# Parameter Calibration in Global Soil Carbon Models Using Surrogate-based Optimization

Haoyu Xu[1], Tao Zhang[1, 2], Yiqi Luo[2, 3], Xin Huang[1, 2], Wei Xue[1, 2]

[1]Department of Computer Science and Technology, Tsinghua University, Beijing 100084

[2]Department of Earth System Science, Ministry of Education Key Laboratory for Earth System Modelling, Tsinghua University, Beijing 100084

[3]Center for Ecosystem Science and Society, Northern Arizona University, USA

*Correspondence to*: Wei Xue (xuewei@tsinghua.edu.cn)

**Abstract.** Soil organic carbon (SOC) has a significant effect on the carbon emission and climate change. However, current SOC prediction accuracy of most models is very low. Most evaluation studies indicate that the prediction error mainly comes from parameter uncertainties, which can be improved by parameter calibration. Data assimilation technique has been successfully employed for parameter calibration of SOC models. However, data assimilation algorithms such as sampling-based Bayesian Markov Chain Monte Carlo (MCMC) generally require a large amount of computation cost and are not appropriate for complex global land models. This study proposes a new parameter calibration method based on surrogate optimization techniques for improving the prediction accuracy of SOC. Experiments on three types of soil carbon cycle models, including Community Land Model with Carnegie-Ames-Stanford Approach biogeochemistry sub-model (CLM-CASA') and two microbial models show that surrogate-based optimization method is more effective and efficient than MCMC on both accuracy and cost. Compared to the predictions using the tuned parameter values through Bayesian MCMC, the root mean squared errors (RMSEs) between the predictions using the calibrated parameter values with surrogate-base optimization and the observations could be reduced up to 12% for different SOC models. Meanwhile, the corresponding computation cost required is only one thousandth of that with Bayesian MCMC.

## 1 Introduction

Soil organic carbon (SOC) is the largest pool of global land carbon (Todd Brown et al., 2013; Luo et al., 2015). The emission of CO2, the most important greenhouse gas, from the land ecosystems greatly depends on the amount of carbon stored in soils. Simultaneously, more emitted CO2 increases the climate warming (Houghton et al., 2001) and the climate warming intensifies soil carbon release, resulting in a positive feedback cycle between the carbon cycle and climate warming (Melillo et al., 2003;

Friedingstein et al., 2006; Luo, 2007). In the fifth Coupled Model Intercomparison Project (CMIP5), the outputs of 11 Earth system models (ESMs) show great uncertainty in the SOC predictions and simulations. Despite the similarity in model structures (Huang et al., 2017), simulated soil carbon content varies six-fold among the models with the simulation results ranging from 510 to 3040 PgC (Todd-Brown et al., 2013). There are only half of 11 models whose the predicted global total SOC falls within the estimated range of the Harmonized World Soil Database (HWSD) and the highest correlation coefficient between the model output and the observation is even lower than 0.4 (Todd Brown et al., 2013; Luo et al., 2015).

Considering the high similarity in the carbon cycle component structures of the 11 ESMs, the difference of SOC simulations mainly comes from the parameterizations (Todd Brown et al., 2013); thus parameter calibration can improve the simulation of carbon cycle obviously (Luo et al., 2016). However, the parameter calibration with global observations has not been widely applied owing to the high computational cost. Take an example, the Bayesian Markov chain Monte Carlo (MCMC) algorithm has ever been used for parameter calibration of SOC simulation and microbial process successfully (Harauk et al., 2014 and 2015). Bayesian MCMC is a sampling-based approach and usually requires a large number of simulations for building an acceptable parameter chain. For instance, over 500,000 simulations are required during parameter calibration of soil carbon models (Harauk et al. 2014 and 2015). Even using high performance computer to provide the computation power, some carbon-enabled complex models, like the latest version of Community Land Model (CLM), require a very long spin-up time for carbon cycle simulation, leading to several hours or days for one simulation. Therefore, Bayesian MCMC cannot be extended to expensive global land models. More effective and efficient parameter calibration algorithms are intensively demanding.

Parameter calibration of SOC models can be formulated as an optimization problem that aims to minimize the output of a metric function. This metric function evaluates the difference between the outputs of model simulation and the corresponding observations and returns a single value (e.g. RMSE) to represent the model error. Global optimization algorithms are introduced to find the minimum value of the non-linear, non-convex and black-box problems (Hapuarachchi et al., 2001; Ma H, et al., 2006; Rocha H, 2008). Unfortunately, the number of required simulations of most global optimization is still very large.

To reduce the number of simulations and decrease the computational cost, we for the first time present the surrogate-based optimization method for calibrating the soil carbon models. Surrogate models serve as computationally cheap approximations of expensive simulation models (Booker et al., 1999), such as complex geoscientific models. During the optimization process, the surrogate model can be used to determine the new promising point in the parameter space at which originally the expensive simulation model has to be evaluated. With the help of surrogate model, many unnecessary simulations with bad parameter values, which lead to high prediction errors, are avoided. Surrogate-based optimization has been proved to be able to find the near-optimal parameter values within only few hundred simulations for different problems (Aleman et al., 2009; Giunta et al., 1997; Regis, 2011; Simpson et al., 2001).

Most studies on both global and surrogate optimizations focus on the mathematical function benchmarks like Comparing Continuous Optimisers, abbreviated as COCO (Hansen et al., 2010; Wang and Duan, 2014). However, the optimization of the mathematical functions may extremely different from the parameter calibration of complex real-world models. In this paper, for the first time, we try to exploit state-of-the-art surrogate optimization method for the parameter calibration of three types of soil organic carbon (SOC) models and compare the performance of surrogate-based optimization to advanced global optimization algorithms and the data assimilation method. The evaluation and analysis based on these representative SOC models prove that surrogate-based optimization has potential to be extended to other complex SOC models or even Community Land Model (CLM).

This paper is organized as follows. Section 2 presents the structure and parameters of three representative SOC models. In Section 3, we introduce the algorithm design of surrogate-based optimization. The parameter calibration results and the analysis of different parameter calibration algorithms are presented in Section 4. Section 5 discusses the calibrated results by using the surrogate-based optimization. Finally, we draw conclusions in Section 6.

## 2 Global Land Carbon Models and Metrics

Earth system models (ESMs) are the fundamental tools for simulating climate impacts on carbon cycle at the global scale and there are many similarities among structures of different ESMs. They define different carbon pools such as soil and litter pools. Carbon transfer among these pools by respirations (Todd Brown et al., 2013; Weng and Luo, 2011). In this study, we selected three types of SOC models. These models are summarized and extracted from global land models. They keep the key equations and structures of carbon transferring and can be regarded as the representative models in this field. The first model is the soil carbon component of the Community Land Model coupled with Carnegie-Ames-Stanford Approach biogeochemistry submodel (CLM-CASA') (Oleson et al., 2004, 2008). The CLM is the land model for the Community Earth System Model (CESM), which is a widely used earth system model in the world. It is also a collaborative project between scientists in the Terrestrial Sciences Section (TSS) and the Climate and Global Dynamics Division (CGD) at the National Center for Atmospheric Research (NCAR) and the CESM Land Model Working Group. The other two SOC models are microbial models. These two models consider microbial biomass dynamics explicitly which most conventional SOC models like CLM-CASA' don't take into account. The calibrated models explain more variability of the observed SOC (Hararuk et al., 2015). Considering the similarity in model structures, these three types of SOC models can represent the structures of the SOC components in most current land models (Luo and Weng, 2011).

### 2.1 CLM-CASA' C-only Version Model

The CLM-CASA' includes biogeophysics and biogeochemistry sub-models based on the CLM3.5. Carbon transferring among various plants, litter, and soil pools are simulated in the biogeochemistry sub-models (Parton et al., 1993). The influx and efflux of each pool determine the carbon content of that pool. Carbon influx into the whole system is partitioned into three live biomass pools. Carbon efflux is heterotrophic respiration which is determined by the decomposition rate of organic carbon in each pool. Heterotrophic respiration is influenced by environmental conditions (especially, temperature and soil moisture), soil texture, tissue lignin and available tissue nitrogen content.

The CLM-CASA' model simulates soil carbon decomposition as a first-order decay process (Todd-Brown et al., 2013b). Based on theoretical analysis, carbon cycle of most ESMs can be summarized to differential equations with linear coefficients (Luo and Weng, 2011; Xia et al., 2013).

$$\frac{dX(t)}{dt} = A\xi(t)KX(t) + BU(t) \quad (1)$$

Where $X(t)$ is the carbon content of different pools; $\frac{dX(t)}{dt}$ is the change of the carbon content; $A$ is a matrix of partitioning coefficients among different pools; $\xi(t)$ and $K$ are both diagonal matrixes, representing environmental factors and baseline carbon exit rates, respectively; $U(t)$ is the carbon influx into the whole system and $B$ represents the partitioning coefficients of the carbon influx. The steady state solution of equation is given by Eq.2 (Xia et al. 2012):

$$X_{ss} = -(A\underline{\xi}K)^{-1}\underline{B}\,\underline{U} \quad (2)$$

Where $\underline{\xi}$, $\underline{B}$, and $\underline{U}$ are long-term averages of the environmental scalars, C partitioning among the three live pools, and NPP, respectively. The structure details of CLM-CASA' C-only model are presented in Fig. 1a and parameters are described in Table 1. The CLM-CASA' C-only version is the CLM-CASA' when we only consider the C processes of CLM 3.5. The steady state soil C generated by this C-only version agreed largely with that simulated by original CLM-CASA' model (Xia et al., 2012).

### 2.2 The Microbial Models

Microbial process has various kinds of effects on the land carbon cycle, such as the real priming effects and temperature increase caused by soil microbial biomass (Kuzyakov et al., 2000; Luo et al., 2001; Peng et al., 2009). However, most conventional SOC models including CLM-CASA' do not explicitly represent microbial processes. Considering the microbial processes, the SOC decomposition rate is controlled by extracellular enzyme concentrations rather than simple decay constants in the CLM-CASA' and other traditional SOC models (Schimel and Weintraub, 2003). In this study, we focused on two enzyme driven decomposition models, one has two pools (Fig. 1b) introduced by German et al. (2012) and Hararuk et al. (2015), and the other has 4 pools (Fig. 1c) introduced by Allison et al. (2010). We call these two models 2-pool microbial

model and 4-pool microbial model, respectively. The 2-pool microbial model is described as the following equations (Hararuk et al., 2014 and 2015).

$$\frac{dMIC}{dt} = CUE \times V_{max} \times MIC \frac{SOC}{K_m + SOC} - r_d \times MIC \quad (3)$$

$$\frac{dSOC}{dt} = Input_{soil} + r_d \times MIC - V_{max} \times MIC \frac{SOC}{K_m + SOC} \quad (4)$$

Where

$$CUE = CUE_{slope} \times T_s - CUE_0 \quad (5)$$

$$V_{max} = V_{max_0} \times exp\left(-\frac{E_a}{R \times (T_s + 273)}\right) \times exp(-par_{clay} \times clay) \quad (6)$$

$$Km = Km_{slope} \times T_s + Km_0 \times exp(par_{lig} \times lignin) \quad (7)$$

MIC represents the microbial biomass and SOC represents the soil organic carbon pool. $Input_{soil}$ is the carbon influx to soil. The other parameters listed in Table 2 are to be calibrated. Only the first eight ones in Table 2 are the parameters of the 2-pool microbial model.

