Using process-oriented model output to enhance machine learning-based soil organic carbon prediction in space and time

Science of the Total Environment

Zhang, Lei; Heuvelink, Gerard B.M.; Mulder, Vera L.; Chen, Songchao; Deng, Xunfei et al

https://doi.org/10.1016/j.scitotenv.2024.170778

Science of the Total Environment 922 (2024) 170778

Contents lists available at ScienceDirect

Science of the Total Environment

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

Using process-oriented model output to enhance machine learning-based soil organic carbon prediction in space and time

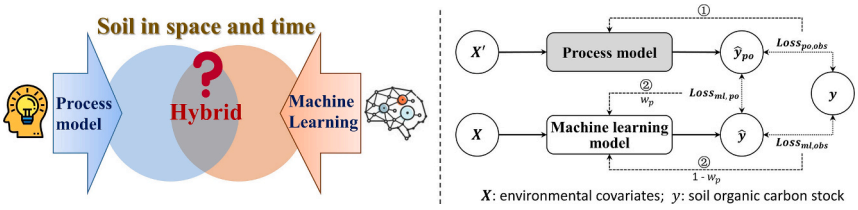
Lei Zhang [a,b], Gerard B.M. Heuvelink [b,c], Vera L. Mulder [b], Songchao Chen [d], Xunfei Deng [e], Lin Yang [a,f,*]

[a] School of Geography and Ocean Science, Nanjing University, Nanjing, China
[b] Soil Geography and Landscape Group, Wageningen University, Wageningen, the Netherlands
[c] ISRIC – World Soil Information, Wageningen, the Netherlands
[d] ZJU-Hangzhou Global Scientific and Technological Innovation Center, Zhejiang University, Hangzhou, China
[e] Institute of Digital Agriculture, Zhejiang Academy of Agricultural Sciences, Hangzhou, Zhejiang, China
[f] Frontiers Science Center for Critical Earth Material Cycling, Nanjing University, Nanjing, China

HIGHLIGHTS

- Clarifies the advantages and disadvantages of process-oriented (PO) models and machine learning (ML) models in space-time soil carbon modelling
- Proposes a general framework (POML method) for integrating PO and ML models
- Evaluates and compares POML model with single PO and ML model in spatial patterns, temporal trends and prediction accuracies.
- Discusses applicability and future horizons on space-time modelling of soil carbon using POML framework

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Paulo Pereira

Keywords:
Hybrid modelling
Mechanistic knowledge-guided machine learning
RothC
Random forest
Digital soil mapping
Soil carbon dynamics

ABSTRACT

Monitoring and modelling soil organic carbon (SOC) in space and time can help us to better understand soil carbon dynamics and is of key importance to support climate change research and policy. Although machine learning (ML) has attracted a lot of attention in the digital soil mapping (DSM) community for its powerful ability to learn from data and predict soil properties, such as SOC, it is better at capturing soil spatial variation than soil temporal dynamics. By contrast, process-oriented (PO) models benefit from mechanistic knowledge to express physiochemical and biological processes that govern SOC temporal changes. Therefore, integrating PO and ML models seems a promising means to represent physically plausible SOC dynamics while retaining the spatial prediction accuracy of ML models. In this study, a hybrid modelling framework was developed and tested for predicting topsoil SOC stock in space and time for a regional cropland area located in eastern China. In essence, the hybrid model uses predictions of the PO model in unsampled years as additional training data of the ML model, with a weighting parameter assigned to balance the importance of SOC values from the PO model and real measurements. The results indicated that temporal trends of SOC stock modelled by PO and ML models were largely different, while they were notably similar between the PO and hybrid models. Cross-validation showed that the hybrid model had the best performance (RMSE = 0.29 kg m^{-2}), with a 19 % improvement compared

* Corresponding author at: School of Geography and Ocean Science, Nanjing University, Nanjing, China.
E-mail addresses: lei.zhang.geo@outlook.com (L. Zhang), yanglin@nju.edu.cn (L. Yang).

https://doi.org/10.1016/j.scitotenv.2024.170778
Received 29 September 2023; Received in revised form 23 January 2024; Accepted 5 February 2024
Available online 8 February 2024

with the ML model. We conclude that the proposed hybrid framework not only enhances space-time soil carbon mapping in terms of prediction accuracy and physical plausibility, it also provides insights for soil management and policy decisions in the face of future climate change and intensified human activities.

1. Introduction

Soil organic carbon (SOC) is a key component of soil health and plays an important role in regulating the global carbon cycle (Crowther et al., 2016; Lehmann et al., 2020; Friedlingstein et al., 2022). Monitoring and modelling SOC in space and time can help understand SOC changes and thus support predicting and mitigating future climate change (Rumpel, 2019; Bossio et al., 2020; Padarian et al., 2022b). However, temporal variation of SOC is usually small compared to its spatial variation, leading to difficulties in capturing SOC dynamics without sufficient monitoring data across time. In addition, since carbon stocks in soils have complex variations influenced by natural and anthropogenic factors (Zhao et al., 2018; Huang et al., 2022; Q. Wang et al., 2023a), space-time modelling of soil carbon using digital soil mapping (DSM) methods is challenging. Modelling is particularly challenging in agricultural lands that are strongly impacted by human activities (Lal, 2002; Wadoux et al., 2021; Padarian et al., 2022a; Huang et al., 2022).

DSM builds on the concept that soil variation is a resultant of variation in soil forming factors, that in turn are represented by environmental covariates. DSM has become the most commonly used approach to predict the spatial distribution of soil properties (McBratney et al., 2003; Minasny and McBratney, 2016). While geostatistical methods have been widely used in DSM for many years (Goovaerts, 1999; Heuvelink and Webster, 2001; Heuvelink et al., 2016), in the last decade, machine learning (ML) methods have attracted more attention because of the advantage of not requiring models with strict statistical assumptions and its powerful learning ability to adaptively fit the data distribution (Brungard et al., 2015; Heung et al., 2016; Lamichhane et al., 2019; Wadoux et al., 2020; Zhang et al., 2022). ML-based DSM is a data-driven approach that typically requires large training datasets, but has also demonstrated its validity with limited sample sizes (Zhang et al., 2021). Most ML-based DSM methods are applied to static cases, that is, predicting the spatial variation of soil properties at one point in time or predicting it at multiple times by separately modelling each time event. Despite recent studies showed that space-time modelling and mapping of soil properties with ML is possible (e.g. Ivushkin et al., 2019; Heuvelink et al., 2021; Helfenstein et al., 2022), the prediction uncertainties are often too large to obtain statistically significant results about the temporal variation (Heuvelink et al., 2021).

In spite of the powerful learning and predictive ability of ML models in the field of DSM, a common criticism is that it is a purely data-driven approach, in which it is difficult to embed pedological knowledge or soil-related physiochemical laws. This hampers the effective representation of existing pedological knowledge in soil predictions (Hendriks et al., 2021) and acknowledgement of physical laws in modelling temporal variation of soil properties, such as SOC (Heuvelink and Webster, 2001). By contrast, some existing process-oriented (PO) models (e.g. RothC, Century and Millennial) explicitly describe the accumulation and decomposition of SOC over time (Parton et al., 1988; Coleman and Jenkinson, 1996; Abramoff et al., 2018, 2022). These models incorporate the effects of physical, chemical and biological processes that influence soil carbon dynamics and turnover rates (Parton et al., 2015; Sierra and Müller, 2015; Smith et al., 2020). A comprehensive understanding of soil carbon change processes through PO models contributes to a better generalization and transferability of models (Abramoff et al., 2022). PO models are specifically designed to simulate soil dynamics, so that they are more suited for capturing temporal changes and can extrapolate over time under different scenarios. Although we can calibrate a ML model based on multi-year soil samples with static and dynamic environmental covariates to model soil variation in space and time (Heuvelink et al., 2021; Helfenstein et al., 2022), this approach has not yet been able to explicitly and accurately capture the temporal variation as well as PO models can. It is therefore desirable to use PO models to compensate the shortcoming of ML models in modelling soil temporal variation.

While PO models have important advantages on modelling temporal variation of soil, they also have limitations compared with ML models. Firstly, using a PO model often requires substantial effort from users to understand the detailed processes involved and how these are represented in the model. Secondly, the number of input variables (e.g. plant residues, soil temperature and moisture in case of SOC) of PO models are limited and usually fixed by the design of the model structure, while ML models can be calibrated with a wide range of environmental covariates. Thirdly, while there is a wide array of ML model types which can be trained by a unified learning framework, PO models are generally restricted to express kinetic equations that simulate soil carbon dynamics, leading to the modelling procedures becoming less flexible. Therefore, PO models are often found to be less accurate than ML models when the goal is to map the spatial distribution of soil properties (Hendriks et al., 2021; Abramoff et al., 2022; Xie et al., 2022).

Considering the benefits and limitations of PO and ML models reviewed above, integrating these two modelling approaches may advance our capacity to make more robust soil carbon predictions in both space and time. Some efforts have been made to show the potential of combining pedological knowledge and process-oriented models with ML models. For example, structural equation modelling (Angelini et al., 2016) and Bayesian belief networks (Taalab et al., 2015) have been used to convert a conceptual soil-landscape model or predefined rules of expert knowledge into a statistically explicit model. Recently, Xie et al. (2022) proposed an integrated approach that includes the RothC output as a covariate of a geographically weighted regression kriging model. Zhang et al. (2023) used a similar idea that used the output of two PO models as additional covariates for a ML model. Although these studies showed improved prediction accuracy by using PO model outputs as additional covariates, in general, the development of hybrid models combining PO and ML methods for spatiotemporal modelling of SOC is still scarce. To date, a tangible framework that formalizes the integration of the PO model into the ML-based soil carbon modelling in space and time is still underdeveloped. Therefore, this study aims to provide a novel solution to combine the two types of models, thus addressing one of the ten challenges for the future of pedometrics (Wadoux et al., 2021).

