Geosci. Model Dev., 11, 3027–3044, 2018
https://doi.org/10.5194/gmd-11-3027-2018

Geoscientific Model Development
Open Access
EGU

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], **and Wei Xue**[1,2]

[1]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[2]Department of Earth System Science, Ministry of Education Key Laboratory for Earth System Modelling, Tsinghua University, Beijing 100084, China
[3]Center for Ecosystem Science and Society, Northern Arizona University, Flagstaff, AZ, USA

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

Received: 28 February 2017 – Discussion started: 5 April 2017
Revised: 2 May 2018 – Accepted: 9 May 2018 – Published: 27 July 2018

Abstract. Soil organic carbon (SOC) has a significant effect on carbon emissions and climate change. However, the 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 techniques have been successfully employed for the parameter calibration of SOC models. However, data assimilation algorithms, such as the sampling-based Bayesian Markov chain Monte Carlo (MCMC), generally have high computation costs and are not appropriate for complex global land models. This study proposes a new parameter calibration method based on surrogate optimization techniques to improve the prediction accuracy of SOC. Experiments on three types of soil carbon cycle models, including the Community Land Model with the Carnegie–Ames–Stanford Approach biogeochemistry submodel (CLM-CASA') and two microbial models show that the surrogate-based optimization method is effective and efficient in terms of both accuracy and cost. Compared to 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 by up to 12 % for different SOC models. Meanwhile, the corresponding computational cost is lower than other global optimization algorithms.

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 CO_2, the most important greenhouse gas, from land ecosystems greatly depends on the amount of carbon stored in soils. Moreover, anthropogenic CO_2 emission leads to climate warming (Houghton et al., 2001), which further stimulates soil carbon release, forming a positive feedback between the carbon cycle and climate warming (Melillo et al., 2002; Friedingstein et al., 2006; Luo, 2007). In the fifth phase of the Coupled Model Intercomparison Project (CMIP5), the outputs of 11 Earth system models (ESMs) show great uncertainty in the SOC predictions. Despite the similarity in model structures (Huang et al., 2017), simulated soil carbon content varies six-fold, ranging from 510 to 3040 Pg of carbon, among the models (Todd-Brown et al., 2013). Only half of the 11 models have a predicted global total SOC which falls within the estimated range of the (HWSD). Modeled SOC is hardly agrees with the observation (Luo et al., 2015).

Considering the high similarity in the structures of the 11 ESMs, the difference in the SOC simulations mainly results from parameterizations (Todd-Brown et al., 2013). Thus, parameter calibration is among the top priorities regarding the improvement of the prediction of global land carbon cycle dynamics (Luo et al., 2016). However, parameter calibration with global observations has not been widely applied, owing to the high computational cost. A matrix approach has recently been developed to reorganize the carbon balance equations in the original ESMs into one matrix equation without changing any modeled C (carbon) cycle processes and

mechanisms (Luo et al., 2003, 2017; Huang et al., 2018). The matrix land carbon cycle model can be semi-analytically saved to obtain steady-state solutions and save more than 90 % computational time compared to original models (Xia et al., 2012). As a consequence, the matrix approach makes parameter estimation and calibration possible. The matrix approach has been successfully used for parameter calibration to constrain SOC turnover and microbial process using the Bayesian Markov chain Monte Carlo (MCMC) algorithm (Harauk et al., 2014, 2015; Shi et al., 2018).

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 the parameter calibration of soil carbon models (Xu et al., 2006). Even using high-performance computers, complex land models, like the latest version of Community Land Model (CLM5.0), require a very long spin-up time for carbon cycle simulation, leading to several hours or days for one simulation. Although the matrix approach has been developed to enable data assimilation of global land carbon cycle models (Harauk et al., 2014, 2015; Shi et al., 2018), Bayesian MCMC is still very computationally expensive for calibrating global land models. More effective and efficient parameter calibration algorithms are urgently needed.

The parameter calibration of SOC models can be formulated as an optimization problem that aims to minimize the output of a cost function. This cost 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 nonlinear, non-convex, and black-box problems (Hapuarachchi et al., 2001; Ma et al., 2006; Rocha, 2008). Unfortunately, the number of required simulations for most global optimizations is very large.

To reduce the number of simulations and decrease the computational cost, we, for the first time, present a surrogate-based optimization method (SBO) for calibrating soil carbon models. Surrogate models serve as computationally inexpensive 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 a new promising point in the parameter space at which the expensive simulation model originally has to be evaluated. With the help of the surrogate model, many unnecessary simulations with bad parameter values, which lead to high prediction errors, are avoided. SBO has been shown to find the near-optimal parameter values within only a 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 Optimizers", abbreviated as COCO (Hansen et al., 2010; Wang et al., 2014). However, the optimization of the mathematical functions may be extremely different from the parameter calibration of complex real-world models. In this paper, we explore the state-of-the-art surrogate optimization method for parameter calibration of three SOC models: CLM coupled with Carnegie–Ames–Stanford Approach biogeochemistry submodel (CLM-CASA'), and two microbial models such as those used in studies by Hararuk et al. (2014, 2015). Although the three models are computationally attainable for parameter calibration, we compare the performance of surrogate-based optimization to advanced global optimization algorithms and the data assimilation method to examine the potential of SBO. The SBO method may be extended to other complex global land models.

In this paper, we present the structure and parameters of three SOC models in Sect. 2. Section 3 introduces the algorithm design of SBO. The parameter calibration results and the analysis of different parameter calibration algorithms are presented in Sect. 4. Section 5 discusses the calibrated results using SBO. Finally, we draw conclusions in Sect. 6.

2 Global land carbon models, data and cost function

Earth system models (ESMs) are a fundamental tool for simulating climate impacts on the carbon cycle at the global scale. There are many common properties among structures of different global land carbon modules of ESMs (Luo et al., 2016). Almost all models have multiple carbon pools and carbon is transferred among these pools (Weng and Luo, 2011). In this study, we selected three SOC models, which have previously been calibrated for their parameters with the Bayesian MCMC algorithm (Hararuk et al., 2015). The first model is the soil carbon component of CLM coupled with the 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). The other two SOC models are microbial models, which consider the nonlinear regulation of SOC dynamics with microbial biomass.

2.1 CLM-CASA' C-only version model

The CLM-CASA' is embedded in CLM3.5. The latter includes biogeophysics and biogeochemistry submodels. CLM-CASA' inputs carbon through net primary productivity (NPP), which is partitioned to three live biomass pools (wood, leaves and fine roots) (Fig. 1a). Dead plant materials become litter and are transferred separately to four litter pools. Litter decomposition results in part of the carbon being released to the atmosphere as heterotrophic respiration and part of the carbon being stabilized into soil carbon pools. Organic carbon in the soil pools is partially decomposed to be released as CO_2 via microbial respiration and partially incorporated into other soil carbon pools. One of the key model

outputs to indicate SOC dynamics is the total soil organic carbon content, which is the sum of carbon in soil microbial (or active) pools, slow pools and passive pools (Fig. 1).

The CLM-CASA' model simulates soil carbon decomposition as a first-order decay process (Oleson et al., 2004, 2008). Based on theoretical analysis, the carbon cycle of most ESMs can be summarized with one matrix equation (Luo et al., 2001; Luo and Weng, 2011; Xia et al., 2013) as follow:

$$\frac{\mathrm{d}X(t)}{\mathrm{d}t} = \mathbf{A}\boldsymbol{\xi}(t)\mathbf{K}X(t) + BU(t). \quad (1)$$

In the abovementioned equation $X(t)$ is the carbon content of different pools; $\frac{\mathrm{d}X(t)}{\mathrm{d}t}$ is the change of the carbon content; $\mathbf{A}$ is a matrix of transfer coefficients among different pools; $\boldsymbol{\xi}(t)$ and $\mathbf{K}$ are both diagonal matrixes, representing environmental scaling factors and baseline carbon decomposition rates, respectively; $U(t)$ is NPP, the carbon influx into the whole system; and B represents the partitioning coefficients of the carbon influx among plant pools. The steady-state solution of equation is given by Xia et al. (2012) as follows:

$$X_{\mathrm{ss}} = -(\mathbf{A}\underline{\boldsymbol{\xi}}K)^{-1}\underline{BU}, \quad (2)$$

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

2.2 The microbial models

Microorganisms catalyze various processes of the land carbon cycle, such as the decomposition and stabilization of SOC (Kuzyakov et al., 2000; Luo et al., 2001; Peng et al., 2009). However, most conventional SOC models, such as CLM-CASA', do not explicitly represent microbial processes. Microbially explicit models usually represent SOC decomposition by considering extracellular enzyme activities rather than simple decay constants as in the CLM-CASA' and other traditional SOC models (Schimel and Weintraub, 2003). In this study, we focus on two enzyme-driven decomposition models; one has two pools (Fig. 1b) introduced by German et al. (2012) and the other has four pools (Fig. 1c) introduced by Allison et al. (2010). We call these two models the two-pool microbial model and the four-pool microbial model, respectively. C inputs for the two models are NPP and the outputs are the carbon content of each pool at a steady state.

