Geoderma 469 (2026) 117789

Contents lists available at ScienceDirect

Geoderma

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

Monitoring the spatio-temporal changes of topsoil organic carbon content in Mollisol croplands using Landsat time series (2000–2023) and ensemble learning

Yayu Yang [a,b,c,1], Linya Zhao [a,b,c,1], Renjie Ji [a,b,c,1], Huimin Dai [d,e], Xue Wang [a,b,c], Chao Niu [a,b,c], Kun Tan [a,b,c,*]

[a] School of GeoAI and Hindon STAI Institute, East China Normal University, Shanghai 200241, China
[b] Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai 200241, China
[c] Key Laboratory of Spatial-temporal Big Data Analysis and Application of Natural Resources in Megacities, Ministry of Natural Resources, East China Normal University, Shanghai 200241, China
[d] Shenyang Center of China Geological Survey, Shenyang 110034, China
[e] Key Laboratory of Black Soil Evolution and Ecological Effect, Ministry of Natural Resources, Shenyang 110034, China

ARTICLE INFO

Handling Editor: Jingyi Huang

Keywords:
Soil organic carbon (SOC)
Mollisols
Ensemble learning
Spatio-temporal dynamics
Remote sensing

ABSTRACT

Mollisols play a crucial role in global sustainable development due to their high fertility and large carbon stocks. However, the high spatio-temporal resolution dynamic monitoring of soil organic carbon (SOC) content in global Mollisol croplands remains limited, particularly regarding its driving factors and regional variations. In this study, we utilized 35,760 Landsat satellite images from 2000 to 2023 to develop an ensemble learning model (R_c^2 = 0.71, $RMSE_c$ = 4.25 g/kg, R_v^2 = 0.61, $RMSE_v$ = 4.98 g/kg) with 14 features, including spatial location, topography, and spectral indices, to estimate the annual topsoil SOC content dynamics in the global Mollisol croplands. The results showed that the topsoil SOC content averaged 21.30 g/kg, with higher levels in Eurasia than in the Americas. From 2000 to 2023, global topsoil SOC content exhibited a significant fluctuating increase, rising by 3.17% overall, with an average annual percentage change of 0.04 g/kg/year. Considerable regional variation was observed, with a sustained 5.66% increase in the Americas but fluctuating declines in Eurasia, including a 2.21% decrease in Northeast China. These regional disparities reflect the coupled effects of vegetation dynamics, soil–water-atmosphere interactions, and human activities. Further analyses reveal the dual sensitivity of SOC dynamics to agro-environmental controls and socio-economic drivers, including cultivation practices, policy shifts, and socio-political stability. The findings of this study represent a new baseline for precision agricultural management and global soil carbon monitoring.

1. Introduction

Mollisols (USDA Soil Taxonomy system (Liu et al., 2012; Smith, 2014)) are a type of soil with high fertility and strong carbon storage capacity. They store approximately 8.2% of the global soil organic carbon (SOC) stocks (FAO, 2022a), and the quality changes of the soil are not only related to the sustainable utilization of arable land, but also directly affect the global carbon cycle process. As a key indicator of soil quality and ecosystem functioning, SOC content plays an essential role in regulating soil structure, nutrient cycling, and crop productivity. In recent years, more and more studies have reported a significant decrease in SOC content in the global Mollisols, compared to the 20th century, due to a combination of natural changes and anthropogenic disturbances (Meng et al., 2025). The continuous decline of SOC is often accompanied by soil structural degradation, reduced fertility, and diminished carbon sequestration capacity, potentially posing risks to agricultural productivity and regional carbon balance. Pursuant to the 2030 United Nations Sustainable Development Goals (SDGs), the end of hunger (Goal 2) and the preservation of terrestrial ecosystems (Goal 15) are closely dependent on the rational use and conservation of soil resources (Zhang et al., 2021). Therefore, long-term, high-resolution, and high-precision SOC content monitoring of Mollisols facilitates the early

* Corresponding author at: School of GeoAI and Hindon STAI Institute, East China Normal University, Shanghai 200241, China.
E-mail address: tankuncu@gmail.com (K. Tan).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.geoderma.2026.117789
Received 12 August 2025; Received in revised form 19 March 2026; Accepted 20 March 2026
Available online 7 April 2026

identification of land degradation signals, and also provides a scientific basis for agricultural management and global carbon reduction (Wang et al., 2023a).