The objectives of this study are to: (i) propose a general hybrid modelling framework that integrates PO and ML models; and (ii) evaluate the performance and applicability of the hybrid method in modelling SOC stocks in space and time. The methodology was tested in a case study located in a cropland (paddy soils) area in eastern China, from 1980 to 2000. The hybrid model was compared with the two individual models and each model was evaluated by analyzing modelling results in terms of spatial patterns, temporal trends and prediction accuracies.

2. Materials and methods

2.1. Study area and datasets

2.1.1. Study area

The study area is located in the Hang-Jia-Hu region, which is the largest plain in Zhejiang province, China, and is an integral part of the Yangtze Delta (Fig. 1a,b). The main cities in the study area are Hangzhou, Huzhou and Jiaxing. There are twelve counties in this area,

covering an area about 7500 km^2. The area has a subtropical humid monsoon climate with abundant precipitation of about 1300 mm per year (Yang et al., 2021). The average annual temperature is around 16 °C, with hot humid summers and cool winters. It is a coastal and lacustrine alluvial plain with a low and flat topography in the eastern part. Rice (*Oryza sativa*) is the dominant crop in this area, so most of the study area has paddy soils. Most regions in the plain are dominated by single cropping rice, the growing season of rice is roughly from May to October (Xiao et al., 2021; S. Wang et al., 2023b).

2.1.2. Soil sample data

The soil dataset consists of 856 field topsoil samples (0–20 cm) from 428 locations that were sampled both in 1980 and in 2000 (Fig. 1c,d). The SOC content (g kg^{-1}) was measured for each sample and converted to SOC stock (SOCS) (kg m^{-2}) by:

$$SOCS = SOC_{content} \bullet BD \bullet (1 - CF) \bullet 0.2 \quad (1)$$

where BD is soil bulk density (g cm^{-3}), and CF is the proportion of coarse fragments in the whole soil. As the sample data do not have BD and CF values, we extracted these variables from the latest high-resolution National Soil Information Grids of China (Liu et al., 2020, 2022). Fig. 2a shows histograms of the SOCS values at sample locations in the two sampling years. Fig. 2b shows that SOCS at most sampling sites increased from 1980 to 2000, with about 0.5 kg m^{-2} (~10 %) increase on average (Table S1).

2.1.3. Covariate data

The environmental covariates selected in this study represent topography, climate, vegetation, soil and human activity (Table 1). Elevation (ELEV), slope (SLP) and stream power index (SPI) were obtained from the Geomorpho90m dataset (Amatulli et al., 2020). Monthly mean annual temperature (TMP), precipitation (PRE) and evapotranspiration (PET) from the TerraClimate dataset (Abatzoglou et al., 2018) represent the long-term dynamic climate in the area. The satellite-based normalized difference vegetation index (NDVI) was adopted for representing the vegetation factor in the area. The monthly NDVI database created by Ma et al. (2022) was used. Fine-scale NDVI spatial information was extracted from high-resolution MODIS images and integrated with long-term temporal observations from the AVHRR database (Ma et al., 2022). A map of soil clay content was obtained from Liu et al. (2020). For representing human activity, we used fertilizer input data as key agricultural management information, considering that the cropland area experienced a rapid increase in fertilizer use during the 1980–2000 period (Zhao et al., 2018; Yu et al., 2022). Annual county-level data of fertilizer input were obtained and compiled from the National Bureau of Statistics of China (Chinese Statistical Yearbook, CSY).

Topography and soil clay content were assumed to be static covariates, while all other covariate data were adopted as dynamic variables, thus supporting the temporal component of the space-time SOCS modelling. Topographic covariates were only used as input for the ML model, while all other covariates were used for both the PO and ML models. In case of NDVI we used monthly values for the PO model, while seasonal means were used for ML model. This was done to avoid using an excessive number of vegetation covariates compared to other covariates for ML model training.

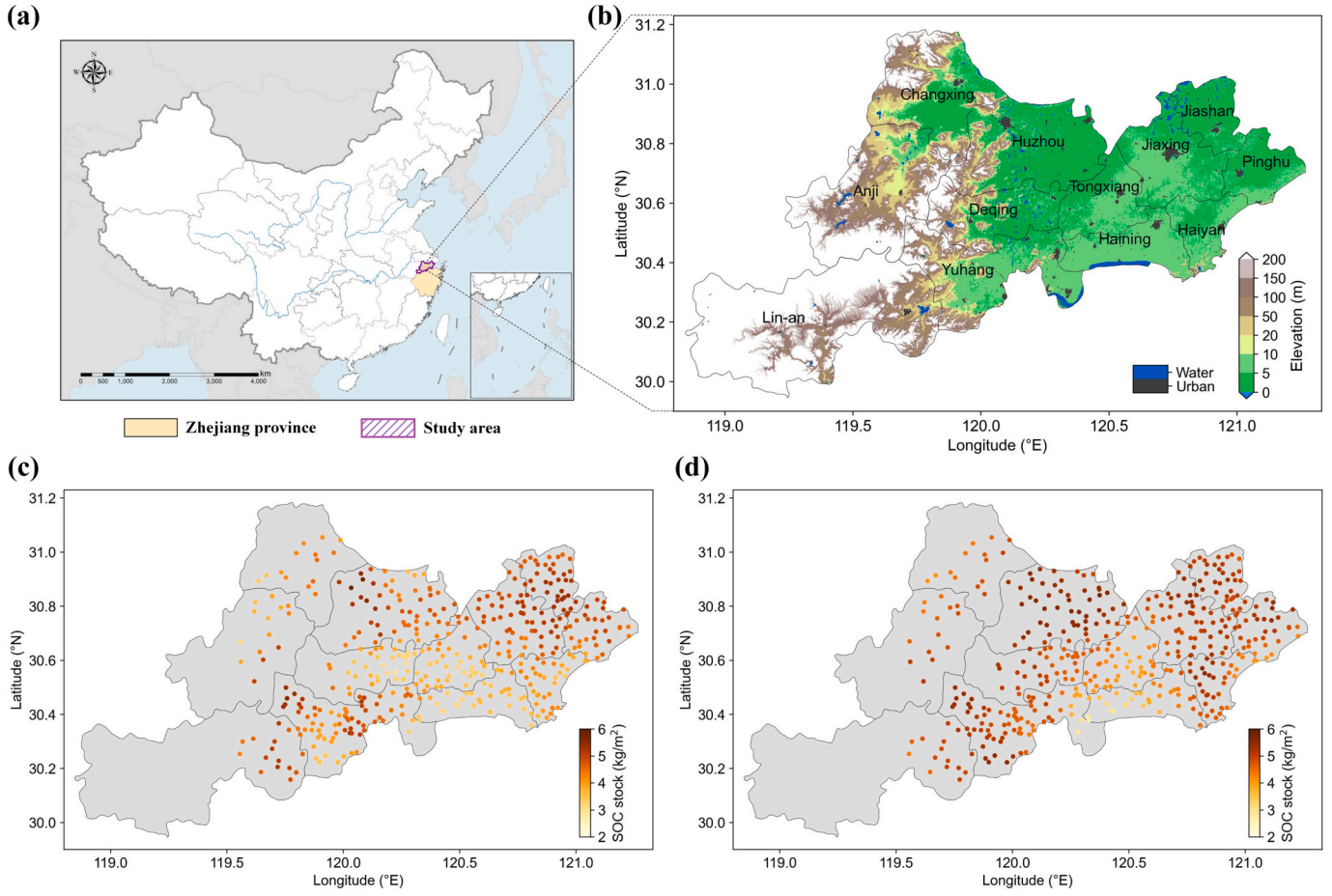


Fig. 1. Location (a) and elevation map (b) of the study area. Soil organic carbon stocks at sample locations collected in 1980 and 2000 are shown in (c) and (d), respectively.

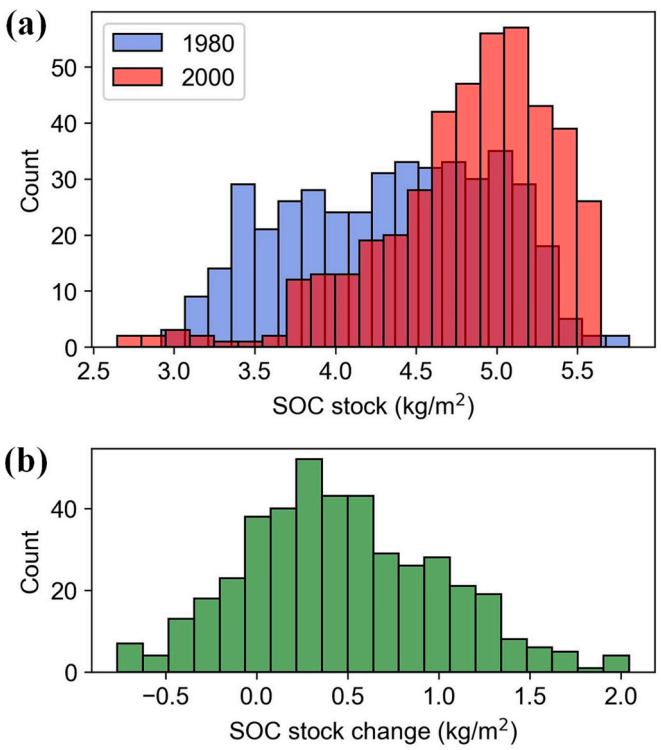


Fig. 2. (a) Histograms of soil organic carbon stock (SOCS) data from 1980 and 2000; (b) Histogram of SOCS change between 1980 and 2000.

2.2. Process-oriented model (RothC)

A process-oriented soil carbon model provides a mathematical description of SOCS dynamics. The commonly adopted concept in these models is to divide the total carbon storage into multiple pools. The varying turnover or residence times of organic carbon in different pools is represented by using different decay rates of carbon in each pool. Mathematically, the decomposition of carbon is usually described by first-order kinetics, which means that the decomposition rate of carbon stock is proportional to its size (Parton et al., 2015). Many PO models (e.g. RothC and Century) were designed based on this structure. In this study, we used the RothC model as an example to conduct SOC predictions, it being one of the most widely used soil carbon stock PO models and which has also been adopted in paddy soils (Jiang et al., 2013).

