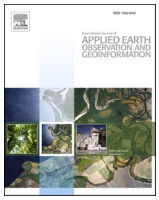
International Journal of Applied Earth Observation and Geoinformation 139 (2025) 104542

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



A novel approach of generating pseudo revisited soil sample data based on environmental similarity for space-time soil organic carbon modelling

Wenkai Cui [a], Lin Yang [a,b,*], Lei Zhang [a], Chenconghai Yang [a], Chenxu Zhu [a], Chenghu Zhou [a,c]

[a] School of Geography and Ocean Science, Nanjing University, Nanjing 210023, China
[b] Frontiers Science Center for Critical Earth Material Cycling, Nanjing University, Nanjing 210023, China
[c] State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

ARTICLE INFO

Keywords:
Soil organic carbon (SOC)
Pseudo revisited samples
Third law of geography (similar principle)
Environmental similarity
Rothamsted carbon model (RothC)

ABSTRACT

Soil Organic Carbon (SOC) is vital for the global carbon cycle, agricultural sustainability, and climate change. Process-based models like Rothamsted carbon model (RothC) simulate SOC dynamics, but their accuracy relies on revisited soil samples for calibration, which are often scarce, especially in environmentally heterogeneous regions. This study proposes a novel approach to generate pseudo revisited samples using environmental similarity, addressing the lack of revisited samples. For each intervening-year sample, pseudo SOC stocks in unsampled years were constructed by calculating environmental similarity with existing samples and applying weighted averaging. These pseudo SOC stocks served as revisited samples for model calibration. Bayesian optimization was used to adjust RothC's microbial activity parameters. Validation results showed the approach significantly improved predictive accuracy, with an RMSE of 5.28 t/ha (31.6 % lower than global parameter optimization and 10.7 % lower than spatiotemporal Random Forest) and an R2 improved from 0.319 (spatiotemporal Random Forest) to 0.456. The approach enhances SOC model accuracy by leveraging environmental similarity and parameter optimization, offering a practical solution for regions lacking revisited samples and improving long-term SOC dynamics simulations. This approach not only addresses data scarcity but also provides more reliable predictions for climate and agricultural management.

1. Introduction

Soil carbon is the largest carbon pool in terrestrial ecosystem, which plays an important role in global carbon cycle, soil health and crop yield (Hamzehpour et al., 2019; Lal, 2004). Soil Organic Carbon (SOC) serves as a critical indicator of soil fertility and plays an essential role in the global carbon cycle and climate change (Hamzehpour et al., 2019). The spatial distribution of SOC is influenced by a combination of factors such as climate, topography, vegetation types, and land use (Grinand et al., 2017; Lehmann et al., 2020). Temporal SOC changes reflect the dynamic responses of soil ecosystems to natural processes (e.g., vegetation succession) and human activities (e.g., tillage and fertilization) (McBratney et al., 2014). When SOC changes, it not only affects the soil's carbon storage capacity but also regulates atmospheric carbon dioxide concentrations. These changes, in turn, exert significant effects on global climate, carbon balance of terrestrial ecosystem and ecosystem services (Hisse et al., 2022; Shen et al., 2023). Therefore, accurately estimating the spatiotemporal change of SOC is not only crucial for understanding soil health and carbon stocks but also holds great significance for predicting future trends in climate change.

Process-based models such as Rothamsted carbon model (RothC) (Coleman and Jenkinson, 1996), CENTURY (Kelly et al., 1997), and Millennial (Abramoff et al., 2018) have been widely used to simulate the temporal change of SOC. These models categorize SOC into different pools with varying stability and decomposition rates. They predict SOC changes by describing the transformation processes between these pools and the inputs and outputs of carbon (Coleman and Jenkinson, 1996). The application of these models has expanded from site to regional scale studies as well as assessing the impacts of climate change (Viscarra Rossel et al., 2019), land use changes (Qingsong et al., 2023; Zhou et al., 2019), and agricultural management practices (Meng et al., 2024) on SOC stocks. Although process-based models have general applicability and broad scope, their effectiveness largely depends on the appropriateness of the parameter settings (Paramesh et al., 2022). Key

* Corresponding author.
E-mail address: yanglin@nju.edu.cn (L. Yang).

https://doi.org/10.1016/j.jag.2025.104542
Received 22 February 2025; Received in revised form 27 March 2025; Accepted 14 April 2025

parameters in the models, such as the decomposition rates of carbon pools, carbon input, and temperature and humidity correction factors, are typically set based on the model's default values. However, due to significant differences in soil types (Jebari et al., 2021), climatic conditions (Farina et al., 2013; Jebari et al., 2021), vegetation cover, and land management practices such as carbon inputs (Nemo et al., 2017) across different regions, using default parameters may not accurately reflect the conditions of the study area. Parameter optimization methods, such as Bayesian optimization (Dechow et al., 2019), can be employed to improve the calibration. During the parameter optimization process, most studies optimize parameters and generate a single set of parameters across the entire study area, which may overlook the environmental heterogeneity of the whole area.

Using process-based models typically requires revisited soil samples (samples at the same location at different times) within a study area to calibrate the model parameters (Ugbaje et al., 2024). In practical research, due to high sampling costs, lack of historical data, or differing sampling objectives, the number of revisited soil samples is often insufficient or absent, making it impossible to directly calibrate the parameters of process-based models (Smith et al., 2020). Yet, many study areas have non-revisited soil samples at multiple time points. In case of non-revisited samples, researchers have attempted to use spatiotemporal Digital Soil Mapping (DSM) methods such as Random Forest (RF) and other machine learning (ML) methods, which establish soil-environment relationships based on samples at multiple time points and generate soil maps by inputting environmental covariates at the target years. These methods have demonstrated high predictive capabilities at the spatial dimension, compensating for the limitations posed by non-revisited samples (Radočaj, 2024; Zhou et al., 2020). However, the modeling of spatiotemporal RF doesn't consider the temporal change of SOC in mechanism, so it is unable to perfectly capture the change trend of SOC at time dimension (Zhang et al., 2024a).

Although revisited soil samples are scarce, samples collected at different years/periods often exist in regions. These non-revisited soil samples, while spatially distinct, may demonstrate environmental similarities within the environmental factor space (Zhu et al., 2018). By calculating the environmental similarity between samples from different years, one can identify environmentally similar samples at both the starting and ending years for any given sample at the intervening years, and pseudo revisited soil samples can be constructed. Using the SOC stock (SOCS) values of these pseudo revisited soil samples (hereinafter referred to as pseudo SOCS), one can calibrate model parameters to simulate changes over time, providing a potential solution for the common situation of non-revisited samples. Based on this idea, this study proposes a novel approach that matches environmentally similar soil samples across multiple years using environmental similarity, constructs pseudo revisited soil samples for model calibration. Our case study uses the widely applicable RothC model (Liu et al., 2011) as an example to perform calibration in a study area in Jiangsu Province, China. Independent samples from intervening years are used for validation, and results are compared with global parameter optimization approach and spatiotemporal RF in terms of prediction accuracy and SOC change over years at validation samples. The study aims to provide a reliable soil carbon spatiotemporal change modelling approach in the absence of revisited soil samples.

2. Study area and dataset

2.1. Study area

The study area is located in Jiangsu Province, China, encompassing the Huaihe River Basin in the central Jiangsu region, the Yangtze River Basin and the Taihu Lake Basin in the southern Jiangsu region (Shen et al., 2023), covering an area of approximately 75,000 square kilometers. The area is situated in the eastern coastal region of China, with a predominantly flat plain topography. It falls within the transitional zone between the subtropical monsoon climate and the warm temperate monsoon climate, characterized by distinct seasons, cold winters and hot summers, with rain and heat occurring in the same season. The annual average temperature ranges from 13 °C to 17 °C, with annual precipitation around 1000 mm. Major cities in the region include Nanjing, Yangzhou, Suzhou, and Wuxi. The region has fertile soil, abundant water resources, and a mild climate, with rice as the main crop. Machine transplanting occurs in mid-May, and harvest takes place in early November (https://nynct.jiangsu.gov.cn/art/2019/5/30/art_13466_8349710.html). The location of the study area with samples is shown in Fig. 1.

2.2. Dataset

2.2.1. Soil samples

The soil dataset covers a total of 186 non-revisited surface soil (0–20 cm) samples collected 48 samples in 1980, 72 samples in 2010, and 66 samples in 2020. During sampling, the SOC concentration (g/kg) at each sample was originally measured at different depths. The SOC concentration and soil bulk density data for the 1980 samples were sourced from *Soil Species of China*, based on the Second National Soil Survey of China (Office for the Second National Soil Survey of China, 1993). The SOC concentration and soil bulk density data for the 2010 samples were from *Soil Series of China* (Huang and Pan, 2017). The SOC concentration data for the 2020 samples was derived from field sampling using the Latin Hypercube Sampling method (Minasny and McBratney, 2006; Wang et al., 2024) and then subsequent processed, while the soil bulk density data came from the Chinese Soil Database Used for Land Surface Modeling (Shi et al., 2024). The SOCS (t/ha) was then calculated using Equation (1) (Liu et al., 2011):

$$SOCS = \sum_{i=1}^{n} (soc_i \cdot BD_i \cdot D_i \cdot 0.1) \quad (1)$$

where soc_i is the SOC concentration (g/kg) in depth interval i; BD_i is the soil bulk density (g/cm^3) in depth interval i; D_i is the corresponding depth interval (cm).

2.2.2. Environmental covariates

Due to that climate, topography, and agricultural factors influence SOC in this area, this study selected six environmental covariates for environmental similarity calculation and SOC mapping, including three static covariates: Digital Elevation Model (DEM), slope, and soil sand content, and three dynamic covariates: annual average temperature, summer Normalized Difference Vegetation Index (NDVI), and nitrogen fertilizer. DEM data was obtained from the Shuttle Radar Topography Mission (SRTM) with a 90 m resolution. Slope was calculated from DEM using ArcGIS. Soil sand content data was derived from the Chinese Soil Characteristic Dataset with a 1 km resolution (Shangguan et al., 2012). Annual average temperature data was sourced from the National Tibetan Plateau / Third Pole Environment Data Center with a 1 km resolution (Ding and Peng, 2020; Peng, 2019; Peng et al., 2019, 2018, 2017). NDVI data was provided by Li et al. (2024) with a 5 km resolution. Furthermore, as most of the study area is agricultural land, the impact of agricultural activities on SOCS needed to be considered (Zhao et al., 2018). Therefore, nitrogen fertilizer data (Yu et al., 2022) was included as one of the covariates to reflect the impact of human activities like agricultural practices on SOC (Padarian et al., 2022). Since the original nitrogen fertilizer data is available until 2018, data for 2019 and 2020 for the study area was represented by the data in 2018.