The 4-pool microbial model from Allison et al. (2010) is described as follws:

$$\frac{dMIC}{dt} = V_{maxup} \times MIC \frac{DOC}{Kmup + DOC} \times CUE - r_d \times MIC - r_{EnzProd} \times MIC \quad (8)$$

$$\frac{dDOC}{dt} = a_{lit\text{-}to\text{-}DOC} \times Input_{soil} + r_d \times MIC \times \left(1 - a_{MIC\text{-}to\text{-}SOC}\right) + V_{max} \times ENZ \frac{SOC}{Km + SOC} + r_{EnzLoss} \times ENZ - V_{maxup} \times MIC \frac{DOC}{Kmup + DOC} \quad (9)$$

$$\frac{dSOC}{dt} = a_{lit\text{-}to\text{-}SOC} \times Input_{soil} + r_d \times MIC \times a_{MIC\text{-}to\text{-}SOC} - V_{max} \times ENZ \frac{SOC}{Km + SOC} \quad (10)$$

$$\frac{dENZ}{dt} = r_{EnzProd} \times MIC - r_{EnzLoss} \times ENZ \quad (11)$$

Where

$$CUE = CUE_{slope} \times T_s - CUE_0 \quad (12)$$

$$V_{maxup} = V_{maxup_0} \times exp\left(-\frac{E_{aup}}{R \times (T_s + 273)}\right) \quad (13)$$

$$Kmup = Kmup_{slope} \times T_s + Kmup_0 \quad (14)$$

$$V_{max} = V_{max_0} \times exp(-\frac{E_a}{R\times(T_s+273)}) \times exp(-par_{clay} \times clay) \quad (15)$$

$$Km = Km_{slope} \times T_s + Km_0 \times exp(par_{lig} \times lignin) \quad (16)$$

*ENZ* and *DOC* are enzyme and dissolved organic carbon pools, respectively. Compared to the 2-pool version, the 4-pool microbial model has additional 7 parameters to be calibrated. Total 15 parameters of the 4-pool microbial model are described in Table 2.

### 2.3 Data and Metrics

Microbial models and CLM-CASA' C-only models divide the world into 64*128 grid cells and output SOC content at each grid (Fig. 2). The observed SOC data for parameter calibration comes from the International Geosphere Biosphere Programme – Data and Information System (IGBP-DIS) dataset (Global Soil Data Task Group, 2000). The IGBP-DIS dataset includes a 1-km resolution global land carbon data set and the dataset has been widely used in many studies to evaluate and improve models (Zhou et al., 2009; Smith et al., 2013).

The goal of parameter calibration is to improve SOC predictions to better fit the observations. Therefore, we use the root mean squared errors (RMSEs) between the model SOC predictions and the observations at all grid cells as the metric function. This metric function can be described as the following formula:

$$r = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - O_i)^2} \quad (17)$$

Where $N$ denotes the total number of grid cells, $X_i$ and $O_i$ are the SOC of model prediction and IGBP-DIS observation, respectively. To avoid overfitting and evaluate the calibrated parameters more fairly, we separate all grid cells into training set and validation set. The training set is used to guide the parameter calibration process and the validation set is used to evaluate the calibrated results. Hararuk et al. (2014 and 2015) also used this method when calibrating SOC parameters with the Bayesian MCMC approach. The experiment results in Section 3 and 4 refer to the results for the validation set.

## 3 Surrogate-Based Optimization Algorithm Design

### 3.1 Introduction to Surrogate-Based Optimization Algorithm

The parameters of most soil carbon models and land models are tuned manually or based on gradient searching algorithms. Manual tuning method might be effective but highly depends on expert experience. Moreover, complex models may consist

of various components from different disciplines and then require different experts working collaboratively during tuning process.

Considering the difficulty of manually tuning scheme, different parameter calibration algorithms have been developed based on optimization theory. The gradient search algorithms like the quasi-Newton method are introduced to search a set of parameters with better performance in the parameter domain. The gradient search algorithms are usually efficient and fast. However, the gradient search algorithms are designed for finding the local optimum. Essentially it cannot be used to solve the multimodal problems derived from complex earth system models. In addition, the gradient search algorithms are based on the gradient information, which is unavailable for most soil carbon and land models. These models are too complex to get the gradient information, and thus the parameter calibration usually leads to solving a black-box optimization problem. Global optimization algorithms, such as genetic algorithms, particle swarm optimization algorithms, are based on parameter generation and selection strategies and keep gradient independent, which can be easily used for parameter calibration of complex earth system models. Global optimization algorithms are designed to find the global minimum. However, for complex models with large number of parameters, the number of samples (model runs) might be too large to comfortably afford (Jones et al., 1998). Moreover, complex earth system models, for example CLM, require several hours over hundreds of cores for only one sample run and pose a special challenge to the feasibility of automatic parameter calibration.

The surrogate-based optimization is an efficient and effective automatic parameter calibration framework. The surrogate-based optimization fits a surrogate model (or response surface) based on the previous samples and use this surrogate model to emulate the output behaviours of original models with an acceptable accuracy. The main idea of the surrogate-based optimization is to save computational cost during global optimization process by using the surrogate model instead of the original model, and to continuously improve the surrogate model by exploiting new sample runs with the original model. With the surrogate model, the algorithm can make full use of previous samples information and reduce the sample size, time-to-solution as well as the computation cost. The surrogate-based optimization is proved to be successful in solving parameter calibration of computationally expensive black-box problems (Vu et al., 2016).

### 3.2 Key Components of Surrogate-Based Optimization Algorithm

The flowchart of the surrogate-based optimization (referred as SBO hereafter) is presented in Fig. 3. Firstly, initial sets of parameter values are generated using a sampling method. These sets are then used as inputs to run the real simulation model. Secondly, a surrogate model is constructed by fitting the outputs of these sample runs. The surrogate model serves as a computationally cheap approximation of the expensive simulation model (Booker et al., 1999). Then in each iteration, new sample points simulated by the real model are generated according to specific strategy. This strategy can make use of the information gained from the surrogate model and only exploits the avoidable real model runs to meet the accuracy requirement. The new sample points and their simulation outputs are used to update the surrogate model at the same time. Finally, when

some stop criteria (usually the maximum number of simulations allowed) are met, the algorithm return the optimized parameter values. During the surrogate-based optimization process, quite a few sample runs are generated based on the evaluation of the surrogate model and most meaningless simulations with terrible parameter values are avoided. As a result, the computationally expensive model is simulated at only a few selected promising parameter points and the surrogate model will replace the real model during the calibration process; thus the computation cost is reduced substantially.

There are three key components in the surrogate-based optimization algorithm: the surrogate model, the initial sampling and the parameter point generation for real model run. There are various surrogate models such as multivariate adaptive regression splines (Friedman, 1991), polynomial regression models (Myers and Montgomery, 1995), radial basis functions (RBFs) (Gutmann, 2001; Müller et al., 2013; Powell, 1992; Regis and Shoemaker, 2007, 2009; Wild and Shoemaker, 2013), kriging (Davis and lerapetritou, 2009; Forrester et al., 2008; Jones et al., 1998). Many machine learning regression models are also introduced like support vector regression (Zhang et al., 2009), artificial neural network (Behzadian et al., 2009) and random forest (Breiman, 2001).

As for the initial sampling, the Monte Carlo sampling and Latin Hypercube sampling (LHS for short) are two main sampling methods (McKay et al., 1979; Iman et al., 1981). The Monte Carlo sampling samples values from a probability distribution, which is usually a uniform distribution unless we have additional knowledge about the model and the parameters. During LHS procedure, the range of each parameter is divided into $M$ equally probable intervals. $M$ sample points are selected to cover all intervals of each parameter. Compared to the random sampling, LHS ensures that the ensemble of random numbers is representative of the real variability of the parameters.

The strategies of parameter point generation are iterative algorithms that use data acquired from previous iterations to guide new parameter point generation. Most strategies convert the parameter point generation to optimization problems using an evaluation criterion (Fig. 3). There are many different generation strategies, including Minimizing an Interpolating Surface (MIS) (Jones, 2001) and Maximizing Expected Improvement (MEI) (Schonlau et al., 1997; Picheny et al., 2013). In MIS, the minimum of the surrogate model response surface is found and treated as the new parameter point to evaluate the real simulation model and then update the surrogate model. MEI introduces the “expected improvement” criterion. This criterion estimates the uncertainty of the surrogate model and balances the exploration and exploitation. Another parameter generation strategy is candidate point approach (CAND) (Regis and Shoemake, 2007). In the CAND strategy, the criterion for exploitation is MIS and the criterion for exploration is the distance of the candidate point to the set of sampled parameter points from previous iterations. The previous sampled points represent the explored region and we can estimate the uncertainty with the distance to the explored region. A weighted sum of these two criteria is used to determine the new parameter point during the surrogate-based optimization.

### 3.3 Design of Surrogate-Based Optimization Algorithm for Soil Carbon Models

Based on the previous introduction of SBO, the detailed procedure of SBO can be found as follows. It is worthy noted that the best parameter set of the real model can be iteratively searched in Step 4 during looking for the new sample points.

Step 1: Generate an initial sampling set $S_0$.

Step 2: Run the real model and calculate the output error of the parameter points of $S_0$.

Step 3: Build the surrogate model using the parameters and the outputs generated in Step 2.

Step 4: Predict the output errors of those points which do not belong to $S_0$ using the surrogate model and determine the points at which to run the real model.

Step 5: Run the real model again for the new parameter points of Step 4 and calculate the output errors of these selected points.

Step 6: Update the surrogate model with the new data of Step 5.

Step 7: Iterate through Steps 4 to 6 until the end condition has been met.

The surrogate-based optimization scheme mentioned in previous sections is a parameter calibration framework, and the key components introduced in Section 3.2 have to be selected during tuning the parameters of soil carbon models. We assume that the soil carbon models (especially together with land model) are computationally expensive and at most several hundred samples can be afforded. The LHS can cover the whole parameter space with limited number of sample points while Monte Carlo sampling usually requires much larger number of samples. Therefore, we choose the LHS as the initial sampling strategy. As mentioned in the previous section, many kinds of surrogate-based models have been introduced and developed. The Machine learning regression models perform not so well as RBF and kriging models according to the evaluation on similar cases (Wang et al., 2014). In this study, we use the RBF surrogate model (RBF-SBO) as our default choice because it has been proved to perform better than other surrogate model types (Müller and Shoemaker, 2014) and has easy-to-use implementation. Moreover, we also implement other surrogate models including Kriging and Mars in our algorithm framework and can also introduce other advanced surrogate models later.

The soil carbon models are usually complex, nonlinear and not smooth and the surrogate model are not accurate when the SBO starts. The MIS can be very efficient but easy to trap into local optima, since the strategy does not consider the uncertainty of the surrogate model and only select the optimum of the surrogate model. The MEI eliminates the disadvantage of MIS but can only be used for the kriging surrogate model because the calculation of the expected improvement requires the standard error at the parameter point and only the kriging (Gaussian Process) surrogate model can provide the standard error (Jones et al., 1995). Finally, we use the CAND strategy as the parameter generation strategy in our algorithm, which has the advantage of balancing the exploitation and exploration of uncertain region.