RothC partitions the total organic carbon over five pools, including four active pools (decomposable plant material, DPM; resistant plant material, RPM; microbial biomass, BIO; humified organic matter, HUM) and one inactive pool, the inert organic matter (IOM) (Fig. S1). The decomposition of carbon in each active pool is defined by a first-order process with its own decay rate. By also including the carbon input (C_{input}), the carbon stock (C) dynamics in the five pools is thus given by:

$$\boldsymbol{C}(t+\Delta t) = \boldsymbol{C}(t) + \boldsymbol{C}_{input}(t) - m(t) \bullet \boldsymbol{A} \bullet \boldsymbol{C}(t) \bullet \Delta t \quad (2)$$

where $\boldsymbol{C}(t) = [C_{DPM}(t), C_{RPM}(t), C_{BIO}(t), C_{HUM}(t), C_{IOM}(t)]^T$, $\boldsymbol{C}_{input}(t) = [C_{input_{DPM}}(t), C_{input_{RPM}}, 0, 0, 0]^T$, Δt is a time step (we used monthly time steps). $C_{input_{DPM}}$ and $C_{input_{RPM}}$ are determined from C_{input} by a DPM/RPM ratio, for which a default value can be used. $\boldsymbol{A}$ is a matrix determining the transition of carbon between pools, and given by:

$$\boldsymbol{A} = \begin{bmatrix} k_{DPM} & 0 & 0 & 0 & 0 \\ 0 & k_{RPM} & 0 & 0 & 0 \\ -\alpha\, k_{DPM} & -\alpha\, k_{RPM} & (1-\alpha)\, k_{BIO} & -\alpha\, k_{HUM} & 0 \\ -\beta\, k_{DPM} & -\beta\, k_{RPM} & -\beta\, k_{BIO} & (1-\beta)\, k_{HUM} & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (3)$$

where k_{DPM}, k_{RPM}, k_{BIO} and k_{HUM} are the decomposition rate constants for each active carbon pool, default values of which have been derived from the original field experiments (Jenkinson et al., 1987, 1992). Fractions α and β, determined by the clay content of the soil, represent the proportion of decomposed carbon that goes to BIO and HUM. The remaining part $(1-\alpha-\beta)$ is released as CO_2 and lost from the system.

In Eq. 2, $m(t)$ is a rate-modifying factor that depends on external variables. In the original RothC model, it includes modifiers a, b and c for considering the effects of temperature, moisture and soil cover, respectively. These three factors can be determined by using the equations described in Coleman and Jenkinson (1996) (see Table S2 for details). As microbial processes have gradually gained widespread attention in soil process models (Allison et al., 2010; Woolf and Lehmann, 2019), in our study, we added a factor μ to additionally consider the biological effect (beyond first-order dynamics) on carbon decay rates. Thus, we defined the modifying factor m and the factor μ as:

$$m(t) = a(t) \bullet b(t) \bullet c(t) \bullet \mu(t) \quad (4)$$

$$\mu(t) = \mu_{max} \frac{\mathrm{MB}(t)}{K_m + \mathrm{MB}(t)} \quad (5)$$

where μ_{max} is the maximum value of this rate-modifying factor, MB is the microbial biomass, which can be represented by the size of BIO pool in RothC, and K_m is the Michaelis constant, which represents the value of MB at which the reaction rate is at half-maximum ($\frac{\mu_{max}}{2}$). All values or equations of abovementioned model parameters are presented in Table S2.

The general modelling procedure of RothC consists of a spin up phase and a forward phase. The spin up phase initializes the model by computing an equilibrium model state under static carbon input and a known constant climate and total carbon stock (Gottschalk et al., 2012; Smith et al., 2005, 2007). In other words, it divides the carbon stock over the five pools. While it is not difficult to compute the equilibrium analytically from Eq. 2, it is often computed numerically by running the model over a long time period under constant climatic and carbon input conditions. In our study, the model was run at each sampling location for 10,000 years, with average climate data for the two decades prior to the

Table 1
Environmental covariates used in the study.

Category	Covariate name	Abbreviation	Spatial resolution	Used for ML/PO model	Reference
Topography	Elevation	ELEV	90 m	ML	Amatulli et al. (2020)
	Slope	SLP		ML	
	Stream power index	SPI		ML	
Climate	Temperature	TMP	1/24°	ML, PO	Abatzoglou et al. (2018)
	Precipitation	PRE		ML, PO	
	Evapotranspiration	PET		ML, PO	
Vegetation	Normalized Difference Vegetation Index	NDVI	250 m	ML (seasonal), PO (monthly)	Ma et al. (2022)
Soil	Clay content	CLAY	90 m	ML, PO	Liu et al. (2020)
Human activity	Fertilizer input	FER	County-level	ML, PO	Chinese Statistical Yearbook – CSY

Note: ML and PO represent the machine learning and process-oriented model, respectively. CSY dataset is available from: https://data.stats.gov.cn

first sampling period (i.e. 1960—1980). During the spin up phase, we used the standard RothC set-up assuming first-order dynamics and set the rate-modifying factor μ to 1. The annual carbon input can be initially assumed to be an arbitrary value, such as 0.1 kg C m^{-2} yr^{-1} (Smith et al., 2007). The difference between the simulated and observed carbon stock is then used to adjust the carbon inputs as follows (Smith et al., 2005):

$$C_{input}^{eq} = C_{input}^{'} \times \frac{SOCS_{obs} - C_{IOM}}{SOCS_{sim} - C_{IOM}} \quad (6)$$

where $C_{input}^{'}$ is the initially assumed arbitrary value of total carbon input; C_{input}^{eq} is the carbon input required to reach the observed carbon stock at equilibrium; $SOCS_{obs}$ and $SOCS_{sim}$ are the observed SOC stock in the starting year and the steady-state simulated SOC stock, respectively. C_{IOM} has to be defined separately, for this we used the equation provided in Falloon et al. (1998):

$$C_{IOM} = 0.049 \times C_{obs}^{1.139} \quad (7)$$

Then, the sizes of different SOC pools at equilibrium can be estimated using pedotransfer functions (Weihermüller et al., 2013).

After having calculated the carbon inputs and carbon stocks in different pools at the starting year, the model can be run to simulate SOC stock and stock change for the time period of interest (from 1980 to 2000) using climatic data and agricultural practice data from that period. The carbon input for each year is ideally derived from time series of plant residues and farm manure, but detailed information on these variables were not available in our study. Facing the same problem, many previous studies assumed that the carbon inputs to the soil is proportional to the net primary production (NPP) (Smith et al., 2005; Gottschalk et al., 2012; Zhang et al., 2023). Some recent studies pointed out that deriving carbon input from remote sensing-based NPP products is not realistic, especially in cropland areas (e.g. Minasny et al., 2022). Considering that our study area witnessed a marked increase of fertilizer use and an increase of agricultural production during 1980–2000 (Zhao et al., 2018; Yu et al., 2022; Pu et al., 2024), we therefore assumed that carbon input was proportional to fertilizer application, as also noted in previous studies in croplands in China (Ge et al., 2015; Zhao et al., 2018; Pu et al., 2024). We obtained the fertilizer input per unit area per year from county-level agriculture management records in CSY. Thus, the carbon inputs during the simulation period were obtained following the same approach as described in Smith et al. (2005), but with NPP replaced by fertilizer input.

Two model parameters, μ_{max} and K_m, were calibrated using training data. We used the quasi-Newton Broyden–Fletcher–Goldfarb–Shanno (BFGS) method (Nocedal and Wright, 2006; Shanno, 1970) to adjust the two parameters by minimizing the mean squared error (i.e., the mean squared differences between simulated and observed SOC stocks at the end of the simulation period) of the training data in a cross-validation procedure (see Section 2.4 for the details of model training and validation).

2.3. Combining the process-oriented and machine learning models

When using machine learning to generate the soil-environment relationship, formally, it can learn a functional relationship $f_{ml}: y \leftarrow X$, where X represents the environmental covariates and y is the target soil property (e.g. SOCS). If the covariates vary both in space and time and observations of the target variable are paired with covariate values of the sampling locations and sampling time, we can derive a space-time model. However, a ML model fitted with training data collected in a limited number of sampling years may not reproduce the dynamic behavior well. In the case of SOCS, spatial variation dominates over temporal variation which further complicates reproducing dynamic variation well (Heuvelink et al., 2021). Hence, if we have a PO model that can generate more realistic patterns of soil carbon temporal changes, it would be sensible to integrate PO model simulations into machine learning and thus improve the soil carbon modelling in space as well as in time.

The basic concept of the proposed hybrid method is to incorporate the PO model derived simulations at unsampled years as additional training data into the data-driven machine learning model, to achieve that the training data span the entire time of interest. The framework of this strategy is illustrated in Fig. 3. The overall workflow consists of two main steps:

(1) Calibrate the PO model using the environmental input data $X^{'}$ and observed SOC data y. This involves both the spin up and estimation of μ_{max} and K_m. Next run the forward phases as described in Section 2.2 to get PO predictions for the time period of interest. This yields SOC stock simulations for all sampling locations and all years between the starting and end year.
(2) Calibrate a space-time ML model by using both the sample data and the simulated data from the PO model. Note that the covariates of the ML model (X) is different from the input to the PO model ($X^{'}$), and usually $X^{'} \subset X$, as described in Table 1. Two loss functions were adopted to optimize the model: $Loss_{ml,obs}$ measures the mean squared difference between the predicted values ($\widehat{y}$) and observed values (y) at sampling locations in the sampling years; $Loss_{ml,po}$ is the mean squared difference between $\widehat{y}$ and $\widehat{y}_{po}$ (PO model simulated values) at sampling locations in the years without sampling (i.e., the period 1981–1999). The overall loss function was chosen as a weighted sum of the two loss functions, and the final predictive model $\widehat{f}$ was obtained by minimizing the overall loss function:

$$\widehat{f} = \underset{f}{argmin}\left[w_p \bullet Loss_{ml,po}\left(f(\boldsymbol{X}), \widehat{\boldsymbol{y}}_{po}\right) + \left(1 - w_p\right) \bullet Loss_{ml,obs}\left(f(\boldsymbol{X}), \boldsymbol{y}\right)\right] \quad (8)$$

where w_p is a weighting parameter, ranging from 0 to 1, controlling the importance of the PO modelling results (the simulated data) compared with the observed SOCS. This weighting parameter is a user-defined hyper-parameter that we set to 0.5 as a default, assuming that both losses are equally important. Note that setting $w_p = 0$ reverts to a common RF model that does not make use of PO model simulations, while setting $w_p = 1$ means that all observations are ignored and that the RF model builds a meta-model of the PO model. The effect of choosing different values of the weighting parameter on the modelling result for this study were also analyzed.