When calculating environmental similarity, the dynamic covariate data were from the corresponding years. For the RothC model, in addition to temperature, precipitation and evaporation data were also sourced from the National Tibetan Plateau / Third Pole Environment Data Center (Ding and Peng, 2021, 2020; Peng, 2022, 2020; Peng et al., 2019, 2018, 2017). NDVI data was used to determine whether the soil

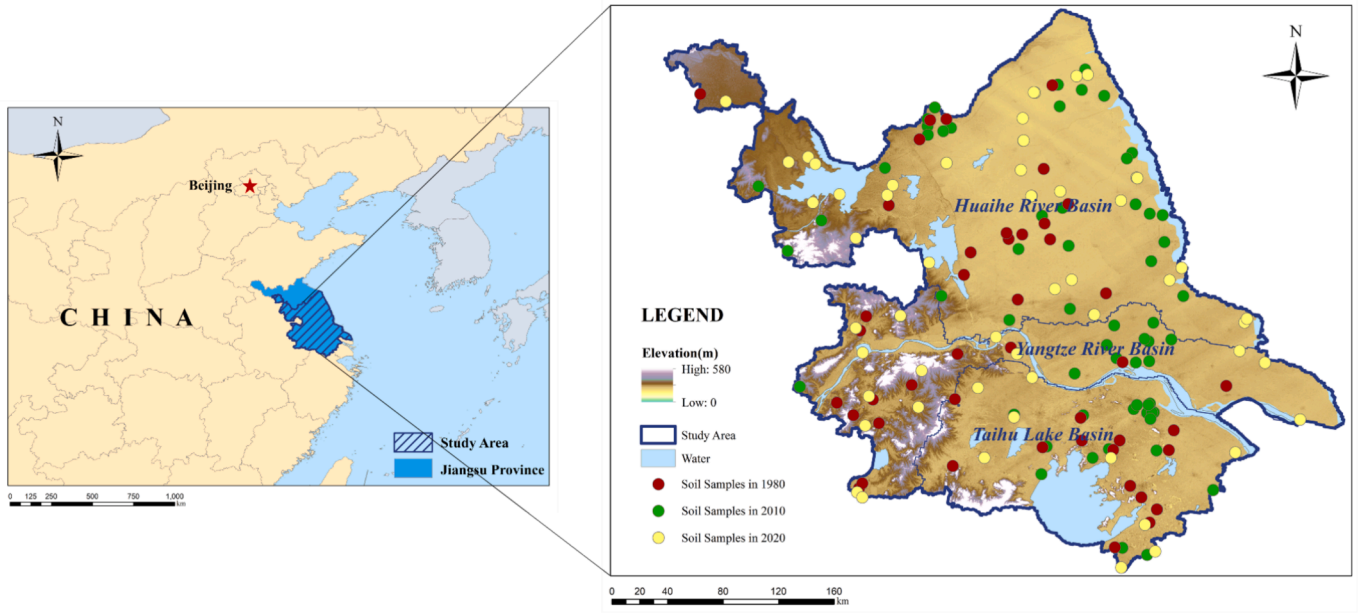


Fig. 1. Study area and distribution of samples.

was vegetated or bare. Soil clay content was obtained from the Chinese Soil Characteristic Dataset. Land use type (Zhang et al., 2024b) was used to determine the monthly carbon input pattern according to Smith et al. (2007). The temperature, precipitation, evaporation and NDVI data used in RothC are all monthly data to ensure the proper functioning of the model. The above covariate data without 1 km as resolution were resampled to a resolution of 1 km using ArcGIS version 10.2.

3. Methodology

3.1. The overview of the approach

According to the third law of geography, also known as the geographic similarity principle or similarity principle, soil properties (here, SOCS) tend to be more similar in more similar environments (Zhu and Turner, 2022; Zhu et al., 2020). The basic idea of the proposed approach to generate the pseudo revisited soil samples in a study area is to select environmentally similar samples in years to be calculated for each intervening-year sample and then calculate pseudo samples based on those environmentally similar samples in corresponding years. Taking the study area with three sampling years as an example—namely, the starting year 1980, intervening year 2010, and ending year 2020. For each intervening-year sample's location, its environmentally similar samples in the starting and ending years could be selected by environmental similarity calculation, and used to generate a pseudo SOCS value (taken as pseudo revisited samples) in the starting and ending year, respectively. Before the environmental similarity calculation, the samples that are not applicable to the similarity principle in each year should be filtered firstly. This means that samples with similar environmental conditions but different SOCS deemed to be of low reliability and should be excluded from the calculation. Then for the remaining samples, we calculate the environmental similarity between the intervening-year samples' locations and all the samples from the starting and ending years, based on both static and dynamic environmental covariates. For each intervening-year sample's location, by setting an environmental similarity threshold, we select the environmentally similar sample pool from the two years respectively. Then according to the SOCS of each sample in the sample pool and their environmental similarity with the intervening-year samples, the pseudo SOCS of the starting and ending years are calculated by a weighted average method, which in turn allows us to derive the pseudo revisited soil samples. The RothC model is calibrated for those samples' locations with the pseudo revisited soil samples in starting and ending years by parameter optimization, then the intervening year SOCS value can be predicted with calibrated model. The experimental workflow is shown in Fig. 2.

3.2. The specific approach

Based on the above basic idea, the proposed approach for calculating pseudo samples based on environmental similarity consists of the following four steps:

(1) Filter samples in each year separately which are not applicable to the similarity principle based on overall reliability assessment;
(2) Calculate the environmental similarity between spatial locations of filtered intervening-year samples and filtered samples in starting and ending year respectively;
(3) Obtain the sets of environmentally similar samples for each intervening-year sample's location by setting an environmental similarity threshold and calculate the pseudo SOCS for the locations at both the starting and ending years;
(4) Use the pseudo SOCS to calibrate the RothC model.

3.2.1. Filtering samples by environmental similarity and overall reliability

The environmental similarity between any two samples can be calculated based on their environmental covariates, using the method outlined in individual predictive soil mapping (iPSM) (Zhu et al., 2015). For each simple, the similarity with other samples is calculated using the method in Equation (2):

$$ES_{i,j} = \sum_{k=1}^{m} \exp[-(V_{i,k} \cdot V_{j,k})^2 \cdot \frac{SD_{k,j}^2}{2SD_k^4}] \quad (2)$$

where $ES_{i,j}$ is the similarity between the sample i and j; $V_{i,k}$ is the k-th environmental covariate value at the i-th sample; $V_{j,k}$ is the k-th environmental covariate value at the j-th sample (where $k = 1, 2, \ldots, m$); SD_k is the standard deviation of the k-th environmental covariate across all grids in the study area; $SD_{k,j}$ is the root mean square deviation for all grids in the study area in relation to the j-th sample for the k-th environmental covariate. Unlike Zhu et al. (2015), this study uses the

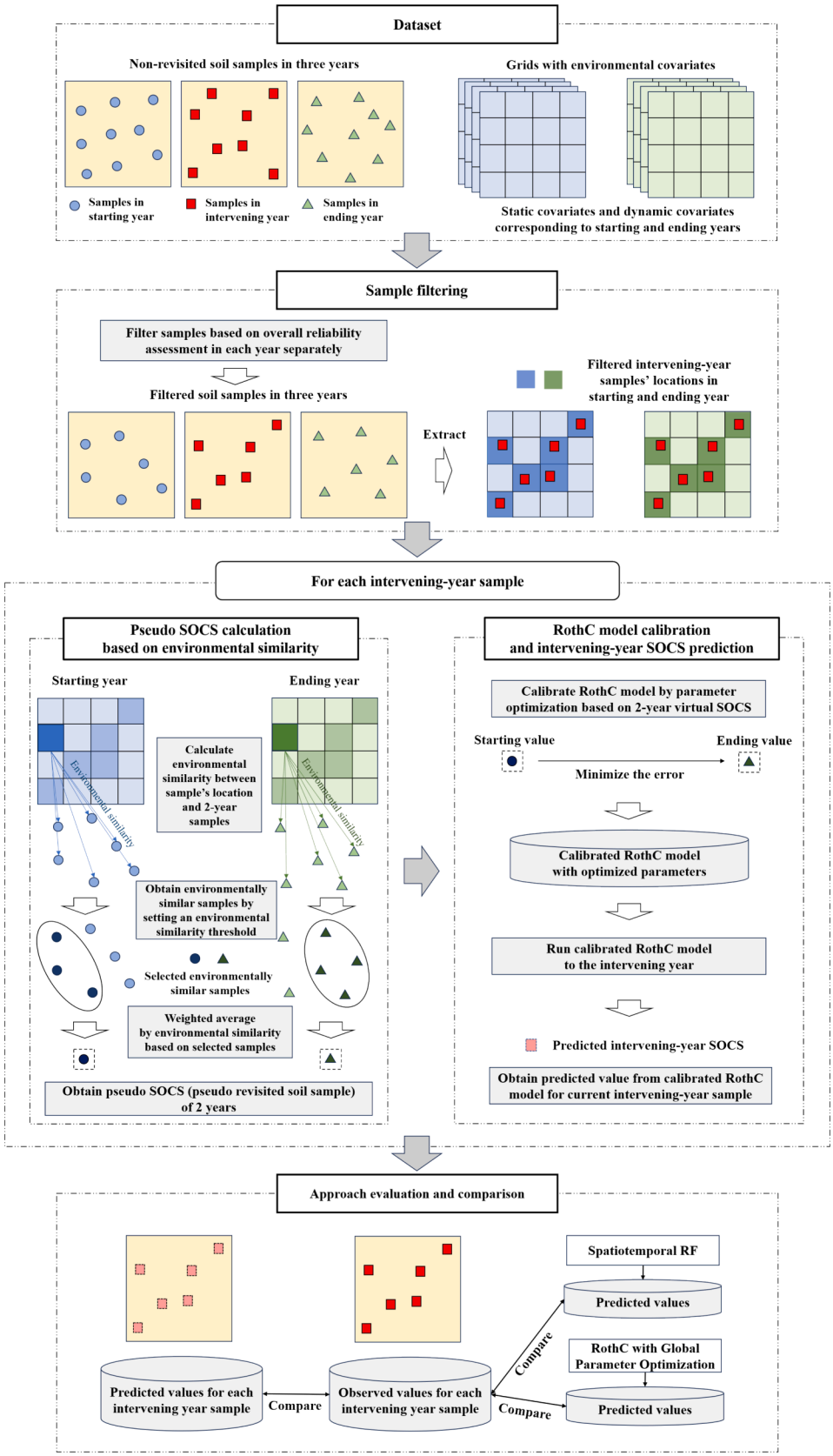


Fig. 2. The flow chart of this study.

average method rather than the minimum operator when calculating $ES_{i,j}$.