The two-pool microbial model is described using the following equations: (Hararuk et al., 2015).

$$\frac{\mathrm{dMIC}}{\mathrm{d}t} = \mathrm{CUE} \times V_{\max} \times \mathrm{MIC}\frac{\mathrm{SOC}}{K_m + \mathrm{SOC}} - r_{\mathrm{d}} \times \mathrm{MIC} \quad (3)$$

$$\frac{\mathrm{dSOC}}{\mathrm{d}t} = \mathrm{Input}_{\mathrm{soil}} + r_{\mathrm{d}} \times \mathrm{MIC} - V_{\max} \times \mathrm{MIC}\frac{\mathrm{SOC}}{K_m + \mathrm{SOC}}. \quad (4)$$

Where

$$\mathrm{CUE} = \mathrm{CUE}_{\mathrm{slope}} \times T_{\mathrm{s}} - \mathrm{CUE}_0 \quad (5)$$

$$V_{\max} = V_{\max_0} \times \exp\left(-\frac{E_{\mathrm{a}}}{R \times (T_{\mathrm{s}} + 273)}\right) \times \exp(-\mathrm{par}_{\mathrm{clay}} \times \mathrm{clay}) \quad (6)$$

$$\mathrm{Km} = \mathrm{Km}_{\mathrm{slope}} \times T_{\mathrm{s}} + \mathrm{Km}_0 \times \exp(\mathrm{par}_{\mathrm{lig}} \times \mathrm{lignin}). \quad (7)$$

In the abovementioned, MIC represents the microbial biomass; $V_{\max}$ is the temperature adjusted rate of SOC decomposition; K_m is the half-saturation constant for the substrate-limited SOC decomposition rate; r_{d} is the microbial death rate; CUE is the microbial carbon use efficiency; $\mathrm{Input}_{\mathrm{soil}}$ is the carbon influx to soil, a 30-year averages of soil C input produced by CLM-CASA' (Hararuk et al., 2015); T_{s} is soil temperature; R is the gas constant ($8.31\,\mathrm{J\,K^{-1}\,mol^{-1}}$); CUE_0 and $\mathrm{CUE}_{\mathrm{slope}}$ are the baseline microbial carbon use efficiency and its dependency on temperature, respectively; $V_{\max 0}$ is the maximum rate of microbial carbon uptake; E_{a} is the activation energy of SOC decomposition; Km_0 and $\mathrm{Km}_{\mathrm{slope}}$ are the baseline half-saturation constant and its dependency on temperature, respectively; lignin is lignin content; and $\mathrm{par}_{\mathrm{lig}}$ is a parameter to regulate the lignin-dependent correction factor. See Table 2 for a more comprehensive description of those parameters.

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

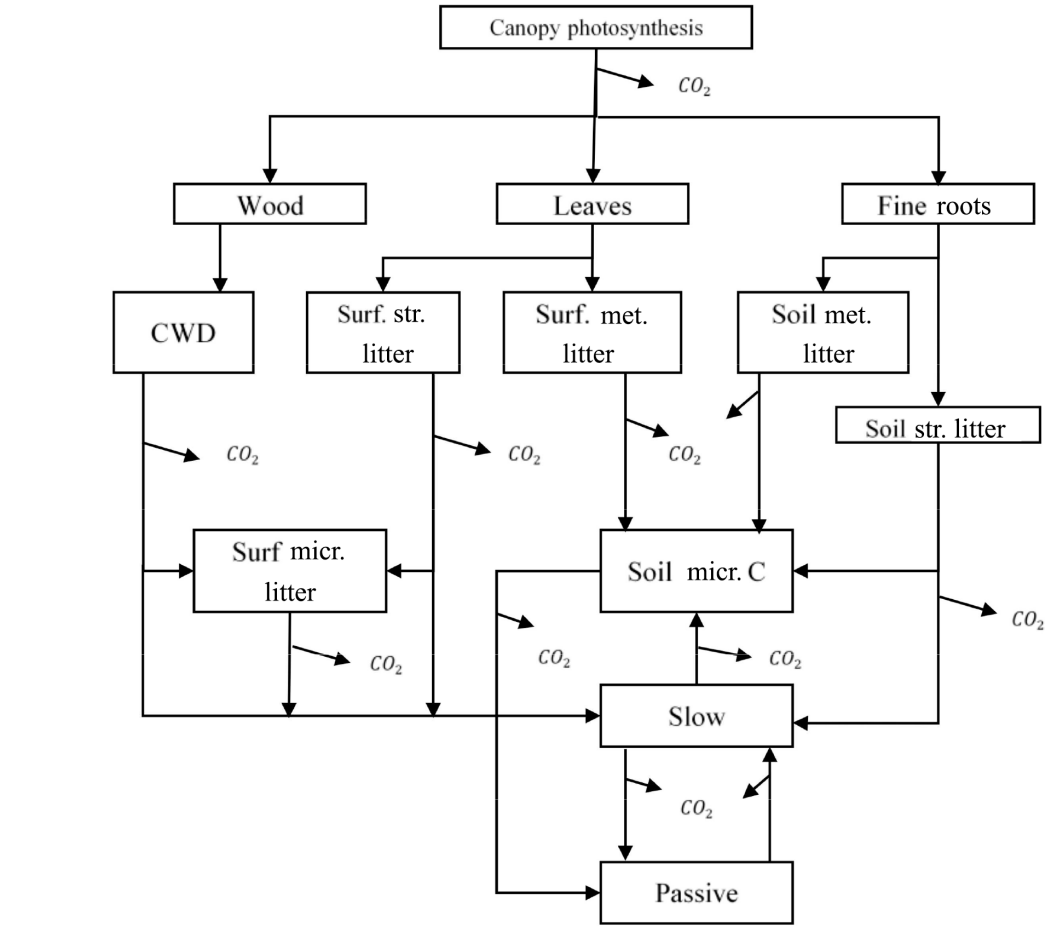
$$\frac{\mathrm{dMIC}}{\mathrm{d}t} = V_{\mathrm{maxup}} \times \mathrm{MIC}\frac{\mathrm{DOC}}{\mathrm{Kmup} + \mathrm{DOC}} \times \mathrm{CUE} - r_{\mathrm{d}} \times \mathrm{MIC} - r_{\mathrm{EnzProd}} \times \mathrm{MIC} \quad (8)$$

$$\frac{\mathrm{dDOC}}{\mathrm{d}t} = a_{\mathrm{lit-to-DOC}} \times \mathrm{Input}_{\mathrm{soil}} + r_{\mathrm{d}} \times \mathrm{MIC} \times (1 - a_{\mathrm{MIC-to-SOC}}) + V_{\max} \times \mathrm{ENZ}\frac{\mathrm{SOC}}{\mathrm{Km} + \mathrm{SOC}} + r_{\mathrm{EnzLoss}} \times \mathrm{ENZ} - V_{\mathrm{maxup}} \times \mathrm{MIC}\frac{\mathrm{DOC}}{\mathrm{Kmup} + \mathrm{DOC}} \quad (9)$$

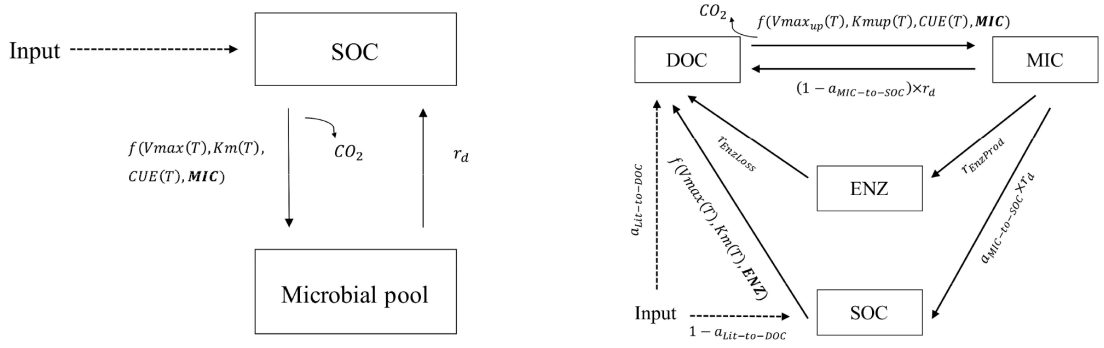
$$\frac{\mathrm{dSOC}}{\mathrm{d}t} = a_{\mathrm{lit-to-SOC}} \times \mathrm{Input}_{\mathrm{soil}} + r_{\mathrm{d}} \times \mathrm{MIC} \times a_{\mathrm{MIC-to-SOC}} - V_{\max} \times \times \mathrm{ENZ}\frac{\mathrm{SOC}}{\mathrm{Km} + \mathrm{SOC}} \quad (10)$$

$$\frac{\mathrm{dENZ}}{\mathrm{d}t} = r_{\mathrm{EnzProd}} \times \mathrm{MIC} - r_{\mathrm{EnzLoss}} \times \mathrm{ENZ}, \quad (11)$$

where ENZ and DOC are enzyme and dissolved organic carbon pools, respectively; V_{maxup} is the temperature-adjusted rate of DOC uptake by microbes; Kmup is a half-saturation



