gene. The value of mutation probability ( $\rho_m$ ) also requires to be assigned to determine how many the crossover particles are selected for mutation operation. In the mutation process, the mutated gene (i.e., state variable) in the chromosomes (i.e., particle) is altered randomly by the following formula:

$$x_{t-1}^{k'} = x_{t-1}^k + \eta \quad x_{t-1}^k \in \{x_{t-1}^{i'}, x_{t-1}^{j'}\} \quad \eta \sim N(0, \psi Var(x_{t-1}^{k-})) \quad (23)$$

where  $N(0, \psi Var(x_{t-1}^{k-}))$  denotes a random sample following the Gaussian distribution with mean zero and variance  $\psi Var(x_{t-1}^{k-})$ .  $\psi$  is a small tuning parameter that is set to 0.01 through ad hoc process.  $Var(x_{t-1}^{k-})$  is the variance of the states at time  $t - 1$ .

Up to this point, the new proposal state  $x_{t-1}^{i,p}$  is generated through the crossover and mutation processes. The number of the proposal states is the same as the particle size ( $N_{ens}$ ) of parents. The next step is to determine which particles should remain or be replaced with the old corresponding particle. A single MCMC step is performed to decide which offspring should be accepted or rejected. This acceptance/rejection step is essential because doing so can obtain an appropriate prior state distribution in each time step, and thus, estimating a more informative posterior distribution. To know whether to accept or reject the proposed states, the metropolis acceptance ratio  $\alpha$  is calculated as follows:

$$\alpha = \min \left( 1, \frac{p(x_t^{i,p}, \theta_t^{i-} | y_{1:t})}{p(x_t^{i-}, \theta_t^{i-} | y_{1:t})} \right) = \min \left( 1, \frac{p(y_{1:t} | x_t^{i,p}, \theta_t^{i-}) p(x_t^{i,p} | \theta_t^{i-}, y_{1:t-1})}{p(y_{1:t} | x_t^{i-}, \theta_t^{i-}) p(x_t^{i-} | \theta_t^{i-}, y_{1:t-1})} \right) \quad (24)$$

**Table 2**  
*Summary of Perturbation Parameters for the Atmospheric Forcing Dataset*

Forcing variables	Noise	Function distribution	Mean value	Standard diversion	Forcing cross correlation
Precipitation	Multiplicative	Lognormal	1	0.5	[1.0, -0.8, 0.5, 0.0]
Shortwave radiation	Multiplicative	Lognormal	1	0.3	[-0.8, 1.0, -0.5, 0.4]
Longwave radiation	Additive	Normal	0	20 W/m <sup>2</sup>	[0.5, -0.5, 1.0, 0.4]
Air temperature	Additive	Normal	0	1 K	[0.0, 0.4, 0.4, 1.0]

where  $p(x_t^{i,p}, \theta_t^{i-} | y_{1:t})$  is the proposed joint state-parameter probability density function.

$$p(x_t^{i,p}, \theta_t^{i-} | y_{1:t}) \propto p(y_t | x_t^{i,p}, \theta_t^{i-}) \cdot p(x_t^{i,p} | \theta_t^{i-}, y_{1:t-1}) \cdot p(\theta_t^{i-} | y_{1:t-1}) \quad (25)$$

$$x_t^{i,p} = M(x_{t-1}^{i,p}, u_t^i, \theta_t^{i-}) \quad (26)$$

where  $x_t^{i,p}$  and  $x_{t-1}^{i,p}$  represent the proposal states of  $i$  th particle at time  $t$  and  $t-1$ , respectively.  $u_t^i$  and  $\theta_t^{i-}$  are the forcing data for  $i$  th particle and the prior parameters for  $i$  th ensemble member, respectively.  $p(y_t | x_t^{i,p}, \theta_t^{i-})$  is computed via the same likelihood function shown in Equation 18.  $p(x_t^{i,p}, \theta_t^{i-} | y_{1:t})$  is calculated based on an assumption that the proposal states fit the marginal Gaussian distributions with a mean of  $\mu_t$  and variance of  $\sigma_t^2$ .  $\mu_t$  and  $\sigma_t^2$  are calculated as follows:

$$x_t^{i-} = f(x_{t-1}^{i+}, u_t^i, \theta_t^{i-}) \quad (27)$$

$$\mu_t = \sum \omega_{t-1}^{i+} x_t^{i-} \quad (28)$$

$$\sigma_t^2 = \sum \omega_{t-1}^{i+} (x_t^{i-} - \mu_t)^2 \quad (29)$$

where  $x_t^{i-}$  is the prior states for  $i$  th particle at time  $t$ .  $x_{t-1}^{i+}$  is the posterior states for  $i$  th particle at time  $t-1$ .  $u_t^i$  and  $\theta_t^{i-}$  are the same as the meanings in Equation 26.  $\omega_{t-1}^{i+}$  is the posterior weight for  $i$  th particle at time  $t-1$ . The GA-MCMC framework ensures that an appropriate prior state distribution is constructed in each time step before the parameter updating process. This process does not adjust the state variables so preserves the water-energy balance in modeling. To sum up, the EPFM algorithm results in a more informative estimation of the posterior distribution and improves the accuracy and reliability of DA.

In this study, we use an ensemble size of 50 to implement DA by the EPFM algorithm. Multiple atmospheric forcing variables, including precipitation, shortwave radiation, longwave radiation, and air temperature, are perturbed to represent the forcing uncertainties in the ensemble simulations of CLM. The perturbation of energy-related forcing variables is essential because the region of interest has a warming climate and it would to a certain extent influence the land surface hydrological processes such as the variations of SM. Table 2 summarizes the perturbation parameters for the forcing data used in this study. Their values are chosen based on Reichle et al. (2007, see Table 1 therein) and Han et al. (2014, see Table 1 therein). The perturbation is carried out by adding the spatially correlated noises generated through the Fast Fourier Transform approach (Park & Xu, 2009). The spatial correlation rule maintains the physical consistency of atmospheric balance. For example, the positive perturbation of incoming longwave radiation corresponds to the negative perturbation of incoming shortwave radiation and air temperature.

To characterize the uncertainties in initial conditions, we add a Gaussian white noise with a relative error of 25% to the initial SM values. It is also assumed that the error of SMAP-observed SM follows a Gaussian distribution with a relative error of 15%, which is consistent with our previous studies (Xu et al., 2020). A crossover probability ( $\rho_c$ ) is set to 80%, meaning that 80% of the particles are used for crossover operation and the remaining particles are kept unchanged to ensure stability. A mutation probability ( $\rho_m$ ) is set to 10%, indicating that 10% of the newly generated particles after crossover operation are randomly selected for mutation operation (Abbaszadeh et al., 2018). These parameters are set by a trial and error process. It is noteworthy that the EPFM algorithm is used here only to reduce the state uncertainty while the parameter uncertainty is calibrated by the aforementioned ASMO algorithm in a batch-processing mode.

### 3.3. Experiment Design

The CLM is run at the temporal and spatial resolutions of 3 hr and 0.25° grids, respectively. The daily outputs of simulated surface SM in the top two soil layers with a total thickness of ~5 cm are used for the analysis in this study. Considering that the parameterization scheme of CLM does not include the simulation of the impact of lake/wetland expansion on the surrounding SM in the study area of TP, the grid cells whose plant functional types contain lake or wetland are pre-masked and will not participate in the simulation experiment of this study. To assess the usefulness and effectiveness of ASMO and EPFM algorithms when they are used independently or jointly to improve the SM simulation skill of CLM, we designed four experiments as follows:

1. *CTRL experiment*: This represents the control (CTRL) simulation where the CLM is run by using the default parameter values. The simulation results can be used as a reference standard to determine the effectiveness of the following experiments
2. *ASMO experiment*: The ASMO algorithm is performed to optimize the CLM parameters. The optimum parameters are obtained when the simulated SM well matches the SMAP satellite observation
3. *EPFM experiment*: The EPFM algorithm is performed in combination with the un-calibrated CLM. To assess the performance of the EPFM algorithm, the posterior SM estimates are compared against the SMAP satellite observation through multiple deterministic and probabilistic performance measures
4. *ASMO + EPFM experiment*: The EPFM algorithm is performed in combination with the CLM whose parameters are calibrated with the ASMO algorithm. The performance of the ASMO + EPFM is assessed by using the similar procedure explained above for the EPFM experiment. How effective such a combined framework is will be shown in this experiment

Note that, in the CTRL experiment, the default model parameters are used and the model is spun-up by cycling the 2-yr (from 1 January 2016 to 31 December 2017) meteorological forcing dataset 10 times, until all state variables reach an equilibrium. Then, the initial state variables for the last loop are used as the initial conditions to conduct the ASMO and EPFM experiments. However, in the ASMO + EPFM experiment, since the default model parameter values are replaced by the optimal parameter values, the initial conditions are obtained by re-spinning up for the model with optimal parameter values.