According to the similarity principle, the higher the environmental similarity is, the more similar the SOCS should be. However, there are cases where some samples may have similar environments but significantly different SOCS. This discrepancy may be influenced by human activities and soil management practices (e.g., tillage, fertilization, grazing), biological activities (e.g., microorganisms, plant roots, soil fauna), and historical soil usage (e.g., historical soil erosion, salinization, or other natural disasters), which could reduce the reliability of these samples. Therefore, such samples should be filtered out. To filter out samples that may exhibit high uncertainty, it is necessary to calculate the overall reliability of each sample in steps. First, using a fuzzy similarity measure, we assume that the similarity between soil attributes of the samples follows a bell-shaped distribution (Burrough et al., 1992; *Environmental Soil-Landscape Modeling*, n.d.). This can be calculated using the following Equation (3):

$$SS_i = \exp[-\frac{(S_i - \mu_S)^2}{2\sigma_S^2}] \tag{3}$$

where the value of SS_i is the similarity of sample i for different values of the selected soil attribute; S_i is the current soil attribute value of sample i; μ_S is the mean value of all samples for the selected soil attribute; and σ_S^2 is the variance of the values of all samples for the selected soil attribute. In other words, for the sample i in the sample set, the similarity relationship between its soil attribute value and others follows a bell-shaped curve function $f_i = SS_i$, where each sample has its own bell-shaped curve. For two samples, the closer the bell-shaped curves are to each other and the larger the intersection area, the stronger the consistency between the soil attribute values of the two samples. According to the similarity principle, this implies that the environments corresponding to these two samples should be more similar. The method for calculating the consistency between samples is called Consistency Measure (CM) (Zhu, 1999), as shown in Fig. 3 and Equation (4):

$$SCM_{i,j} = \frac{2(A_{i\cap j})}{(A_i + A_j)} \tag{4}$$

where $SCM_{i,j}$ is the CM value between sample i and sample j; A_i is the area under the soil attribute similarity distribution curve SS_i for sample i; A_j is the area under the soil attribute similarity distribution curve SS_j for sample j; $A_{i\cap j}$ is the area of intersection between the two curves. After calculating the CM values between samples, it is necessary to evaluate the environmental conditions of the samples. Specifically, if a sample has a high overall reliability, it is considered applicable to the similarity principle (Liu et al., 2020); otherwise, it is classified as an outlier.

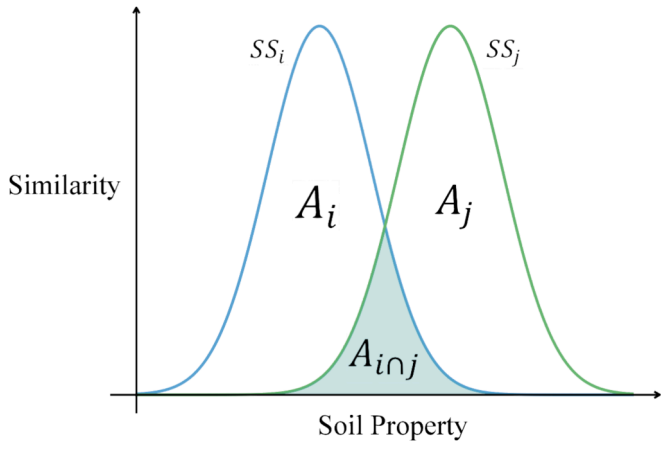


Fig. 3. Soil attribute similarity distribution curve of samples (Liu et al., 2020).

Overall reliability refers to the average CM value of a sample with other environmentally similar samples within the same time period. By setting an environmental similarity threshold, samples with the environmental similarity ($ES_{i,j}$) exceeding the threshold are considered as environmentally similar to the current sample. The average of these samples' CM values is then calculated to determine the overall reliability of the current sample (Equation (5)):

$$OR_i = \begin{cases} \frac{\sum_{j=1}^{n} SCM_{i,j}}{n}, n > 0 \\ NA, n = 0 \end{cases} \tag{5}$$

where OR_i is the overall reliability of sample i; n is the number of other samples for sample i with $SCM_{i,j}$ exceeding the environmental similarity threshold (the setting of the environmental similarity threshold is detailed in Section 3.2.4). Samples with a low OR_i are considered outliers and not used in this study.

After calculation, under the condition of considering both the number of samples and overall reliability, samples with OR_i below 0.5 are filtered out. As a result, 45 samples remain for 1980, 60 samples remain for 2010, and 60 samples remain for 2020. In total, 165 samples are selected from the original 186. The spatial distribution of soil samples across the three years is shown in Fig. 1.

3.2.2. Calculation of environmental similarity between the intervening-year samples' locations and samples in starting and ending year after filtering

To obtain pseudo revisited samples, the environmental covariates in starting and ending years are extracted to the locations of filtered intervening-year simples firstly. Then the environmental similarity between locations and filtered starting-and-ending-year simples is calculated respectively. The calculation method is the same as Equation (2).

3.2.3. Obtaining the sets of environmentally similar samples for each intervening-year sample's location and calculating the pseudo SOCS for the locations

In the same year, for each intervening-year sample's location, the similarity to each sample is calculated, a higher value indicates that the location is more similar to the current sample's environment, while a lower value indicates less similarity. Samples with low similarity are considered to interfere with the calculation for the current location. Therefore, by setting an environmental similarity threshold (detailed in Section 3.2.5), only samples with a similarity exceeding the threshold are included in the SOCS calculation for the current location. In other words, for an intervening-year sample's location, it may not be possible to calculate the SOCS if no sample with similarity exceeding the threshold is found. For each location with environmentally similar samples, the method for calculating the pseudo SOCS for the current year is given by Equation (6):

$$SOCS_i = \frac{\sum_{j=1}^{n} ES_{i,j} \cdot SOCS_j}{\sum_{j=1}^{n} ES_{i,j}} \tag{6}$$

where $SOCS_i$ is the pseudo SOCS at the i-th intervening-year sample's location; $SOCS_j$ is the SOCS at the j-th sample.

3.2.4. Approach parameter determination: Environmental similarity threshold and sample reliability threshold

There are two conditions that the environmental similarity threshold is used. The first is to be used for calculating the overall reliability of samples. The second is to be used for obtaining the set of environmentally similar samples for each intervening-year sample's location and calculating the pseudo SOCS.

When calculating the overall reliability of samples, it is necessary to consider the environmental similarity between samples based on the environmental similarity threshold to determine the number of samples (n) that can participate in the calculation (Eq. (5)). In the study by Liu

et al. (2020), the impact of different environmental similarity thresholds was discussed. It was found that the higher the environmental similarity threshold, the stronger the relationship of similarity between the environment and soil property is, but the number of samples (n) was small. In other words, the environmental similarity threshold plays an important role in the interpretability of the environment and soil property when calculating the overall reliability of samples.

The environmental similarity threshold is not only used to calculate the overall reliability of samples, also needs to be considered when calculating the pseudo SOCS using Equation (2) and (6). Specifically, if the environmental similarity between some samples and an intervening-year sample's is too low, those samples should not be included in the calculation of the pseudo SOCS for that location, as they may otherwise introduce interference. By setting an environmental similarity threshold, the calculating result will be more accurate. However, if the environmental similarity threshold is set too high, some locations will not have any samples exceeding the threshold. It is important to note that for a location (and also for the corresponding validation sample) to be considered available, it must meet the condition that there are environmentally similar samples for both 1980 and 2020.

The sample reliability threshold is used when, after calculating the overall reliability, it is obvious that a sample with low overall reliability has the weak relationship between the environment and soil property, which will interfere with the results, so it is necessary to delete these samples by setting a sample reliability threshold. An increase in the sample reliability threshold directly leads to a decrease in the number of available samples for further calculation (Liu et al., 2020). Therefore, to ensure a sufficient number of sample points, a threshold of 0.5 for overall reliability is used in this study, and samples with overall reliability below 0.5 are considered as interfering samples.

In this study, we selected environmental similarity thresholds of 0.5, 0.6, 0.7, 0.8, and 0.9 to test the sensitivity and observe the effect. The first test was that when calculating the overall reliability of samples, we used different environmental similarity thresholds. The second test was that when calculating the environmental similarity of filtered validation samples' locations to the filtered samples in starting and ending years, we used different environmental similarity thresholds. In the second test, to control the variable and ensure sufficient sample sizes for the calculation of environmental similarity and pseudo SOCS, the environmental similarity threshold for calculating the 1980, 2010 and 2020 samples' overall reliability was uniformly set at 0.7. Details will be presented in Section 4.3.

3.3. Modelling of SOCS change and validation

To evaluate the approach of model calibration by using pseudo revisited soil samples, the optimized parameters are used to rerun the RothC model for the year corresponding to each validation sample. The predicted values for each validation sample are compared with the observed values to assess prediction accuracy. The results are then compared using two approaches:

(1) Applying the parameter optimization approach across all validation samples;
(2) Using the spatiotemporal RF model.