(a) The CLM-CASA' model



(b) Two-pool microbial model **(c) Four-pool microbial model**

Figure 1. Schematic representations of **(a)** a CLM-CASA' model, **(b)** a two-pool microbial model and **(c)** a four-pool microbial model. Surf.: surface; str.: structural; met.: metabolic; micr.: microbial.

constant limiting the microbial uptake of DOC; r_{EnzProd} is a rate of enzyme production; $\mathrm{Input}_{\mathrm{soil}}$ is C transferred from litter to soil; $a_{\mathrm{lit-to-DOC}}$ is the fraction of $\mathrm{Input}_{\mathrm{soil}}$ that is transferred to DOC; $a_{\mathrm{MIC-to-SOC}}$ is the fraction of dead microbes transferred to soil; and r_{EnzLoss} is the rate of enzyme

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

Parameter description	Symbol	Default value (×0.001)	Calibrated value by SBO (×0.001)
Decomposition rate from slow pool.	c(11,11)	200	495.6
Decomposition 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 is present 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 is present 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 is present.	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 is present.	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 is present.	t_8	14	0.012
C partitioning from slow to soil microbial pool if no clay is present.	t_9	449	368.8

loss. The temperature-dependent functions are follows:

$$\mathrm{CUE} = \mathrm{CUE}_{\mathrm{slope}} \times T_{\mathrm{s}} - \mathrm{CUE}_0 \quad (12)$$

$$V_{\mathrm{maxup}} = V_{\mathrm{maxup}_0} \times \exp\left(-\frac{E_{\mathrm{aup}}}{R \times (T_{\mathrm{s}} + 273)}\right) \quad (13)$$

$$\mathrm{Kmup} = \mathrm{Kmup}_{\mathrm{slope}} \times T_{\mathrm{s}} + \mathrm{Kmup}_0 \quad (14)$$

$$V_{\mathrm{max}} = V_{\mathrm{max}_0} \times \exp\left(-\frac{E_{\mathrm{a}}}{R \times (T_{\mathrm{s}} + 273)}\right) \times \exp(-\mathrm{par}_{\mathrm{clay}} \times \mathrm{clay}) \quad (15)$$

$$\mathrm{Km} = \mathrm{Km}_{\mathrm{slope}} \times T_{\mathrm{s}} + \mathrm{Km}_0 \times \exp(\mathrm{par}_{\mathrm{lig}} \times \mathrm{lignin}), \quad (16)$$

where V_{maxup0} is the maximum rate of microbial DOC uptake; E_{aup} is the activation energy of DOC uptake; Kmup_0 and $\mathrm{Kmup}_{\mathrm{slope}}$ are baseline half-saturation constants for substrate limitation of DOC uptake and its dependency on temperature, respectively.

Fifteen parameters of the four-pool microbial model are also described in Table 2.

2.3 Data and cost function

Microbial models and CLM-CASA' C-only models divide the world into 64 × 128 grid cells and output SOC content for 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 that has been widely used in many studies to evaluate and improve models (Zhou et al., 2009; Smith et al., 2013).

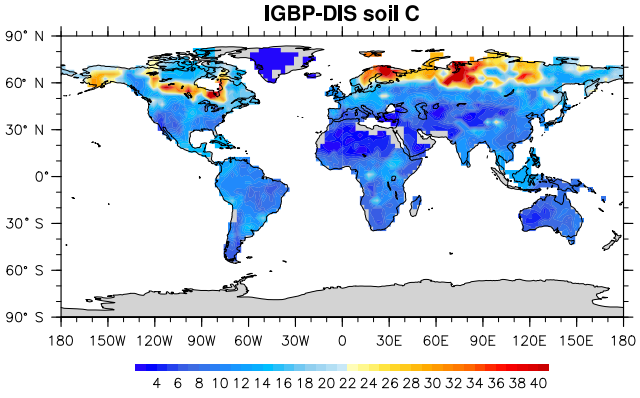


Figure 2. IBGP-DIS soil carbon distribution. Soil carbon varies from 0 kg m^{-2} in deserts to 60 kg m^{-2} in boreal regions.

The goal of parameter calibration is to improve the 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 cost function. This cost function can be described by the following formula:

Table 2. Parameter and description of the four-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.	500 000	498 467
Km_{slope}	Km_0 dependency on temperature.	5000	9751
E_a	Activation energy of SOC decomposition.	47 000	36 669
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.	47 000	34 811

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

where N denotes the total number of grid cells, and X_i and O_i are the SOC of the model prediction and the IGBP-DIS observation, respectively. To avoid overfitting and evaluate the calibrated parameters more fairly, we separate all grid cells into a training set and a 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, 2015) also used this method when calibrating SOC parameters with the Bayesian MCMC approach. The experiment results in Sects. 3 and 4 refer to the results for the validation set.

3 Surrogate-based optimization algorithm design

3.1 Introduction to the surrogate-based optimization algorithm

The parameters of most soil carbon models and land models have traditionally been tuned manually (Luo et al., 2001, 2016). The manual tuning method might be effective for simple models but still highly depends on expert experience. Complex models may consist of various components from different disciplines and have hundreds or thousands of parameters. In these cases manual tuning consequently becomes impractical.

Different parameter calibration algorithms are available, which have been developed based on optimization theory. Gradient search algorithms like the quasi-Newton method are introduced to search for a set of parameters with better performance in the parameter domain. These algorithms are usually efficient and fast. However, gradient search algorithms are designed for finding the local optimum. They cannot be used to solve multimodal problems derived from complex Earth system models. In addition, they are based on the gradient information which is unavailable for most soil carbon and land models, as these models are too complex to obtain such details. Thus, parameter calibration usually becomes a black-box optimization problem. Global optimization algorithms, such as genetic algorithms and particle swarm optimization algorithms, are based on parameter generation and selection strategies. They are basically still gradient independent but can be easily used for parameter calibration of complex Earth system models. Global optimization algorithms are designed to find the global minimum. However, the number of samples (model runs) might still be too large to be applicable to complex models with large numbers of parameters (Jones et al., 1998). Moreover, complex Earth system models, for example CLM, require several hours over hundreds of cores for just one sample run and pose a special challenge regarding the feasibility of automatic parameter calibration.

SBO is an efficient and effective automatic parameter calibration framework. It fits a surrogate model (or response surface) based on the previous samples and uses this surrogate model to emulate the output behaviors of original models with an acceptable level of accuracy. The main advantage of SBO is to save computational costs during the global optimization by using the surrogate model instead of the original model. Furthermore the surrogate model can be continuously improved 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, the time-to-solution, as well as the computation cost. SBO has been successfully used to solve the pa-

rameter calibration of computationally expensive black-box problems (Vu et al., 2016).

3.2 Key components of the surrogate-based optimization algorithm

A flowchart describing the SBO is presented in Fig. 3.

First, initial sets of parameter values are generated using a sampling method. These sets are then used as inputs to run the real simulation model. Second, a surrogate model is constructed by fitting the outputs of these sample runs. The surrogate model serves as a computationally inexpensive 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 a 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 (typically the maximum number of simulations allowed) are met, the algorithm returns the optimized parameter values. During the SBO process, quite a few sample runs are generated based on the evaluation of the surrogate model and most meaningless simulations with bad parameter values are avoided. As a result, the computationally expensive model is only simulated at a few selected promising parameter points, and the surrogate model replaces the real model during the calibration process. Thus, the computation cost is substantially reduced.

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

- 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 (these data are called the training set).
- How to decide the new points at which to run the real model in each iteration.

For the initial sampling, Monte Carlo sampling and Latin hypercube sampling (LHS) are two main sampling methods (McKay et al., 1979; Iman et al., 1981). In Monte Carlo sampling values are sampled from a probability distribution, which is generally a uniform distribution unless we have additional knowledge about the model and the parameters. During the 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 random sampling, LHS ensures that the ensemble of random numbers is representative of the real variability of the parameters. As a result, we use LHS to generate the initial set S_0 (Iman et al., 1981).

There are various surrogate models, such as multivariate adaptive regression splines (MARS) (Friedman, 1991), polynomial regression models (Myers et al., 2016), radial basis functions (RBFs) (Gutmann, 2001; Müller and Shoemaker, 2014; Powell, 1992; Regis and Shoemaker, 2007, 2009; Wild and Shoemaker, 2013), and kriging (Davis and Lerapetritou, 2009; Forrester et al., 2008; Jones et al., 1998).