### 3.4. Performance Measures

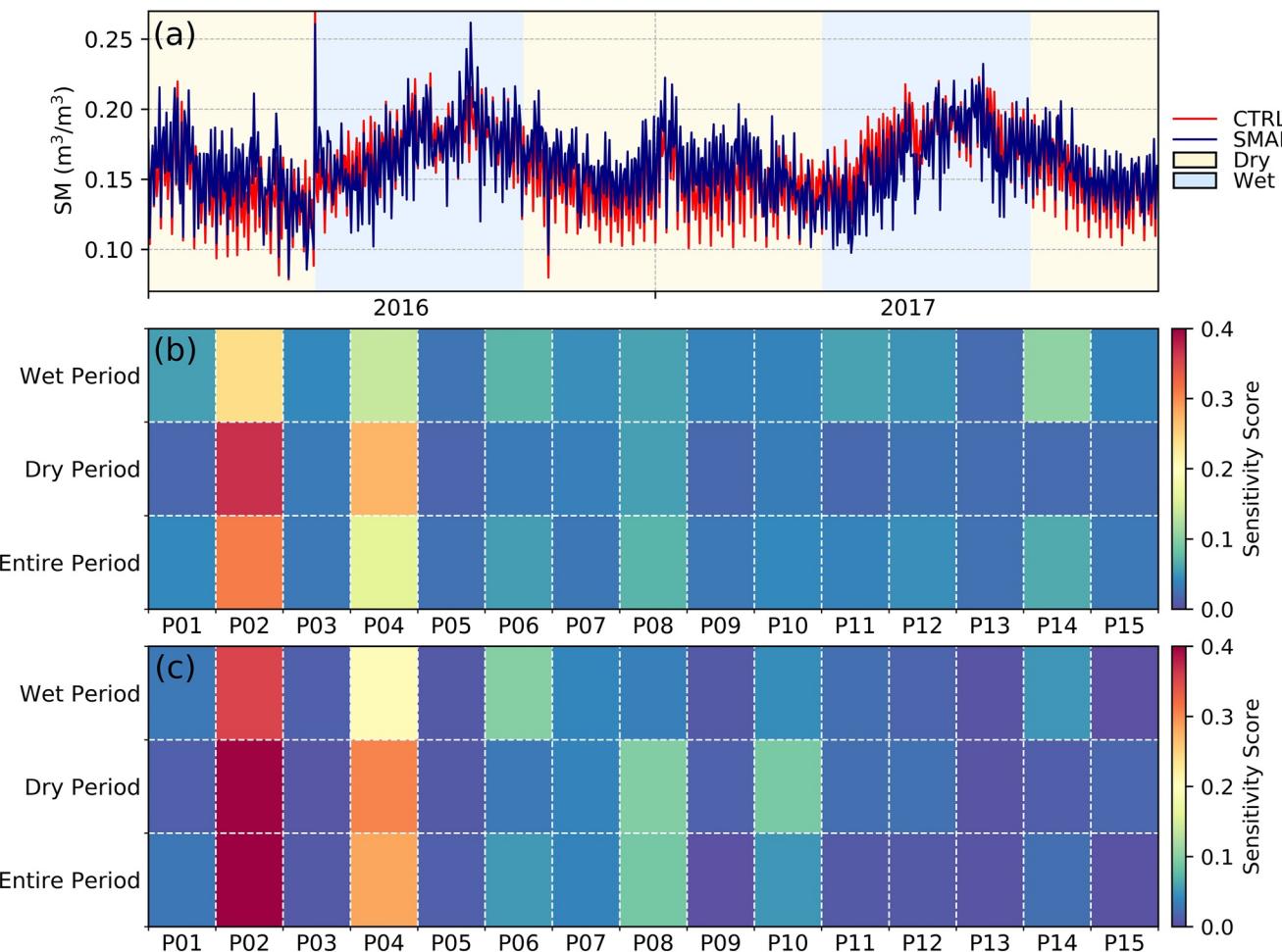
To assess the usefulness and effectiveness of the PO and DA algorithms used in this study, we employed two deterministic (i.e., Anomaly Correlation Coefficient [ACC] and Root-Mean-Square Error [RMSE]) and one probabilistic (i.e., Normalized RMSE Ratio [NRR]) performance measures as follows (Moradkhani, Sorooshian, Gupta, & Houser, 2005):

$$ACC = \frac{\sum_{t=1}^T (y_t - y_c)(y'_t - y_c)}{\sqrt{\sum_{t=1}^T (y_t - y_c)^2 \sum_{t=1}^T (y'_t - y_c)^2}} \quad (30)$$

$$RMSE = \sqrt{\frac{1}{T} \sum_{t=1}^T (y_t - y'_t)^2} \quad (31)$$

$$NRR = \frac{\frac{1}{T} \sum_{t=1}^T \sqrt{\left[ \left( \frac{1}{n} \sum_{i=1}^n y'_{t(n)} \right) - y_t \right]^2}}{\frac{1}{n} \sum_{i=1}^n \sqrt{\frac{1}{T} \sum_{t=1}^T (y'_{t(n)} - y_t)^2} \times \sqrt{\frac{(n+1)}{2n}}} \quad (32)$$

where  $y_t$  and  $y'_t$  are the observed and simulated SM values, respectively, at a certain time  $t$  in the study period of  $T$ .  $y_c$  is the climatological mean of SM calculated by averaging the SM values of each day from a long-term simulation (from January 1979 to December 2017).  $n$  is the ensemble size in DA. ACC reflects the temporal similarity of the observed and simulated anomalies, with the values ranging from -1 (best negative correlation) to +1 (perfect positive correlation). RMSE is an indicator that measures the difference between the simulated and observed values. It ranges from 0 (best) to  $+\infty$  (worst). NRR measures the ensemble spread and evaluates how confidently



**Figure 4.** (a) Comparison of the surface SM estimated and retrieved, respectively, through model CTRL experiment and SMAP satellite at daily time scale from 2016 to 2017. These time series highlight the difference between the two SM data for the wet (April–September) and dry (October–March) periods. The sensitivity scores of parameters to surface SM simulation during the wet, dry, and entire periods are calculated by (b) sum-of-tree and (c) multivariable adaptive regression spline, respectively.

the ensemble mean is statistically distinguishable from the ensemble spread. The ideal value of NRR is 1. However,  $\text{NRR} > 1$  and  $\text{NRR} < 1$  indicate that the ensemble has a too narrow and large spread, respectively. For more detailed information about NRR, we refer the readers to Moradkhani, Sorooshian, Gupta, & Houser (2005).