In this study, the samples in 1980 and 2020 are used to calculate the pseudo SOCS, while the samples in 2010 are used as independent validation samples.

3.3.1. Parameter optimization for calibration of RothC model based on the generated pseudo revisited samples

The RothC model is based on first-order kinetics and divides soil organic carbon into five functional carbon pools, each representing different stabilities and decomposition characteristics of organic matter components, thereby refining the dynamic process of soil carbon cycling (Coleman and Jenkinson, 1996). These five carbon pools include decomposable plant material (DPM), resistant plant material (RPM), microbial biomass (BIO), humus (HUM), and inert organic matter (IOM). DPM represents organic matter that is chemically simple and easy to decompose, with the highest decomposition rate, while RPM corresponds to more chemically complex plant residues, decomposing more slowly and being more stable. BIO refers to metabolic products generated by soil microbial activity, with a medium decomposition rate, and HUM represents more stable organic matter in the soil, typically composed of complex organic compounds with a longer decomposition cycle. IOM is an extremely stable inert carbon pool, usually associated with mineral compounds, and its decomposition is assumed to be unmeasurable or non-existent in the model. After each carbon input, the carbon is divided between the DPM and RPM pools according to their ratio, and then decomposes into carbon dioxide, BIO, and HUM, with the decomposition rates determined by clay content. The BIO and HUM pools further decompose dynamically into more carbon dioxide, BIO, and HUM. For each pool, the amount of carbon decomposed during a carbon input and decomposition cycle (one month) is described by Equation (7).

$$\Delta C_t = C_{t-1}\left(1 - \exp[-\frac{1}{12}\cdot abck]\right) \tag{7}$$

where ΔC_t is the carbon decomposition in the current month; C_{t-1} is the carbon content in the pool at the end of the previous month; a is the rate modifying factor for temperature;b is the rate modifying factor for moisture; c is the rate modifying factor for soil cover; k is the decomposition rate constant for different pools.

In the RothC model, the equilibrium refers to the state where the SOC pools reach a dynamic equilibrium under long-term conditions. During this state, the total carbon input equals the total carbon decomposition, and the SOCS no longer undergoes significant changes. With changes in land use or management practices (such as tillage, fertilization, or vegetation changes), the current equilibrium of the soil may be disrupted and gradually shift toward a new equilibrium. In this experiment, the climate data for the spin-up phase was set as the average from 1970 to 1980, running for 10,000 years to reach equilibrium. During this long spin-up phase, the original first-order dynamics were maintained, and the effects of microbial factors were not considered. The initial annual carbon input value was set to 1 t/ha (Smith et al., 2007), and after reaching the equilibrium for the first time, the new carbon input was adjusted according to Equation (8).

$$C_{eq} = C_{init}\cdot\frac{SOCS_{obs} - IOM}{SOCS_{sim} - IOM} \tag{8}$$

where C_{eq} is the adjusted carbon input; C_{init} is the initial carbon input; $SOCS_{obs}$ is the observed SOCS in the starting year; $SOCS_{sim}$ is the simulated result after the initial run; and IOM refers to the inert organic matter pool. The calculation method for IOM is given by Equation (9) (Falloon et al., 1998):

$$IOM = 0.049 \times SOCS_{obs}^{1.139} \tag{9}$$

In the forward prediction state of the model, carbon input needs to be calculated based on actual conditions. However, in the absence of data such as plant residues, many studies adjust the annual carbon input ratio based on equilibrium carbon input, using Net Primary Production (NPP) (Gottschalk et al., 2012; Paul and Polglase, 2004) or fertilizer application (Dechow et al., 2019). Considering that the majority of the study area consists of farmland and that agricultural development has been rapid since 1980, with a positive correlation between fertilizer application and SOC accumulation (Meng et al., 2024; Pu et al., 2024), this study opts to adjust the model's carbon input during the prediction state using annual nitrogen fertilizer variations:

$$CI_t = CI_{t-1} \cdot \frac{Nfer_t}{Nfer_{t-1}} \tag{10}$$

where CI_t is the carbon input for the current year; CI_{t-1} is the carbon input for the previous year; $Nfer_t$ is the total amount of nitrogen fertilizer applied in the current year; and $Nfer_{t-1}$ is the total amount of nitrogen fertilizer applied in the previous year.

Many studies have optimized various parameters in RothC to calibrate the model, such as changing the value of decomposable plant material: resistant plant material (DPM:RPM) ratio (Wang et al., 2013) or changing the decomposition rate for the submergence period and the drained period (Shirato and Yokozawa, 2005), to make the model more applicable to the selected research area. Considering that microbial activity plays a significant role in SOC decomposition, and that quantitative studies on microbial activity already exist (Allison et al., 2010; Hofmeyr and Cornish-Bowden, 1997), this study selects microbial biomass rate correction parameters based on the inverse Michaelis-Menten (MM) equation (Woolf and Lehmann, 2019). The rate of carbon decomposition is adjusted according to the following Equation (11) and (12):

$$\Delta C_t = C_{t-1}\left(1 - \exp[-\frac{1}{12} \cdot abck \cdot \mu]\right) \tag{11}$$

$$\mu = \mu_{max}\frac{[MB]}{K_M + [MB]} \tag{12}$$

where μ is the monthly microbial biomass rate correction parameter; μ_{max} is the maximum microbial enzyme decomposition rate; K_M is the MM constant at which the reaction rate reaches half of its maximum value; [MB] is microbial biomass, here represented by the monthly BIO pool. During the parameter optimization, the Bayesian Optimization method (Fenton et al., 2016; Stone, 2013) is used to explore μ and K_M.

For parameter optimization at each validation sample's location (RothC with Parameter Optimization for Each Intervening-year Sample's Location, RPOEISL), the pseudo SOCS of 1980 is used as the starting observed value for the RothC model, and the pseudo SOCS of 2020 is used as the ending observed value. Bayesian optimization is then applied to optimize the values of μ_{max} and K_M, minimizing the error between the predicted value and the ending observed value, and obtaining the microbial parameters for that location. The calibrated RothC model is then rerun with the optimized parameters, using the pseudo SOCS of 1980 as the starting value and running the model to 2010 (Fig. 2).

3.3.2. The comparison experiments

As a contrast experiment, the starting and ending pseudo SOCS of all available validation samples in 2010 were used as observed values to optimize the microbial parameters using Bayesian optimization. That is, for all validation samples' locations, to minimize the overall mean squared error between all pairs of the predicted and observed values in 2020 and obtain the global microbial parameters μ_{max} and K_M (RothC with Global Parameter Optimization, RGPO). The optimized parameters were then applied to rerun the calibrated RothC model to 2010, and the validation accuracy under the globally unique parameters was assessed.

Meanwhile, a spatiotemporal RF model was used as a comparison model, with samples from 1980 and 2020 as training samples and 2010 as the validation samples. The spatiotemporal RF model utilized dynamic covariates corresponding to the sample year to reflect the temporal scale characteristics and predict the SOCS of the 2010 samples. Selected covariates were the same as when calculating the environmental similarity.

3.3.3. Validation indices

The root mean squared error (RMSE) and R-squared were used as accuracy metrics to evaluate the accuracy of the three experiments with the validation samples (Equation (13) and (14)). The model was run in the R 4.1.3 and Python 3.9 environments.

$$RMSE = \frac{1}{\sqrt{n}}\sqrt{\sum_{i=1}^{n}(y_{pre,i} - y_{obs,i})^2} \tag{13}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_{pre,i} - y_{obs,i})^2}{\sum_{i=1}^{n}(y_{pre,i} - \overline{y_{obs}})^2} \tag{14}$$

where n is the total number of validation samples; $y_{pre,i}$ is the predicted value for the i-th validation sample; $y_{obs,i}$ is the observed value for the i-th validation sample; $\overline{y_{obs}}$ is the mean of the observed values for the validation sample set.

4. Results

4.1. The pseudo SOCS results based on environmental similarity calculation

Fig. 4 displays the pseudo SOCS for 1980 and 2020 across all available grids within the study area. Grids marked with a value of 0 in Figure 4 indicate locations where no environmentally similar samples are available after environmental threshold screening. Grids labeled with NA values result from limitations in the original data resolution and the presence of water bodies, where valid raster values are unavailable. The figure reveals that the mean SOC storage in 2020 (Fig. 4-b) decreased compared to 1980 (Fig. 4-a), especially in the southwestern region of the study area. According to the elevation distribution shown in Fig. 1, the study area generally has low elevations and flat terrain, with most of the available samples located in low-altitude areas. In contrast, fewer samples are available in the higher-altitude, mountainous regions. Consequently, grids at these elevations lack samples with high environmental similarity and become unavailable after the environmental similarity threshold is applied. Then the two-year pseudo SOCS are extracted to the intervening-year samples' locations for RothC calibration.

4.2. The validation results of three approaches

The prediction results for each validation sample's location, obtained using the RothC model with parameters calibrated based on the pseudo SOCS (environmental similarity threshold set to 0.7), the RothC model with global parameter optimization, and the spatiotemporal RF model, are shown in Fig. 5. It can be seen that the RMSE (5.28) for the prediction results after RPOEISL is improved by 31.6 % compared to the RGPO (7.72) and by 10.7 % compared to the spatiotemporal RF (5.91). The R^2 (0.456) is approximately 42.9 % higher than that of spatiotemporal RF (0.319), proving that the calculation with environmental similarity constraints better reflects the SOCS differences caused by environmental characteristics at different locations. We also observed that the RGPO approach exhibits lower accuracy in this study area. As shown in Fig. 5-b, most of the predicted results are distributed in the 23–30 t/ha range. The environmental difference at different locations is significant as shown in Fig. 1. This is also evident from Fig. 4. It implies that global parameters based on overall trends do not effectively reflect the variation at all samples, particularly those where the SOCS change trend differs from most other samples. In contrast, the predicted results with RPOEISL and spatiotemporal RF do not show a clear concentration in a specific range, but a wider and more reasonable range.