The MARS model is an extension of naïve linear models introduced by J. H. Friedman (Friedman, 1991). The form of MARS is presented as follows:

$$\hat{f}(x) = \sum_{i=1}^{m} c_i B_i(x), \tag{18}$$

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

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

$$\hat{f}(x) = \sum_{i=1}^{N} w_i \, \phi(\|x, x_i\|), \tag{19}$$

where 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 of these 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. The x and y should be normalized to satisfy a normalization distribution where the means is zero and the covariance is 1 one before they are used to train the kriging model. The kriging predictor can be found as follows:

$$\hat{f}(x) = \hat{\mu} + \sum_{i=1}^{n} c_i r_i(x), \tag{20}$$

where $\hat{\mu}$ is the estimated mean of the Gaussian process, c_i is a constant representing the weight and $r_i(x) = \text{Corr}(x, x^{(i)})$ 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.

In addition, many machine learning regression models are also introduced, such as support vector regression (Zhang et al., 2009), artificial neural network (Behzadian et al., 2009) and random forest (Breiman, 2001).

The strategies of parameter point generation are iterative algorithms that use data acquired from previous iterations to

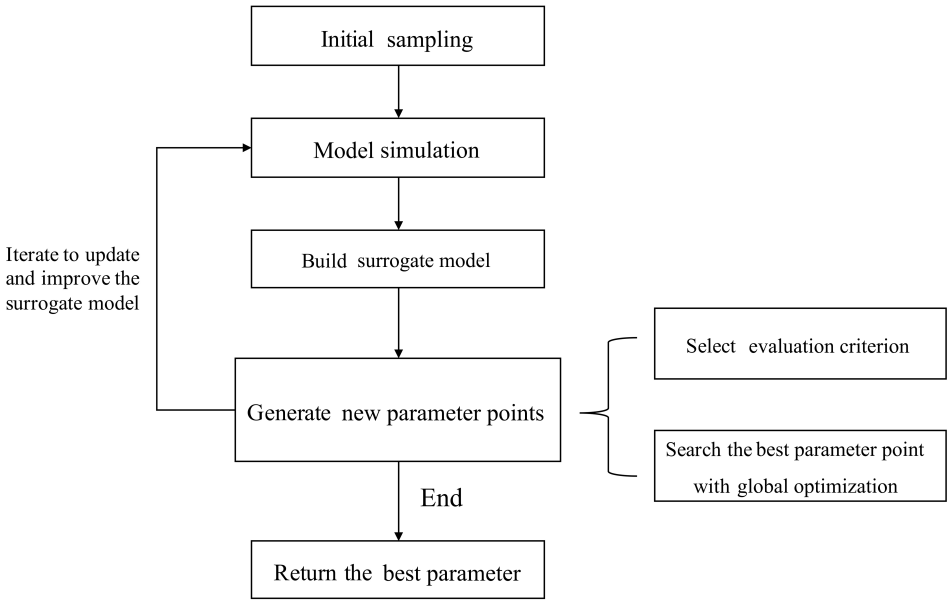


Figure 3. A flowchart of the surrogate-based optimization.

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., 2012). 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. Exploration refers to searching in an unfamiliar area of the parameter space to learn about it and avoiding being trapped in the local optimum. Exploitation means fast convergence in an area. Balancing the exploration and exploitation ensures SBO can find real global optimum and does not waste simulations on meaningless parameter sets and areas. Another parameter generation strategy is the 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 previously sampled points represent the explored region and we can estimate the uncertainty from the distance to the explored region. A weighted sum of these two criteria is used to determine the new parameter point during the SBO.

3.3 Design of the surrogate-based optimization algorithm for soil carbon models

Elaborating on the previous introduction of SBO, a detailed account of the SBO procedure follows.

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 outputs generated in Step 2.

Step 4: Predict the output errors of those points that 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 from Step 4 and calculate the output errors of these selected points.

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

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

The SBO scheme mentioned in previous sections is a parameter calibration framework. The key components introduced in Sect. 3.2 must be selected when calibrating the parameters of soil carbon models. The LHS can cover the whole parameter space with a limited number of sample points while Monte Carlo sampling usually requires a 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 do not perform as well as RBF and kriging models, according to an evaluation of 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 shown to perform better than other surrogate models (Müller and Shoemaker, 2014) and is easily implemented. Our algorithm framework also includes other surrogate models, such as kriging and MARS, and can introduce others in the future.

The surrogate model is not accurate enough to represent complex and nonlinear models when the SBO starts. The MIS can be very efficient but is easy to trap into local optima, as the strategy does not consider the uncertainty of the surrogate model and only selects 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 this (Jones et al., 1998). Finally, we use the CAND strategy as the parameter generation strategy in our algorithm to balance the exploitation and exploration of the 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 for comparison with our proposed SBO method. Our SBO algorithm is implemented based on the "Surrogate Model Optimization Toolbox" toolkit (Müller, 2014). Three SOC models and their cost functions are introduced in Sect. 2. The target of parameter calibration is to find the optimal values of parameters to achieve the minimum value of the cost 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 in terms of both effectiveness and efficiency. The effectiveness refers to the accuracy of the calibrated results and the efficiency can be evaluated by the required number of simulations of the original SOC models.

4.2 Various global optimization algorithms and the Bayesian MCMC approach

The Bayesian MCMC approach and the four following advanced global optimization algorithms: differential evolution (DE), particle swarm optimization (PSO), shuffled complex evolution (SCE-UA) and the covariance matrix adaption evolution strategy (CMA-ES), are compared with our RBF-SBO.

DE (Storn and Price, 1997) and PSO (Kennedy, 2011; 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 et al., 2006; Shi and Eberhart, 2009). SCE-UA is designed for the parameter calibration of hydrologic models and has been successfully applied to various hydrology models such as the TOPMODEL, the Xinanjiang watershed model and short-term load forecasting (Hapuarachchi et al., 2001; Ma et al., 2006; Li et al., 2007). SCE-UA tries to keep both effectiveness and efficiency by combining the local (the simplex method) and global optimization methods. Despite the difference in algorithm details, DE, PSO and SCE-UA all generate new parameter points according to some simple mathematical formulas. Unlike these three algorithms, CMA-ES creates new parameter points based on a multivariate normal distribution (Hansen and Ostermeier, 2001; Hansen and Kern, 2004). The dependencies between parameters are represented by the covariance matrix of a normal distribution. CMA-ES has been shown to be the best global optimization algorithm in the BBOB-2009 comparison study (Hansen, 2009).

The Bayesian MCMC approach is typically 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 consists of two main parts: sampling and parameter estimation. During the sampling part, the adaptive Metropolis (AM) algorithm, a Markov chain Monte Carlo method, is used to conduct sampling from the prior parameter distributions and generate a parameter chain (Haario et al., 2001). The AM algorithm has two steps: the proposing step and the moving step. A new parameter set p^{k+1} is generated from the previously accepted parameter set p^k through a proposal distribution $q(p^{k+1}|p^k)$. In the moving step, the probability of acceptance is calculated according to the Metropolis criterion (Xu et al., 2006). The parameter set that is not accepted is discarded. The AM algorithm repeats the proposing step until the new parameter set is accepted. The accepted new parameter set becomes the p^{k+1} set of accepted parameters in the posterior parameter distribution (Marshall et al., 2004). The proposal step is usually repeated for 50 000 to 1 000 000 times to generate enough accepted parameter sets for the posterior parameter distribution. The posterior distribution is used to estimate the maximum likelihood estimator (MLE). Hararuk et al. (2014, 2015) applied the Bayesian MCMC approach to the parameter calibration problem of the CLM-CASA' C-only model and microbial models. They also conducted experiments in which the proposing step required 50 000 simulations for microbial models and 1 000 000 simulations for the CLM-CASA' model. We used the code from Hararuk et al. (2014, 2015) and repeated the calibration experiments. The detailed

calibration results from the Bayesian MCMC approach are presented in Table 3.

4.3 Results and analysis

4.3.1 Effectiveness and efficiency

Figure 4 presents the calibrated results (RMSE) of the different algorithms we applied. For each algorithm, we only perform 100 simulations to compare the effectiveness if the number of simulations is limited. As the Bayesian MCMC approach requires a large number of samples to reach a stable distribution, over 500 000 simulations were conducted for the algorithm evaluation.

Clearly, the average RMSE of the RBF-SBO is the lowest ($0.6\,kg\,m^{-2}$ better than the Bayesian MCMC algorithm) for the two microbial models among all the algorithms (Fig. 4b, c). For the CLM-CASA' model, our RBF-SBO algorithm is still superior to the global optimization algorithms. The Bayesian MCMC approach performs slightly better (about $0.02\,kg\,m^{-2}$) but requires many more simulations to achieve the results (Fig. 4a).

The results of RBF-SBO also indicate less variation among the 50 repeated experiments than the global optimization algorithms for the three models. For the same reason previously mentioned, the Bayesian MCMC approach has less variation than our RBF-SBO algorithm for the two microbial models. For the CLM-CASA' model, our RBF-SBO is still promising to get stable results. Among the global optimization algorithms, CMA-ES shows a very significant fluctuation (Fig. 4b, c), indicating that it 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, RBF-SBO is the most effective and stable algorithm when the number of simulations is limited.