Random forest (RF) regression (Breiman, 2001) was adopted as the

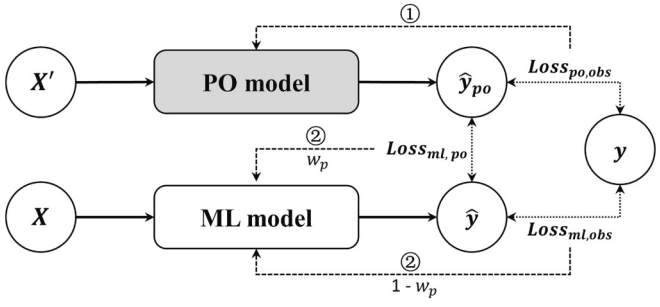


Fig. 3. Framework of the proposed hybrid model that incorporates the process-oriented (PO) model into the machine learning (ML) model. X and $X^{'}$ represent input covariates to the ML and PO model, respectively; $\widehat{y}$ and $\widehat{y}_{po}$ represent predicted values from the ML and PO model, respectively; y represents observed values. Solid arrows represent the computing direction of data; dotted arrows represent the differences (e.g. the mean squared error) among the predicted values from the PO and ML model and the observed values, which forms the loss functions; dashed arrows show how loss functions were used to optimize or calibrate the model. Numbers ① and ② above two dashed arrows indicate the order of the model calibration. w_p is a weighting parameter which controls the importance of two loss functions.

ML model in this study. The *scikit-learn* (Pedregosa et al., 2011) package in the Python programming language (Pérez et al., 2011) was used to apply the model. Training an RF model or any other ML model with this package requires training data, with the option to supply a vector containing the weights of each observation. This allowed us to weigh the importance of SOCS observations and the PO simulated SOCS. Note that since the number of observed and simulated data is not equal, the weights assigned to individual simulated points were multiplied by the ratio of the number of observed to simulated data (i.e., $w_p \times 2/19$ in our case study). For the hyper-parameters in RF, '*n_estimators*' (the number of trees in the forest) was set to 200, since previous studies showed that this is sufficient to obtain stable results (Wadoux, 2019; Zhang et al., 2021). Hyper-parameter '*max_features*' (the number of covariates that are randomly selected for each tree building process) was set to the default value of the square root of the total number of covariates.

2.4. Model validation

The PO, ML and hybrid (POML) model for predicting SOCS were validated using five-fold cross-validation (CV). Accuracy metrics were calculated to assess individual model accuracy and evaluate the model performance. The CV procedure for three models is illustrated in Fig. 4. The SOCS observations and PO model simulated data were each divided into five equally sized subsets, and the division was performed on each year of data to ensure that the five split sets remained consistent within each year (i.e., both the observations and simulations from a sampling location always ended up in the same fold). Four folds were used as the training data to calibrate the model, and the prediction was validated on the remaining fold, using only the observations. This procedure was carried out five times, each time using a different fold for validation. The PO model was trained using data from the four folds in t_1 and t_2 (i.e., the years1980 and 2000 in this study), and only validated on the validation set in t_2 because the SOCS at all sample locations in t_1 were used for initializing (spinning up) the PO model. The ML model was trained and validated based on the sample data in t_1 and t_2. The hybrid model was trained based on the sample data in t_1 and t_2 and the simulated data belonging to the training set between the two years, while the model performance was evaluated on the validation set in sampling years using the CV procedure.

The accuracy metrics of the root mean squared error (RMSE) and the concordance correlation coefficient (CCC) were computed from the validation samples to assess and compare the performance of all three models.

3. Results

3.1. Modelling accuracies

The prediction accuracies of the PO, ML and hybrid models are shown in Fig. 5. Density scatter plots show that the predictions are generally unbiased when three models were fitted using sample data from 1980 and 2000 (Fig. 5a,c,d). However, for the case where we limit ourselves to solely using sample data from 1980 to fit a ML model and then predicting in 2000, the predictions becomes strongly biased (Fig. 5b). When using sample data from both years, the ML model performance is better than that of the PO model (0.36 kg m^{-2} versus 0.53 kg m^{-2} in RMSE; 0.80 versus 0.63 in CCC). Importantly, our results show that the hybrid model has the highest accuracy (RMSE = 0.29 kg m^{-2} and CCC = 0.88). In our study area, the improvement of the hybrid POML model compared to the ML model was 10 % in terms of CCC and 19 % in terms of RMSE. Our cross-validation results thus demonstrate that the proposed POML modelling approach can effectively integrate the outputs of the PO model into ML to achieve an improved spatio-temporal prediction accuracy.

3.2. Temporal trends

The temporal trends of the modelled SOCS by three models are shown in Fig. 6. Although the three models show a general increasing trend of SOCS from 1980 to 2000, the temporal variations between two years derived from PO and ML models are largely different (Fig. 6a,b). The PO model shows a trend of smooth variation across time, with a slight decline during the first five years and a continuous increase until tending to be stable after 1996. The county-level results also show a similar general trend (Fig. S2). In contrast, the predictions of the ML model show a fluctuating trend of SOCS change. This suggests that only using sample data in two sampling years cannot accurately fit a ML model to generate a smooth trend of SOCS between the observation years, similar to the trend derived from the PO model. It should also be noted that the standard deviation of the ML SOCS predictions in each year is smaller than that of the PO model simulations. This might be because the soil-environment relationship extracted from the ML model mostly reflects spatial patterns, thus narrowing down the extrapolation ability in the time dimension. However, our proposed POML hybrid model generated a temporal trend which is similar to the PO model outputs (Fig. 6c). These results indicate that the hybrid approach not only improved the modelling accuracy according to the validation on observed sample data, but can also adjust to the temporal trend as derived for the PO model.

3.3. The impact of w_p on modelling accuracy

The weighting parameter w_p, which controls the importance of simulated PO data compared with observed sample data in sampling years, has an important impact on the modelling result. Fig. 7 shows the cross-validation accuracies of the POML model with different values of w_p, ranging from 0 to 1 with discrete increases of 0.1 units. The hybrid model ($0 < w_p < 1$) always had a better performance than the pure ML ($w_p = 0$) or PO ($w_p = 1$) model. Fig. 7 also shows that, in this study, the prediction accuracy first increases with increasing values of w_p, reaches an optimum at $w_p = 0.3$, and then decreases as w_p further increases. This suggests that the hybrid model can be further improved from the default value $w_p = 0.5$ that we had used, by assigning lower weights to the PO model simulations compared with that of observations. The reason of this recommendation on w_p is probably related to the relatively larger size of PO derived simulated SOC data compared to the size of observed data, and the validation data are only available in the sampling years.

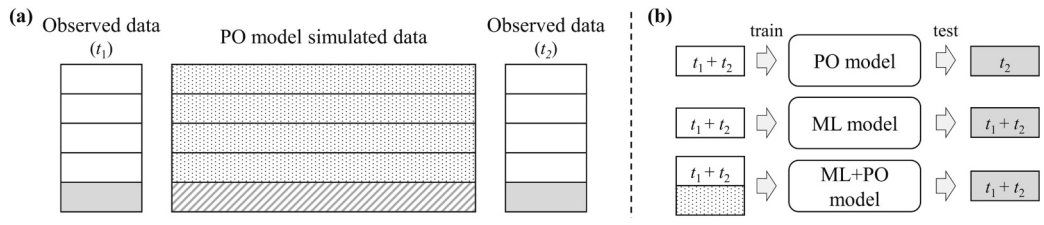


Fig. 4. Illustration of the cross-validation (CV) procedure for evaluating the performance of three models. The figure shows one of five steps of the five-fold CV. The same process is carried out five times by setting a different fold for validation in each step.