## 4 Parameter Calibration Experiments

### 4.1 Experiment Configuration

In this study, we select the Bayesian MCMC approach and four advanced global optimization algorithms to compare with our proposed surrogate-based optimization method. Three types of SOC models and their metric functions are introduced in Section 2. The target of parameter calibration is to find the optimal values of parameters to achieve the minimum value of the metric function (average RMSE). Moreover, we repeat the parameter calibration process of each algorithm 50 times and use the average results for algorithm evaluation. We compare the performance of algorithms from both the effectiveness and efficiency. The effectiveness refers to the accuracy of the calibrated results and the efficiency can be evaluated by the required simulation times of the original SOC models.

### 4.2 Various Global Optimization Algorithms and the Bayesian MCMC Approach

The Bayesian MCMC approach and four advanced global optimization algorithms, including Differential Evolution (DE), Particle Swarm Optimization (PSO), Shuffled Complex Evolution (SCE-UA) and Covariance Matrix Adaption Evolution Strategy (CMA-ES), are used to compare with our RBF surrogate-based optimization.

DE (Storn and Price, 1997) and PSO (Kennedy, 1995; Shi and Eberhart, 2009) are the representative algorithms of the evolution strategy and swarm intelligence, respectively. They both have the ability to converge quickly and outperform many genetic algorithms and simulated annealing algorithms (Price and Storn, 2006; Shi and Eberhart, 2009). SCE-UA is designed for the parameter calibration of hydrologic models and has gained success in various hydrology models such as the TOPMODEL, the Xinanjiang watershed model and short-term load forecasting (Hapuarachchi et al., 2001; Ma H, et al., 2006; Li G, et al., 2007). SCE-UA ensures the effectiveness and efficiency by combining the local (the simplex method) and global optimization methods. Despite the difference in detail, DE, PSO and SCE-UA all generate new parameter points according to some simple mathematical formulas. Different from these three algorithms, CMA-ES (Hansen and Ostermeier, 2001; Hansen and Kern, 2004) creates new parameter points based on a multivariate normal distribution. The dependencies between parameters are represented by the covariance matrix of a normal distribution. CMA-ES has been proven to be the best global optimization algorithm in the BBOB-2009 comparison study (Hansen, 2009).

The Bayesian MCMC approach is usually designed to obtain the posterior distributions of model parameters but it can also be used to calibrate parameters to reduce the prediction error. The Bayesian MCMC approach consists of two steps: the proposing step and the moving step. In the proposing step, the parameter covariance matrix is estimated from a series of parameter sets. A new parameter set is generated from the last accepted parameter set through a uniform proposal distribution (Xu et al., 2006). In the moving step, a probability of acceptance determined by prediction error is calculated (Marshall et al., 2004). The final calibrated parameter set is estimated by Maximum likelihood estimator (MLE) with an accepted parameter chain. Hararuk et

al. (2014, 2015) applies the Bayesian MCMC approach using the Metropolis algorithm to parameter calibration of the CLM-CASA' C-only model and microbial models. During the experiments of Hararuk et al. (2014, 2015), the proposing step requires 50,000 simulations and the moving step requires 500,000 simulations for microbial models and 1,000,000 simulations for the CLM-CASA' model. We have got the code from Hararuk and repeated the calibration experiments. The detailed calibration results with the Bayesian MCMC approach are presented in Table 3.

### 4.3 Results and Analysis

#### 4.3.1 Effectiveness and Efficiency

Fig. 4 presents the calibrated results (RMSE) of different algorithms we applied. For each algorithm, we only perform 100 simulations for comparison. As the Bayesian MCMC approach requires a large number of samples to reach a stable distribution, over 500,000 simulations have been conducted for algorithm evaluation.

Obviously, the average RMSE of the RBF-SBO is the lowest (0.6 $kg/m^2$ better than the Bayesian MCMC algorithm) for two microbial models among all the algorithms (Fig. 4b, c). For the CLM-CASA' model, our RBF-SBO algorithm is still the best one compared to the global optimization algorithms. And the Bayesian MCMC approach gets a little better result ( about 0.02 $kg/m^2$) since it has to conduct a lot more simulations for better result (Fig. 4a).

With respect to stability, the results of RBF-SBO also shows a lower variation among the 50 repeated experiments compared with the global optimization algorithms on three types of models. For the same reason mentioned before, the Bayesian MCMC approach gets lower variation than our RBF-SBO algorithm on two microbial models. For the most complicated CLM-CASA' model, our RBF-SBO is still promising on stability. Among the global optimization algorithms, the CMA-ES shows a very significant fluctuation (Fig. 4b, c), indicating CMA-ES is unreliable when the number of simulations is as small as 100. This is because the CMA-ES requires quite a few simulations on the exploration of the parameter domain and the construction of the parameter covariance matrix. Therefore, the RBF surrogate-based optimization is the most effective and stable one when the number of simulations is limited.

Fig. 5 shows the results in terms of average validation RMSE. The Bayesian MCMC is not shown here because it requires at least 50000 simulations in the proposing step. The average validation RMSE of RBF-SBO is lower than other four global optimization algorithms before the number of simulations increases to 600 for two microbial models and to 200 for the CLM-CASA' model, respectively. The number of simulations our RBF-SBO requires is fewer than other global optimization algorithms when they reach the same RMSE value and accuracy range. Thus, our RBF surrogate optimization is also the most efficient algorithm which requires the minimum simulation times and computation cost. The reason why our SBO outperforms other global optimization algorithms is that our SBO can enhance the searching efficiency by using a surrogate model to simulate the real model and avoiding bad parameter points ('bad' means the high prediction error). Moreover, the setup scheme

of our SBO also contributes to the superiority of our algorithm, which conducts several sample runs and selects the good parameter sets for use.

Another important observation is that the difference between the results of our RBF-SBO and other global optimization algorithms decreases as the number of simulation increases (Fig. 5). Moreover, the CMA-ES outperforms the RBF-SBO when the number of simulations exceeds 200 for the CLM-CASA' model (Fig. 5a). Our SBO can build the surrogate model with relatively good accuracy quickly, which contributes to finding the near-optimal solution with less computation cost. However, the surrogate model is only an approximation of the real model and the accuracy might be limited due to the strong nonlinearity and the high complexity of the real model. After gaining enough knowledge of the real model by lots of simulations, the excellent global optimization algorithms, such as CMA-ES, may achieve a similar performance or even outperform our SBO, which suggests that our SBO is better to use for parameter calibration problem of cost-expensive models such as CLM.

#### 4.3.2 Impact of the Model Complexity

Compared to the 2-pool and 4-pool microbial models, the CLM-CASA' model has 13 carbon pools and 20 parameters and thus it is significantly more complex than two microbial models. Despite increasing complexity of the CLM-CASA' model, the SBO gets better results before conducting 200 simulations of the real model (Fig. 5a). Moreover, our SBO is always the best parameter calibration method for the 2-pool and 4-pool microbial models before conducting 600 simulations (Fig. 5b, c). In addition, only one global optimization algorithm, CMA-ES, shows better performance compared to our SBO on the CLM-CASA' model after 200 simulations. Considering the high variance of CMA-ES on two microbial models (Fig. 4b, c), our SBO is more effective and more reliable on average.

#### 4.3.3 Impact of Different Types of Surrogate Models

We select the RBF as the surrogate model in the former experiments because the RBF is the promising choice in many surrogate-based optimization algorithms (Müller and Shoemaker, 2014). In this section, we also test two other typical surrogate models, kriging and the multivariate adaptive regression splines (Mars). The Mars model is simple and has almost no requirements to the sample quality. Mars is very quick to train and predict. The kriging, also Gaussian process regression, is a method of interpolation for which the interpolated values are modelled by a Gaussian process governed by prior covariance. Kriging gives the best linear unbiased prediction of the intermediate values under suitable assumptions on the priors.

Figure 6 presents the results of kriging, Mars and RBF in terms of average validation RMSE. The performance of the three surrogate models is similar. The three surrogate models all get reasonable performance in the parameter calibration of the three types of SOC models and perform better than global optimization algorithms, indicating that our surrogate-based optimization is robust.

## 5 Analysis of Parameter Calibration Results

### 5.1 Analysis of CLM-CASA' Model

The steady state global SOC simulations (Eq. 2) using CLM-CASA' with the default and calibrated parameter values are presented in Fig. 7a and b, which are also compared to the observed SOC pools provided by the IGBP-DIS dataset. The SOC simulation result using the calibrated parameter values obtained by the surrogate-based optimization matches the observation better than that with the default parameter values (Fig. 7c) with a relatively lower RMSE. By using the calibrated parameter values, the SOC simulations are significantly improved almost all over the world, except some grid cells in the west of Canada and the east of Russia (Fig. 7a, b and c). As a result, the CLM-CASA' simulation result with the default parameter values can only explain 33% of variation in the observed soil C, whereas that with the calibrated parameter values can explain an improved ratio (42%) of variation in the observed soil C. The unexplained variation is partly due to uncertainty in observations. To further improve the model accuracy, we need to gain more understanding of uncertainty sources from data, model structure, parameters, and forcing.

Figure 8 presents the frequency distributions of the 20 calibrated parameters based on MCMC and the calibrated parameter values by using the proposed surrogate-based optimization (the blue lines in Fig. 8). Narrow posterior distributions indicate highly sensitive parameters, agreeing with the conclusions of Hararuk et al. (2014) and Post et al., (2008). The calibrated parameter values of the surrogate-based optimization are close to the responding parameter values at the peaks of posterior distributions for most highly sensitive parameters such as temperature sensitivity of heterotrophic respiration ($Q_{10}$) and clay effect on C partitioning from slow to passive pools ($t_7$). The parameter calibration results (RMSE) of the surrogate-based optimization and Bayesian MCMC are similar, agreeing with the parameter calibration results listed in Table 3.

Some calibrated parameter values are very close to the assigned bounds of the parameters in Fig. 8, which is usually related to the correlations among parameters. Further investigation on the covariance among parameters is necessary to explain this issue. In addition, the unreasonable setting of those bounds might be another possible reason. For instance, the calibrated c(12,12) value ($1.01 \times 10^{-3}$) reaches its lower bound, indicating that passive SOC residence time almost approaches 1000 years.

As listed in Table 1, the calibrated temperature sensitivity ($Q_{10}$) decreases from 2 to 1.74. The size of soil microbial and passive pools increase due to longer residence time of passive pool and lower temperature sensitivity ($Q_{10}$). The size of the slow pool, on the contrary, decreases due to the increase in exit rate from slow pool or the decrease of its residence time. Comprehensively, the size of SOC, which is the sum of carbon capacity in passive pools, slow pools and soil microbial pools, increases and more approximates to the observation.

### 5.2 Analysis of Microbial Models

According to the calibrated RMSE and $r^2$, the SOC simulation of the 2-pool and 4-pool microbial models are very similar. Without loss of generality, we only analyze the parameter calibration results of the 4-pool microbial model in this section. After parameter calibration using the surrogate-based optimization, the global SOC produced by the 4-pool microbial model are improved, especially in the regions of China, Russia, Europe and North America (as shown in Fig. 9). Overall, the microbial models explain a higher fraction of global variability of the observed SOC data and have lower spatial RMSEs than the CLM-CASA' model (as listed in Table 3).