Figure 5 shows the results in terms of average validation RMSE. We don't compare the efficiency of Bayesian MCMC since it is in nature a sampling algorithm, not an optimization algorithm. The average validation RMSE of RBF-SBO is lower than the four global optimization algorithms before the number of simulations increases to 600 for the two microbial models and to 200 for the CLM-CASA' model, respectively. Our RBF-SBO requires fewer simulations than the global optimization algorithms when they reach the same RMSE value and accuracy range. Thus, our RBF surrogate optimization is the most efficient algorithm, requires the minimum number of simulations and has the lowest computational cost. Compared to the global optimization algorithms, SBO has two main advantages. Firstly, SBO samples an initial parameter set to build a surrogate model; this building process is a learning process which can better understand the parameter space and thus help conduct better optimization. Secondly, SBO can avoid some bad parameter points ("bad" means high prediction error), which are not supposed to be evaluated with the help of the surrogate model.

Another important observation is that the difference between the results of our RBF-SBO and the 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 quickly build the surrogate model with relatively good accuracy, which helps find a near-optimal solution with lower computation cost. However, the surrogate model is only an approximation of the real model and accuracy might be limited due to the strong nonlinearity and high complexity of the real model. After gaining sufficient knowledge of the original model through many simulations, the excellent global optimization algorithms, such as CMA-ES, may achieve similar performance to or even outperform our SBO, which suggests that our SBO is better fitted for use regarding the parameter calibration of cost-intensive models.

4.3.2 Impact of the model complexity

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

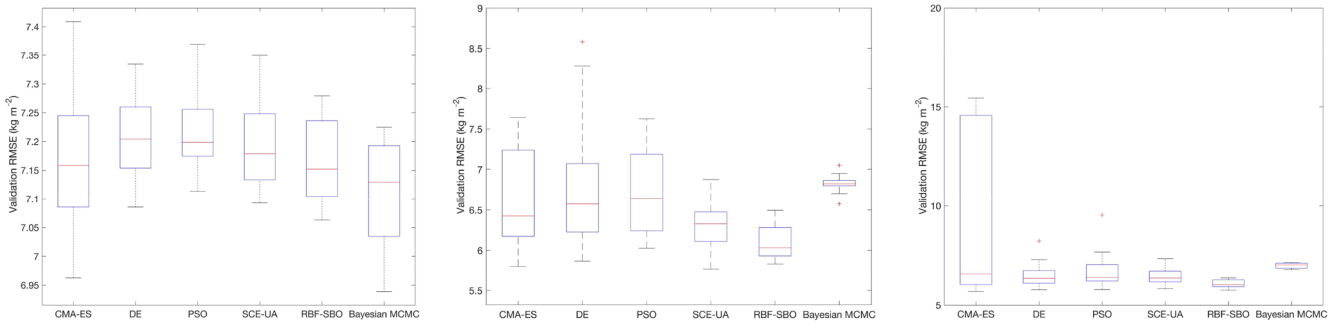
4.3.3 Impact of different types of surrogate models

We select the RBF as the surrogate model in the experiments because the RBF is widely adopted in many SBO algorithms (Müller and Shoemaker, 2014). In this section, we also test two other typical surrogate models, kriging and MARS. The MARS model is simple and has almost no requirements for the sample quality. MARS is very quick to train and predict. Kriging, also known as Gaussian process regression, is a method of interpolation for which the interpolated values are modeled by a Gaussian process governed by prior covariance. Kriging provides 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 have reasonable performance in the parameter calibration of the three types of SOC models and perform better than

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

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



(a) CLM-CASA' model **(b) Two-pool microbial model** **(c) Four-pool microbial mod**

Figure 4. The RMSEs of the different optimization algorithms: **(a)** CLM-CASA' model; **(b)** two-pool microbial model and **(c)** four-pool microbial model. The box plots show the means and the quartiles spreading over a total of 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 and the red crosses represent the outliers. The number of simulations of the five algorithms is 100 and the number of simulations of the Bayesian MCMC is presented in Table 3.

global optimization algorithms, indicating that our SBO is robust.

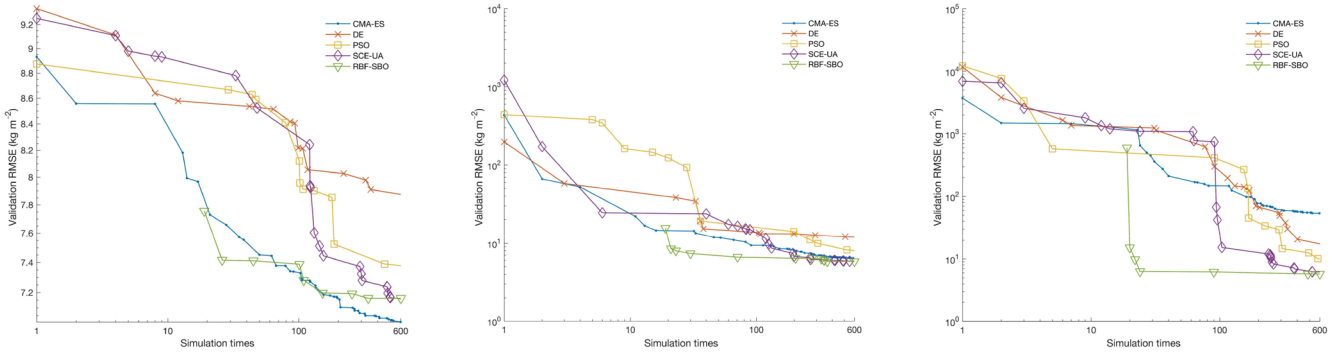
5 Analysis of parameter calibration results

5.1 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 results, using the calibrated parameter values from the SBO, match the observation better than the simulation results with the default parameter values (Fig. 7c) and display a relatively lower RMSE. By using the calibrated parameter values the SOC simulations are significantly improved in most parts of the world, except for 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 the simulation 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's accuracy, we need to gain a more in-depth understanding of uncertainty sources from the data, model structure, parameters and forcing.

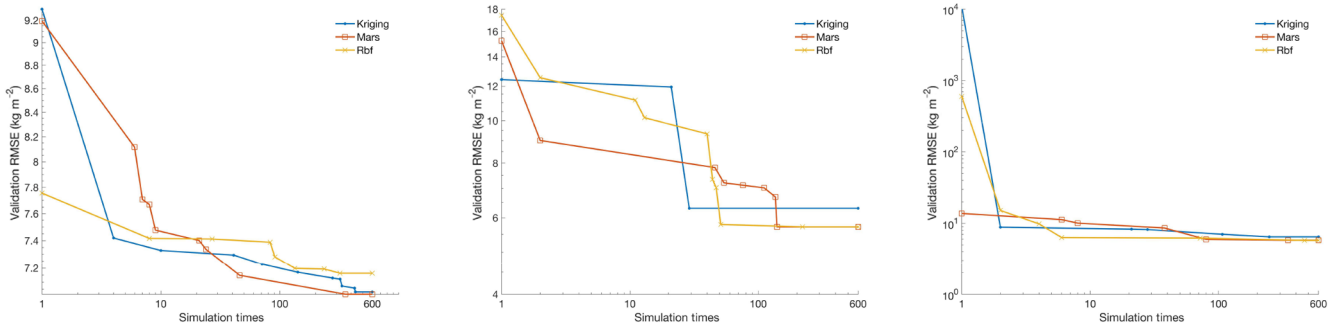
Figure 8 presents the frequency distributions of the 20 calibrated parameters based on MCMC and the calibrated parameter values using the proposed SBO (the blue lines in Fig. 8). Narrow posterior distributions indicate highly sensitive parameters, consistent with the conclusions of Hararuk et al. (2014). The calibrated parameter values of SBO are quite similar to the responding parameter values at the peaks of posterior distributions for most highly sensitive parameters, such as the temperature sensitivity of heterotrophic respiration (Q_{10}) and the clay effect on C partitioning from slow pools to passive pools (t_7). The parameter calibration results (RMSE) of the SBO and the Bayesian MCMC are similar, which means that both are consistent 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



(a) The RMSEs for CLM-CASA' model (b) The RMSEs for two-pool microbial model (c) The RMSEs for four-pool microbial model

Figure 5. The average RMSEs with the increase of number of simulations and different optimization algorithms: **(a)** the CLM-CASA' model, **(b)** the two-pool microbial model and **(c)** the four-pool microbial model. Since RBF-SBO requires some initial simulations to start the optimization process, RBF-SBO starts when the x axis value is 19.



(a) The RMSEs for CLM-CASA' model (b) The RMSEs for two-pool microbial model (c) The RMSEs for four-pool microbial model

Figure 6. The average RMSEs with the increase of number of simulations and different surrogate models.