4.3. Impact of environmental similarity threshold when calculating pseudo SOCS

As mentioned in Section 3.2.4, in the first test, the environmental similarity thresholds were set to 0.6, 0.7, 0.8, and 0.9, and the count of samples with overall reliability exceeding 0.5 under each environmental

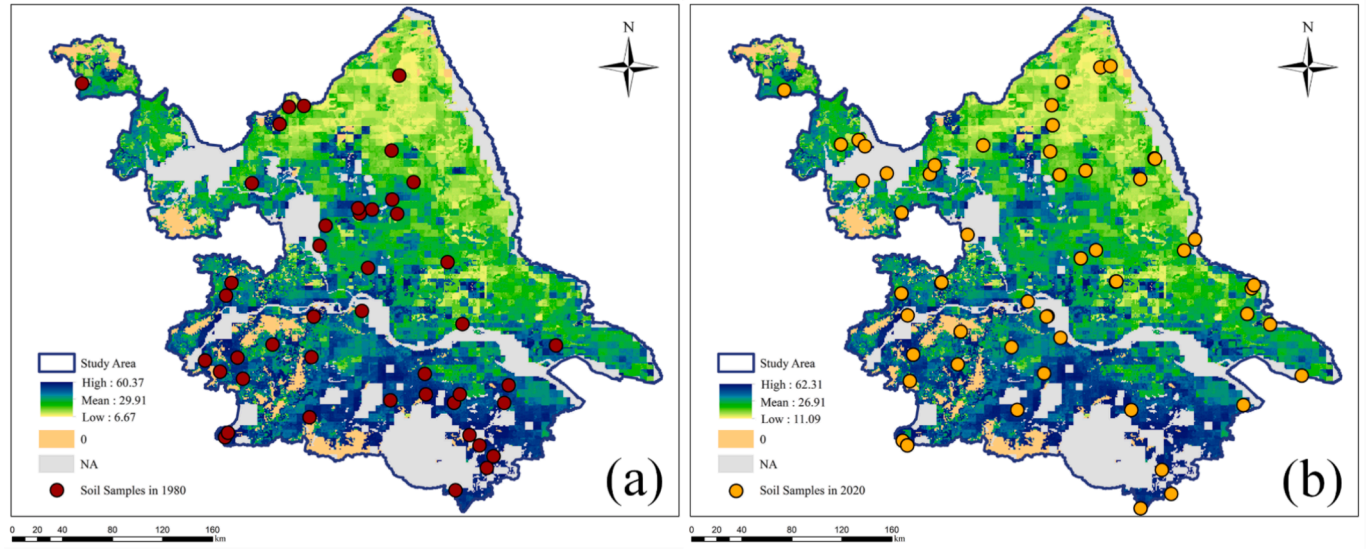


Fig. 4. Pseudo SOCS calculation results for all raster grids for 1980 (a) and 2020 (b).

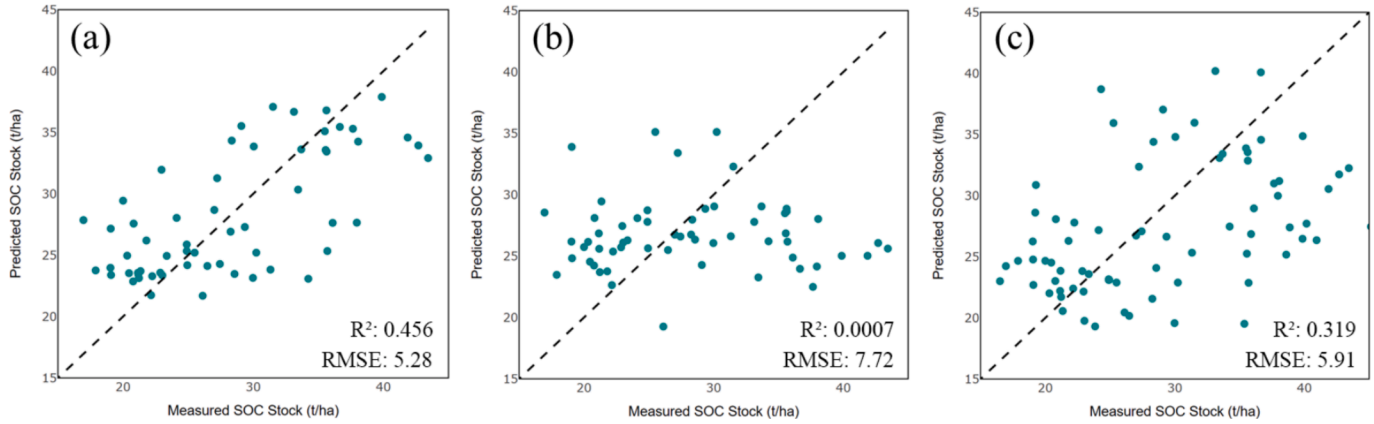


Fig. 5. Scatter plots of predictions from three approaches: (a) Predictions after RothC with Parameter Optimization for Each Intervening-year Sample's Location; (b) Predictions after RothC with Global Parameter Optimization; (c) Predictions from the spatiotemporal RF model.

similarity threshold was calculated (Table 1). It is obvious that with high environmental similarity threshold, the counts of samples in three years are small. We finally selected 0.7 when calculating the overall reliability.

The results of the second test are shown in Table 2, with the increase of the environmental similarity threshold, the number of available validation samples gradually decreased. The RMSE initially decreased, then increased, showing a U-shaped trend. When the environmental similarity threshold reached 0.9, the number of available validation samples significantly decreased, which reduced the statistical significance of the validation samples. This led to a rise in the overall RMSE due to large errors from a small number of samples. However, as the environmental similarity threshold increased, the R^2 value improved significantly compared to lower thresholds. Yet, the R^2 value also fluctuated due to the changes in the samples. For instance, when the threshold was set to 0.8, the R^2 was lower than at thresholds of 0.7 and 0.9. Considering these points, to ensure that the 1980 and 2020 samples reflect the regional environmental conditions through their similarity relations, and that the number of validation samples is sufficient, we ultimately selected an environmental similarity threshold of 0.7 after balancing these considerations.

Table 1
Number of samples exceeding the overall reliability threshold under different environmental similarity thresholds for each year.

Environmental Similarity Threshold	Count of 1980 Samples	Count of 2010 Samples	Count of 2020 Samples
0.6	47	67	64
0.7	45	60	60
0.8	36	56	55
0.9	22	41	40

Table 2
Number of available validation samples and validation accuracy in the RothC with Parameter Optimization for Each Intervening-year Sample's Location approach with different environmental similarity thresholds.

Environmental Similarity Threshold	Count of Available Validation Samples	RMSE	R^2
0.5	63	6.19	0.154
0.6	61	5.64	0.252
0.7	60	5.28	0.456
0.8	53	6.35	0.394
0.9	35	8.01	0.421

4.4. The dynamic changes of SOCS at the validation samples from 1980 to 2020

The simulation results of all validation samples from 1980 to 2020 using the three approaches are shown in Fig. 6. It can be seen that the overall trend of SOCS change from 1980 to 2020 presents a fluctuating downward pattern. In the RothC model results (Fig. 6-a and 6-b), the average SOCS changes from 1980 to around 1995 with an initial decline followed by a rise, and it starts to gradually decline after 2000. With the RPOEISL approach (Fig. 6-a), the mean line aligns well with the mean of the 2010 validation samples, which performs better than RGPO approach and spatiotemporal RF.

From Fig. 6-a and Fig. 6-b, it can be seen that the RothC model produces smoother simulations over time, while the spatiotemporal RF simulation results in Fig. 6-c show much more pronounced variability over the time scale, which indicates that when conducting analyses on a temporal scale, the results derived from spatiotemporal RF may not align closely with actual conditions, also as shown in Fig. 5-c.

5. Discussion

5.1. Advantages of the proposed approach

This study calculates pseudo SOCS based on the environmental similarity and calibrate the RothC model for the intervening-year SOCS prediction. According to the result, our approach is more accurate than the spatiotemporal RF model. This indicates that using environmentally similar samples is a practical and effective approach to generate pseudo samples with SOC calculated based on environmental similarity for carbon model calibration. Yet, using RothC model calibrated with global parameter optimization was not effective for our study area at a regional scale. In some studies, model typically runs based on default parameters or globally optimized parameter (Zhang et al., 2020), but relying on these generalized parameters may affect the model's accuracy under varying environmental conditions (Tao et al., 2023). This means that the accuracy of this approach is strongly related to the heterogeneity and existing samples of the selected study area. Thus, there has been many studies calibrate model for each subarea or sample (Li and Jun 2018; Paramesh et al., 2022; Wilcox et al., 2023). In this study, to optimize parameters on each point to better reflect the influence of the site-specific environment condition also performs better than for globally.

To more accurately capture the changes in SOCS over time, a sufficient number of revisited samples are usually required. But in fact, it is somewhat challenging to require a large number of revisited samples in different study areas. From a temporal perspective, non-revisited samples hinder the analysis of soil property changes over time. Our approach addresses this issue by selecting validation samples with relatively high simulation accuracy. These samples, with calculated values for both the starting and ending years, are treated as expanded revisited samples. This approach allows the simulation of SOC change across different years with process-based models.

RF model primarily specializes in constructing the relationship between environment and soil property over space (Fathizad et al., 2020; Grimm et al., 2008; Guio Blanco et al., 2018). The spatiotemporal RF builds spatial–temporal relationships based on existing sample soil properties over multiple time points and predicts unobserved locations at the target time point with the corresponding environmental covariates. However, it is not like process-based models which simulate based on the soil temporal change knowledge (Talebi et al., 2022; Zhang et al., 2024a). As shown in Fig. 6, our study reveals that simple spatiotemporal RF models show significant fluctuations when predicting SOCS for different years. This highlights that when only two years of training samples are available, the RF method for simulating long-term SOCS changes needs further investigation. For comparison, process-based models such as RothC analyses carbon change on time series under the influence of environment (Coleman and Jenkinson, 1996; Peltre et al., 2012), results from process-based models show smoother temporal variations, and are closer to the real change patterns compared to the sharp annual fluctuations in the RF results.