The microbial models achieve better SOC predictions than that of the calibrated CLM-CASA' model in terms of the prediction of C capacity in the low-temperature regions (Russia, Europe, North America) and in the regions with small soil C inputs (Fig. 7b and 9). The SOC contents are determined by two main factors: the soil carbon inputs and the SOC residence time (Luo et al., 2003). Considering the same soil carbon inputs of the CLM-CASA' and the microbial models, the improvement is mostly induced by the differences of SOC residence time. In all the three models, the SOC residence time is essentially controlled by temperature (Xia et al. 2013). As a result, the temperature sensitivity ($Q_{10}$) contributes to the difference across three models. The temperature sensitivity keeps constant in the CLM-CASA'. However, both of two microbial models calculate spatially variable $Q_{10}$ with higher values in the low-temperature regions and lower $Q_{10}$ in the high-temperature regions, which reflects the impact of the temperature to the microbial activity. In addition, the SOC residence time can also be affected by the quality of SOC inputs and is related to the microbial decomposition processes. Fresh C input stimulates the microbial dynamics growth, resulting in the increase of old SOC decomposition rate (i.e., priming effect) (Kuzyakov et al, 2000; Fontaine et al., 2004, 2007). Therefore, the microbial models simulate lower SOC residence times than the CLM-CASA' in the regions with high SOC input and high SOC residence time and the regions with low SOC input. This is due to the nonlinearity of substrate limitation in the microbial models (Eq. 8 and 10), as well as the dependency of residence time in microbial dynamics. Comprehensively, the introduction of microbial dynamics helps the microbial models predict SOC better than the CLM-CASA' model.

Figure 10 presents the posterior distributions of the parameters calculated by Bayesian MCMC and the calibrated parameter values by our surrogate-based optimization. According to the posterior distribution, $r_d$, $CUE_{slope}$, $CUE_0$, $E_a$, $par_{lig}$ and $par_{clay}$ are the most constrained and sensitive parameters. The calibration results of the surrogate-based optimization agree with the posterior distributions of these highly sensitive parameters (Fig. 10) except $CUE_{slope}$ and $CUE_0$. $CUE_{slope}$ and $CUE_0$ are highly sensitive owing to their influence on temperature sensitivity. Due to the difference between $CUE_{slope}$ and $CUE_0$, the RMSE of the surrogate-based optimization is 1.4 $kg/m^2$ and 0.8 $kg/m^2$ lower than those with Bayesian MCMC for 4-pool and 2-pool microbial models respectively (as listed in Table 3). The mismatch of $CUE_{slope}$ and $CUE_0$ may be mainly due to the different targets of the parameter selection between two methods.

## 6 Conclusions

Parameter calibration becomes more and more challenging for SOC model development, especially for the computationally-expensive global land models owing to large number of simulations. In this study, we introduce a surrogate-based optimization algorithm to the parameter calibration of computationally expensive SOC models. The main findings are:

1) Compared to advanced global optimization algorithms, the surrogate-based optimization is more effective and more efficient on average. Our RBF surrogate-based optimization outperforms other parameter calibration algorithms when the number of simulations is no more than 200.
2) The parameter optimization based on RBF surrogate model gains more accurate calibration results than those of the Bayesian MCMC approach in the three soil carbon models. Moreover, the computation cost of the surrogate-based optimization is only 0.1% of that of the Bayesian MCMC.
3) The surrogate-based optimization scheme is robust. Various types of surrogate models have the similar performance in parameter calibration tasks of SOC models.
4) Although the surrogate-based optimization is only guided by a single metric function, it still can find better parameter values compared to the default ones. We carefully analyze the spatial SOC distributions produced by the SOC models with the calibrated parameters using our surrogate-based optimization, which indicates that the surrogate-based optimization truly improves the model prediction and simulation capability.

Nowadays, more and more complex simulation models present challenges to the surrogate-based optimization algorithm. To improve the accuracy of the surrogate-based optimization, better surrogate models are expected. Current surrogate models including our implementation for soil carbon models most employ only one surrogate model, which may limit the successful use for different kinds of models. We will focus on the application of multiple surrogate models using ensemble learning in the future.

## 7 Code and data availability

The code and data of three models and the related algorithm implementations can be found in the supplement. If you have any problem when using the code and repeating the experiments, please feel free to contact the first author of this paper: Haoyu Xu (ocean920329@gmail.com).

## Appendix: The detailed description of the proposed Surrogate-Based Optimization algorithm

The parameter calibration of the soil carbon models can be formulated to the following optimization problem:

$min_{x \in D} \{f(x)\}$

In which, $f: R^d \rightarrow R$ is a continuous black-box function, representing the prediction error of the real model. And $D$ is a subset of $R^d$, representing the legal ranges of parameters. A parameter can be represented as a point in $R^d$. The surrogate-based optimization algorithm conducts several steps as follows.

1. Generate an initial sampling set $S_0$.
2. Run the real model and calculate the output error of the parameter points of $S_0$.
3. Build the surrogate model using the parameters and the outputs generated in Step 2.
4. Predict the output errors of those points which do not belong to $S_0$ using the surrogate model and determine the points at which to run the real model.
5. Run the real model again for the new parameter points of Step 4 and calculate the output errors of these selected points.
6. Update the surrogate model with the new data of Step 5.
7. Iterate through Steps 4 to 6 until the end condition has been met.

Different surrogate-based optimization algorithms may have different choices with respect to the follows:

- ✧ The sampling method to generate the initial set $S_0$.
- ✧ The surrogate model, which predicts the output $y$ using the given data point $x$. Before prediction, some $(x, y)$ data pairs should be given to train the model and the data is called training set.
- ✧ How to decide the new points at which to run the real model in each iteration.

As introduced in Section 3, we use Latin hypercube sampling (LHS) to generate the initial set $S_0$ (Iman et al., 1981). As for the surrogate model, we compare three kinds of surrogate models including Kriging, Mars and RBF.

The Mars model (abbreviation of multivariate adaptive regression splines) is an extension of naïve linear models, which introduced by Friedman J H. (Friedman J H., 1991). The form of Mars is presented as follows:

$$\hat{f}(x) = \sum_{i=1}^{m} c_i B_i(x)$$

Where $\hat{f}(x)$ represents the prediction of $y$ at the point $x$, and $c_i$ is a constant coefficient to be trained. The $B_i(x)$ is the basis function which can take one of the three forms: a constant, a hinge function like $max\ (0, x - const)$ and a product of more than one hinge function.

The RBF model (abbreviation of radial basis function) is a real-valued function. The prediction at a point $x$ using RBF model only depends on the distance between $x$ and other points in the training set, whose outputs have been already given. The distance $r = ||x, c||$ is usually Euclidean distance. The radial function is the function satisfies the property $\phi(x, c) = \phi(||x, c||) = \phi(r)$. The prediction at point x with RBF model is formulated as:

$$\hat{f}(x) = \sum_{i=1}^{N} w_i \, \phi(\|x, x_i\|)$$

Where the $x_i$ represents the point of the training set which has $N$ points in total. Many different radial functions have been introduced and some commonly-used ones are Gaussian $\phi(r) = e^{-(\varepsilon r)^2}$, Multiquadric $\phi(r) = \sqrt{1 + (\varepsilon r)^2}$, and Polyharmonic spline: $\phi(r) = r^k ln(r)$. In our experiments, we choose the Gaussian radial function.

Both the Kriging model and the Gaussian process regression model predict the output using a Gaussian process governed by prior covariance. Before used to train the kriging model, the $x$ and $y$ should be normalized to satisfy a normalization distribution where the means is 0 and the covariance is 1. The Kriging predictor can be found as follows:

$$\hat{f}(x) = \hat{\mu} + \sum_{i=1}^{n} c_i r_i(x)$$

Where $\hat{\mu}$ is the estimated mean of the gaussian process, $c_i$ is a constant representing the weight and $r_i(x) = Corr(x, x^{(i)})$ is the correlation between the $x$ and the $ith$ point $x^{(i)}$ in the training set. $\hat{\mu}$ and $c_i$ can be trained with the training set.

**References**


Aleman D M, Romeijn H E, Dempsey J F. A response surface approach to beam orientation optimization in intensity-modulated radiation therapy treatment planning[J]. INFORMS Journal on Computing, 2009, 21(1): 62-76.

Huang YY, XJ Lu, Z Shi, D Lawrence, C Koven, JY Xia, ZG Du, E Kluzek, YQ Luo. Matrix approach to land carbon cycle modeling: A case study with Community Land Model[J]. Global Change Biology, 2017.

Allison S D, Wallenstein M D, Bradford M A. Soil-carbon response to warming dependent on microbial physiology[J]. Nature Geoscience, 2010, 3(5): 336-340.

Booker A J, Dennis Jr J E, Frank P D, et al. A rigorous framework for optimization of expensive functions by surrogates[J]. Structural optimization, 1999, 17(1): 1-13.

Chen H, Tian H Q. Does a General Temperature-Dependent Q10 Model of Soil Respiration Exist at Biome and Global Scale[J]. Journal of Integrative Plant Biology, 2005, 47(11): 1288-1302.

Corzo G, Solomatine D. Special Issue: Knowledge-based modularization and global optimization of artificial neural network models in hydrological forecasting[J]. Neural Networks the Official Journal of the International Neural Network Society, 2007, 20(4):528-536.

Duan Q, Sorooshian S, Gupta V K. Optimal use of the SCE-UA global optimization method for calibrating watershed models[J]. Journal of hydrology, 1994, 158(3): 265-284.

Duan Q, Sorooshian S, Gupta V. Effective and efficient global optimization for conceptual rainfall-runoff models[J]. Water resources research, 1992, 28(4): 1015-1031.

Fontaine S, Bardoux G, Abbadie L, et al. Carbon input to soil may decrease soil carbon content[J]. Ecology Letters, 2004, 7(4):314-320.

Fontaine S, Barot S, Barré P, et al. Stability of organic carbon in deep soil layers controlled by fresh carbon supply[J]. Nature, 2007, 450(7167):277-80.

Friedlingstein, P., et al., Climate-carbon cycle feedback analysis: Results from the C4MIP model intercomparison[J]. Climate, 2006, 19(14), 3337–3353.

Friedman J H. Multivariate adaptive regression splines[J]. The annals of statistics, 1991, 19(1): 1-67.

German D P, Marcelo K R B, Stone M M, et al. The Michaelis–Menten kinetics of soil extracellular enzymes in response to temperature: a cross-latitudinal study[J]. Global Change Biology, 2012, 18(4): 1468-1479.

Giunta A A. Aircraft multidisciplinary design optimization using design of experiments theory and response surface modeling methods[D]. Virginia polytechnic institute and state university, 1997.

Global Soil Data Task Group. Global Gridded Surfaces of Selected Soil Characteris- tics (IGBP-DIS). [Global Gridded Surfaces of Selected Soil Characteristics (International Geosphere-Biosphere Programme - Data and Information System)]. Data set, 2000, Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, TN.