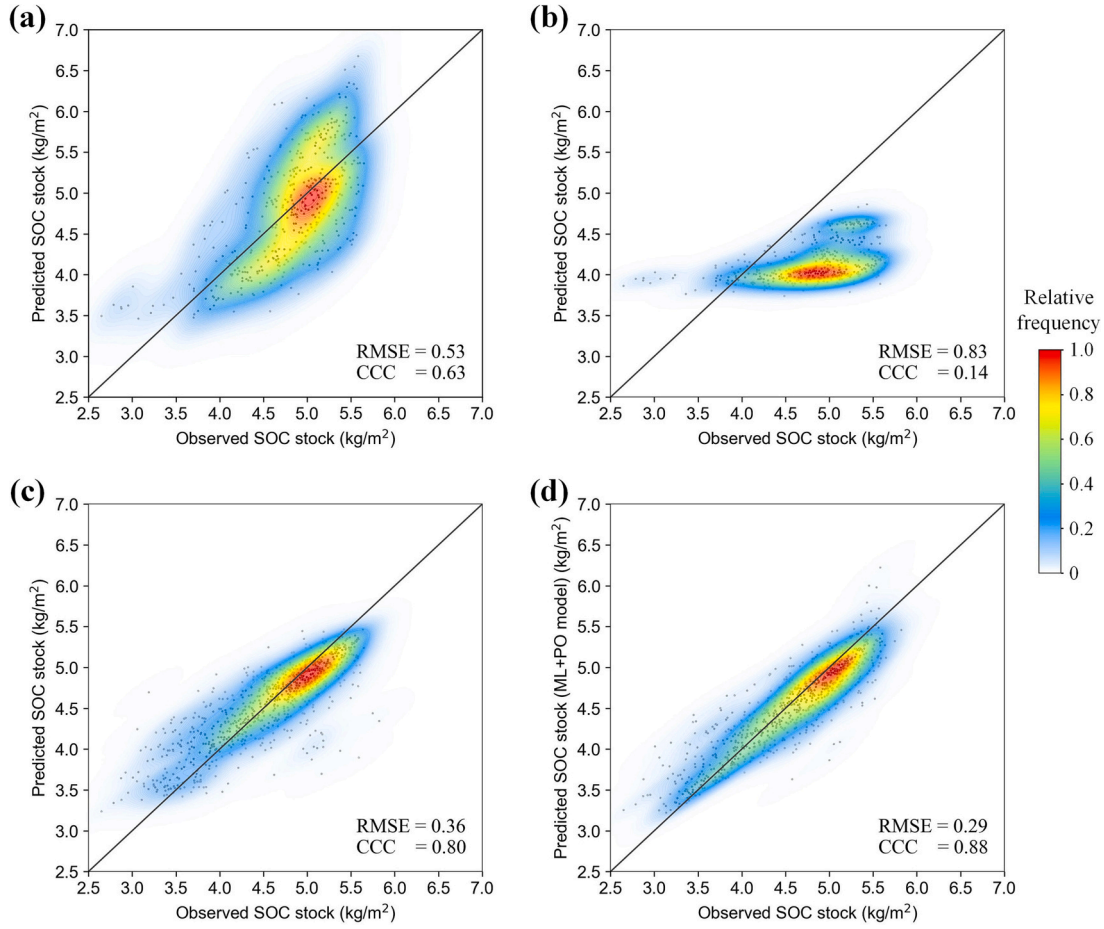


Fig. 5. Plots of prediction accuracies based on different models. (a) Observed against predicted soil organic carbon stocks (SOCS) for all sample data in the year 2000 based on the process-oriented (PO) model (RothC); (b) Observations against predictions for validation sets in the cross-validation (including 1980 and 2000 samples) based on machine learning (ML) model (random forest); (c) Observations against predictions for all sample data in the year 2000 based on the ML model fitted by all sample data in 1980. (d) Simulated SOC stocks between 1980 and 2000 based on PO model versus the corresponding predicted values based on the ML model fitted by sample data in 1980 and 2000. Colors indicate the proportion (relative frequency) of data points. Solid line represents the 1:1 line. RMSE, root mean squared error; CCC, concordance correlation coefficient.

3.4. Mapping results of SOC stocks

The spatial distributions of topsoil SOC stocks obtained with the POML model from 1980 to 2000 are shown in Fig. 8. An animated GIF of annual prediction maps is provided in the Supplementary Materials. The spatial pattern of the prediction maps generally matches the sample-level SOC distribution. Relatively large SOC stocks are found in the eastern two counties (Jiashan and Pinghu counties) and in regions with higher elevation in the western part, near to the Yellow Mountain area. The lower SOC stocks are observed in southern regions in Tongxiang and Haining counties. Comparing the prediction maps based on POML model and ML model (Fig. S3), it can be seen that their spatial patterns of SOC stocks between the two sampling years are different. The POML model derived SOC stock maps show more details with higher spatial variation, while the maps generated by the ML model fail to depict such spatial variation. This is also reflected in the temporal trend comparison shown in Fig. 6, which reveals that the ML model results show a more homogeneous effect between the observation years. This is probably explained by the SOC-environment relationships for the time period between the observation years are hardly learned for the pure ML model that only considers the data in sampling years.

4. Discussion

4.1. Advantages of incorporating process-oriented model into machine learning for space-time modelling

Although ML models have been successfully applied in DSM during past decades, most cases were aimed at solving static DSM tasks that usually predict or map the soil information for a fixed time or assume that soil properties do not vary over time. However, given the increasing magnitude of climate change and the intensified anthropogenic impacts on natural environments (Steffen et al., 2015; IPCC, 2022), spatiotemporal DSM should be one of the critical research topics in the present and near future (Wadoux et al., 2021; Chen et al., 2022; Huang et al., 2022; Pu et al., 2024). As the sample data are usually collected in limited time periods in an area, the density of soil sample data is often denser in space than in time. Therefore, the soil-environment relationship generated by machine learning models might be highly inclined to represent the spatial variation during the sampling time. This implies that a purely data-driven approach, such as ML-based DSM, cannot reliably be extrapolated to other time periods and cannot easily be used for space-time modelling. Our results indicate that the ML model fitted based on observations in one time period cannot be directly applied to make predictions in another time period (Fig. 5b,c), which is mainly because soil-environment relationships can change over time. This implies that space-for-time substitution, which also has been criticized in ecology

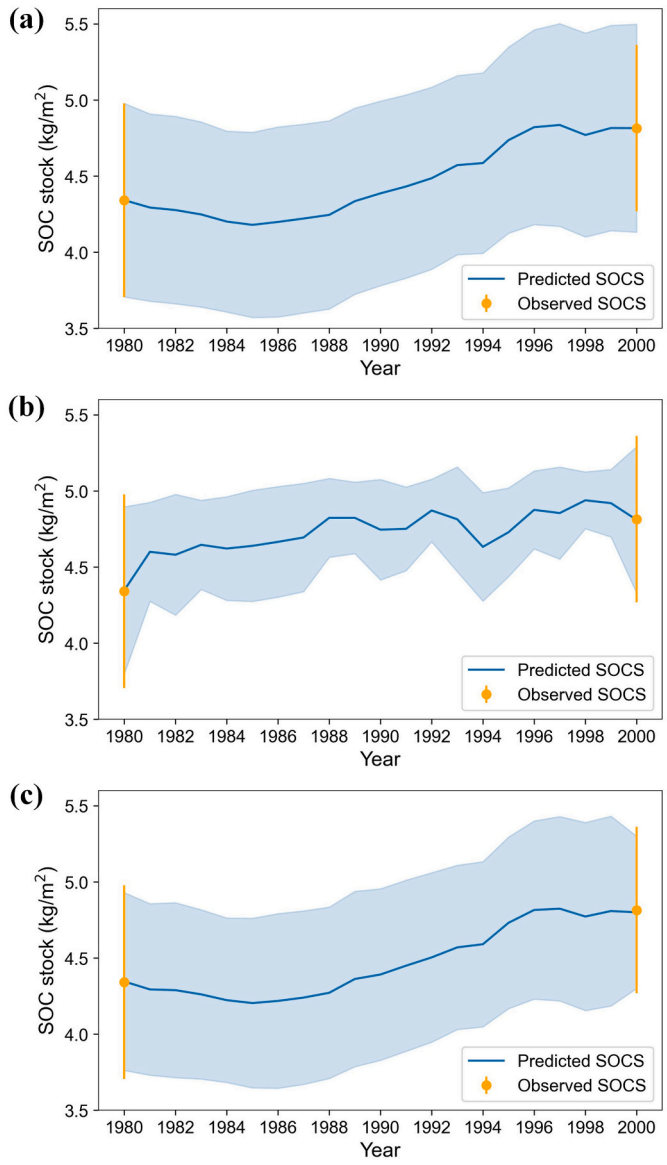


Fig. 6. Temporal trends of SOCS predicted by the process-oriented model (a), machine learning (ML) model (b), and the proposed hybrid method combining PO and ML models (c). The solid blue line is the mean value of SOCS across years; the shaded blue area represents the standard deviation (±1 SD) around the predicted SOCS values at all sampling locations in each year; the yellow point and its error bar represent the mean and SD of observed SOCS in two sampling years.

(Damgaard, 2019), is probably not the appropriate way in solving space-time DSM tasks.

A recent study showed that large prediction uncertainties were obtained when applying a ML model for space-time SOC mapping, indicating that temporal changes of SOC could not realistically be assessed (Heuvelink et al., 2021). Our findings showed that the ML-based modelling result did not agree well with the modelled temporal trend from the PO model. This discrepancy gives rise to the need of combining machine learning with process models. The dynamic patterns of SOCS modelled by the PO model reflect the mechanistic responses of soil carbon to a changing climate and human activities. If the sample data are collected in only few sampling years, such as only two years in our case study, ML model calibration will likely fail to capture the dynamic processes affecting changes in SOCS. The fluctuating trend of SOCS derived by the ML model also reveals that a purely data-driven approach has difficulties capturing the soil variation across time (Fig. 6b). By

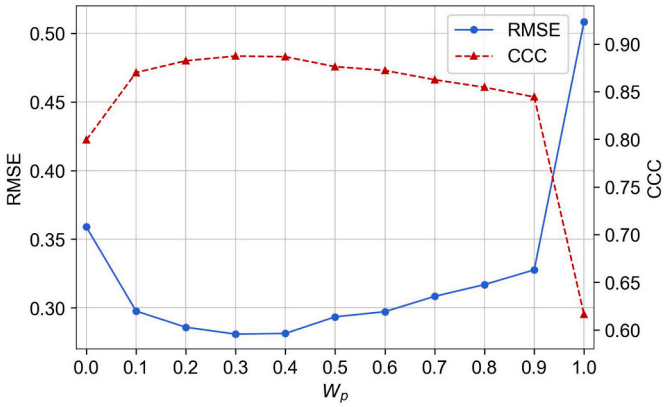


Fig. 7. Cross-validation accuracies (RMSE and CCC) of the hybrid model with different sizes of weighting parameter (w_p) that controls the importance of simulated data derived from PO model compared with observed sample data.

comparison, the temporal variation of SOCS generated from the PO model (Fig. 6a) is more realistic and can be confirmed with evidence from previous studies. The small decreasing trend of carbon stocks during the first years might result from the rising temperature and the decreasing precipitation during those years (Fig. S4), which historically had a negative impact on agriculture (Piao et al., 2010). After that, the increasing trend of SOCS mainly results from the continuous increase in fertilizer input before 2000 (Zhao et al., 2018; Pu et al., 2024), as application of fertilizer increases crop dry matter production and therefore increases carbon inputs to soils (Schlesinger, 1999). As the temporal variation of SOCS predicted by the PO model is more in line with real changes, our proposed hybrid framework used simulated outputs of the PO model to extend the training data for calibration of a machine learning model, which had mainly an effect in the time dimension. Thus, the results of this study confirmed that extending the training data with PO model simulations substantially improved the modelling accuracy, so that it was able to learn the dynamic evolution of SOCS in a changing environment (Figs. 5 and 6).