## 4. Results

### 4.1. Parameter Sensitivity Analysis

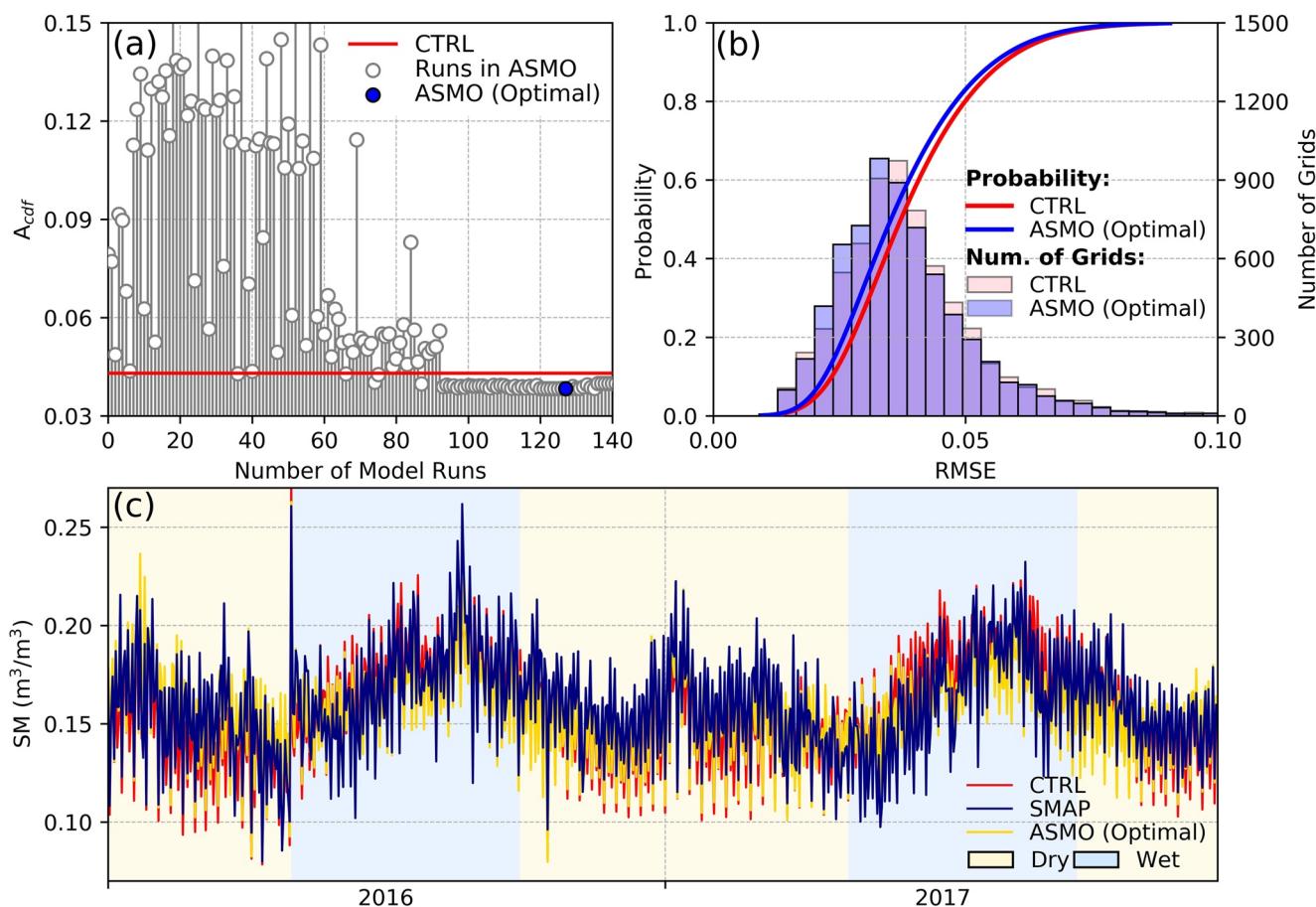
To assess the effectiveness of the model, Figure 4a first compares the 2016–2017 daily time series of regional mean surface SM estimated by CTRL simulation and SMAP observation in TP. The two deterministic performance measures, that is,  $\text{ACC} = 0.78$  and  $\text{RMSE} = 0.0107 \text{ m}^3/\text{m}^3$ , clearly indicate that the simulated SM closely follows the SMAP-observed SM during the wet period (from April to September). However, more discrepancies between the simulated and observed SM are detected during the dry period (from October to March) when the CTRL tends to underestimate SM versus SMAP observation (Chen et al., 2017; Li et al., 2018), with  $\text{ACC}$  and  $\text{RMSE}$  of 0.62 and  $0.0171 \text{ m}^3/\text{m}^3$ , respectively. This implies that reducing the SM discrepancies during the dry period is the key to improve the performance of the model. It should be noted that failure to identify the most dominant parameters may result in suboptimal performance of the PO algorithm (Gan et al., 2014; Gou et al., 2020). To address this problem, the parameter sensitivity analysis is performed using SOT and MARS techniques, and the results including the sensitivity scores of 15 parameters (Table 1) during the wet, dry, and entire

periods are reported in Figures 4b and 4c, respectively. As seen, depending on the type of the used sensitivity analysis algorithm (i.e., SOT and MARS) and the study period (i.e., dry, wet, and entire) for which the analysis is performed, the sensitivity score of each parameter can be different. The results showed that P02 and P04 are the most sensitive parameters, while P06, P08, P10, and P14 are marginally sensitive ones.

P02 ( $f_{drai}$ , decay factor of subsurface flow) and P04 ( $S_y$ , the average specific yield for aquifer) are identified as the most sensitive parameters to SM, consistent with the findings of previous studies (Göhler et al., 2013; Hou et al., 2012; Zhang et al., 2020). According to the CLM technical description document (Oleson et al., 2010), SM is closely related to subsurface discharge ( $q_{drai}$ ) that is calculated by the parameterization:  $q_{drai} = (1 - f_{imp}) \times Q_{dm} \times e^{-f_{drai} \times z_v}$ , where  $f_{imp}$  is the impermeable area fraction in frozen soil and  $z_v$  is the groundwater table depth. A larger value of  $f_{drai}$  means less  $q_{drai}$  generation given the same groundwater table depth and soil state. The decreasing  $q_{drai}$  then reduces the draining of deep layer SM, in turn decelerating the water infiltration from shallow to deep soil layers in the soil column (Göhler et al., 2013; Hou et al., 2012; Zhang et al., 2020). Similarly, a smaller value of P04 ( $S_y$ ) corresponds to a shallower  $z_v$  given an equal volume of groundwater stored in aquifers. It thereby makes the groundwater table possible to reach the upper soil column, which results in less  $q_{drai}$  and more surface SM according to the aforementioned parameterization (Niu et al., 2007; Zhang et al., 2020). However, if the groundwater table is within the soil column, the groundwater storage is set to a constant level that halts the water exchange between the soil column and aquifer (Oleson et al., 2010). In this case,  $S_y$  plays an insignificant role in SM simulation. This also indicates a relatively small sensitivity score of P04 ( $S_y$ ) against P02 ( $f_{drai}$ ; Figures 4b and 4c).

Among four marginally sensitive parameters, P06 is used for soil porosity calculation and its sensitivity score during the wet period is higher than that during the dry period (see Figures 4b and 4c). This may be due to that soil porosity tends to limit SM only when it increases to saturation state and such a case is more likely to occur during the wet period accompanied by abundant precipitation (Figure 4a). P08 is a parameter that is used to calculate the soil hydraulic conductivity and water potential (Li et al., 2013). A decrease in P08 gives a decrease in exponent  $B$  (Clapp & Hornberger, 1978) and an increase in volumetric SM at the wilting point for vegetation. These changes are adverse to the absorption of SM by vegetation. This is especially noticeable during the dry period where vegetation requires SM to live and grow (Oleson et al., 2010). P10 is related to the saturated matric potential that represents the force of binding water molecules to solid particles (Clapp & Hornberger, 1978). A larger value of P10 corresponds to a smaller saturated matrix potential, which reduces the soil resistance, and therefore, facilitates the SM infiltration and vegetation water absorption processes (Oleson et al., 2010). P14 is a parameter for calculating saturated thermal conductivity. An increase in P14 value renders an increase in ground heat flux, thus affecting the energy balance of the land surface. Although ground heat flux is a smaller term in the energy budget than latent and sensible heat fluxes, it may influence SM in the form of evaporation particularly over the TP region with the warming climate. This explains the relatively large sensitivity score of P14 during the wet period (Figures 4b and 4c) because evaporation during the wet period is commonly stronger than that during the dry period (Zhang et al., 2020).