Gutmann H M. A radial basis function method for global optimization[J]. Journal of Global Optimization, 2001, 19(3): 201-227.

Haario H, Saksman E, Tamminen J. An adaptive Metropolis algorithm[J]. Bernoulli, 2001, 7(2): 223-242.

Hansen N, Kern S. Evaluating the CMA evolution strategy on multimodal test functions[C]. Proceedings of International Conference on Parallel Problem Solving from Nature. Springer Berlin Heidelberg, 2004: 282-291.

Hansen N, Ostermeier A. Completely de-randomized self-adaptation in evolution strategies[J]. Evolutionary Computation, 2001, 9(2): 159-195.

Hansen N. Benchmarking the Nelder-Mead downhill simplex algorithm with many local restarts[C]. Proceedings of the 11th Annual Conference Companion on Genetic and Evolutionary Computation Conference: Late Breaking Papers. ACM, 2009: 2403-2408.

Hapuarachchi H A P, Li Z, Wang S. Application of SCE-UA method for calibrating the Xinanjiang watershed model[J]. Journal of lake sciences, 2001, 13(4): 304-314.

Hararuk O, Smith M J, Luo Y. Microbial models with data-driven parameters predict stronger soil carbon responses to climate change[J]. Global change biology, 2015, 21(6): 2439-2453.

Hararuk O, Xia J, Luo Y. Evaluation and improvement of a global land model against soil carbon data using a Bayesian Markov chain Monte Carlo method[J]. Journal of Geophysical Research: Biogeosciences, 2014, 119(3): 403-417.

Houghton, J. T., Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden, X. Dai, K. Maskell, and C. Johnson. Climate Change 2001: the scientific basis[B]. Cambridge University Press, Cambridge.

Iman R L, Campbell J, Helton J. An approach to sensitivity analysis of computer models[J]. Journal of Quality Technology, 1981, 13.

Ise, T., and P. R. Moorcroft. The global-scale temperature and moisture dependencies of soil organic carbon decomposition: An analysis using a mechanistic decomposition model[J], Biogeochemistry, 2006, 80(3), 217–231.

Jones D R, Schonlau M, Welch W J. Efficient global optimization of expensive black-box functions[J]. Journal of Global optimization, 1998, 13(4): 455-492.

Jones, D.R. A taxonomy of global optimization methods based on response surfaces[J]. Journal of Global Optimization, 2001,21(4): 345-383.

Kennedy J. Particle swarm optimization[M]. Encyclopedia of machine learning. Springer US, 2011: 760-766.

Kim K J. Toward Global Optimization of Case-Based Reasoning Systems for Financial Forecasting[J]. Applied Intelligence, 2004, 21(3):239-249.

Kuzyakov Y, Friedel J K, Stahr K. Review of mechanisms and quantification of priming effects.[J]. Soil Biology & Biochemistry, 2000, 32(11–12):1485-1498.

Li G, Cheng C, Lin J, et al. Short-term load forecasting using support vector machine with SCE-UA algorithm[C]. Third International Conference on Natural Computation (ICNC 2007). IEEE, 2007, 1: 290-294.

Luo Y, A Ahlström, SD Allison, NH Batjes, V Brovkin, N Carvalhais, A Chappell, P Ciais, EA Davidson, A Finzi, K Georgiou, B Guenet, O Hararuk, JW Harden, YJ He, F Hopkins, LF Jiang, C Koven, RB Jackson, CD Jones, MJ Lara, JY Liang, AD McGuire, W Parton, CH Peng, JT Randerson, A Salazar, CA Sierra, MJ Smith, HQ Tian, KEO Todd-Brown, M Torn, KJ van Groenigen, YP Wang, TO West, YX Wei, WR Wieder, JY Xia, X Xu, XF Xu, T Zhou.. Towards More Realistic Projections of Soil Carbon Dynamics by Earth System Models[J]. Global Biogeochemical Cycles, 2016, 30(1): 40-56.

Luo Y. Terrestrial carbon-cycle feedback to climate warming[J]. Annual Review of Ecology, Evolution, and Systematics, 2007: 683-712.

Luo YQ, TF Keenan, M Smith, Predictability of the terrestrial carbon cycle[J]. Global Change Biology 2015, 1737-1751.

Luo, Y., L. W. White, J. G. Canadell, E. H. DeLucia, D. S. Ellsworth, A. Finzi, J. Lichter, and W. H. Schlesinger, Sustainability of terrestrial carbon sequestration: A case study in Duke Forest with inversion approach, [J] Global Biogeochem. Cycles, 2003, 17(1), 1021.

Luo, Y., L. Wu, J. A. Andrews, L. White, R. Matamala, K. V. R. Schäfer, and W. H. Schlesinger, Elevated CO2 differentiates ecosystem carbon processes: Deconvolution analysis of Duke Forest FACE data, Ecol[J]. Monogr., 2001, 71(3), 357–376.

Ma H, Dong Z, Zhang W, et al. Application of SCE-UA algorithm to optimization of TOPMODEL parameters [J]. Journal of Hohai University (Natural Sciences edition), 2006, 4:001.

Madescu G, Boldea I, Miller T J E. The optimal lamination approach (OLA) to induction machine design global optimization[C]// Ias Meeting, Ias '96. Conference Record of the. IEEE, 1996:574-580 vol.1.

Maranas C D, Androulakis I P, Floudas C A, et al. Solving long-term financial planning problems via global optimization[J]. Journal of Economic Dynamics & Control, 1997, 21(8-9):1405-1425.

Mckay M D, Beckman R J, Conover W J. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code[M]. American Society for Quality Control and American Statistical Association, 2000.

Melillo, J. M., P. A. Steudler, J. D. Aber, K. Newkirk, H. Lux, F. P. Bowles, C. Catricala, A. Magill, T. Ahrens, and S. Morrisseau, Soil warming and carbon-cycle feedbacks to the climate system [J], Science, 2002, 298(5601), 2173–2176.

Müller J, Shoemaker C A. Influence of ensemble surrogate models and sampling strategy on the solution quality of algorithms for computationally expensive black-box global optimization problems[J]. Journal of Global Optimization, 2014, 60(2): 123-144.

Myers R H, Montgomery D C, Anderson-Cook C M. Response surface methodology: process and product optimization using designed experiments[M]. John Wiley & Sons, 2016.

Oleson, K. W., G. Niu, Z. Yang, D. Lawrence, P. Thornton, P. Lawrence, R. Stockli, R. Dickinson, G. Bonan, and S. Levis, Improvements to the Community Land Model and their impact on the hydrological cycle [J], J. Geophys. Res., 2008, 113.

Oleson, K. W., Y. Dai, G. Bonan, M. Bosilovich, R. Dickinson, P. Dirmeyer, F. Hoffman, P. Houser, S. Levis, and G.-Y. Niu Technical description of the community land model (CLM) [J]. NCAR Tech. 2004. .

Parton, W. J., et al., Observations and modelling of biomass and soil organic matter dynamics for the grassland biome worldwide [J], Global Biogeochem. Cycles, 1993, 7(4), 785–809.

Peng S, Piao S, Wang T, et al. Temperature sensitivity of soil respiration in different ecosystems in China[J]. Soil Biology and Biochemistry, 2009, 41(5): 1008-1014.

Picheny V, Ginsbourger D, Richet Y, et al. Quantile-Based Optimization of Noisy Computer Experiments With Tunable Precision[J]. Technometrics, 2012, 55(1): 2-9.

Price K, Storn R M, Lampinen J A. Differential evolution: a practical approach to global optimization[M]. Springer Science & Business Media, 2006.

Regis R G, Shoemaker C A. A stochastic radial basis function method for the global optimization of expensive functions[J]. INFORMS Journal on Computing, 2007, 19(4): 497-509.

Regis R G, Shoemaker C A. Parallel stochastic global optimization using radial basis functions[J]. INFORMS Journal on Computing, 2009, 21(3): 411-426.

Regis R G. Stochastic radial basis function algorithms for large-scale optimization involving expensive black-box objective and constraint functions[J]. Computers & Operations Research, 2011, 38(5): 837-853.

Schimel J P, Weintraub M N. The implications of exoenzyme activity on microbial carbon and nitrogen limitation in soil: a theoretical model[J]. Soil Biology and Biochemistry, 2003, 35(4): 549-563.

Schonlau M, Welch W J, Jones D R. Global versus local search in constrained optimization of computer models[J]. Lecture Notes-Monograph Series, 1998: 11-25.

Shi Y, Eberhart R C. Empirical study of particle swarm optimization[J]. Frontiers of Computer Science in China, 2009, 3(1):31-37.

Simpson T W, Mauery T M, Korte J J, et al. Kriging models for global approximation in simulation-based multidisciplinary design optimization[J]. AIAA journal, 2001, 39(12): 2233-2241.

Smith, M. J., D. W. Purves, M. C. Vanderwel, V. Lyutsarev, and S. Emmott, The climate dependence of the terrestrial carbon cycle, including parameter and structural uncertainties [J], Biogeosciences, 2013, 10(1), 583–606.

Sorooshian S, Duan Q, Gupta V K. Calibration of rainfall-runoff models: application of global optimization to the Sacramento soil moisture accounting model[J]. Water resources research, 1993, 29(4): 1185-1194.

Storn R, Price K. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces[J]. Journal of global optimization, 1997, 11(4): 341-359.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, An overview of CMIP5 and the experiment design [J], Bull. Am. Meteorol. Soc., 2011, 93(4), 485–498.

Todd-Brown, K. E. O., J. T. Randerson, W. M. Post, F. M. Hoffman, C. Tarnocai, E. A. G. Schuur, and S. D. Allison, Causes of variation in soil carbon simulations from CMIP5 Earth system models and comparison with observations, Biogeosciences, 2013, 10(3), 1717–1736.

Vu K K, D'Ambrosio C, Hamadi Y, et al. Surrogate-based methods for black-box optimization[J]. International Transactions in Operational Research, 2016.

Wang C, Duan Q, Gong W, et al. An evaluation of adaptive surrogate modelling based optimization with two benchmark problems[J]. Environmental Modelling & Software, 2014, 60(76):167-179.

Weng, E., and Y. Luo, Relative information contributions of model vs. data to short- and long-term forecasts of forest carbon dynamics, Ecol. Appl., 2011, 21(5), 1490–1505.

Wieder W R, Bonan G B, Allison S D. Global soil carbon projections are improved by modelling microbial processes[J]. Nature Climate Change, 2013, 3(10): 909-912.