Some previous studies proposed an alternative approach to integrate the PO into DSM (Xie et al., 2022; Zhang et al., 2023). These studies used a geostatistical or ML model to generate a map of SOCS at the starting year, then used a PO model to simulate yearly SOCS maps by forwarding the model from the starting year. The yearly simulated SOCS maps were taken as additional dynamic covariates for training the geostatistical or ML model. Although the results of these studies showed that this improved the spatiotemporal modelling of SOCS, one problem is that this approach did not generate a final model that reflects the space-time soil-environment relationships, due to the fact that the simulated soil carbon data were used as a covariate rather than as a target variable. By contrast, our proposed hybrid strategy takes the simulation outputs of the PO model and environmental covariates of the corresponding year as augmented training data for the fitting of the ML model. This approach allows the information on the variation of soil carbon and corresponding environmental covariates in the spatial and temporal dimensions to be directly used as input into a ML model, thus enabling the space-time soil-environmental relationships to be embedded in the training process of the final hybrid model.

4.2. Applicability, limitations and future horizons on POML modelling of soil carbon

Soil carbon modelling in space and time needs a model that reflects changes in soil that result from spatial and temporal changes in environmental conditions. Although the results of our study demonstrate the effectiveness of the proposed method, it is importance to it is important to recognize its applicability, limitations and possible improvements for

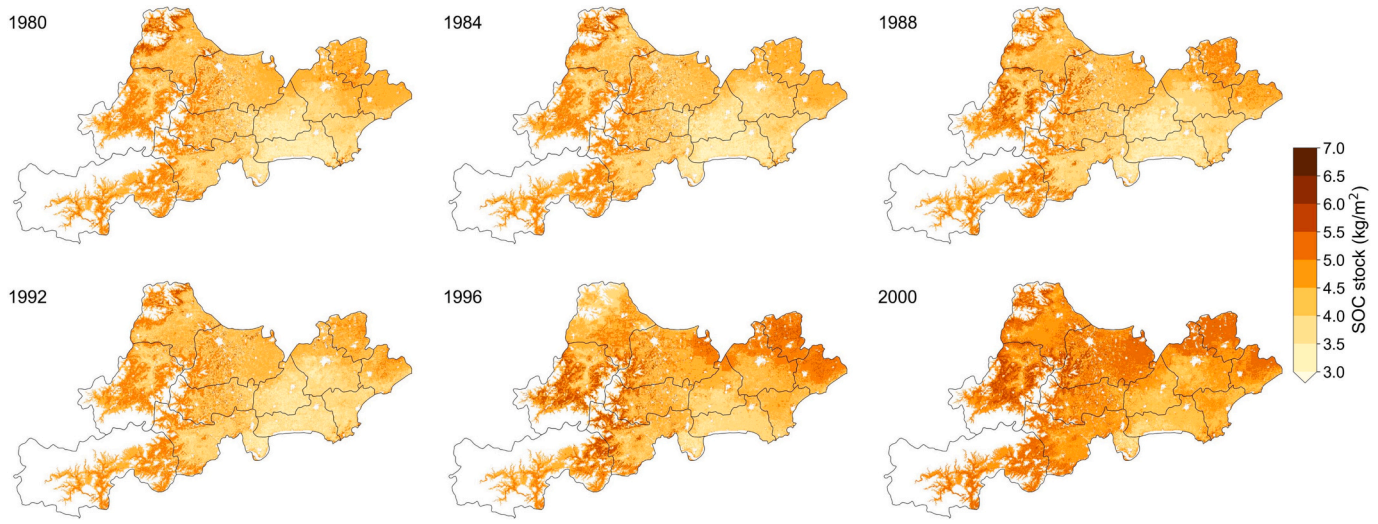


Fig. 8. Prediction maps of SOCS in six selected years from 1980 to 2000 as derived using the hybrid POML model.

future research.

Prior to combining two types of models, soil mappers need to devote careful attention to comprehending variations in SOCS over time within the target area, and how such variation can be well represented in PO models. Given that SOC changes formulized in a PO model are primarily characterized by the turnover rate/time of carbon in soils (Sierra and Müller, 2015), evaluating how to precisely use available datasets to better estimate carbon inputs and outputs is important before integrating models. For calculating the carbon input, our study focused on a cropland region and considered a time period of twenty years before 2000. Given the land management during this period, we decided to use recorded information of fertilizer inputs as an important regulator of carbon inputs. We also verified that the NPP dataset did not correctly reflect the carbon increasing trend in the area, which is inconsistent to the approach taken in previous studies (e.g. Xie et al., 2022; Zhang et al., 2023), but in line with the findings in Yu et al. (2022) and Zhao et al. (2018) and comments in Minasny et al. (2022). For modelling SOC in other areas or different time periods, other appropriate way of adjusting carbon input needs to be carefully considered, since this has a large impact on the PO modelling result. For calculating the carbon output, the decomposition rates for different carbon pools in the soil should be better adjusted to achieve a good fit with the situation in the target area. In our case, the Michaelis-Menten function was added into RothC for a better approximation of the decomposition rate of carbon in response to the biomass increase. In our opinion, this adaptation to the original RothC model is justified for this specific case study and for the years considered. However, more advanced process-based models, such as microbial models (e.g. Luo et al., 2016; Woolf and Lehmann, 2019) and the Millennial model that uses measurable SOC pools (Abramoff et al., 2018, 2022), need more attention to be developed for integration into space-time DSM frameworks.

When integrating PO models into ML, it remains crucial for users to assess whether the assumptions inherent in the chosen PO model align with the conditions in the study area. Especially in croplands, different land management policies would lead to diverse pathways of soil development and carbon dynamics. For instance, the cultivation of crops in submerged or non-waterlogged soils also impacts the sensitivity of carbon flux, and it requires more studies on how to appropriately adjust the models for such different soil conditions. The integrated modelling results could be further improved by fine-tuning the calibration of the PO model, among others using additional detailed agricultural management and land use datasets.

It should be noted that the extended training data from PO model simulations are not substitute for real observations, because a PO model is a simplified representation of reality that does not perfectly describe the complex real-world process of soil dynamics. Thus, quantifying the uncertainty in POML modelling needs to be considered. In this study, the uncertainty in PO simulations between sampling years compared to that in ML predictions in sampling years might be reflected when optimizing the weighting parameter w_p. How to accurately quantify uncertainties in PO and ML models as well as the uncertainty resulting from integrating models is also a worthy future research question. Beyond the RothC and RF models adopted in our case study, the comparison of other models and incorporating multiple PO and ML models can help to improve predictions and quantify the uncertainty in space-time SOC modelling.

In our study we benefited from a sampling design where the same location was sampled again after a period of time. In areas where sampling locations are not revisited, the calibration of a PO model will be much more difficult and will lead to larger uncertainty. For ML models, it is still unclear what accuracy improvement is achieved by having revisited samples and how this influences the space-time modelling. Since soil sampling at different times is often not at the same locations, due to different sampling purposes and land use changes, how to tackle the problem of modelling SOC dynamics with soil observations at different locations across time within the POML framework is a potential research topic.

The space-time soil-environment relationships generated from the hybrid POML model can also be useful for improving the ability to predict soil carbon changes in the future under different climate scenarios or socio-economic pathways. Although we did not have revisited sample data from recent years in the study area to validate future predictions of the POML model, our results suggest that we can be more confident about SOCS predictions for the future generated by the POML model than by the ML model. However, it is recommended that this is rigorously tested, also in other study areas.

5. Conclusions

In this paper, we proposed a hybrid POML model to provide a new approach for space-time SOCS modelling. Since the density of soil sample data is often much sparser in time than in space, it is beneficial to constrain ML models to follow the temporal dynamic pattern of a PO modelling result. The proposed general hybrid framework takes PO model simulated SOCS data at sample points as extended training data for a machine learning model, and adapts the loss function so that the training process can consider both the sample data and the PO model simulations. The results of our case study show that a hybrid approach can effectively combine the advantages of process-oriented soil carbon

models and machine learning models. This approach not only constrains the ML-based modelling to be aligned to the SOC dynamic pattern generated by the PO model, but also makes up for the deficiency of a PO model in spatial prediction. We highlight that integrating PO and ML models is a promising future avenue for dynamic soil mapping. The proposed general framework also allows future studies to incorporate more advanced models or fuse multiple models to achieve more robust soil organic carbon predictions in space and time.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2024.170778.

CRediT authorship contribution statement

Lei Zhang: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Gerard B.M. Heuvelink:** Investigation, Methodology, Resources, Supervision, Writing – review & editing. **Vera L. Mulder:** Investigation, Supervision, Writing – review & editing. **Songchao Chen:** Data curation, Writing – review & editing. **Xunfei Deng:** Data curation, Resources. **Lin Yang:** Funding acquisition, Investigation, Resources, Supervision, Writing – review & editing.

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.

Data availability

The authors do not have permission to share data.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (grant no. 41971054), the Fundamental Research Funds for the Central Universities (0209-14380115), and the Research Funds for the Frontiers Science Center for Critical Earth Material Cycling, Nanjing University. L.Z. acknowledges the support from the Postgraduate Research and Practice Innovation Program of Jiangsu Province (KYCX22_0109) and the financial support provided by the China Scholarship Council (grant no. 202206190058), which supports a one-year research stay at Wageningen University. The authors express sincere gratitude to soil experts Rose Z. Abramoff, Gerard H. Ros and Dominic Woolf, whose valuable comments and suggestions on the soil process modelling have greatly improved the quality of the paper.

References

Abatzoglou, J.T., Dobrowski, S.Z., Parks, S.A., Hegewisch, K.C., 2018. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. Sci. Data 5, 170191. https://doi.org/10.1038/sdata.2017.191.