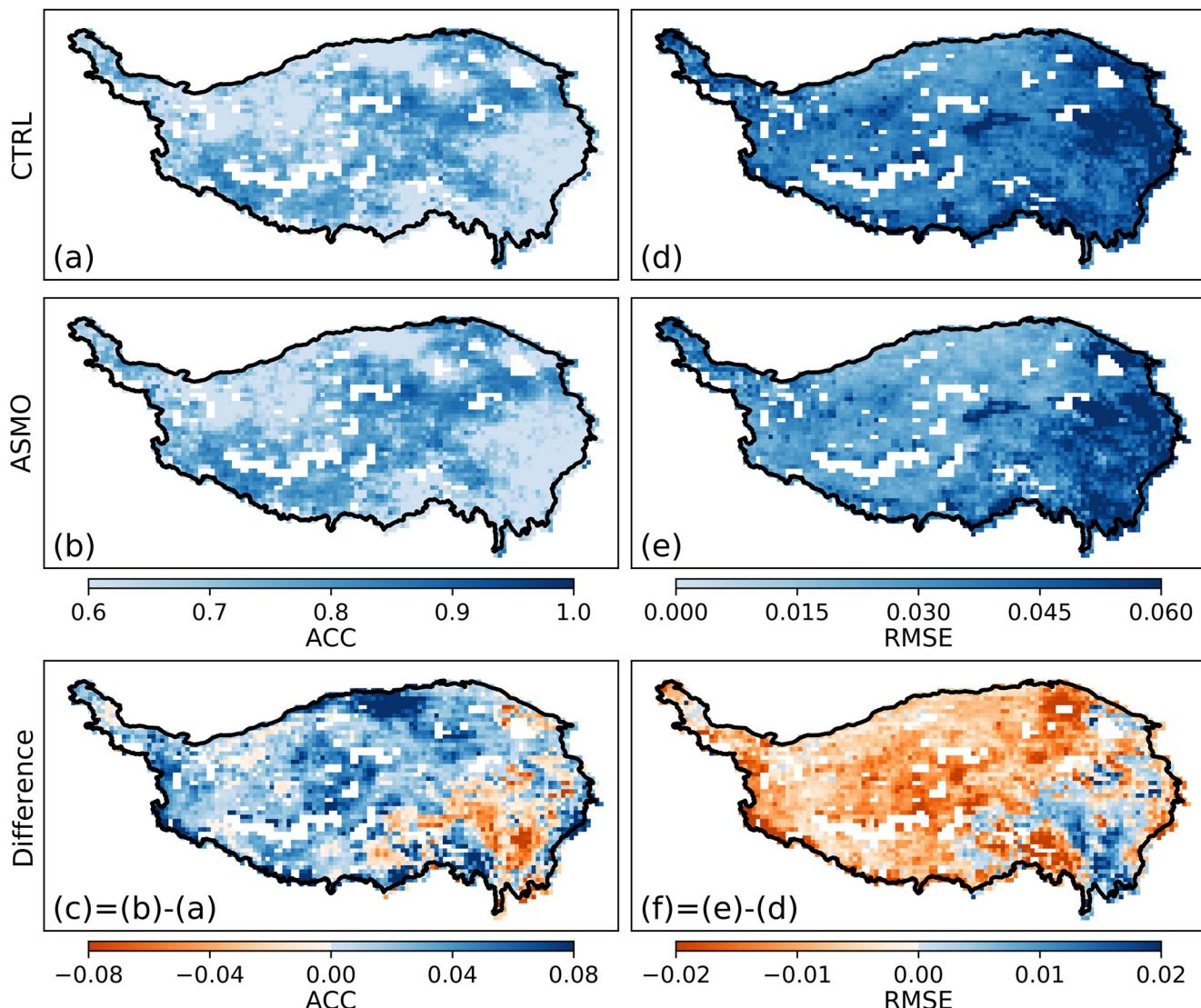
It is important to note that out of 11 parameters (from P05 to P15) derived from the pedotransfer functions, only P06, P08, P10, and P14 are identified as marginally sensitive parameters to SM in the modeling process. These parameters mostly act as the intercept-related parameters in the pedotransfer functions while others are slope-related parameters. Since the independent variables in the pedotransfer functions are the percentages of sand and clay contents whose values are <1 (Oleson et al., 2010), the equivalent changes of intercept-related parameter values should have a larger impact on the dependent variables (i.e., soil hydraulic and thermal properties) than those of slope-related parameter values. As noted in Hou et al. (2012) and Göhler et al. (2013), parameter sensitivity is not only time-varying but also spatially varying with different soil types and plant functional types in CLM modeling. For instance, a larger value of P02 could cause more evapotranspiration from shallow soil layers in an area covered by grasslands (i.e., the major land cover type in TP, Figure 1b), while in an area covered by barren the effect may not be distinguishable due to the water limitation (Hou et al., 2012; Ren et al., 2016). Among 15 parameters, thus, optimizing P02, P04, P06, P08, P10, and P14 is expected to improve the performance of CLM for simulating SM in TP. For more information about the functionality of the parameters, we refer the interested readers to the technical document of CLM (Oleson et al., 2010).



**Figure 5.** (a) Convergence of  $A_{cdf}$  values during the ASMO process. (b) The cumulative distribution functions and histogram statistics of RMSE for the simulated surface SM from optimal ASMO and CTRL experiments against the SMAP observation. (c) Time series of surface SM data obtained from the CTRL experiment, SMAP satellite, and ASMO algorithm at daily time scale throughout years 2016 and 2017.

#### 4.2. PO Performance

Figure 5a shows the  $A_{cdf}$  value convergence during the ASMO process over 140 simulations, including 60 initial samples and 80 adaptive optimizations. The reference value of  $A_{cdf}$  from CTRL experiment is 0.0431.  $A_{cdf}$  values range from 0.0429 to 0.1777 in the initial 60 model simulations, where a few of which outperformed the reference value. However,  $A_{cdf}$  value frequently falls below the reference value and gradually tends to be stable in the following 80 model simulations, which corresponds to the adaptive optimizations. This indicates that ASMO is effective in searching for optimal solutions (Duan et al., 2017; Wang et al., 2014). The optimal ASMO run is achieved in the 128th simulation where the  $A_{cdf}$  value (0.0383) becomes less than the reference value (0.0431). For more clarity, Figure 5b illustrates the histogram statistics and cumulative probability curves of RMSE values estimated by CTRL and ASMO experiments, respectively. ASMO experiment results in decreasing grid cells with RMSE values ranging  $0.03\text{--}0.07\text{ m}^3/\text{m}^3$  and increasing grid cells with RMSE values ranging  $0.01\text{--}0.03\text{ m}^3/\text{m}^3$  (Figure 5b). The cumulative probability curve of ASMO bulges toward the left relative to that of CTRL. When the daily time series of regional mean surface SM estimated by CTRL and ASMO is compared against the SMAP SM observation (Figure 5c), the results show that the simulated SM by ASMO experiment is more accurate than that estimated by the CTRL experiment. Both ACC and RMSE values indicate that the discrepancies between the SM estimated by CTRL experiment and SMAP observation ( $\text{ACC} = 0.62$  and  $\text{RMSE} = 0.0171\text{ m}^3/\text{m}^3$ ) during the dry period have been significantly alleviated in the ASMO experiment ( $\text{ACC} = 0.70$  and  $\text{RMSE} = 0.0130\text{ m}^3/\text{m}^3$ ). It is important to note that although the RMSE value during the wet period has slightly increased from  $0.0107$  to  $0.0111\text{ m}^3/\text{m}^3$ , during the entire study period the PO resulted in better model performance ( $\text{ACC} = 0.73$  and  $\text{RMSE} = 0.0121\text{ m}^3/\text{m}^3$ ).



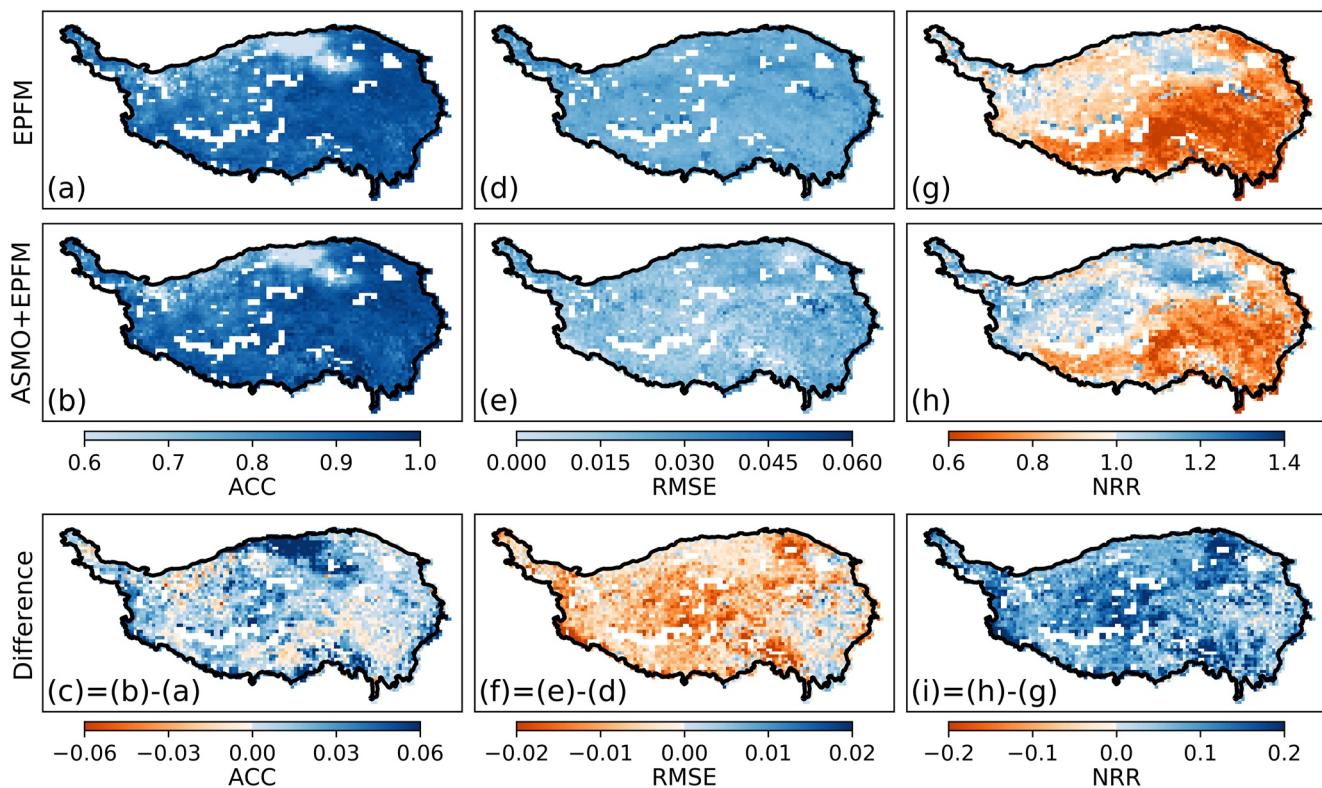
**Figure 6.** Spatial patterns of (a and b) ACC and (d and e) RMSE by comparing the surface SM estimated by CTRL and ASMO experiments against SMAP observation. The differences of (c) ACC and (f) RMSE are plotted for identifying the improvement of ASMO over the CTRL experiment.