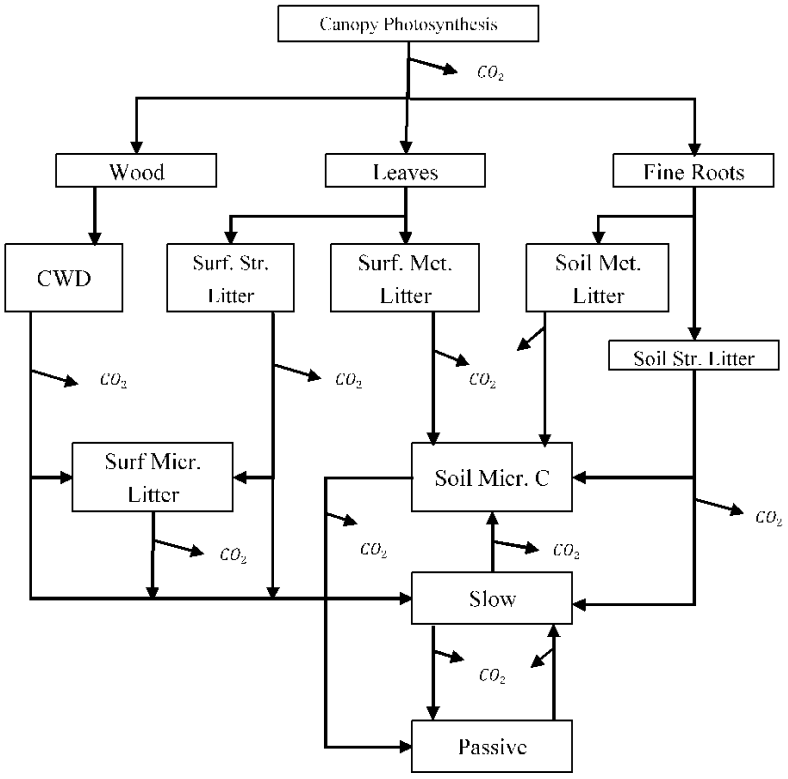
Xia, J., Y. Luo, Y. P. Wang, and O. Hararuk, Traceable components of terrestrial carbon storage capacity in biogeochemical models [J], Global Change Biol., 2013, 19, 2104–2116.

Xia, J., Y. Luo, Y. Wang, E. Weng, and O. Hararuk, A semi-analytical solution to accelerate spin-up of a coupled carbon and nitrogen land model to steady state [J], Geosci. Model Dev., 2012, 5(5), 1259–1271.

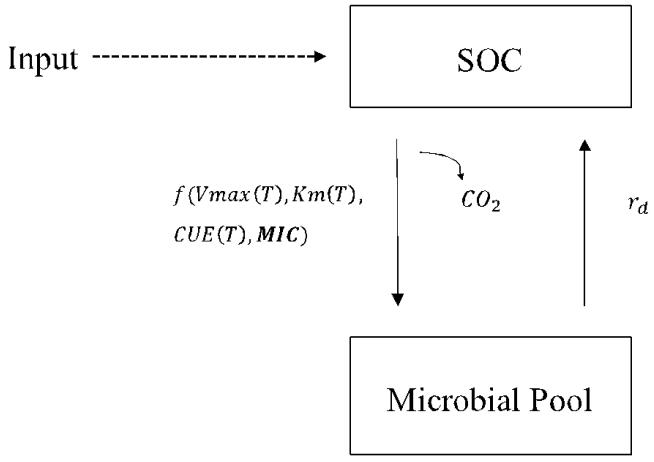
Xu, T., L. White, D. Hui, and Y. Luo, Probabilistic inversion of a terrestrial ecosystem model: Analysis of uncertainty in parameter estimation and model prediction [J], Global Biogeochem. Cycles, 2006, 20, GB2007.

Zhou, T., P. Shi, D. Hui, and Y. Luo, Global pattern of temperature sensitivity of soil heterotrophic respiration (Q10) and its implications for carbon-climate feedback [J], J. Geophys. Res., 2009, 114.

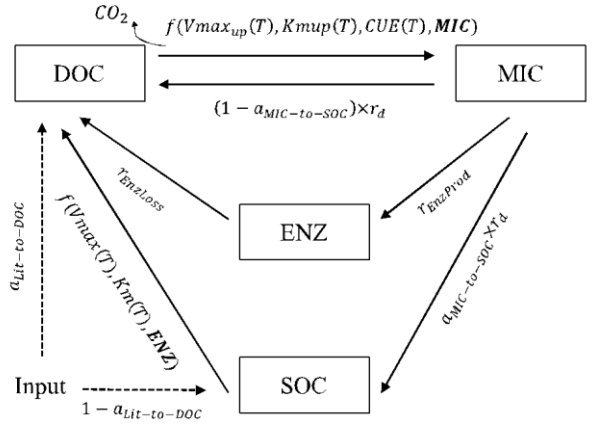
Friedman J H. Multivariate Adaptive Regression Splines[J]. Annals of Statistics, 1991, 19(1): 1-67.




(a) **The CLM-CASA' model**




(b) **2-pool microbial model**




(c) **4-pool microbial model**

**Figure 1. Schematic representations of (a) CLM-CASA model, (b) 2-pool microbial model and (c) 4-pool microbial models**

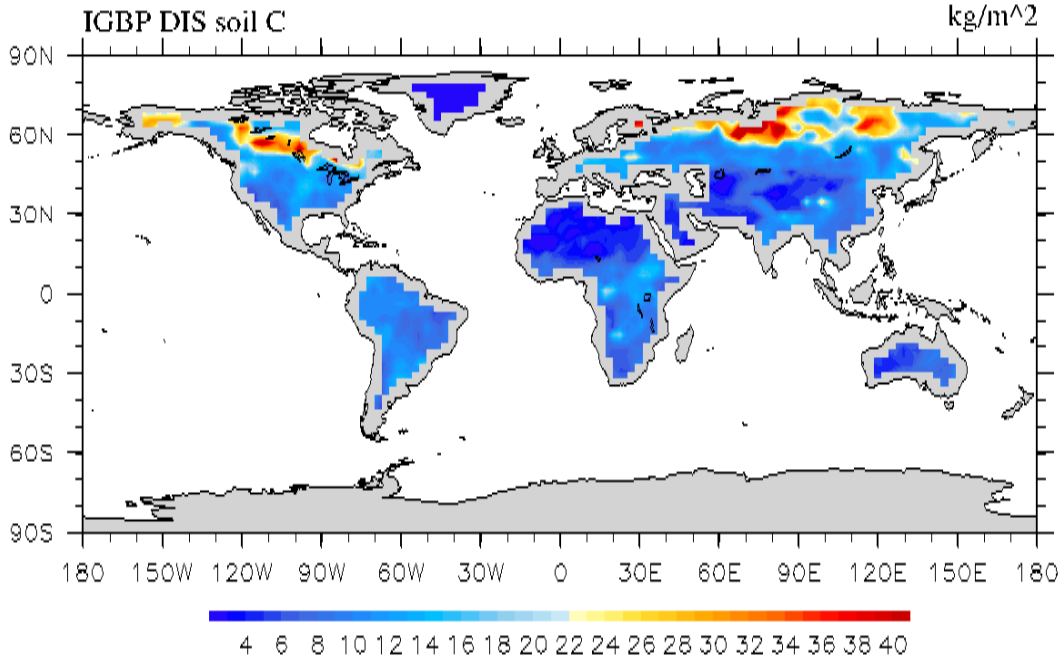



**Figure 2. IBGP-DIS soil carbon distribution. Soil carbon varies from 0 $kg/m^2$ in deserts to 60 $kg/m^2$ in boreal regions**

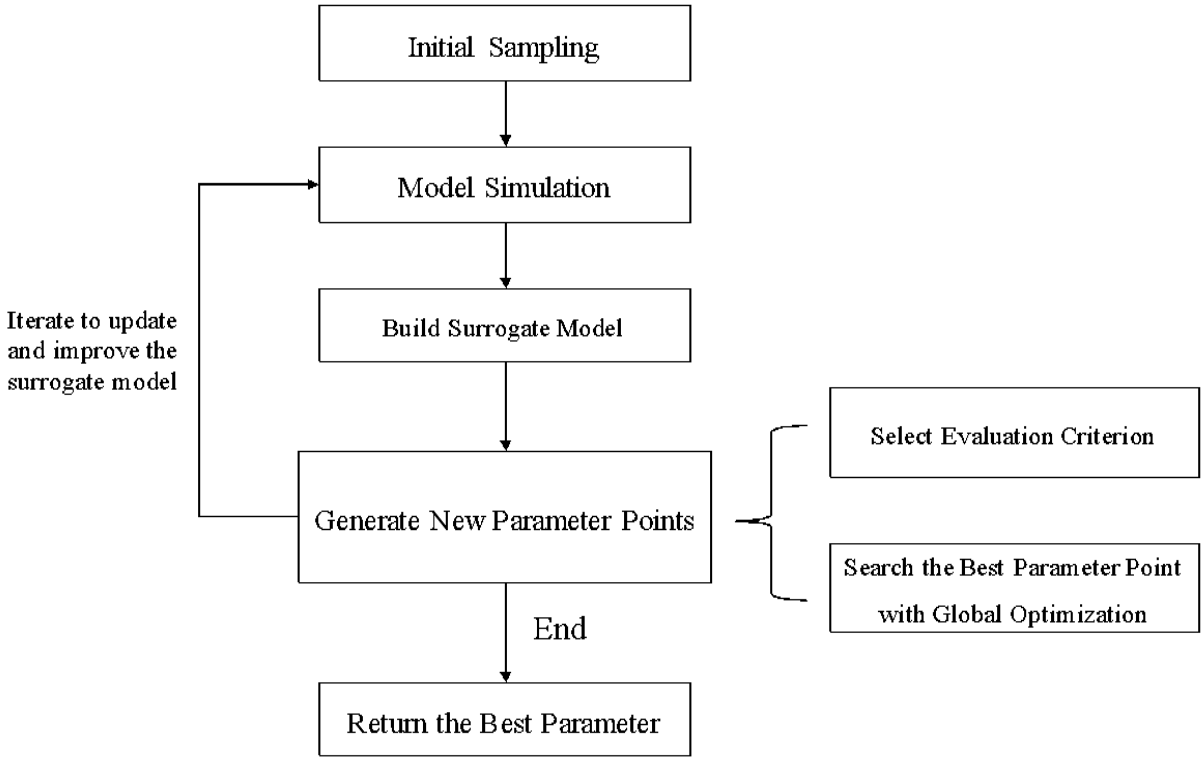
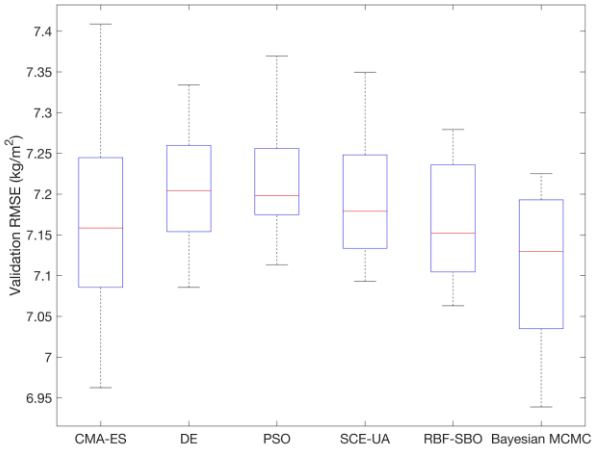
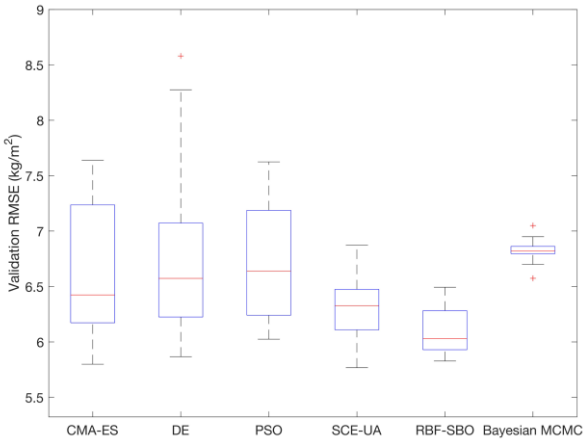