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.

Abramoff, R.Z., Guenet, B., Zhang, H., Georgiou, K., Xu, X., Viscarra Rossel, R.A., Yuan, W., Ciais, P., 2022. Improved global-scale predictions of soil carbon stocks with Millennial Version 2. Soil Biol. Biochem. 164, 108466 https://doi.org/10.1016/j.soilbio.2021.108466.

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.

Amatulli, G., McInerney, D., Sethi, T., Strobl, P., Domisch, S., 2020. Geomorpho90m, empirical evaluation and accuracy assessment of global high-resolution geomorphometric layers. Sci. Data 7, 162. https://doi.org/10.1038/s41597-020-0479-6.

Angelini, M.E., Heuvelink, G.B.M., Kempen, B., Morrás, H.J.M., 2016. Mapping the soils of an Argentine Pampas region using structural equation modelling. Geoderma 281, 102–118. https://doi.org/10.1016/j.geoderma.2016.06.031.

Bossio, D.A., Cook-Patton, S.C., Ellis, P.W., Fargione, J., Sanderman, J., Smith, P., Wood, S., Zomer, R.J., von Unger, M., Emmer, I.M., Griscom, B.W., 2020. The role of soil carbon in natural climate solutions. Nat. Sustain. 3, 391–398. https://doi.org/10.1038/s41893-020-0491-z.

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/a:1010933404324.

Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., Edwards, T.C., 2015. Machine learning for predicting soil classes in three semi-arid landscapes. Geoderma 239, 68–83. https://doi.org/10.1016/j.geoderma.2014.09.019.

Chen, S., Arrouays, D., Leatitia Mulder, V., Poggio, L., Minasny, B., Roudier, P., Libohova, Z., Lagacherie, P., Shi, Z., Hannam, J., Meersmans, J., Richer-de-Forges, A. C., Walter, C., 2022. Digital mapping of GlobalSoilMap soil properties at a broad scale: a review. Geoderma 409, 115567. https://doi.org/10.1016/j.geoderma.2021.115567.

Coleman, K., Jenkinson, D., 1996. RothC-26.3 - a model for the turnover of carbon in soil. In: Evaluation of Soil Organic Matter Models Using Existing, Long-term Datasets, pp. 237–246. https://doi.org/10.1007/978-3-642-61094-3_17.

Crowther, T.W., Todd-Brown, K.E.O., Rowe, C.W., Wieder, W.R., Carey, J.C., Machmuller, M.B., Snoek, B.L., Fang, S., Zhou, G., Allison, S.D., Blair, J.M., Bridgham, S.D., Burton, A.J., Carrillo, Y., Reich, P.B., Clark, J.S., Classen, A.T., Dijkstra, F.A., Elberling, B., Emmett, B.A., Estiarte, M., Frey, S.D., Guo, J., Harte, J., Jiang, L., Johnson, B.R., Kröel-Dulay, G., Larsen, K.S., Laudon, H., Lavallee, J.M., Luo, Y., Lupascu, M., Ma, L.N., Marhan, S., Michelsen, A., Mohan, J., Niu, S., Pendall, E., Peñuelas, J., Pfeifer-Meister, L., Poll, C., Reinsch, S., Reynolds, L.L., Schmidt, I.K., Sistla, S., Sokol, N.W., Templer, P.H., Treseder, K.K., Welker, J.M., Bradford, M.A., 2016. Quantifying global soil carbon losses in response to warming. Nature 540, 104–108. https://doi.org/10.1038/nature20150.

Damgaard, C., 2019. A critique of the space-for-time substitution practice in community ecology. Trends Ecol. Evol. 34, 416–421. https://doi.org/10.1016/j.tree.2019.01.013.

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.

Friedlingstein, P., Jones, M.W., O'Sullivan, M., Andrew, R.M., Bakker, D.C.E., Hauck, J., Le Quéré, C., Peters, G.P., Peters, W., Pongratz, J., Sitch, S., Canadell, J.G., Ciais, P., Jackson, R.B., Alin, S.R., Anthoni, P., Bates, N.R., Becker, M., Bellouin, N., Bopp, L., Chau, T.T.T., Chevallier, F., Chini, L.P., Cronin, M., Currie, K.I., Decharme, B., Djeutchouang, L.M., Dou, X., Evans, W., Feely, R.A., Feng, L., Gasser, T., Gilfillan, D., Gkritzalis, T., Grassi, G., Gregor, L., Gruber, N., Gürses, Ö., Harris, I., Houghton, R. A., Hurtt, G.C., Iida, Y., Ilyina, T., Luijkx, I.T., Jain, A., Jones, S.D., Kato, E., Kennedy, D., Klein Goldewijk, K., Knauer, J., Korsbakken, J.I., Körtzinger, A., Landschützer, P., Lauvset, S.K., Lefèvre, N., Lienert, S., Liu, J., Marland, G., McGuire, P.C., Melton, J.R., Munro, D.R., Nabel, J.E.M.S., Nakaoka, S.-I., Niwa, Y., Ono, T., Pierrot, D., Poulter, B., Rehder, G., Resplandy, L., Robertson, E., Rödenbeck, C., Rosan, T.M., Schwinger, J., Schwingshackl, C., Séférian, R., Sutton, A.J., Sweeney, C., Tanhua, T., Tans, P.P., Tian, H., Tilbrook, B., Tubiello, F., van der Werf, G.R., Vuichard, N., Wada, C., Wanninkhof, R., Watson, A.J., Willis, D., Wiltshire, A.J., Yuan, W., Yue, C., Yue, X., Zaehle, S., Zeng, J., 2022. Global carbon budget 2021. Earth Syst. Sci. Data 14, 1917–2005. https://doi.org/10.5194/essd-14-1917-2022.

Ge, T., Liu, C., Yuan, H., Zhao, Z., Wu, X., Zhu, Z., Brookes, P., Wu, J., 2015. Tracking the photosynthesized carbon input into soil organic carbon pools in a rice soil fertilized with nitrogen. Plant Soil 392, 17–25. https://doi.org/10.1007/s11104-014-2265-8.

Goovaerts, P., 1999. Geostatistics in soil science: state-of-the-art and perspectives. Geoderma 89, 1–45. https://doi.org/10.1016/s0016-7061(98)00078-0.

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.

Helfenstein, A., Mulder, V.L., Heuvelink, G.B.M., Okx, J.P., 2022. Tier 4 maps of soil pH at 25 m resolution for the Netherlands. Geoderma 410, 115659. https://doi.org/10.1016/j.geoderma.2021.115659.

Hendriks, C.M.J., Stoorvogel, J.J., Álvarez-Martínez, J.M., Claessens, L., Pérez-Silos, I., Barquín, J., 2021. Introducing a mechanistic model in digital soil mapping to predict soil organic matter stocks in the Cantabrian region (Spain). Eur. J. Soil Sci. 72, 704–719. https://doi.org/10.1111/ejss.13011.

Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., Schmidt, M.G., 2016. An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. Geoderma 265, 62–77. https://doi.org/10.1016/j.geoderma.2015.11.014.

Heuvelink, G.B.M., Webster, R., 2001. Modelling soil variation: past, present, and future. Geoderma Dev. Trends Soil Sci. 100, 269–301. https://doi.org/10.1016/S0016-7061(01)00025-8.

Heuvelink, G.B.M., Kros, J., Reinds, G.J., De Vries, W., 2016. Geostatistical prediction and simulation of European soil property maps. Geoderma Reg. 7, 201–215. https://doi.org/10.1016/j.geodrs.2016.04.002.

Heuvelink, G.B.M., Angelini, M.E., Poggio, L., Bai, Z., Batjes, N.H., van den Bosch, R., Bossio, D., Estella, S., Lehmann, J., Olmedo, G.F., Sanderman, J., 2021. Machine learning in space and time for modelling soil organic carbon change. Eur. J. Soil Sci. 72, 1607–1623. https://doi.org/10.1111/ejss.12998.

Huang, H., Yang, L., Zhang, L., Pu, Y., Yang, C., Wu, Q., Cai, Y., Shen, F., Zhou, C., 2022. A review on digital mapping of soil carbon in cropland: progress, challenge, and prospect. Environ. Res. Lett. 17, 123004 https://doi.org/10.1088/1748-9326/aca41e.

IPCC, 2022. Climate Change 2022. Mitigation of Climate Change.

Ivushkin, K., Bartholomeus, H., Bregt, A.K., Pulatov, A., Kempen, B., De Sousa, L., 2019. Global mapping of soil salinity change. Remote Sens. Environ. 231, 111260 https://doi.org/10.1016/j.rse.2019.111260.

Jenkinson, D.S., Hart, P.B.S., Rayner, J.H., Parry, L.C., 1987. Modelling the turnover of organic matter in long-term experiments at Rothamsted. INTECOL Bull. 15, 1–8.

Jenkinson, D.S., Harkness, D.D., Vance, E.D., Adams, D.E., Harrison, A.F., 1992. Calculating net primary production and annual input of organic matter to soil from the amount and radiocarbon content of soil organic matter. Soil Biol. Biochem. 24, 295–308. https://doi.org/10.1016/0038-0717(92)90189-5.

Jiang, G., Shirato, Y., Xu, M., Yagasaki, Y., Huang, Q., Li, Z., Nie, J., Shi, X., 2013. Testing the modified Rothamsted carbon model for paddy soils against the results from long-term experiments in southern China. Soil Sci. Plant Nutr. 59, 16–26. https://doi.org/10.1080/00380768.2012.733923.

Lal, R., 2002. Soil carbon dynamics in cropland and rangeland. Environ. Pollut. 116, 353–362. https://doi.org/10.1016/S0269-7491(01)00211-1.

Lamichhane, S., Kumar, L., Wilson, B., 2019. Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: a review. Geoderma 352, 395–413. https://doi.org/10.1016/j.geoderma.2019.05.031.

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.

Liu, F., Zhang, G.L., Song, X., Li, D., Zhao, Y., Yang, J., Wu, H., Yang, F., 2020. High-resolution and three-dimensional mapping of soil texture of China. Geoderma 361, 114061. https://doi.org/10.1016/j.geoderma.2019.114061.