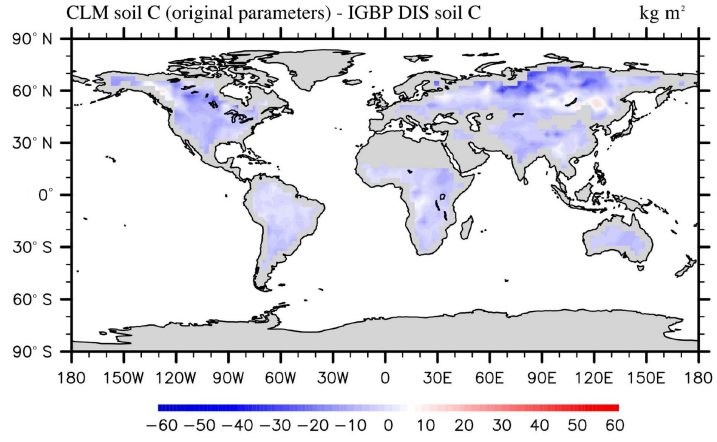
explain this issue. In addition, the unreasonable setting of those bounds might be another possible reason. For instance, the calibrated c(12,12) value (1.01×10^{-3}) reaches its lower bound, indicating that passive SOC residence time almost approaches 1000 years.

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 increases due to the longer residence time of the passive pool and lower temperature sensitivity (Q_{10}). The size of the slow pool, on the contrary, decreases due to the increase in the decomposition rate from the slow pool or the decrease of its residence time. Comprehensively, the size of the SOC, which is the sum of carbon capacity in passive pools, slow pools and soil microbial pools, increases and closely approximates the observation.

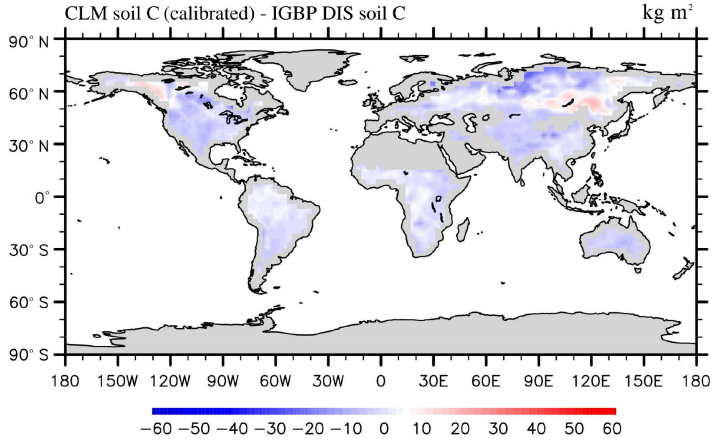
5.2 The microbial models

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

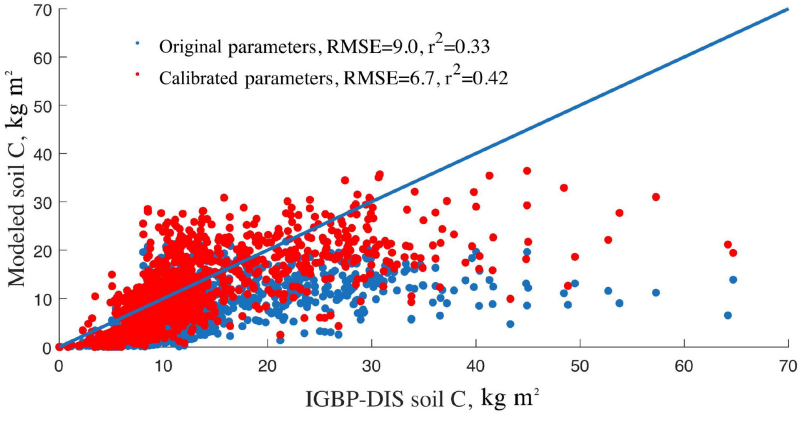
The microbial models achieve better SOC predictions than the calibrated CLM-CASA' model in terms of their prediction of the C capacity in the low-temperature regions (Russia, Europe, North America) and in the regions with small soil C inputs (Figs. 7b and 9). The SOC contents are determined by two main factors: the soil carbon inputs and the SOC residence time (Luo et al., 2003). Considering the same soil carbon inputs in the CLM-CASA' and the microbial models, the improvement is mostly induced by the differences in SOC residence time. In all 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 the three models. The temperature sensitivity remains constant in the CLM-CASA'. However, both



(a) CLM soil C – IGBP-DIS soil C



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



(c) Spatial correspondence between modeled soil and IGBP-DIS soil

Figure 7. Spatial correspondence of SOC produced by CLM-CASA' to SOC reported by IGBP-DIS. Map **(a)** shows the results using the default parameter values, and **(b)** shows the results after parameter calibration using the surrogate-based optimization. The points in **(c)** represent the grid cell values (blue represents the results with default parameter values and red represents the results after parameter calibration). CLM-CASA' with the default parameter values explains 33 % of the variation in the observed soil C, while CLM-CASA' with the calibrated parameter values explains 42 % of the variation in the observed soil C.

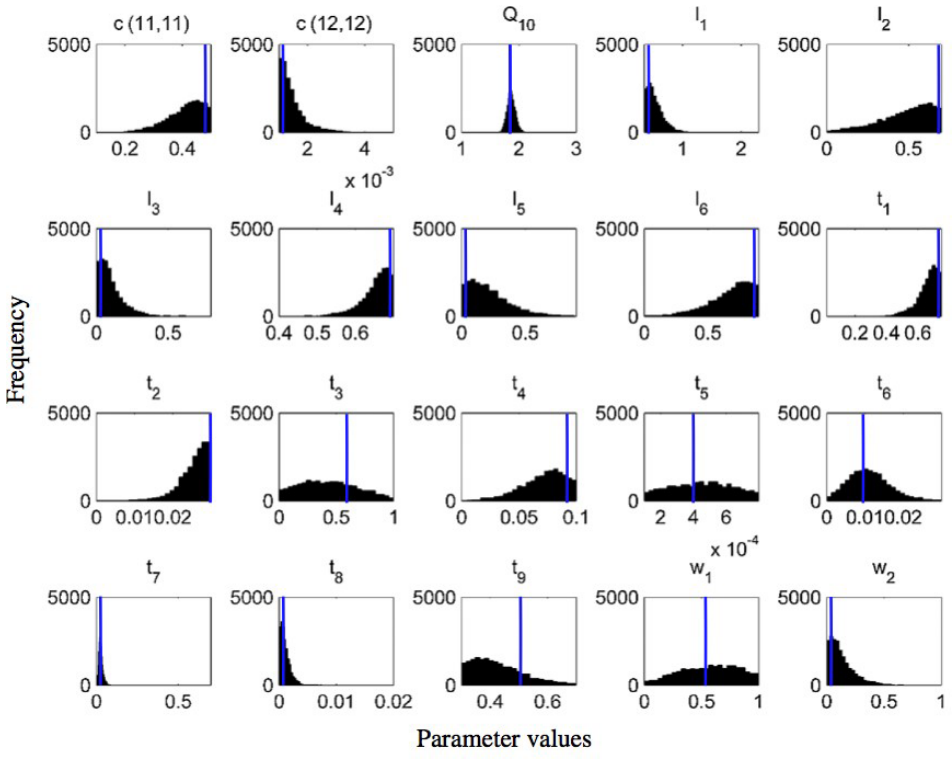


Figure 8. Frequency distributions of 20 calibrated parameters of the CLM-CASA' model using the Bayesian MCMC approach (Harauk et al., 2014) and surrogate-based optimization (blue line in each graph).

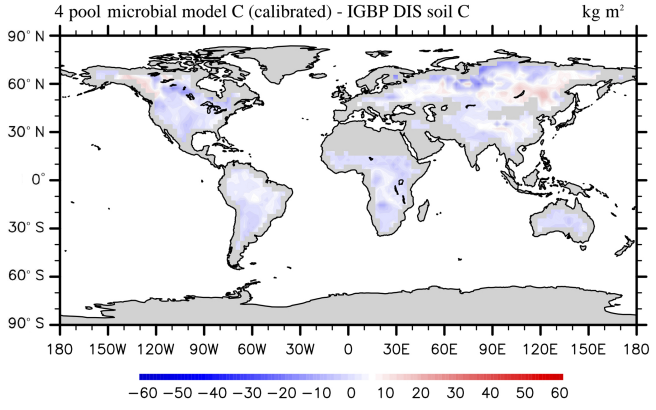


Figure 9. Spatial correspondence of the four-pool microbial model produced SOC to the IGBP-DIS reported SOC.

of the microbial models calculate spatially variable Q_{10} with higher values in the low-temperature regions and lower values in the high-temperature regions, which reflects the impact of temperature on 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 microbial growth dynamics, resulting in an increase in the 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 regions with high SOC input and a high SOC residence time and the regions with low SOC input. This is due to the nonlinearity of the substrate limitation in microbial models (Eqs. 8 and 10), as well as the dependency of the residence time on microbial dynamics. Comprehensively, the introduction of microbial dynamics allows the microbial models to better predict SOC than the CLM-CASA' model.