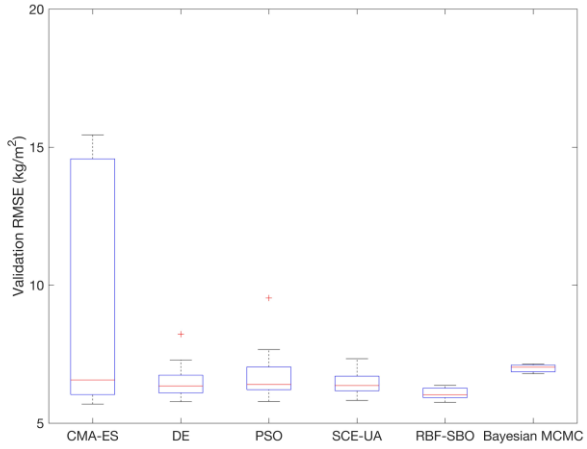

**Figure 3. The flowchart of the surrogate-based optimization**




**(a) CLM-CASA' model**

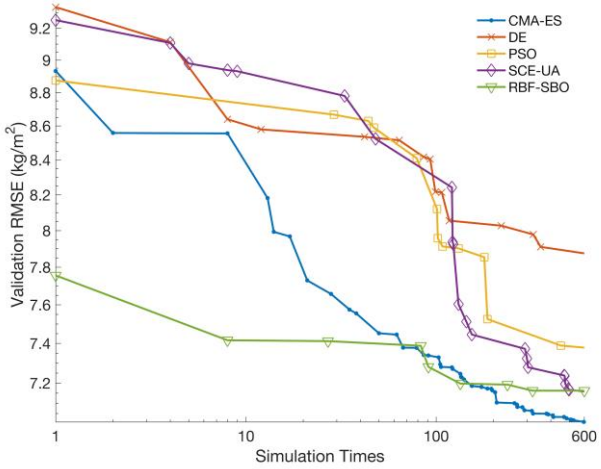



**(b) 2-pool microbial model**

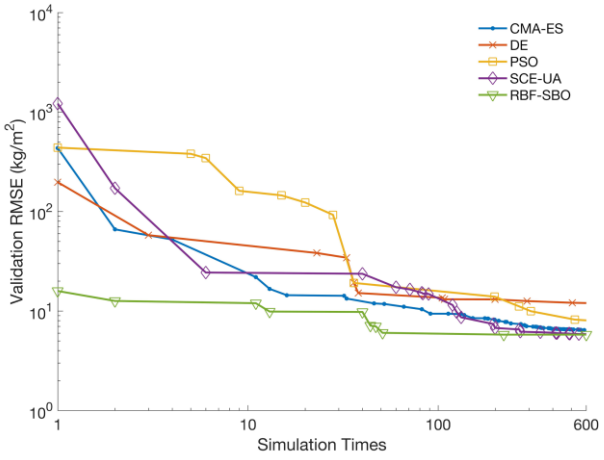



**(c) 4-pool microbial model**

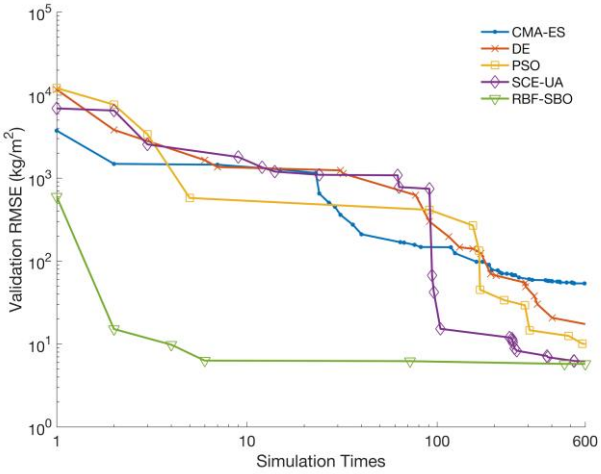
**Figure 4. The RMSEs of different optimization algorithms: (a) CLM-CASA' model; (b) 2-pool microbial model and (c) 4-pool microbial model. The box plots show the means and the quartiles spreading over total 50 calibration runs. The central line indicates the median; the bottom and top of the box are the first and third quartiles; the black bottom and top lines out of the rectangles are the maximum and minimum; the red crosses represent the outliers. The simulation times of former 5 algorithms are 100 and the simulation times of Bayesian MCMC are presented in Table 3.**




**(a) The RMSEs for CLM-CASA' model**

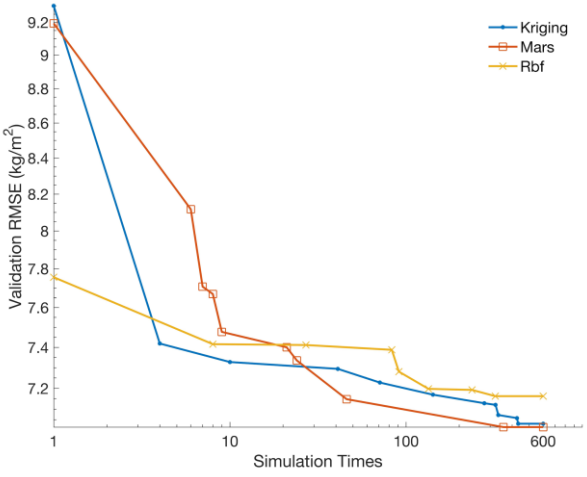



**(b) The RMSEs for 2-pool microbial model**

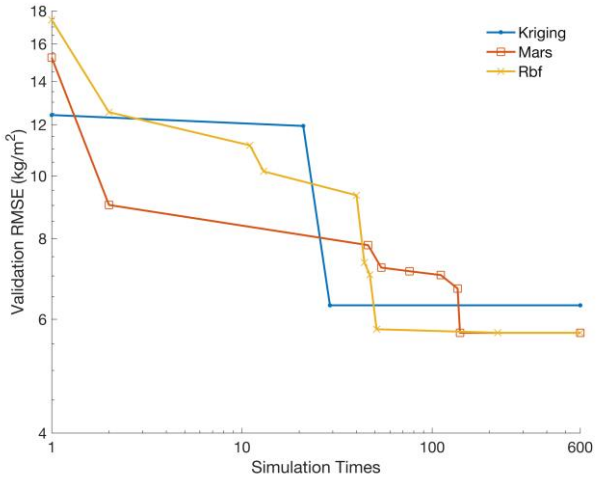



**(c) The RMSEs for 4-pool microbial model**

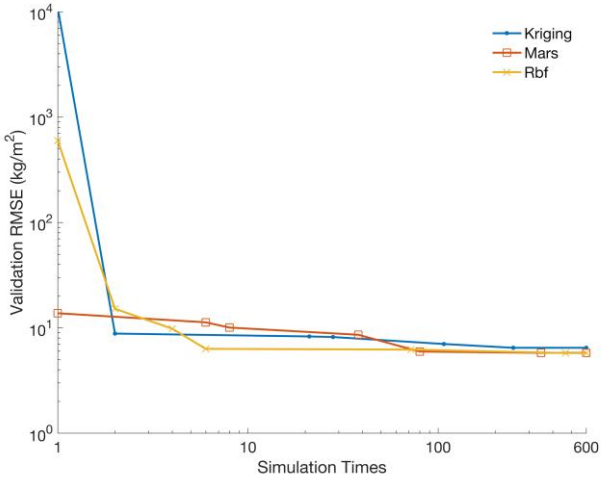
**Figure 5. The average RMSEs with the increase of simulation times and different optimization algorithms: (a) the CLM-CASA' model; (b) 2-pool microbial model and (c) 4-pool microbial model.**




**(a) The RMSEs for CLM-CASA' model**

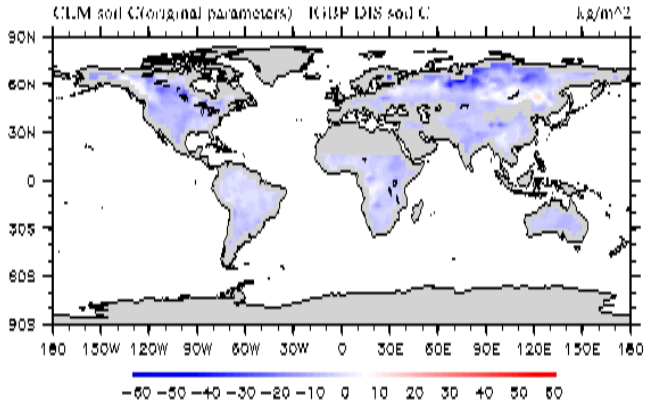



**(b) The RMSEs for 2-pool microbial model**

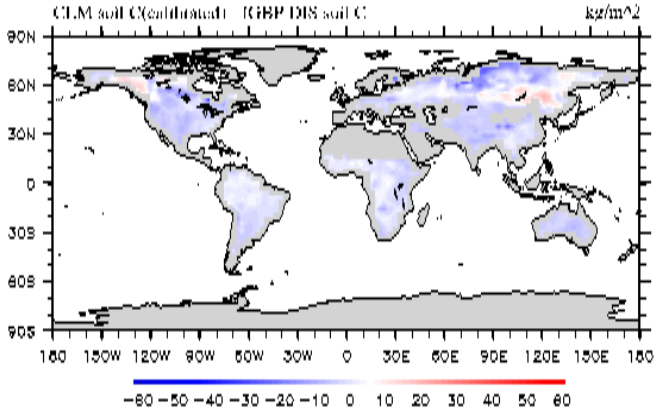



**(c) The RMSEs for 4-pool microbial model**

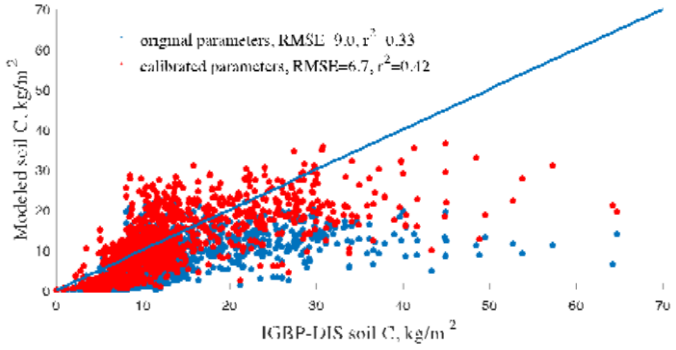
**Figure 6. The average RMSEs with the increase of simulation times and different surrogate models.**




**(a) CLM soil C – IGBP-DIS soil C**




**(b) CLM soil C (calibrated) – IGBP-DIS soil C**




**(c) Spatial correspondence between modelled soil and IGBP-DIS soil**

**Figure 7. Spatial correspondence of SOC produced by CLM-CASA' to SOC reported by IGBP-DIS. The subgraph (a) shows the results using the default parameter values and the subgraph (b) shows the results after parameter calibration using the surrogate-based optimization. The points in Fig. 7c represent the grid cell values (blue ones for the results with default parameter values and**