Liu, F., Wu, H., Zhao, Y., Li, D., Yang, J.-L., Song, X., Shi, Z., Zhu, A.-X., Zhang, G.-L., 2022. Mapping high resolution National Soil Information Grids of China. Sci. Bull. 67, 328–340. https://doi.org/10.1016/j.scib.2021.10.013.

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., Hopkins, F., Jiang, L., Koven, C., Jackson, R.B., Jones, C.D., Lara, M.J., Liang, J., McGuire, A.D., Parton, W., Peng, C., Randerson, J.T., Salazar, A., Sierra, C. A., Smith, M.J., Tian, H., Todd-Brown, K.E.O., Torn, M., van Groenigen, K.J., Wang, Y.P., West, T.O., Wei, Y., Wieder, W.R., Xia, J., Xu, Xia, Xu, Xiaofeng, Zhou, T., 2016. Toward more realistic projections of soil carbon dynamics by earth system models. Glob. Biogeochem. Cycles 30, 40–56. https://doi.org/10.1002/2015GB005239.

Ma, Z., Dong, C., Lin, K., Yan, Y., Luo, J., Jiang, D., Chen, X., 2022. A global 250-m downscaled NDVI product from 1982 to 2018. Remote Sens. 14, 3639. https://doi.org/10.3390/rs14153639.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B., 2003. On digital soil mapping. Geoderma 117. https://doi.org/10.1016/S0016-7061(03)00223-4.

Minasny, B., McBratney, Alex.B., 2016. Digital soil mapping: a brief history and some lessons. In: Geoderma, Soil Mapping, Classification, and Modelling: History and Future Directions, 264, pp. 301–311. https://doi.org/10.1016/j.geoderma.2015.07.017.

Minasny, B., Arrouays, D., Cardinael, R., Chabbi, A., Farrell, M., Henry, B., Koutika, L.-S., Ladha, J.K., McBratney, Alex.B., Padarian, J., Román Dobarco, M., Rumpel, C., Smith, P., Soussana, J.-F., 2022. Current NPP cannot predict future soil organic carbon sequestration potential. Comment on "Photosynthetic limits on carbon sequestration in croplands". Geoderma 424, 115975. https://doi.org/10.1016/j.geoderma.2022.115975.

Nocedal, J., Wright, S.J., 2006. Numerical optimization, 2nd ed. ed. In: Springer Series in Operations Research and Financial Engineering. Springer, New York.

Padarian, J., Minasny, B., McBratney, A., Smith, P., 2022a. Soil carbon sequestration potential in global croplands. PeerJ 10, e13740. https://doi.org/10.7717/peerj.13740.

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

Parton, W.J., Stewart, J.W.B., Cole, C.V., 1988. Dynamics of C, N, P and S in grassland soils: a model. Biogeochemistry 5, 109–131. https://doi.org/10.1007/BF02180320.

Parton, W.J., Del Grosso, S.J., Plante, A.F., Adair, E.C., Lutz, S.M., 2015. Chapter 17 - modeling the dynamics of soil organic matter and nutrient cycling. In: Paul, E.A. (Ed.), Soil Microbiology, Ecology and Biochemistry, Fourth edition. Academic Press, Boston, pp. 505–537. https://doi.org/10.1016/B978-0-12-415955-6.00017-7.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Pérez, F., Granger, B.E., Hunter, J.D., 2011. Python: an ecosystem for scientific computing. Comput. Sci. Eng. 13, 13–21. https://doi.org/10.1109/MCSE.2010.119.

Piao, S., Ciais, P., Huang, Y., Shen, Z., Peng, S., Li, J., Zhou, L., Liu, H., Ma, Y., Ding, Y., Friedlingstein, P., Liu, C., Tan, K., Yu, Y., Zhang, T., Fang, J., 2010. The impacts of climate change on water resources and agriculture in China. Nature 467, 43–51. https://doi.org/10.1038/nature09364.

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.

Rumpel, C., 2019. Soils linked to climate change. Nature 572, 442–443. https://doi.org/10.1038/d41586-019-02450-6.

Schlesinger, W.H., 1999. Carbon sequestration in soils. Science 284, 2095. https://doi.org/10.1126/science.284.5423.2095.

Shanno, D.F., 1970. Conditioning of quasi-Newton methods for function minimization. Math. Comput. 24, 647–656. https://doi.org/10.1090/S0025-5718-1970-0274029-X.

Sierra, C.A., Müller, M., 2015. A general mathematical framework for representing soil organic matter dynamics. Ecol. Monogr. 85, 505–524. https://doi.org/10.1890/15-0361.1.

Smith, J., Smith, P., Wattenbach, M., Zaehle, S., Hiederer, R., Jones, R.J.a., Montanarella, L., Rounsevell, M.D.a., Reginster, I., Ewert, F., 2005. Projected changes in mineral soil carbon of European croplands and grasslands, 1990–2080. Glob. Chang. Biol. 11, 2141–2152. https://doi.org/10.1111/j.1365-2486.2005.001075.x.

Smith, P., Smith, J.U., Franko, U., Kuka, K., Romanenkov, V.A., Shevtsova, L.K., Wattenbach, M., Gottschalk, P., Sirotenko, O.D., Rukhovich, D.I., Koroleva, P.V., Romanenko, I.A., Lisovoi, N.V., 2007. Changes in mineral soil organic carbon stocks in the croplands of European Russia and the Ukraine, 1990–2070; comparison of three models and implications for climate mitigation. Reg. Environ. Chang. 7, 105–119. https://doi.org/10.1007/s10113-007-0028-2.

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. Chang. Biol. 26, 219–241. https://doi.org/10.1111/gcb.14815.

Steffen, W., Richardson, K., Rockström, J., Cornell, S.E., Fetzer, I., Bennett, E.M., Biggs, R., Carpenter, S.R., de Vries, W., de Wit, C.A., Folke, C., Gerten, D., Heinke, J., Mace, G.M., Persson, L.M., Ramanathan, V., Reyers, B., Sörlin, S., 2015. Planetary boundaries: guiding human development on a changing planet. Science 347, 1259855. https://doi.org/10.1126/science.1259855.

Taalab, K., Corstanje, R., Mayr, T.M., Whelan, M.J., Creamer, R.E., 2015. The application of expert knowledge in Bayesian networks to predict soil bulk density at the landscape scale: the application of expert knowledge in Bayesian networks. Eur. J. Soil Sci. 66, 930–941. https://doi.org/10.1111/ejss.12282.

Wadoux, A.M.J.C., 2019. Using deep learning for multivariate mapping of soil with quantified uncertainty. Geoderma 351, 59–70. https://doi.org/10.1016/j.geoderma.2019.05.012.

Wadoux, A.M.J.-C., Minasny, B., Mcbratney, A.B., 2020. Machine learning for digital soil mapping: applications, challenges and suggested solutions. Earth Sci. Rev. 210, 103359 https://doi.org/10.1016/j.earscirev.2020.103359.

Wadoux, A.M.J.-C., Heuvelink, G.B.M., Lark, R.M., Lagacherie, P., Bouma, J., Mulder, V. L., Libohova, Z., Yang, L., McBratney, A.B., 2021. Ten challenges for the future of pedometrics. Geoderma 401, 115155. https://doi.org/10.1016/j.geoderma.2021.115155.

Wang, Q., Le Noë, J., Li, Q., Lan, T., Gao, X., Deng, O., Li, Y., 2023a. Incorporating agricultural practices in digital mapping improves prediction of cropland soil organic carbon content: the case of the Tuojiang River Basin. J. Environ. Manag. 330, 117203 https://doi.org/10.1016/j.jenvman.2022.117203.

Wang, S., Sun, N., Liang, S., Zhang, S., Meersmans, J., Colinet, G., Xu, M., Wu, L., 2023b. SOC sequestration affected by fertilization in rice-based cropping systems over the last four decades. Front. Environ. Sci. 11.

Weihermüller, L., Graf, A., Herbst, M., Vereecken, H., 2013. Simple pedotransfer functions to initialize reactive carbon pools of the RothC model. Eur. J. Soil Sci. 64, 567–575. https://doi.org/10.1111/ejss.12036.

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.

Xiao, W., Xu, S., He, T., 2021. Mapping paddy rice with Sentinel-1/2 and phenology-, object-based algorithm—a implementation in Hangjiahu Plain in China using GEE platform. Remote Sens. 13, 990. https://doi.org/10.3390/rs13050990.

Xie, E., Zhang, X., Lu, F., Peng, Y., Chen, J., Zhao, Y., 2022. Integration of a process-based model into the digital soil mapping improves the space-time soil organic carbon modelling in intensively human-impacted area. Geoderma 409, 115599. https://doi.org/10.1016/j.geoderma.2021.115599.

Yang, L., Li, X., Yang, Q., Zhang, L., Zhang, S., Wu, S., Zhou, C., 2021. Extracting knowledge from legacy maps to delineate eco-geographical regions. Int. J. Geogr. Inf. Sci. 35, 250–272. https://doi.org/10.1080/13658816.2020.1806284.

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, L., Yang, L., Ma, T., Shen, F., Cai, Y., Zhou, C., 2021. A self-training semi-supervised machine learning method for predictive mapping of soil classes with limited sample data. Geoderma 384, 114809. https://doi.org/10.1016/j.geoderma.2020.114809.

Zhang, L., Cai, Y., Huang, H., Li, A., Yang, L., Zhou, C., 2022. A CNN-LSTM model for soil organic carbon content prediction with long time series of MODIS-based phenological variables. Remote Sens. 14, 4441. https://doi.org/10.3390/rs14184441.

Zhang, X., Xie, E., Chen, J., Peng, Y., Yan, G., Zhao, Y., 2023. Modelling the spatiotemporal dynamics of cropland soil organic carbon by integrating process-based models differing in structures with machine learning. J. Soils Sediments. https://doi.org/10.1007/s11368-023-03516-9.

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. U. S. A. 115, 4045–4050. https://doi.org/10.1073/pnas.1700292114.