Figure 6 illustrates the spatial patterns of ACC and RMSE by comparing the surface SM estimates from CTRL and ASMO experiments with SMAP observation. For the case of the CTRL experiment (Figures 6a and 6d), the ACC values are mostly larger than 0.6 and show a gradual increase and then decrease from northwest to southeast. Areas with large RMSE values are mainly located in the central and eastern parts of TP where the land cover types are mostly grassland with some scattered forests (Figures 1b and 6d). Nevertheless, for the case of the ASMO experiment (Figures 6b and 6e), the ACC and RMSE values, respectively, increase and decrease in the western and central areas of TP, which imply the improvement of SM simulation by the ASMO algorithm (Figures 6c and 6f). The results also reveal that in some southeast regions, the SM simulation degrades. It is expected that when ASMO is used together with EPFM, more uniform model performance measures are attained throughout the TP. Table 3 summarizes the spatial mean statistics of ACC and RMSE values for four experiments conducted in this study. The spatial average of ACC (RMSE) in the CTRL experiment is 0.65 ( $0.0389 \text{ m}^3/\text{m}^3$ ), while that in the ASMO experiment is 0.69 ( $0.0352 \text{ m}^3/\text{m}^3$ ), demonstrating that the ASMO algorithm is effective in optimizing the

**Table 3**

*Spatial Mean Statistics of ACC, RMSE, and NRR by Comparing the Surface SM Estimated by CTRL, ASMO, EPFM, and ASMO + EPFM Experiments Against SMAP Observation*

Experiment	ACC	RMSE	NRR
CTRL	0.65	0.0389	–
ASMO	0.69	0.0352	–
EPFM	0.86	0.0246	0.83
ASMO + EPFM	0.89	0.0213	0.92



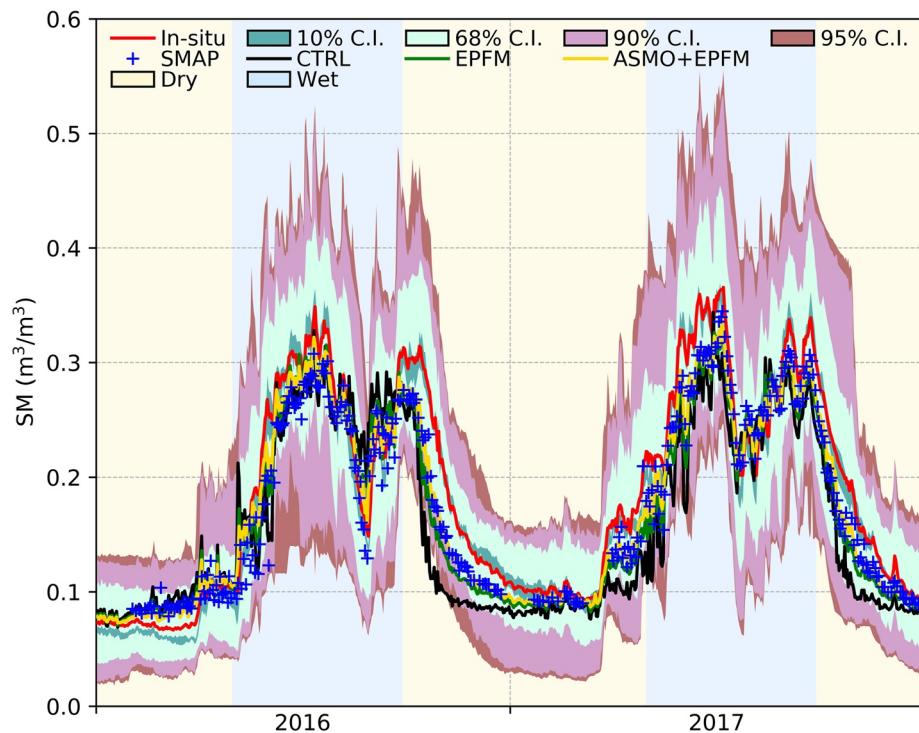
**Figure 7.** Spatial patterns of (a and b) ACC, (d and e) RMSE, and (g and h) NRR, by comparing the surface SM estimated by EPFM and ASMO + EPFM experiments against SMAP observation. (c), (f), and (i) show the improvement made by the ASMO when it is used in combination with the EPFM algorithm.

SM simulation of CLM and the newly used objective function ( $A_{cdf}$ ) is capable of generating more stable optimization performance across the TP.

#### 4.3. Performance of Optimization-Assimilation Framework

Figure 7 shows the spatial patterns of deterministic (ACC and RMSE) and probabilistic (NRR) performance measures by comparing the surface SM estimates from EPFM and ASMO + EPFM experiments with SMAP observation. Their corresponding spatial mean values are listed in Table 3. If compared with the CTRL experiment, the EPFM experiment results in more accurate SM estimates across the TP. Both ACC (Figure 7a) and RMSE (Figure 7d) values indicate the improvement made by the EPFM algorithm. This is especially more discernible across the southern and eastern parts of TP where the ASMO algorithm had resulted in poor SM estimates (with large RMSE values). The spatial average of ACC (RMSE) in EPFM experiment is 0.86 (0.0246 m<sup>3</sup>/m<sup>3</sup>) while that in CTRL experiment is 0.65 (0.0389 m<sup>3</sup>/m<sup>3</sup>; see Table 3). It is therefore concluded that EPFM can significantly improve the SM simulation of CLM throughout the TP.

As expected, the ASMO + EPFM results in a more accurate and reliable SM simulation across the TP compared to the EPFM. As shown in Figures 7a–7c, ASMO + EPFM yields an improved prediction of SM with higher ACC values almost across the entire TP, particularly over the areas where the single EPFM has resulted in sub-optimal performance (i.e., lower ACC values). The spatial pattern of RMSE from the ASMO + EPFM experiment is similar to that from the ASMO experiment (Figures 7d–7f). On average over TP, the ACC (RMSE) value in the ASMO + EPFM experiment is 0.89 (0.0213 m<sup>3</sup>/m<sup>3</sup>), which outperforms the 0.65 (0.0389 m<sup>3</sup>/m<sup>3</sup>) and 0.86 (0.0246 m<sup>3</sup>/m<sup>3</sup>) in the CTRL and EPFM experiments, respectively (Table 3). Although the EPFM and ASMO + EPFM deterministically result in similar model performance throughout the TP, probabilistically the combined framework yields more reliable SM estimates. The ASMO + EPFM algorithm results in the larger areas with NRR values close to 1 than the single EPFM algorithm (Figures 7g–7i), indicating the usefulness of ASMO in improving EPFM performance. This improvement also indicates that when ASMO is used together



**Figure 8.** Comparison of the 2016–2017 daily time series of regional mean surface SM estimated by the CTRL, EPFM, and ASMO + EPFM experiments as well as the SMAP observation and in situ measurements in the region of Naqu network.

with EPFM, it results in more reliable SM simulation (higher NRR values) compared to when EPFM is solely used. The spatial average of NRR values increases from 0.83 (EPFM) to 0.92 (ASMO + EPFM) (Table 3). To sum up, it can be said that the ASMO algorithm by providing pre-optimized parameters for CLM results in better assimilation performance by EPFM and nearly unbiased surface SM simulation.

#### 4.4. Independent Validation Using In Situ Observations

Figure 8 illustrates the daily time series of regional mean surface SM estimated by CTRL, EPFM, and ASMO + EPFM experiments as well as the SM observations from the SMAP satellite and in situ sites across the Naqu region in TP. Compared to in situ SM observations, the CTRL experiment shows a significant underestimation of SM values with the ACC and RMSE of 0.46 and  $0.0501 \text{ m}^3/\text{m}^3$ , respectively. This discrepancy is especially more evident in the dry period ( $\text{ACC} = 0.41$  and  $\text{RMSE} = 0.0593 \text{ m}^3/\text{m}^3$ ). The results also reveal that the SMAP satellite although underestimates the SM over this region, compared to SM estimates from the CTRL experiment, it is more consistent with the in situ observations ( $\text{ACC} = 0.73$ ,  $\text{RMSE} = 0.0311 \text{ m}^3/\text{m}^3$ ). The results show that the SM estimates from ASMO + EPFM are much closer to the in situ observations than those obtained from the EPFM. The ACC (RMSE) value between EPFM and in situ observation is 0.76 ( $0.0329 \text{ m}^3/\text{m}^3$ ) while that between ASMO + EPFM and observation is 0.81 ( $0.0249 \text{ m}^3/\text{m}^3$ ). The improvement is more discernible in the dry period where the ASMO + EPFM increases the ACC value of EPFM from 0.79 to 0.88 and decreases the RMSE value from 0.0331 to  $0.0243 \text{ m}^3/\text{m}^3$ . Considering the percentile ranges of site-observed SM displayed by the shaded areas in Figure 8, the SM values estimated by ASMO + EPFM fall more within the 68% confidence interval than those estimated by EPFM. Therefore, the validation against in situ observations suggests that the assimilation of remotely sensed SMAP SM observations into the CLM when its parameters are calibrated with ASMO algorithm results in better SM simulation compared to any other modeling configurations.