5.2. Applicability, limitations and uncertainties

When applying our approach, there are several issues to be considered. First, although there are three years of sampling points in our case study, our approach is not only applicable to calculate the pseudo SOCS in starting and ending two years. If samples from multiple years are available, any number of them can be selected to anchor the trend of SOCS changes. The more anchor years used, the model's ability to reflect the temporal changes in SOCS and the more accurate the predictions for future years may be better. Similarly, for example, if samples exist for two different years (years A and B) but are both non-revisited, the year-B samples' pseudo revisited samples (in year A) can be calculated directly by existing samples in year A. Moreover, for each year for calculating the pseudo SOCS, the result will be more accurate if more existing available samples can be used.

When calculating the pseudo SOCS, it is crucial to select relevant environmental covariates. That is, the relationship between the environment and the soil property may be reflected by different environmental covariates according to the study area, not all the accessible environmental covariates are suitable. For example, if the study area is small, the climate covariates may not differ so much, then climatic covariates can be not used (Zhu et al., 2018, Zhu et al., 2015).

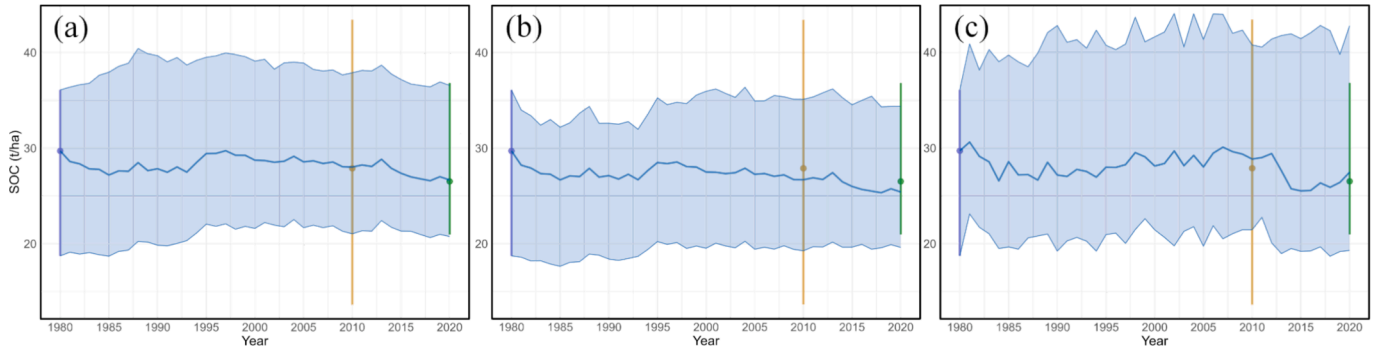


Fig. 6. Time trend of validation samples under different approaches: (a) Time trend after RothC with RothC with Parameter Optimization for Each Intervening-year Sample's Location; (b) Time trend after RothC with Global Parameter Optimization; (c) Time trend of spatiotemporal RF prediction. The blue solid line represents the mean value; the blue shaded area indicates the prediction range for all validation samples each year; the purple vertical lines and dots represent the pseudo SOCS and the mean of all validation samples in 1980; the green vertical lines and dots represent the pseudo SOCS and the mean of all validation samples in 2020; the yellow vertical lines and dots represent the observed values and the mean of all validation samples. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

A suitable environmental similarity threshold needs to be determined, balancing the number of candidate environmentally similar samples selected from the starting and ending years with the representativeness of the selected samples as mentioned in Section 3.2.4. If a high environmental similarity threshold is chosen, the overall reliability of the samples increases, meaning these samples are more applicable to the similarity principle (Liu et al., 2020). However, the reduced number of samples may not represent the environmental characteristics of the majority of locations in the study area. Additionally, some locations may lack sufficiently environmentally similar samples from starting and ending years, reducing the model's generalizability. In other words, for regions with different environment condition, selecting appropriate environmental covariates and adjusting the threshold may be necessary to make this approach more applicable to the selected study area.

The uncertainty in our proposed method primarily arises from two components. First, during the generation of pseudo revisited soil sample data, the selected environmental covariates, the environmental similarity thresholds and the variations in the number of sample points, can influence the experimental outcomes, as discussed earlier and illustrated in Tables 1 and 2. Second, when running the RothC model, uncertainty is predominantly associated with carbon inputs and model parameter settings (Farina et al., 2013; Jebari et al., 2021). However, during the RothC model operation, if the carbon inputs, model parameters, and climatic data during the process are fixed, the resulting outputs are deterministic, meaning that no uncertainty emerges at the operational stage (Davoudabadi et al., 2024).

5.3. Future perspectives

Generating pseudo samples by our approach is used for process-based models in this study. Yet, it may also be used in other ML methods in DSM, or causal model such as structural equation models (Grace and Keeley, 2006; Yang et al., 2021) to analyze SOCS change by further integrating soil science principles (Minasny et al., 2024). When selecting the environmental covariates for ML models, some methods like SHAP (SHapley Additive exPlanations) can help to identify key drivers of soil properties (Beucher et al., 2022; Wadoux et al., 2023). Moreover, if appropriately resolved farmland management data are available in future, they can also be used as environmental covariates. After obtaining the revisited samples, ML methods can be combined to explore SOCS changes across different years, spatial grids, and environmental conditions, serving as supplementary training samples to expand the sample set (Zhang et al., 2024b). Additionally, not only for SOCS, our approach can also simulate other soil properties as long as conforming to the relationship between environment and soil property.

After calculating and filtering environmental similarities, there are some girds without environmentally similar points matched. This means that the existing samples cannot represent all the area well. Thus, supplementary sampling can be conducted in these regions to enhance the representativeness of the samples in capturing regional spatiotemporal variations in the soil property.

It would help enhance accuracy and interpretability across diverse environmental conditions by further integrating process models with pseudo observations and machine learning models in case of lacking revisiting samples. For example, Minasny et al. (2024) proposed methodologies for integrating soil science knowledge into ML models, including enhancing training datasets with observational priors, designing model structures that reflect soil science principles, and supervising model training using loss functions informed by soil science. Zhang et al. (2024a) introduced a process-oriented machine learning model that employed RothC to generate additional training points for years without samples, which were then incorporated into the RF model to enhance its accuracy.

6. Conclusion

This study calculates the environmental similarity and pseudo SOCS, in order to obtain pseudo revisited soil samples and calibrate process-based model. Compared to RothC with global parameter optimization and spatiotemporal RF, this approach achieves higher accuracy in simulating the temporal dynamics of SOCS at validation points. The results confirm that calculating pseudo SOCS based on environmental similarity and subsequently optimizing the model parameters significantly are an effective approach for modeling SOC change. This approach provides the possibility of obtaining revisited soil samples and conducting SOC analysis based on them in areas without revisited soil samples, in a relatively less complicated and accurate way. Furthermore, attention must still be paid to determining environmental similarity thresholds for the model's prediction accuracy. Future studies could incorporate samples from more years and further optimize parameters for model calibration through environmental covariate analysis. Additionally, integrating this approach with other process-based models or ML techniques may improve computational efficiency and model applicability, providing broader prospects for large-scale dynamic monitoring and prediction of SOCS or other soil properties. Based on this, establishing long-term SOC monitoring networks and integrating real-time environmental data (e.g., meteorological data, vegetation indices) could enable dynamic model calibration, capturing both short-term fluctuations and long-term trends in SOC dynamics, thereby providing more timely scientific support for climate change mitigation and agricultural management.

CRediT authorship contribution statement

Wenkai Cui: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Lin Yang:** Writing – review & editing, Supervision, Resources, Methodology, Funding acquisition, Conceptualization. **Lei Zhang:** Writing – review & editing, Validation, Supervision, Methodology. **Chenconghai Yang:** Writing – review & editing, Visualization, Software. **Chenxu Zhu:** Writing – review & editing, Investigation, Data curation. **Chenghu Zhou:** Writing – review & editing, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This study is supported by the National Natural Science Foundation of China (Project No. 42471468), and the Leading Funds for the First Class Universities (0209-1480604109).

Data availability

Data will be made available on request.

References

Abramoff, R., Xu, X., Hartman, M., O'Brien, S., Feng, W., Davidson, E., Finzi, A., Moorhead, D., Schimel, J., Torn, M., Mayes, M.A., 2018. The Millennial model: in search of measurable pools and transformations for modeling soil carbon in the new century. Biogeochemistry 137, 51–71. https://doi.org/10.1007/s10533-017-0409-7.

Allison, S.D., Wallenstein, M.D., Bradford, M.A., 2010. Soil-carbon response to warming dependent on microbial physiology. Nat. Geosci. 3, 336–340. https://doi.org/10.1038/ngeo846.

Beucher, A., Rasmussen, C.B., Moeslund, T.B., Greve, M.H., 2022. Interpretation of convolutional neural networks for acid sulfate soil classification. Front. Environ. Sci. 9, 809995. https://doi.org/10.3389/fenvs.2021.809995.

Burrough, P.A., Macmillan, R.A., Van Deursen, W., 1992. Fuzzy classification methods for determining land suitability from soil profile observations and topography. J. Soil Sci. 43, 193–210. https://doi.org/10.1111/j.1365-2389.1992.tb00129.x.

Coleman, K., Jenkinson, D.S., 1996. RothC-26.3 - A Model for the turnover of carbon in soil. In: Powlson, D.S., Smith, P., Smith, J.U. (Eds.), Evaluation of Soil Organic Matter Models. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 237–246. https://doi.org/10.1007/978-3-642-61094-3_17.

Davoudabadi, M.J., Pagendam, D., Drovandi, C., Baldock, J., White, G., 2024. Innovative approaches in soil carbon sequestration modelling for better prediction with limited data. Sci. Rep. 14, 3191. https://doi.org/10.1038/s41598-024-53516-z.

Dechow, R., Franko, U., Kätterer, T., Kolbe, H., 2019. Evaluation of the RothC model as a prognostic tool for the prediction of SOC trends in response to management practices on arable land. Geoderma 337, 463–478. https://doi.org/10.1016/j.geoderma.2018.10.001.