Figure 10 presents the posterior distributions of the parameters calculated by Bayesian MCMC and the parameter values calibrated using our SBO. 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 SBO are consistent 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 SBO is 1.4 and 0.8 kg m^{-2} lower than those found with Bayesian MCMC for four-pool and two-pool microbial models respectively (as listed in Table 3). The mismatch of CUE_{slope} and CUE_0 may mainly be due to the different targets of the parameter selection between the two methods.

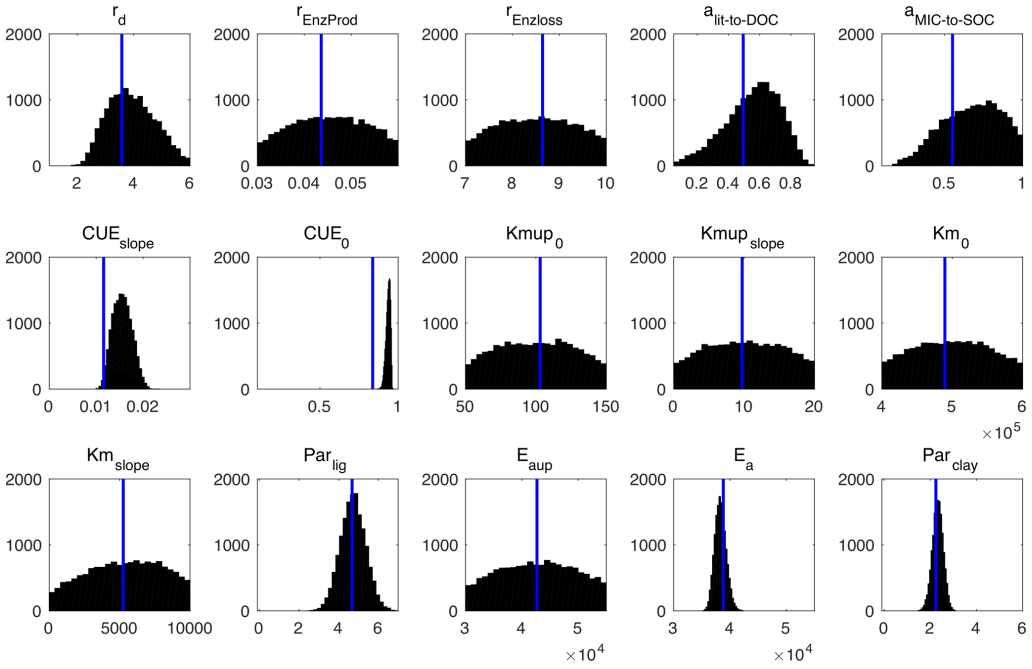


Figure 10. Posterior probability density functions of the four-pool microbial model parameters (generated by Bayesian MCMC). The blue vertical lines are the final calibrated parameter values from our surrogate-based optimization.

6 Conclusions

Parameter calibration is becoming more and more challenging for SOC model development, especially for the computationally expensive global land models, owing to the large number of simulations. In this study, we introduce an SBO algorithm to the parameter calibration of three SOC models. The main findings are as follows:

1. Compared to advanced global optimization algorithms, SBO is more effective and more efficient on average. Our RBF-SBO outperforms other parameter calibration algorithms when the number of simulations does not exceed 200.

2. The parameter optimization based on the RBF surrogate model gains more accurate calibration results than those of the Bayesian MCMC approach in the three soil carbon models.

3. The SBO scheme is robust. Various types of surrogate models have similar performance in the parameter calibration tasks of SOC models.

4. Although SBO is only guided by a single cost function, it can still result in better parameter values than the default algorithms. We carefully analyze the spatial SOC distributions produced by the models with the calibrated parameters using SBO, which indicate that SBO truly improves the model's prediction and simulation capability.

Although the three SOC models used in this analysis are not computationally unattainable for parameter calibration, what we have learned about SBO from this study can potentially be applied to more complex models. Currently, more and more complex simulation models present challenges to the SBO algorithm. To improve the accuracy of SBO, better surrogate models are expected. Current surrogate models including our implementation for soil carbon models mainly employ only one surrogate model, which may limit successful use for different kinds of models. In the future we will focus on the application of multiple surrogate models using ensemble learning.

Code and data availability. The code and data from the three models and the related algorithm implementations can be found in the Supplement. If you have any problem using the code and repeating the experiments, please feel free to contact the corresponding author of this paper: Wei Xue (xuewei@tsinghua.edu.cn).

Supplement. The supplement related to this article is available online at: https://doi.org/10.5194/gmd-11-3027-2018-supplement.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We would like to thank all the reviewers for their insightful comments and suggestions. This work is partially supported by the National Key R&D Program of China (grant no. 2016YFA0602100 and 2017YFA0604500) and National Natural Science Foundation of China (grant no. 91530323 and 41776010).

Edited by: Christoph Müller
Reviewed by: Maarten Braakhekke and one anonymous referee

References

Aleman, D. M., Romeijn, H. E., and Dempsey, J. F.: A response surface approach to beam orientation optimization in intensity-modulated radiation therapy treatment planning, INFORMS J. Comput., 21, 62–76, 2009.

Allison, S. D., Wallenstein, M. D., and Bradford, M. A.: Soil-carbon response to warming dependent on microbial physiology, Nat. Geosci., 3, 336–340, 2010.

Behzad, M., Asghari, K., Eazi, M., and Palhang, M.: Generalization performance of support vector machines and neural networks in runoff modeling, Expert Syst. Appl., 36, 7624–7629, 2009.

Booker, A. J., Dennis Jr., J. E., Frank, P. D., Serafini, D. B., Torczon, V., and Trosset, M. W.: A rigorous framework for optimization of expensive functions by surrogates, Struct. optimization, 17, 1–13, 1999.

Breiman, L.: Statistical modeling: The two cultures (with comments and a rejoinder by the author), Stat. Sci., 16, 199–231, 2001.

Davis, E. and Ierapetritou, M.: A kriging based method for the solution of mixed-integer nonlinear programs containing black-box functions, J. Global Optim., 43, 191–205, 2009.

Fontaine, S., Bardoux, G., Abbadie, L., and Mariotti, A.: Carbon input to soil may decrease soil carbon content, Ecol. Lett., 7, 314–320, 2004.

Fontaine, S., Barot, S., Barré, P., Bdioui, N., Mary, B., and Rumpel, C.: Stability of organic carbon in deep soil layers controlled by fresh carbon supply, Nature, 450, 277–280, 2007.

Forrester A. I. J., Sóbester, A., and Keane, A. J.: Engineering Design via Surrogate Modelling: A Practical Guide, John Wiley & Sons, Chichester, 2008.

Friedlingstein, P., Cox, P., Betts, R., Bopp, L., von Bloh, W., Brovkin, V., and Bala, G.: Climate-carbon cycle feedback analysis: Results from the C4MIP model intercomparison, Climate, 19, 3337–3353, 2006.

Friedman, J. H.: Multivariate adaptive regression splines, Ann. Stat., 19, 1–67, 1991.

German, D. P., Marcelo, K. R. B., Stone, M. M., and Allison, S. D.: The Michaelis–Menten kinetics of soil extracellular enzymes in response to temperature: a cross-latitudinal study, Glob. Change Biol., 18, 1468–1479, 2012.

Giunta, A. A.: Aircraft multidisciplinary design optimization using design of experiments theory and response surface modeling methods, 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, Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, TN, 2000.

Gutmann, H. M.: A radial basis function method for global optimization, J. Global Optim., 19, 201–227, 2001.

Haario, H., Saksman, E., and Tamminen, J.: An adaptive Metropolis algorithm, Bernoulli, 7, 223–242, 2001.

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

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

Hansen, N. and Ostermeier, A.: Completely de-randomized self-adaptation in evolution strategies, Evol. Comput., 9, 159–195, 2001.

Hansen, N., Auger, A., Ros, R., Finck, S., and Pošík, P. Comparing results of 31 algorithms from the black-box optimization benchmarking BBOB-2009, Proceedings of the 12th annual conference companion on Genetic and evolutionary computation, ACM, 1689–1696, 2010.

Hapuarachchi, H. A. P., Li, Z., and Wang, S.: Application of SCE-UA method for calibrating the Xinanjiang watershed model, Journal of Lake Sciences, 13, 304–314, 2001.

Hararuk, O., Xia, J., and Luo, Y.: Evaluation and improvement of a global land model against soil carbon data using a Bayesian Markov chain Monte Carlo method, J. Geophys. Res.-Biogeo., 119, 403–417, 2014.

Hararuk, O., Smith, M. J., and Luo, Y.: Microbial models with data-driven parameters predict stronger soil carbon responses to climate change, Glob. Change Biol., 21, 2439–2453, 2015.

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

Huang, Y. Y., Lu, X. J., Shi, Z., Lawrence, D., Koven, C., Xia, J. Y., Du, Z. G., Kluzek, E., and Luo, Y. Q.: Matrix approach to land carbon cycle modeling: A case study with Community Land Model, Glob. Change Biol., 24, 1394–1404, 2018.