## 5. Discussion

Nowadays LSMs are becoming increasingly complex due mostly to the advances in parameterized representation of physical processes (Lawrence et al., 2011, 2019), which in turn increases the difficulty of treating the uncertainties in LSM simulations (Gong et al., 2015, 2016b). Previous studies in the field of PO often assume that all the sources of uncertainties in the modeling process can be attributed to parameter errors (Duan et al., 1994; Gong et al., 2015; Huang et al., 2016). Yet this is not necessarily true in the field of DA because the modeling can be influenced by many other uncertainties from various sources, including the model structural deficiencies, the errors in meteorological forcing data, and the description of initial boundary conditions (Duan et al., 2019; Liu et al., 2012; Moradkhani et al., 2018). Although some advanced dual state-parameter algorithms take these uncertainties into account (Abbaszadeh et al., 2020; Moradkhani, Hsu, Gupta, & Sorooshian et al., 2005; Moradkhani, Sorooshian, Gupta, & Houser, 2005), they may result in sub-optimal model performance. The underlying cause is that LSM generally includes a multitude number of parameters and accounting for their uncertainties jointly with model state variables would be very costly and even impractical when it is performed over a large study area with highly heterogeneous soil and vegetation features (Huang et al., 2016; Ren et al., 2016; Vrugt et al., 2005). To compensate for such a computational burden, researchers in the hydrologic DA community either ignore the parameter uncertainties in assimilation experiments, or carry out the studies at point scales (or over small regions), or simplify the parameterization scheme of the model in experiment design (Lievens et al., 2016; Moradkhani, Hsu, Gupta, & Sorooshian et al., 2005; Pathiraja et al., 2018). To address this issue, in this study, the ASMO-based PO technique is combined with the EPFM-based land surface DA system for practical and cost-effective monitoring of soil conditions over a large region.

The ASMO algorithm adopts a surrogate model to emulate the behavior of the physical model and then improves the efficiency of surrogate modeling-based optimization through an adaptive sampling strategy (Wang et al., 2014). The novel contribution of the ASMO algorithm lies mainly in the significant reduction of physical model runs required for parameter estimation of LSMs (Gong et al., 2016b; Zhang et al., 2020). However, since the optimization is performed on the surrogate model, the optimal solution obtained can only be taken as the approximate optimal solution of the physical model (Duan et al., 2017; Razavi et al., 2012). This can be partly attributed to the fact that the error response surface between the input and output data of LSM might be very bumpy and has many local optima (Gong et al., 2015; Huang et al., 2016). It is noteworthy that several factors may influence the optimization performance of the ASMO algorithm. First, the inaccurate identification of sensitive parameters could influence the efficacy of the optimization process (Gan et al., 2014; Saltelli et al., 2008). The sensitive parameters for the same model outputs of a LSM may differ spatially due to the varying responses of model outputs to spatially heterogeneous land surface features and climatic conditions (Hou et al., 2012; Ren et al., 2016). The idea of cell-to-cell sensitivity analysis followed by cell-to-cell parameter calibration has the potential to further improve the optimization performance of the ASMO algorithm. Second, the selection of a surrogate model to some extent may affect the optimization performance (Huang et al., 2016; Razavi et al., 2012). Although the GPR is used as the surrogate model in this study based on the findings of our previous studies (Gong et al., 2015; Wang et al., 2014), it is noted that GPR may not be a generic surrogate model for any problem and the best surrogate can be problem-dependent (Gong et al., 2016b; Huang et al., 2016; Razavi et al., 2012). Third, the objective function should be carefully selected in large-scale optimization studies (Duan et al., 2019; Gupta et al., 2003). A simple average of the objective function values of all grid cells in the study area may not be appropriate for the large-scale optimization experiment as it could result in smoothing out the spatial variability of model behavior. We recommend the new objective function ( $A_{cdj}$ ) used in this study as it is proved to be a better metric to generate spatially uniform optimization effects (Huo, 2019).

The EPFM algorithm adopts the evolutionary concept of the GA combined with the MCMC algorithm to effectively shuffle the particles before the resampling step to generate a more informative prior distribution that results in a more accurate representation of posterior distribution (Abbaszadeh et al., 2018, 2019). The main strength of the EPFM algorithm is the mitigation of particle degeneracy and sample impoverishment that exist in conventional particle filtering. Although the effectiveness of the EPFM algorithm has been proved in several assimilation experiments (Abbaszadeh et al., 2018, 2019, 2020; Gavahi et al., 2020; Xu et al., 2020), it is subject to some limitations. First, the EPFM algorithm does not explicitly account for the model structural error (Abbaszadeh et al., 2019)—this is indeed one main challenge in using particle filtering for hydrologic assimilation experiment (Liu et al., 2012; Liu & Gupta, 2007). However, our previous study suggests that this problem can be mitigated

when EPFM is integrated with a variational assimilation algorithm where the model structural error can be explicitly accounted for within the weak-constrained objective function of the variational approach (Abbaszadeh et al., 2019). Second, the EPFM algorithm may require large computational resources depending on the size and complexity of the model. For example, if the model state variables and parameters are simultaneously updated within the assimilation process, depending on the number of parameter sampling, the total number of physical model runs could increase exponentially compared to when state variables are solely updated (Moradkhani, Hsu, Gupta, & Sorooshian et al., 2005; Moradkhani, Sorooshian, Gupta, & Houser, 2005). This may be a huge simulation experiment for optimizing a complex LSM (i.e., CLM) and may require an unaffordable amount of computational resources. This is also the main motivation of the current study that utilizes the ASMO algorithm to reduce the parameter uncertainty and accounts for the uncertainties associated with model state variables and forcing data by the EPFM algorithm.

To sum up, this study provides a prototype for reducing the uncertainties of large-scale complex LSMs through using the proposed optimization-assimilation framework (i.e., a combination of ASMO and EPFM algorithms). It is expected that the results of this study can be improved if the CLM version 5 is used because it has made several updates in the model parameterization scheme, including not only the varying soil thickness in space but also the introduction of a dry surface layer-based soil evaporation resistance parameterization (Lawrence et al., 2019). We also would like to note that the other hydrologic states and fluxes, such as streamflow and evapotranspiration, have not been validated in our analysis since collecting the corresponding in situ measured data are not feasible in the TP of China due to the harsh plateau environment (Luo et al., 2020; Yang et al., 2013). The predicted SM values in this study can also be used for probabilistic agricultural drought monitoring in the TP region. Since this work is only performed over two years, it is not possible to use the simulated SM for drought monitoring, unless the present study would have been conducted over a longer period. We plan to do this in our future studies.

## 6. Conclusions

This study aims to propose a combined framework by jointly using the PO algorithm—Adaptive Surrogate Modeling Based Optimization (ASMO) and the DA algorithm—Evolutionary PF with MCMC (EPFM) to improve the predictive skill of the CLM modeling. We conducted this study in TP and used the SMAP SM dataset as observational data to optimize the model parameters and perform the DA. Two sensitivity analysis approaches, SOT and MARS, are first used to identify the most dominant CLM parameters for SM simulation in TP. A new objective function ( $A_{adj}$ ) is adopted in the ASMO implementation to make the optimization effects as spatially consistent as possible across the study area. After the optimal parameter values are obtained from the ASMO optimization in a batch-processing mode, the EPFM is implemented to assimilate SMAP SM observation into the CLM with and without optimization by ASMO to improve its predictive skill.

The main findings of this study are as follows: (a) out of 15 model parameters in CLM, 6 of them are identified as the most sensitive parameters to SM simulation in TP; (b) when ASMO is used together with EPFM, it results in better CLM SM simulation compared to any other modeling configurations; (c) the new objective function ( $A_{adj}$ ) is effective to make the improvement effects spatially consistent; (d) the assimilation of SMAP SM observations into the CLM improves its predictive skill. This study provides practical guidance to use the optimization and assimilation algorithms to improve the accuracy and reliability of complex LSM simulations over a large region. It is also expected that the proposed approach is beneficial for enhancing the effectiveness and usefulness of other LSMs.

## Data Availability Statement

The source codes of CLM are publicly available at <https://www.cesm.ucar.edu/models/cesm2/land/>. The datasets of China meteorological forcing and SMAP soil moisture retrievals are provided by the national Tibetan Plateau Data Center in China (<https://data.tpdc.ac.cn/en/data/8028b944-daaa-4511-8769-965612652c49/>) and the National Snow & Ice Data Center in United States (<https://nsidc.org/data/SPL3SMP/versions/6>), respectively. We thank Dr. Kun Yang for providing the in situ observed soil moisture dataset in Naqu network (<http://www.tpdc.ac.cn/zh-hans/data/ef949bb0-26d4-4cb6-acc2-3385413b91ee/?q=%E9%98%B3%E5%9D%A4>). The computational resources for this study are provided by the Tianhe-2 supercomputer in China.