Ding, Y., Peng, S., 2021. Spatiotemporal change and attribution of potential evapotranspiration over China from 1901 to 2100. Theor. Appl. Climatol. 145, 79–94. https://doi.org/10.1007/s00704-021-03625-w.

Ding, Y., Peng, S., 2020. Spatiotemporal trends and attribution of drought across China from 1901–2100. Sustainability 12, 477. https://doi.org/10.3390/su12020477.

Falloon, P., Smith, P., Coleman, K., Marshall, S., 1998. Estimating the size of the inert organic matter pool from total soil organic carbon content for use in the Rothamsted carbon model. Soil Biol. Biochem. 30, 1207–1211. https://doi.org/10.1016/S0038-0717(97)00256-3.

Farina, R., Coleman, K., Whitmore, A.P., 2013. Modification of the RothC model for simulations of soil organic C dynamics in dryland regions. Geoderma 200–201, 18–30. https://doi.org/10.1016/j.geoderma.2013.01.021.

Fathizad, H., Ardakani, A.H., Sodaiezadeh, H., Kerry, R., Taghizadeh-Mehrjardi, R., 2020. Investigation of the spatial and temporal variation of soil salinity using random forests in the central desert of Iran. Geoderma 365, 114233. https://doi.org/10.1016/j.geoderma.2020.114233.

Fenton, N., Neil, M., Berger, D., 2016. Bayes and the Law. Annu. Rev. Stat. Its Appl. 3, 51–77. https://doi.org/10.1146/annurev-statistics-041715-033428.

Gottschalk, P., Smith, J.U., Wattenbach, M., Bellarby, J., Stehfest, E., Arnell, N., Osborn, T.J., Jones, C., Smith, P., 2012. How will organic carbon stocks in mineral soils evolve under future climate? Global projections using RothC for a range of climate change scenarios. Biogeosciences 9, 3151–3171. https://doi.org/10.5194/bg-9-3151-2012.

Grace, J.B., Keeley, J.E., 2006. A structural equation model analysis of postfire plant diversity in california shrublands. Ecol. Appl. 16, 503–514. https://doi.org/10.1890/1051-0761(2006)016[0503:ASEMAO]2.0.CO;2.

Grimm, R., Behrens, T., Märker, M., Elsenbeer, H., 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island — Digital soil mapping using Random Forests analysis. Geoderma 146, 102–113. https://doi.org/10.1016/j.geoderma.2008.05.008.

Grinand, C., Maire, G.L., Vieilledent, G., Razakamanarivo, H., Razafimbelo, T., Bernoux, M., 2017. Estimating temporal changes in soil carbon stocks at ecoregional scale in Madagascar using remote-sensing. Int. J. Appl. Earth Obs. Geoinformation 54, 1–14. https://doi.org/10.1016/j.jag.2016.09.002.

Guio Blanco, C.M., Brito Gomez, V.M., Crespo, P., Ließ, M., 2018. Spatial prediction of soil water retention in a Páramo landscape: Methodological insight into machine learning using random forest. Geoderma 316, 100–114. https://doi.org/10.1016/j.geoderma.2017.12.002.

Hamzehpour, N., Shafizadeh-Moghadam, H., Valavi, R., 2019. Exploring the driving forces and digital mapping of soil organic carbon using remote sensing and soil texture. CATENA 182, 104141. https://doi.org/10.1016/j.catena.2019.104141.

Hisse, I.R., Biganzoli, F., Peper, A.M., Poggio, S.L., 2022. Annual productivity of cropping sequences: responses to increased intensification levels. Eur. J. Agron. 137, 126506. https://doi.org/10.1016/j.eja.2022.126506.

Hofmeyr, J.-H.-S., Cornish-Bowden, H., 1997. The reversible Hill equation: how to incorporate cooperative enzymes into metabolic models. Bioinformatics 13, 377–385. https://doi.org/10.1093/bioinformatics/13.4.377.

Huang, B., Pan, J., 2017. *Soil Series of China (Jiangsu Volume)*. Science Press, Beijing, China.

Jebari, A., Álvaro-Fuentes, J., Pardo, G., Almagro, M., Prado, A., 2021. Estimating soil organic carbon changes in managed temperate moist grasslands with RothC. PLOS ONE 16, e0256219. https://doi.org/10.1371/journal.pone.0256219.

Kelly, R.H., Parton, W.J., Crocker, G.J., Graced, P.R., Klír, J., Körschens, M., Poulton, P. R., Richter, D.D., 1997. Simulating trends in soil organic carbon in long-term experiments using the century model. Geoderma 81, 75–90. https://doi.org/10.1016/S0016-7061(97)00082-7.

Lal, R., 2004. Soil carbon sequestration impacts on global climate change and food security. Science 304, 1623–1627. https://doi.org/10.1126/science.1097396.

Lehmann, J., Bossio, D.A., Kögel-Knabner, I., Rillig, M.C., 2020. The concept and future prospects of soil health. Nat. Rev. Earth Environ. 1, 544–553. https://doi.org/10.1038/s43017-020-0080-8.

Li, H., Cao, Y., Xiao, J., Yuan, Z., Hao, Z., Bai, X., Wu, Y., Liu, Y., 2024. A daily gap-free normalized difference vegetation index dataset from 1981 to 2023 in China. Sci. Data 11, 527. https://doi.org/10.1038/s41597-024-03364-3.

Li, L., Jun, K.S., 2018. Distributed parameter unsteady flow model for the Han River. J. Hydro-Environ. Res. 21, 86–95. https://doi.org/10.1016/j.jher.2018.08.004.

Liu, D.L., Chan, K.Y., Conyers, M.K., Li, G., Poile, G.J., 2011. Simulation of soil organic carbon dynamics under different pasture managements using the RothC carbon model. Geoderma 165, 69–77. https://doi.org/10.1016/j.geoderma.2011.07.005.

Liu, J., Zhu, A.-X., Rossiter, D., Du, F., Burt, J., 2020. A trustworthiness indicator to select sample points for the individual predictive soil mapping method (iPSM). Geoderma 373, 114440. https://doi.org/10.1016/j.geoderma.2020.114440.

McBratney, A., Field, D.J., Koch, A., 2014. The dimensions of soil security. Geoderma 213, 203–213. https://doi.org/10.1016/j.geoderma.2013.08.013.

Meng, X., Zhang, X., Li, Y., Jiao, Y., Fan, L., Jiang, Y., Qu, C., Filimonenko, E., Jiang, Y., Tian, X., Shi, J., Kuzyakov, Y., 2024. Nitrogen fertilizer builds soil organic carbon under straw return mainly via microbial necromass formation. Soil Biol. Biochem. 188, 109223. https://doi.org/10.1016/j.soilbio.2023.109223.

Minasny, B., Bandai, T., Ghezzehei, T.A., Huang, Y.-C., Ma, Y., McBratney, A.B., Ng, W., Norouzi, S., Padarian, J., Rudiyanto, S., Styc, Q., Widyastuti, M., 2024. Soil science-informed machine learning. Geoderma 452, 117094. https://doi.org/10.1016/j.geoderma.2024.117094.

Minasny, B., McBratney, A.B., 2006. A conditioned Latin hypercube method for sampling in the presence of ancillary information. Comput. Geosci. 32, 1378–1388. https://doi.org/10.1016/j.cageo.2005.12.009.

Nemo, K., Coleman, K., Dondini, M., Goulding, K., Hastings, A., Michael, Jones B., Leifeld, J., Osborne, B., Saunders, M., Scott, T., Teh, Y.A., Smith, P., 2017. Soil organic carbon (SOC) equilibrium and model initialisation methods: an application to the rothamsted carbon (RothC) model. Environ. Model. Assess. 22, 215–229. https://doi.org/10.1007/s10666-016-9536-0.

Office for the Second National Soil Survey of China, 1993. *Soil Species of China*, Vol. 1 [dataset]. Chinese Agriculture Press, Beijing, China.

Padarian, J., Stockmann, U., Minasny, B., McBratney, A.B., 2022. Monitoring changes in global soil organic carbon stocks from space. Remote Sens. Environ. 281, 113260. https://doi.org/10.1016/j.rse.2022.113260.

Paramesh, V., Kumar, P., Nath, A.J., Francaviglia, R., Mishra, G., Arunachalam, V., Toraskar, S., 2022. Simulating soil organic carbon stock under different climate change scenarios: A RothC model application to typical land-use systems of Goa, India. CATENA 213, 106129. https://doi.org/10.1016/j.catena.2022.106129.

Paul, K.I., Polglase, P.J., 2004. Calibration of the RothC model to turnover of soil carbon under eucalypts and pines. Soil Res. 42, 883. https://doi.org/10.1071/SR04025.

Peltre, C., Christensen, B.T., Dragon, S., Icard, C., Kätterer, T., Houot, S., 2012. RothC simulation of carbon accumulation in soil after repeated application of widely different organic amendments. Soil Biol. Biochem. 52, 49–60. https://doi.org/10.1016/j.soilbio.2012.03.023.

Peng, S., 2022. 1-km Monthly Potential Evapotranspiration Dataset in China (1990–2021). Natl. Tibet. Plateau Data Cent, Beijing China.

Peng, S., 2020. 1-km monthly precipitation dataset for China (1901–2020). Natl. Tibet. Plateau Data Cent, Beijing China.

Peng, S., 2019. 1-km monthly mean temperature dataset for china (1901–2023) . National Tibetan Plateau/Third Pole Environment Data Center.

Peng, S., Ding, Y., Liu, W., Li, Z., 2019. 1 km monthly temperature and precipitation dataset for China from 1901 to 2017. Earth Syst. Sci. Data 11, 1931–1946. https://doi.org/10.5194/essd-11-1931-2019.

Pu, Y., Yang, L., Zhang, L., Huang, H., Zhang, G., Zhou, C., 2024. Major contributions of agricultural management practices to topsoil organic carbon distribution and accumulation in croplands of East China over three decades. Agric. Ecosyst. Environ. 359, 108749. https://doi.org/10.1016/j.agee.2023.108749.