Iman, R. L., Campbell, J., and Helton, J.: An approach to sensitivity analysis of computer models, J. Qual. Technol., 13, 174–183, 1981.

Jones, D. R.: A taxonomy of global optimization methods based on response surfaces, J. Global Optim., 21, 345–383, 2001.

Jones, D. R., Schonlau, M., and Welch, W. J.: Efficient global optimization of expensive black-box functions, J. Global Optim., 13 455–492, 1998.

Kennedy, J.: Particle swarm optimization, Encyclopedia of Machine Learning, Springer US, 760–766, 2011.

Kuzyakov, Y., Friedel, J. K., and Stahr, K.: Review of mechanisms and quantification of priming effects, Soil Biol. Biochem., 32, 1485–1498, 2000.

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

Luo, Y.: Terrestrial carbon-cycle feedback to climate warming, Annu. Rev. Ecol. Evol. S., 38, 683–712, 2007.

Luo, Y. and Weng, E.: Dynamic disequilibrium of the terrestrial carbon cycle under global change, Trends Ecol. Evol., 26, 96–104, 2011.

Luo, Y., Wu, L., Andrews, J. A., White, L., Matamala, R., Schäfer, K. V. R., and Schlesinger, W. H.: Elevated CO_2 differentiates ecosystem carbon processes: Deconvolution analysis of Duke Forest FACE data, Ecol. Monogr., 71, 357–376, 2001.

Luo, Y., White, L. W., Canadell, J. G., DeLucia, E. H., Ellsworth, D. S., Finzi, A., Lichter, J., and Schlesinger, W. H.: Sustainability of terrestrial carbon sequestration: A case study in Duke Forest with inversion approach, Global Biogeochem. Cy., 17, 1021, https://doi.org/10.1029/2002gb001923 , 2003.

Luo, Y., Ahlström, A., Allison, S. D., Batjes, N. H., Brovkin, V., Carvalhais, N., Chappell, A., Ciais, P., Davidson, E. A., Finzi, A., Georgiou, K., Guenet, B., Hararuk, O., Harden, J. W., He, Y. J., Hopkins, F., Jiang, L. F., Koven, C., Jackson, R. B., Jones, C. D., Lara, M. J., Liang, J. Y., McGuire, A. D., Parton, W., Peng, C. H., Randerson, J. T., Salazar, A., Sierra, C. A., Smith, M. J., Tian, H. Q., Todd-Brown, K. E. O., Torn, M., van Groenigen, K. J., Wang, Y. P., West, T. O., Wei, Y. X., Wieder, W. R., Xia, J. Y., Xu, X., Xu, X. F., and Zhou, T.: Towards More Realistic Projections of Soil Carbon Dynamics by Earth System Models, Global Biogeochem. Cy., 30, 40–56, 2016.

Luo, Y. Q., Keenan, T. F., and Smith, M.: Predictability of the terrestrial carbon cycle, Glob. Change Biol., 21, 1737–1751, 2015.

Ma, H., Dong, Z., Zhang, W. M., and Liang, Z. M.: Application of SCE-UA algorithm to optimization of TOPMODEL parameters, Journal of Hohai University 4, 2006.

Marshall, L., Nott, D., and Sharma, A.: A comparative study of Markov chain Monte Carlo methods for conceptual rainfall-runoff modeling, Water Resour. Res., 40, W02501, https://doi.org/10.1029/2003WR002378, 2004.

McKay, M. D., Beckman, R. J., and Conover, W. J.: Comparison of three methods for selecting values of input variables in the analysis of output from a computer code, Technometrics, 21, 239–245, 1979.

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

Müller, J.: MATSuMoTo: The MATLAB Surrogate Model Toolbox For Computationally Expensive Black-Box Global Optimization Problems, arXiv:1404.4261, 2014.

Müller, J. and 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. Global Optim., 60, 123–144, 2014.

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

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

Oleson, K. W., Niu, G., Yang, Z., Lawrence, D., Thornton, P., Lawrence, P., Stockli, R., Dickinson, R., Bonan, G., and Levis, S.: Improvements to the Community Land Model and their impact on the hydrological cycle, J. Geophys. Res., 113, https://doi.org/10.1029/2007jg000563, 2008.

Peng, S., Piao, S., Wang, T., Sun, J., and Shen, Z.: Temperature sensitivity of soil respiration in different ecosystems in China, Soil Biol. Biochem., 41, 1008–1014, 2009.

Picheny, V., Ginsbourger, D., Richet, Y., and Caplin, G.: Quantile-Based Optimization of Noisy Computer Experiments With Tunable Precision, Technometrics, 55, 2–9, 2012.

Powell, M. J. D.: The theory of radial basis function approximation in 1990, in: Advances in Numerical Analysis II: Wavelets, Subdivision Algorithms and Radial Functions, edited by: Light, W. A., Oxford University Press (Oxford), 105–210, 1992.

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

Regis, R. G.: Stochastic radial basis function algorithms for large-scale optimization involving expensive black-box objective and constraint functions, Comput. Oper. Res., 38, 837–853, 2011.

Regis, R. G. and Shoemaker, C. A.: A stochastic radial basis function method for the global optimization of expensive functions, INFORMS J. Comput., 19, 497–509, 2007.

Regis, R. G. and Shoemaker, C. A.: Parallel stochastic global optimization using radial basis functions, INFORMS J. Comput., 21, 411–426, 2009.

Rocha, H.: Model parameter tuning by cross validation and global optimization: application to the wing weight fitting problem, Structural and Multidisciplinary Optimization, 37, 197 pp., 2008.

Schimel, J. P. and Weintraub, M. N.: The implications of exoenzyme activity on microbial carbon and nitrogen limitation in soil: a theoretical model, Soil Biol. Biochem., 35, 549–563, 2003.

Schonlau, M., Welch, W. J., and Jones, D. R.: Global versus local search in constrained optimization of computer models, Lecture Notes-Monograph Series, 11–25, https://doi.org/10.1214/lnms/1215456182, 1998.

Shi, Y. and Eberhart, R. C.: Empirical study of particle swarm optimization, Front. Comput. Sci. Chi., 3, 31–37, 2009.

Shi, Z., Crowell, S., Luo, Y. Q., and Moore III, B.: Uncertainty in soil carbon projection constrained by data but amplified by model structures, Nat. Commun., https://doi.org/10.1038/s41467-018-04526-9, 2018.

Simpson, T. W., Mauery, T. M., Korte, J. J., and Mistree, F.: Kriging models for global approximation in simulation-based multidisciplinary design optimization, AIAA J., 39, 2233–2241, 2001.

Smith, M. J., Purves, D. W., Vanderwel, M. C., Lyutsarev, V., and Emmott, S.: The climate dependence of the terrestrial carbon cycle, including parameter and structural uncertainties, Biogeosciences, 10, 583–606, https://doi.org/10.5194/bg-10-583-2013, 2013.

Storn, R. and Price, K.: Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces, J. Global Optim., 11, 341–359, 1997.

Todd-Brown, K. E. O., Randerson, J. T., Post, W. M., Hoffman, F. M., Tarnocai, C., Schuur, E. A. G., and Allison, S. D.: Causes of variation in soil carbon simulations from CMIP5 Earth system models and comparison with observations, Biogeosciences, 10, 1717–1736, https://doi.org/10.5194/bg-10-1717-2013, 2013.

Vu, K. K., D'Ambrosio, C., Hamadi, Y., and Liberti, L.: Surrogate-based methods for black-box optimization, International Transactions in Operational Research, 24, 393–424, 2016.

Wang, C., Duan, Q., Gong, W., Ye, A., Di, Z., and Miao, C.: An evaluation of adaptive surrogate modelling based optimiza-

tion with two benchmark problems, Environ. Modell. Softw., 60, 167–179, 2014.

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

Wild, S. M. and Shoemaker, C.: Global convergence of radial basis function trust-region algorithms for derivative-free optimization, SIAM Rev., 55, 349–371, 2013.

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

Xia, J. Y., Luo, Y. Q., Wang, Y.-P., Weng, E. S., and Hararuk, O.: A semi-analytical solution to accelerate spin-up of a coupled carbon and nitrogen land model to steady state, Geosci. Model Dev., 5, 1259–1271, https://doi.org/10.5194/gmd-5-1259-2012, 2012.

Xu, T., White, L., Hui, D., and Luo, Y.: Probabilistic inversion of a terrestrial ecosystem model: Analysis of uncertainty in parameter estimation and model prediction, Global Biogeochem. Cy., 20, GB2007, https://doi.org/10.1029/2005gb002468, 2006.

Zhang, X., Srinivasan, R., and Van Liew, M.: Approximating SWAT model using artificial neural network and support vector machine, J. Am. Water Resour. As., 45, 460–474, 2009.

Zhou, T., Shi, P., Hui, D., and Luo, Y.: Global pattern of temperature sensitivity of soil heterotrophic respiration (Q10) and its implications for carbon-climate feedback, J. Geophys. Res., 114, https://doi.org/10.1029/2008jg000850, 2009.