## Acknowledgments

This study was jointly supported by the National Natural Science Foundation of China (51979004), the Second Tibetan Plateau Scientific Expedition and Research Program (2019QZKK0207), and the China Scholarship Council (201906040156).

## References

- Abbaszadeh, P., Gavahi, K., & Moradkhani, H. (2020). Multivariate remotely sensed and in-situ data assimilation for enhancing community WRF-Hydro model forecasting. *Advances in Water Resources*, 145, 103721. <https://doi.org/10.1016/j.advwatres.2020.103721>
- Abbaszadeh, P., Moradkhani, H., & Daescu, D. N. (2019). The quest for model uncertainty quantification: A hybrid ensemble and variational data assimilation framework. *Water Resources Research*, 55, 2407–2431. <https://doi.org/10.1029/2018WR023629>
- Abbaszadeh, P., Moradkhani, H., & Yan, H. (2018). Enhancing hydrologic data assimilation by evolutionary particle filter and Markov Chain Monte Carlo. *Advances in Water Resources*, 111, 192–204. <https://doi.org/10.1016/j.advwatres.2017.11.011>
- Arulampalam, M. S., Maskell, S., & Gordon, N. (2002). A tutorial on particle filters for online nonlinear/nonGaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50, 174–188. <https://doi.org/10.1109/78.978374>
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth International Group.
- Chen, Y., Yang, K., Qin, J., Cui, Q., Lu, H., La, Z., et al. (2017). Evaluation of SMAP, SMOS, and AMSR2 soil moisture retrievals against observations from two networks on the Tibetan Plateau. *Journal of Geophysical Research*, 122(11), 5780–5792. <https://doi.org/10.1002/2016JD026388>
- Clapp, R. B., & Hornberger, G. M. (1978). Empirical equations for some soil hydraulic properties. *Water Resources Research*, 14(4), 601–604. <https://doi.org/10.1029/WR014i004p00601>
- Clark, M. P., Fan, Y., Lawrence, D. M., Adam, J. C., Bolster, D., Gochis, D. J., et al. (2015). Improving the representation of hydrologic processes in Earth System Models. *Water Resources Research*, 51, 5929–5956. <https://doi.org/10.1029/2015WR017096>
- DeChant, C. M., & Moradkhani, H. (2012). Examining the effectiveness and robustness of sequential data assimilation methods for quantification of uncertainty in hydrologic forecasting. *Water Resources Research*, 48(4), W04518. <https://doi.org/10.1029/2011WR011011>
- Di, Z., Ao, J., Duan, Q., Wang, J., Gong, W., Shen, C., et al. (2019). Improving WRF model turbine-height wind-speed forecasting using a surrogate-based automatic optimization method. *Atmospheric Research*, 226, 1–16. <https://doi.org/10.1016/j.atmosres.2019.04.011>
- Di, Z., Duan, Q., Shen, C., & Xie, Z. (2020). Improving WRF typhoon precipitation and intensity simulation using a surrogate-based automatic parameter optimization method. *Atmosphere*, 11(1), 89. <https://doi.org/10.3390/atmos11010089>
- Doucet, A., Gordon, N., & Krishnamurthy, V. (2001). Particle filters for state estimation of jump Markov linear systems. *IEEE Transactions on Signal Processing*, 49(3), 613–624. <https://doi.org/10.1109/78.905890>
- 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., Pappenberger, F., Wood, A., Cloke, H., & Schaake, J. (2019). *Handbook of hydrometeorological ensemble forecasting*. Springer. <https://doi.org/10.1007/978-3-642-39925-1>
- 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>
- Duan, Q., Sorooshian, S., & Gupta, V. (1994). Optimal use of the SCE-UA global optimization method for calibrating watershed models. *Journal of Hydrology*, 158(3), 265–284. [https://doi.org/10.1016/0022-1694\(94\)90057-4](https://doi.org/10.1016/0022-1694(94)90057-4)
- Entekhabi, D., Njoku, E. G., O'Neill, P. E., Kellogg, K. H., Crow, W. T., Edelstein, W. N., et al. (2010). The soil moisture active passive (SMAP) mission. *Proceedings of the IEEE*, 98(5), 704–716. <https://doi.org/10.1109/JPROC.2010.2043918>
- Friedman, J. H. (1991). Multivariate adaptive regression splines. *Annals of Statistics*, 19(1), 1–67. <https://doi.org/10.1214/aos/1176347963>
- Gan, Y., Duan, Q., Gong, W., Tong, C., Sun, Y., Chu, W., et al. (2014). A comprehensive evaluation of various sensitivity analysis methods: A case study with a hydrological model. *Environmental Modelling & Software*, 51, 269–285. <https://doi.org/10.1016/j.envsoft.2013.09.031>
- Gavahi, K., Abbaszadeh, P., Moradkhani, H., Zhan, X., & Hain, C. (2020). Multivariate assimilation of remotely sensed soil moisture and evapotranspiration for drought monitoring. *Journal of Hydrometeorology*, 21(10), 2293–2308. <https://doi.org/10.1175/JHM-D-20-0057.1>
- Göhler, M., Mai, J., & Cuntz, M. (2013). Use of eigendecomposition in a parameter sensitivity analysis of the Community Land Model. *Journal of Geophysical Research*, 118, 904–921. <https://doi.org/10.1002/jgrc.20072>
- 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. (2016a). An intercomparison of sampling methods for uncertainty quantification of environmental dynamic models. *Journal of Environmental Informatics*, 28(1), 11–24. <https://doi.org/10.3808/jei.201500310>
- Gong, W., Duan, Q., Li, J., Wang, C., Di, Z., Ye, A., et al. (2016b). Multiobjective adaptive surrogate modeling-based optimization for parameter estimation of large, complex geophysical models. *Water Resources Research*, 52(3), 1984–2008. <https://doi.org/10.1029/2015WR018230>
- Gou, J., Miao, C., Duan, Q., Tang, Q., Di, Z., Liao, W., et al. (2020). Sensitivity analysis-based automatic parameter calibration of the VIC model for streamflow simulations over China. *Water Resources Research*, 56, e2019WR025968. <https://doi.org/10.1029/2019WR025968>
- Gupta, H. V., Thieman, M., Trosset, M., & Sorooshian, S. (2003). Reply to comment by K. Beven and P. Young on “Bayesian recursive parameter estimation for hydrologic models”. *Water Resources Research*, 39(5). <https://doi.org/10.1029/2002wr001405>
- Han, X., Franssen, H.-J. H., Montzka, C., & Vereecken, H. (2014). Soil moisture and soil properties estimation in the Community Land Model with synthetic brightness temperature observations. *Water Resources Research*, 50(7), 6081–6105. <https://doi.org/10.1002/2013WR014586>
- He, J., Yang, K., Tang, W., Lu, H., Qin, J., Chen, Y., & Li, X. (2020). The first high-resolution meteorological forcing dataset for land process studies over China. *Scientific Data*, 7(1), 25. <https://doi.org/10.1038/s41597-020-0369-y>
- Hou, Z., Huang, M., Leung, L. R., Lin, G., & Ricciuto, D. M. (2012). Sensitivity of surface flux simulations to hydrologic parameters based on an uncertainty quantification framework applied to the Community Land Model. *Journal of Geophysical Research*, 117, D15108. <https://doi.org/10.1029/2012JD017521>
- 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*, 121, 7548–7563. <https://doi.org/10.1002/2015JD024339>
- Huo, X. (2019). *Parameter uncertainty analysis for large scale Noah-MP Land Surface Model (in Chinese) [PhD thesis]*. Beijing Normal University.
- Jones, D. R., Schonlau, M., & Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13, 455–492. <https://doi.org/10.1023/A:1008306431147>
- Korobov, N. M. (1959). Computation of multiple integrals by the method of optimal coefficients. *Vestnik Moskovskogo Universiteta Seriya 1 Matematika Mekhanika*, 4, 19–25.
- Lawrence, D. M., Fisher, R. A., Koven, C. D., Oleson, K. W., Swenson, S. C., Bonan, G., et al. (2019). The Community Land Model version 5: Description of new features, benchmarking, and impact of forcing uncertainty. *Journal of Advances in Modeling Earth Systems*, 11, 4245–4287. <https://doi.org/10.1029/2018MS001583>