**red ones for the results after parameter calibration). CLM-CASA' with the default parameter values explains 33% of variation in the observed soil C, while CLM-CASA' with the calibrated parameter values explains 42% of variability in the observed soil C.**

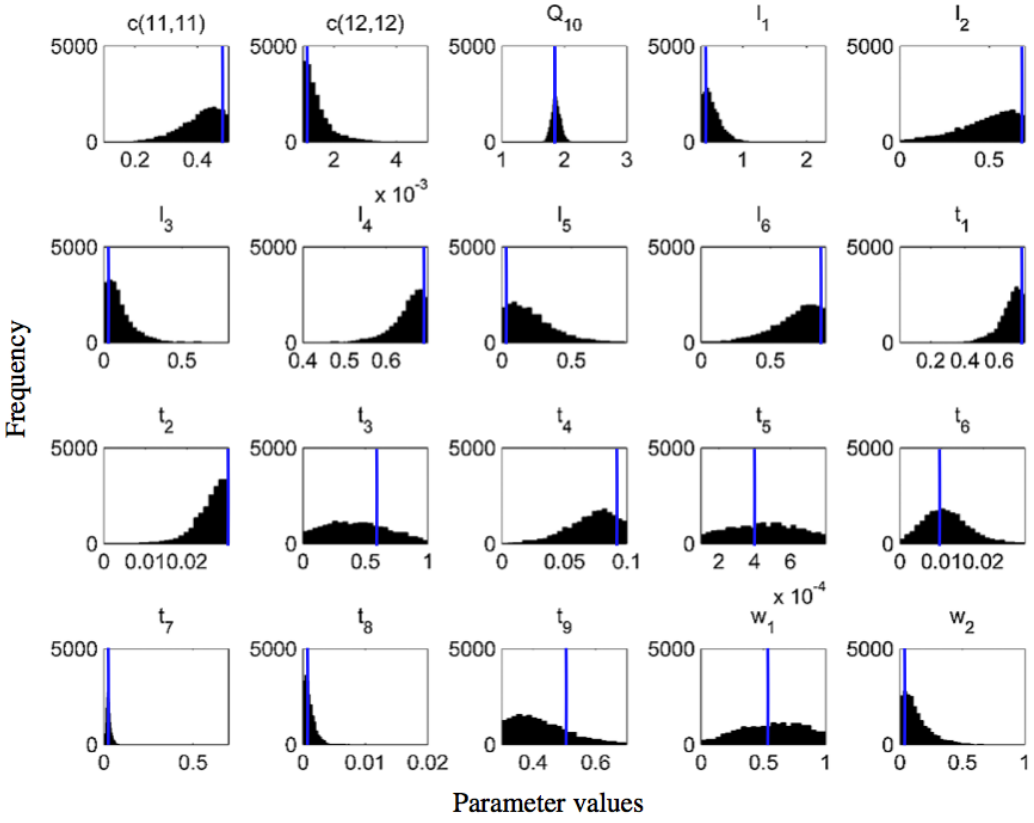



**Figure 8. Frequency distributions of 20 calibrated parameters of CLM-CASA' model by Bayesian MCMC approach (Harauk, 2014) and surrogate-based optimization (blue line in each subgraph).**

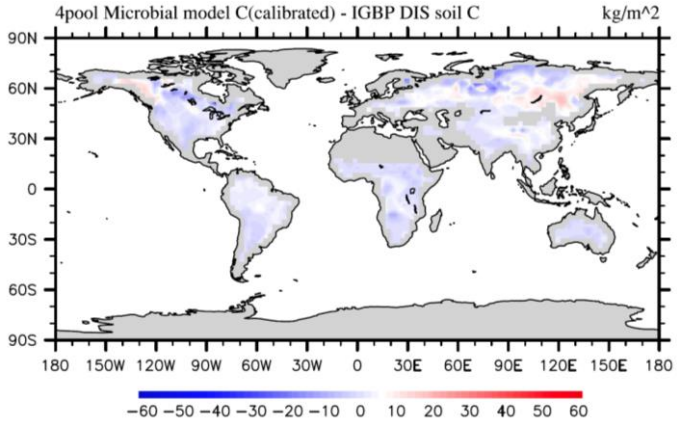



**Figure 9. Spatial correspondence of 4-pool microbial model produced SOC to the IGBP-DIS reported SOC.**

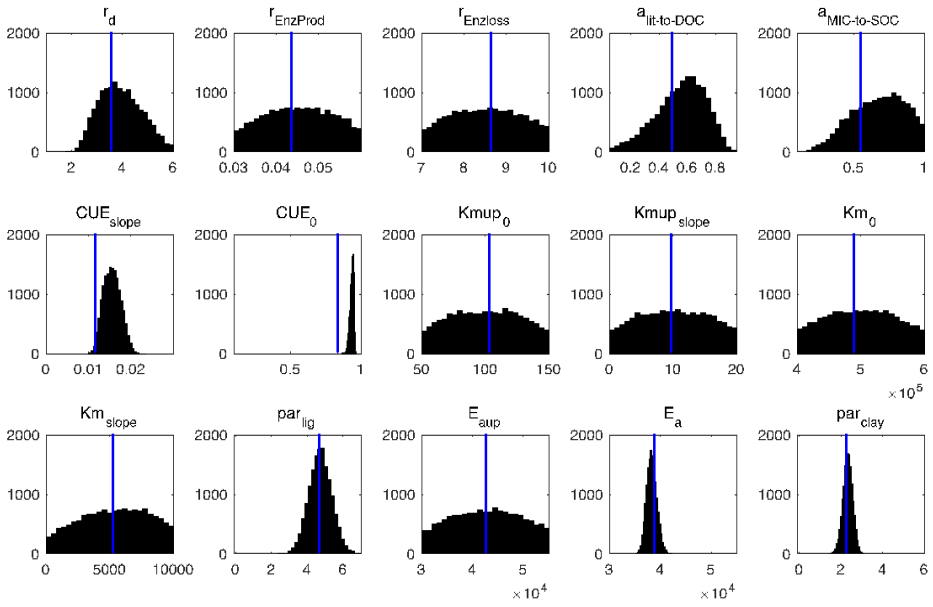



**Figure 10. Posterior probability density functions of the 4-pool microbial model parameters (generated by Bayesian MCMC). The blue vertical lines are the final calibrated parameter values by our surrogate-based optimization.**

**Table 1. Parameter description of CLM-CASA' C-only model**

| Parameter Description | Symbol | Default Value (x0.001) | Calibrated Value by SBO (x0.001) |
|---|---|---|---|
| Exit rate from slow pool | c(11,11) | 200 | 495.6 |
| Exit rate from passive pool | c(12,12) | 4.5 | 1.01 |
| Temperature sensitivity of C decomposition | $Q_{10}$ | 2000 | 1737 |
| Labile C fraction effect on C partitioning from leaves to surface metabolic litter | $w_1$ | 1000 | 589.04 |
| Labile C fraction effect on C partitioning from roots to soil metabolic litter | $w_2$ | 200 | 4.52 |
| Partitioning from surface structural to surface microbial pool if no lignin in surface structural litter | $l_1$ | 400 | 384.5 |
| Lignin effect of partitioning from surface structural litter to surface microbial litter | $l_2$ | 400 | 689 |
| Lignin effect on partitioning from surface structural litter to soil slow pool | $l_3$ | 700 | 7.499 |
| Partitioning from soil structural to soil microbial pool if no lignin in soil structural litter | $l_4$ | 450 | 697.7 |
| Lignin effect on partitioning from soil structural litter to soil microbial pool | $l_5$ | 450 | 54.46 |
| Lignin effect on partitioning from soil structural litter to soil slow pool | $l_6$ | 700 | 871.5 |
| C partitioning from soil microbial pool to slow pool if no sand or clay | $t_1$ | 169 | 747.7 |
| Clay effect on C partitioning from soil microbial pool | $t_2$ | 5.44 | 29.6 |
| Sand effect on C partitioning from soil microbial to slow pool | $t_3$ | 678 | 636.8 |
| Combined effect of sand and clay on C partitioning from soil microbial pool | $t_4$ | 22 | 99.5 |
| C partitioning from soil microbial to passive pool if no sand or clay | $t_5$ | 0.51 | 0.152 |
| Sand effect on C partitioning from soil microbial to passive pool | $t_6$ | 2.04 | 12.99 |
| Clay effect on C partitioning from slow pool to passive pool | $t_7$ | 4.05 | 24.2 |
| C partitioning from slow to passive pool if no clay | $t_8$ | 14 | 0.012 |
| C partitioning from slow to soil microbial pool if no clay | $t_9$ | 449 | 368.8 |

**Table 2. Parameter and description of the 4-pool microbial models**

| Parameter Name | Parameter Description | Default Value | Calibrated Value by SBO |
|---|---|---|---|
| $r_d$ | Microbial death rate | 4.38 | 4.89 |
| $CUE_0$ | Baseline microbial carbon use efficiency | 0.63 | 0.965 |
| $CUE_{slope}$ | $CUE_0$ dependency on temperature | 0.016 | 0.00853 |
| $Km_0$ | Baseline half saturation constant | 500000 | 498467 |
| $Km_{slope}$ | $Km_0$ dependency on temperature | 5000 | 9751 |
| $E_a$ | Activation energy of SOC decomposition | 47000 | 36669 |
| $par_{clay}$ | Clay limitation | 0 | 2.41 |
| $par_{lig}$ | Lignin-dependent correction factor | 0 | 6.23 |
| $r_{EnzProd}$ | Rate of enzyme production | 0.0438 | 0.0361 |
| $r_{EnzLoss}$ | Rate of enzyme loss | 8.76 | 8.08 |
| $a_{lit-to-DOC}$ | Fraction of $Input_{soid}$ that is transferred to soil | 0.3 | 0.832 |
| $a_{MIC-to-SOC}$ | Fraction of dead microbes transferred to soil | 0.5 | 0.716 |
| $Kmup_0$ | Baseline half-saturation constants for substrate limitation of DOC uptake | 100 | 134 |
| $Kmup_{slope}$ | $Kmup_0$ dependency on temperature | 10 | 4.62 |
| $E_{aup}$ | Activation energy of DOC uptake | 47000 | 34811 |

**Table 3 Calibration results of Bayesian MCMC and our surrogate-based optimization**

| SOC model | Detail | Method | Lowest RMSE ($kg \cdot m^{-2}$) | Variance Explained | Number of Simulations |
|---|---|---|---|---|---|
| 2-pool microbial | 8 parameters | Bayesian MCMC | 6.609 | 51.6% | 50,000+500,000 |
| | 2 carbon pools | RBF-SBO | 5.785 | 51.6% | 221 |
| 4-pool microbial | 15 parameters | Bayesian MCMC | 7.142 | 51.3% | 50,000+500,000 |
| | 4 carbon pools | RBF-SBO | 5.756 | 51.4% | 199 |
| CLM-CASA' | 20 parameters | Bayesian MCMC | 7.000 | 41.0% | 50,000+1,000,000 |
| | 13 carbon pools | RBF-SBO | 7.162 | 42.8% | 321 |