Qingsong, S., Xiaobing, L., Xingyi, Z., 2023. Evaluating soil organic carbon changes after 16 years of soil relocation in Chinese Mollisols by optimizing the input data of the RothC model. Soil Tillage Res. 225, 105561. https://doi.org/10.1016/j.still.2022.105561.

Radočaj, D., 2024. Geospatial prediction of total soil carbon in European agricultural land based on deep learning. Sci. Total Environ.

Shangguan, W., Dai, Y., Liu, B., Ye, A., Yuan, H., 2012. A soil particle-size distribution dataset for regional land and climate modelling in China. Geoderma 171–172, 85–91. https://doi.org/10.1016/j.geoderma.2011.01.013.

Shen, F., Yang, L., Zhang, L., Guo, M., Huang, H., Zhou, C., 2023. Quantifying the direct effects of long-term dynamic land use intensity on vegetation change and its interacted effects with economic development and climate change in Jiangsu, China. J. Environ. Manage. 325, 116562. https://doi.org/10.1016/j.jenvman.2022.116562.

Shi, G., Sun, W., Shangguan, W., Wei, Z., Yuan, H., Zhang, Y., Liang, H., Li, L., Sun, X., Li, D., Huang, F., Li, Q., Dai, Y., 2024. A China dataset of soil properties for land surface modeling (version 2). doi: 10.5194/essd-2024-299.

Shirato, Y., Yokozawa, M., 2005. Applying the rothamsted carbon model for long-term experiments on japanese paddy soils and modifying it by simple tuning of the decomposition rate. Soil Sci. Plant Nutr. 51, 405–415. https://doi.org/10.1111/j.1747-0765.2005.tb00046.x.

Smith, J., Smith, P., Wattenbach, M., Gottschalk, P., Romanenkov, V.A., Shevtsova, L.K., Sirotenko, O.D., Rukhovich, D.I., Koroleva, P.V., Romanenko, I.A., Lisovoi, N.V., 2007. Projected changes in the organic carbon stocks of cropland mineral soils of European Russia and the Ukraine, 1990–2070. Glob. Change Biol. 13, 342–356. https://doi.org/10.1111/j.1365-2486.2006.01297.x.

Smith, P., Soussana, J.-F., Angers, D., Schipper, L., Chenu, C., Rasse, D.P., Batjes, N.H., van Egmond, F., McNeill, S., Kuhnert, M., Arias-Navarro, C., Olesen, J.E., Chirinda, N., Fornara, D., Wollenberg, E., Álvaro-Fuentes, J., Sanz-Cobena, A., Klumpp, K., 2020. How to measure, report and verify soil carbon change to realize the potential of soil carbon sequestration for atmospheric greenhouse gas removal. Glob. Change Biol. 26, 219–241. https://doi.org/10.1111/gcb.14815.

Stone, J.V., 2013. Bayes' Rule: A Tutorial Introduction to Bayesian Analysis. Sebtel Press.

Talebi, H., Peeters, L.J.M., Otto, A., Tolosana-Delgado, R., 2022. A truly spatial random forests algorithm for geoscience data analysis and modelling. Math. Geosci. 54, 1–22. https://doi.org/10.1007/s11004-021-09946-w.

Tao, F., Huang, Y., Hungate, B.A., Manzoni, S., Frey, S.D., Schmidt, M.W.I., Reichstein, M., Carvalhais, N., Ciais, P., Jiang, L., Lehmann, J., Wang, Y.-P.,

Houlton, B.Z., Ahrens, B., Mishra, U., Hugelius, G., Hocking, T.D., Lu, X., Shi, Z., Viatkin, K., Vargas, R., Yigini, Y., Omuto, C., Malik, A.A., Peralta, G., Cuevas-Corona, R., Di Paolo, L.E., Luotto, I., Liao, C., Liang, Y.-S., Saynes, V.S., Huang, X., Luo, Y., 2023. Microbial carbon use efficiency promotes global soil carbon storage. Nature 618, 981–985. https://doi.org/10.1038/s41586-023-06042-3.

Ugbemuna Ugbaje, S., Karunaratne, S., Bishop, T., Gregory, L., Searle, R., Coelli, K., Farrell, M., 2024. Space-time mapping of soil organic carbon stock and its local drivers: potential for use in carbon accounting. Geoderma 441, 116771. https://doi.org/10.1016/j.geoderma.2023.116771.

Viscarra Rossel, R.A., Lee, J., Behrens, T., Luo, Z., Baldock, J., Richards, A., 2019. Continental-scale soil carbon composition and vulnerability modulated by regional environmental controls. Nat. Geosci. 12, 547–552. https://doi.org/10.1038/s41561-019-0373-z.

Wadoux, A.-M.-J.-C., Saby, N.P.A., Martin, M.P., 2023. Shapley values reveal the drivers of soil organic carbon stock prediction. SOIL 9, 21–38. https://doi.org/10.5194/soil-9-21-2023.

Wang, J., Lu, C., Xu, M., Zhu, P., Huang, S., Zhang, W., Peng, C., Chen, X., Wu, L., 2013. Soil organic carbon sequestration under different fertilizer regimes in north and northeast C hina: R oth C simulation. Soil Use Manag. 29, 182–190. https://doi.org/10.1111/sum.12032.

Wang, W., Guo, Y., Yang, L., Adams, J.M., 2024. Methanogen–methanotroph community has a more consistent and integrated structure in rice rhizosphere than in bulk soil and rhizoplane. Mol. Ecol. 33, e17416. https://doi.org/10.1111/mec.17416.

Wilcox, K.R., Chen, A., Avolio, M.L., Butler, E.E., Collins, S., Fisher, R., Keenan, T., Kiang, N.Y., Knapp, A.K., Koerner, S.E., Kueppers, L., Liang, G., Lieungh, E., Loik, M., Luo, Y., Poulter, B., Reich, P., Renwick, K., Smith, M.D., Walker, A., Weng, E., Komatsu, K.J., 2023. Accounting for herbaceous communities in process-based models will advance our understanding of “grassy” ecosystems. Glob. Change Biol. 29, 6453–6477. https://doi.org/10.1111/gcb.16950.

Woolf, D., Lehmann, J., 2019. Microbial models with minimal mineral protection can explain long-term soil organic carbon persistence. Sci. Rep. 9, 6522. https://doi.org/10.1038/s41598-019-43026-8.

Yang, L., Shen, F., Zhang, L., Cai, Y., Yi, F., Zhou, C., 2021. Quantifying influences of natural and anthropogenic factors on vegetation changes using structural equation modeling: A case study in Jiangsu Province. China. J. Clean. Prod. 280, 124330. https://doi.org/10.1016/j.jclepro.2020.124330.

Yu, Z., Liu, J., Kattel, G., 2022. Historical nitrogen fertilizer use in China from 1952 to 2018. Earth Syst. Sci. Data 14, 5179–5194. https://doi.org/10.5194/essd-14-5179-2022.

Zhang, C., Di, Z., Duan, Q., Xie, Z., Gong, W., 2020. Improved land evapotranspiration simulation of the community land model using a surrogate-based automatic parameter optimization method. Water 12, 943. https://doi.org/10.3390/w12040943.

Zhang, L., Heuvelink, G.B.M., Mulder, V.L., Chen, S., Deng, X., Yang, L., 2024a. Using process-oriented model output to enhance machine learning-based soil organic carbon prediction in space and time. Sci. Total Environ. 922, 170778. https://doi.org/10.1016/j.scitotenv.2024.170778.

Zhang, X., Zhao, T., Xu, H., Liu, W., Wang, J., Chen, X., Liu, L., 2024b. GLC_FCS30D: the first global 30 m land-cover dynamics monitoring product with a fine classification system for the period from 1985 to 2022 generated using dense-time-series Landsat imagery and the continuous change-detection method. Earth Syst. Sci. Data 16, 1353–1381. https://doi.org/10.5194/essd-16-1353-2024.

Zhao, Y., Wang, M., Hu, S., Zhang, X., Ouyang, Z., Zhang, G., Huang, B., Zhao, S., Wu, J., Xie, D., Zhu, B., Yu, D., Pan, X., Xu, S., Shi, X., 2018. Economics- and policy-driven organic carbon input enhancement dominates soil organic carbon accumulation in Chinese croplands. Proc. Natl. Acad. Sci. 115, 4045–4050. https://doi.org/10.1073/pnas.1700292114.

Zhou, T., Geng, Y., Chen, J., Pan, J., Haase, D., Lausch, A., 2020. High-resolution digital mapping of soil organic carbon and soil total nitrogen using DEM derivatives, Sentinel-1 and Sentinel-2 data based on machine learning algorithms. Sci. Total Environ. 729, 138244. https://doi.org/10.1016/j.scitotenv.2020.138244.

Zhou, Y., Hartemink, A.E., Shi, Z., Liang, Z., Lu, Y., 2019. Land use and climate change effects on soil organic carbon in North and Northeast China. Sci. Total Environ. 647, 1230–1238. https://doi.org/10.1016/j.scitotenv.2018.08.016.

Zhu, A., Lu, G., Liu, J., Qin, C., Zhou, C., 2018. Spatial prediction based on Third Law of Geography. Ann. GIS 24, 225–240. https://doi.org/10.1080/19475683.2018.1534890.

Zhu, A.-X., 1999. A personal construct-based knowledge acquisition process for natural resource mapping. Int. J. Geogr. Inf. Sci. 13, 119–141. https://doi.org/10.1080/136588199241382.

Zhu, A.X., Lv, G.N., Zhou, C.H., Qin, C.Z., 2020. Geographic similarity: third law of geography? J. Geo-Inf. Sci. 22 (4), 673–679. https://doi.org/10.12082/dqxxkx.2020.200069.

Zhu, A.X., Liu, J., Du, F., Zhang, S.J., Qin, C.Z., Burt, J., Behrens, T., Scholten, T., 2015. Predictive soil mapping with limited sample data. Eur. J. Soil Sci. 66, 535–547. https://doi.org/10.1111/ejss.12244.

Zhu, A.-X., Turner, M., 2022. How is the third law of geography different? Ann GIS 28, 57–67. https://doi.org/10.1080/19475683.2022.2026467.