- Lawrence, D. M., Oleson, K. W., Flanner, M. G., Thornton, P. E., Swenson, S. C., Lawrence, P. J., et al. (2011). Parameterization improvements and functional and structural advances in version 4 of the Community Land Model. *Journal of Advances in Modeling Earth Systems*, 3(1), M03001. <https://doi.org/10.1029/2011MS00045>
- Li, C., Lu, H., Yang, K., Han, M., Wright, J., Chen, Y., et al. (2018). The evaluation of SMAP enhanced soil moisture products using high-resolution model simulations and in-situ observations on the Tibetan Plateau. *Remote Sensing*, 10(4), 535. <https://doi.org/10.3390/rs10040535>
- Li, J., Duan, Q., Gong, W., Ye, A., Dai, Y., Miao, C., et al. (2013). Assessing parameter importance of the Common Land Model based on qualitative and quantitative sensitivity analysis. *Hydrology and Earth System Sciences*, 17(8), 3279–3293. <https://doi.org/10.5194/hess-17-3279-2013>
- Lievens, H., De Lannoy, G. J. M., Al Bitar, A., Drusch, M., Dumedah, G., Hendricks, H.-J., et al. (2016). Assimilation of SMOS soil moisture and brightness temperature products into a Land Surface Model. *Remote Sensing of Environment*, 180, 292–304. <https://doi.org/10.1016/j.rse.2015.10.033>
- Liu, Y., & Gupta, H. V. (2007). Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. *Water Resources Research*, 43(7). <https://doi.org/10.1029/2006WR005756>
- Liu, Y., Weerts, A. H., Clark, M., Hendricks Franssen, H. J., Kumar, S., Moradkhani, H., et al. (2012). Advancing data assimilation in operational hydrologic forecasting: Progresses, challenges, and emerging opportunities. *Hydrology and Earth System Sciences*, 16, 3863–3887. <https://doi.org/10.5194/hess-16-3863-2012>
- Luo, Q., Yang, K., Chen, Y., & Zhou, X. (2020). Method development for estimating soil organic carbon content in an alpine region using soil moisture data. *Science China Earth Sciences*, 63, 591–601. <https://doi.org/10.1007/s11430-019-9554-8>
- Moradkhani, H., DeChant, C. M., & Sorooshian, S. (2012). Evolution of ensemble data assimilation for uncertainty quantification using the particle filter-Markov chain Monte Carlo method. *Water Resources Research*, 48, L22702. <https://doi.org/10.1029/2012WR012144>
- Moradkhani, H., Hsu, K.-L., Gupta, H., & Sorooshian, S. (2005). Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter. *Water Resources Research*, 41, W05012. <https://doi.org/10.1029/2004WR003604>
- Moradkhani, H., Nearing, G., Abbaszadeh, P., & Pathiraja, S. (2018). Fundamentals of data assimilation and theoretical advances. In Q. Duan, F. Pappenberger, A. Wood, H. L. Cloke, & J. C. Schaake (Eds.), *Handbook of Hydrometeorological Ensemble Forecasting* (pp. 1–26). Springer. [https://doi.org/10.1007/978-3-642-40457-3\\_30-1](https://doi.org/10.1007/978-3-642-40457-3_30-1)
- Moradkhani, H., Sorooshian, S., Gupta, H. V., & Houser, P. R. (2005). Dual state-parameter estimation of hydrological models using ensemble Kalman filter. *Advances in Water Resources*, 28(2), 135–147. <https://doi.org/10.1016/j.advwatres.2004.09.002>
- Niu, G. Y., Yang, Z. L., Dickinson, R. E., Gulden, L. E., & Su, H. (2007). Development of a simple groundwater model for use in climate models and evaluation with Gravity Recovery and Climate Experiment data. *Journal of Geophysical Research*, 112(D7). <https://doi.org/10.1029/2006jd007522>
- Oleson, K. W., Lawrence, D. M., Bonan, G. B., Flanner, M. G., Kluzek, E., Lawrence, P. J., et al. (2010). *Technical description of version 4.0 of the community land model (CLM)*, NCAR (Technical note NCAR/TN-478+STR). University Corporation for Atmospheric Research. <https://doi.org/10.5065/D6FB50WZ>
- O'Neill, P., Chan, S., Bindlish, R., Chaubell, M., Colliander, A., Chen, F., et al. (2019). *Calibration and validation for the L2/3\_SM\_P version 6 and L2/3\_SM\_P\_E version 3 data products, SMAP project*, JPL D-56297. Jet Propulsion Laboratory.
- Park, S., & Xu, L. (2009). *Data assimilation for atmospheric, oceanic and hydrologic applications*. Springer.
- Pathiraja, S., Moradkhani, H., Marshall, L., Sharma, A., & Geenens, G. (2018). Data-driven model uncertainty estimation in hydrologic data assimilation. *Water Resources Research*, 54, 1252–1280. <https://doi.org/10.1002/2018WR022627>
- Qiu, J. (2008). The third pole. *Nature*, 454(7203), 393–396. <https://doi.org/10.1038/454393a>
- Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.
- Razavi, S., Tolson, B. A., & Burn, D. H. (2012). Review of surrogate modeling in water resources. *Water Resources Research*, 48, w07401. <https://doi.org/10.1029/2011WR011527>
- Reichle, R. H., Koster, R. D., Liu, P., Mahanama, S. P. P., Njoku, E. G., & Owe, M. (2007). Comparison and assimilation of global soil moisture retrievals from the Advanced Microwave Scanning Radiometer for the Earth Observing System (AMSR-E) and the scanning multichannel microwave radiometer (SMMR). *Journal of Geophysical Research*, 112, D09108. <https://doi.org/10.1029/2006JD008033>
- Ren, H., Hou, Z., Huang, M., Bao, J., Sun, Y., Tesfa, T., et al. (2016). Classification of hydrological parameter sensitivity and evaluation of parameter transferability across 431 US MOPEX basins. *Journal of Hydrology*, 536, 92–108. <https://doi.org/10.1016/j.jhydrol.2016.02.042>
- Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., et al. (2008). *Global sensitivity analysis: The primer*. John Wiley & Sons.
- Song, C., Huang, B., Richards, K., Ke, L., & Phan, V. H. (2014). Accelerated lake expansion on the Tibetan Plateau in the 2000s: Induced by glacial melting or other processes? *Water Resources Research*, 50, 3170–3186. <https://doi.org/10.1002/2013WR014724>
- Vrugt, J. A., Diks, C. G., Gupta, H. V., Bouten, W., & Verstraten, J. M. (2005). Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation. *Water Resources Research*, 41, W01017. <https://doi.org/10.1029/2004WR003059>
- Wang, C., Duan, Q. Y., Gong, W., Ye, A. Z., Di, Z. H., & Miao, C. Y. (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>
- Xu, L., Abbaszadeh, P., Moradkhani, H., Chen, N., & Zhang, X. (2020). Continental drought monitoring using satellite soil moisture, data assimilation and an integrated drought index. *Remote Sensing of Environment*, 250, 112028. <https://doi.org/10.1016/j.rse.2020.112028>
- Xu, T., Valocchi, A. J., Ye, M., & Liang, F. (2017). Quantifying model structural error: Efficient Bayesian calibration of a regional groundwater flow model using surrogates and a data-driven error model. *Water Resources Research*, 53, 4084–4105. <https://doi.org/10.1002/2016WR019831>
- Yang, K., & He, J. (2018). *China meteorological forcing dataset (1979–2018)*. National Tibetan Plateau Data Center. <https://doi.org/10.11888/AtmosphericPhysics.tpe.249369.file>
- Yang, K., Qin, J., Zhao, L., Chen, Y. Y., Tang, W. J., Han, M. L., et al. (2013). A multiscale soil moisture and freeze-thaw monitoring network on the third pole. *Bulletin of the American Meteorological Society*, 94, 1907–1916. <https://doi.org/10.1175/bams-d-12-00203.1>
- Yang, K., Wu, H., Qin, J., Lin, C. G., Tang, W. J., & Chen, Y. Y. (2014). Recent climate changes over the Tibetan Plateau and their impacts on energy and water cycle: A review. *Global and Planetary Change*, 112, 79–91. <https://doi.org/10.1016/j.gloplacha.2013.12.001>
- Yang, K., Ye, B. S., Zhou, D. G., Wu, B. Y., Foken, T., Qin, J., & Zhou, Z. Y. (2011). Response of hydrological cycle to recent climate changes in the Tibetan Plateau. *Climatic Change*, 109(3–4), 517–534. <https://doi.org/10.1007/s10584-011-0099-4>
- Yao, T., Thompson, L., Yang, W., Yu, W., Gao, Y., Guo, X., et al. (2012). Different glacier status with atmospheric circulations in Tibetan Plateau and surroundings. *Nature Climate Change*, 2, 663–667. <https://doi.org/10.1038/nclimate1580>
- Zhang, C., Di, Z., Duan, Q., Xie, Z., & Gong, W. (2020). Improved land evapotranspiration simulation of the Community Land Model using a surrogate-based automatic parameter optimization method. *Water*, 12(4), 943. <https://doi.org/10.3390/w12040943>