The formation and distribution of SOC are jointly controlled by multiple environmental factors, including climate conditions, soil physicochemical properties, parent material, topography, and vegetation cover. Accurate characterization of these factors is fundamental for SOC spatial modeling and change analysis. Many studies of the spatial and temporal variations in SOC content at different regional scales and time periods have been conducted in recent years (Heuvelink et al., 2021; Wang et al., 2021; Yang et al., 2023). However, at large spatial scales, conventional field-based surveys are often insufficient to provide continuous, multi-source environmental information. Consequently, satellite remote sensing technology is becoming a vital tool to support large-scale soil surveying and modeling, because of its wide coverage, cost-effectiveness, and sustainability (Abdulraheem et al., 2023; Burke et al., 2021). The spectral information contained in remote sensing imagery can indirectly capture vegetation status, surface cover types, and certain soil properties, providing essential input variables for SOC modeling. At the regional or field scale, remote sensing data with a high spatial and spectral resolution can significantly improve the accuracy and explanatory ability of SOC content estimation models (Jiang et al., 2024b; Meng et al., 2022; Wu et al., 2023a). However, such data are often faced with the problem of limited historical coverage, which makes it difficult to satisfy the demand for research on the long-term trends of SOC content changes. In contrast, medium-resolution remote sensing data that are publicly available and have a wide coverage (Chinilin and Savin, 2023; Yang et al., 2023), such as Moderate-Resolution Imaging Spectroradiometer (MODIS) data, have provided high-frequency observations since 2000 with good temporal continuity. However, the spatial resolution of 250–1000 m makes it difficult to capture the spatial differences in SOC content at the plot scale. Available with a spatial resolution of up to 10 m, Sentinel-2 data have only had a global coverage since 2015, and their temporal dimension is still limited (Cui et al., 2025; Dvorakova et al., 2023; Urbina-Salazar et al., 2023). Among the many sources of remote sensing data, the Landsat series of satellites have continuously provided multispectral observation data with a 30-m resolution since 1984. As such, Landsat data combine the advantages of a high spatial resolution and long time series, and have become an ideal data source to support large-scale, high temporal resolution SOC content estimation (Meng et al., 2024; Meng et al., 2025; Wang et al., 2023b). However, relying solely on appropriate remote sensing data is often insufficient to fully capture the spatial heterogeneity of SOC. To further improve the spatial adaptability and modeling interpretability of SOC content estimation, the choice of feature design and modeling strategy is also crucial.

Mechanistically, SOC content can be expressed as a function of environmental information (Jenny, 1994; McBratney et al., 2003; Meng et al., 2024), with its spatial distribution jointly controlled by climate, soil properties, parent material, vegetation, and topography. Therefore, the introduction of diverse environmental covariates should be theoretically beneficial for SOC content assessment, despite practical challenges in acquiring these data with consistent spatial and temporal coverage. In contrast, spectral indices serve as a practical alternative to accessing environmental information, can diminish the dependence of estimation models on physical parameters while retaining the interpretability of the effects between spectra and soil properties (Bartholomeus et al., 2008). In areas where bare soil images can be acquired, many studies have been conducted to achieve accurate spatial modeling of SOC content by extracting spectral indices reflecting environmental information (Misebo et al., 2024; Yuan et al., 2024) or constructing spectral indices using all the available spectral bands (Jin et al., 2016; Zhang et al., 2024b). In recent years, time-series spectral indices have been introduced to reflect the indirect response of crop growth to soil properties (Meng et al., 2024; Santillano Cázares et al., 2019), but their dependence on high-frequency remote sensing data increases the complexity of the data preprocessing and computation. Consequently, how to extract more stable and adaptive input variables while ensuring the representativeness of the spectral features and spatio-temporal consistency remains a challenge for large-scale SOC content remote sensing estimation. In soil attribute estimation, various modelling strategies have been extensively explored to address characteristics such as the pronounced spatial heterogeneity of SOC and the complex environmental control mechanisms. Local strategies employing rich features and deep learning models effectively isolates complex soil component interactions by enhancing the spectral distinctions through fine-grained classification (Liu et al., 2019; Poggio et al., 2021) and accurately characterizing soil properties at local scales. However, the generalizability of the local strategy is constrained by the model complexity and inherent locality limitations. In contrast, the global strategy typically exhibits a lower modeling accuracy, due to environmental noise interference (Moura-Bueno et al., 2020; Padarian et al., 2022), but has the advantages of fewer parameters, higher efficiency, and consistency, suggesting that the global strategy has the potential to be used for effective soil property assessment. Moreover, agricultural management practices, such as fertilization, crop rotation, and straw return, can significantly influence the spatiotemporal dynamics of SOC by altering soil carbon inputs, including vegetation-derived carbon inputs and organic fertilizer inputs. These processes are often indirectly reflected in remotely sensed vegetation growth and surface characteristics, providing a key entry point for elucidating SOC evolution mechanisms from a remote sensing perspective.

Within the context of environmental drivers and remote sensing–based modeling frameworks, Mollisols, as a representative soil type with high SOC content and strong anthropogenic disturbances, exhibit spatiotemporal dynamics that reflect the combined effects of natural processes and human activities. Globally, Mollisols are primarily distributed across the North American Great Plains, the Eurasian Chernozem belt, and northeastern China. These regions are characterized by intensive agricultural practices and frequent land-use changes, resulting in pronounced temporal dynamics and spatial non-stationarity of SOC. Most previous studies have focused on regional scales or short-term observations, and long-term, continuous estimation of SOC at a global scale remains challenging due to sparse sampling, limited model generalization, and high computational demands (Hengl et al., 2017; Stockmann et al., 2013). Previous studies have employed traditional statistical methods such as multiple linear regression (MLR) and partial least squares regression (PLSR) to estimate SOC in Mollisols (Liu et al., 2019). These approaches offer advantages including high interpretability and straightforward implementation. However, they demonstrate limited capability when addressing complex nonlinear relationships among remote sensing input features (Gomez et al., 2008; Meersmans et al., 2008). In recent years, non-parametric machine learning methods, such as *k*-nearest neighbors (KNN), support vector regression (SVR), random forest (RF), extreme gradient boosting (XGBoost), and categorical boosting (CatBoost), have been increasingly applied to SOC modeling in Mollisols. These methods offer strong capabilities for capturing nonlinear relationships and demonstrate high tolerance to high-dimensional input features, thereby improving prediction accuracy and enhancing model robustness to a certain extent (Li et al., 2025; Mansuy et al., 2014; Zhang et al., 2024a; Zhang et al., 2022). In addition, shallow neural networks, such as the multilayer perceptron (MLP) and radial basis function network (RBFN), further enhance the capacity to capture complex interactions among soil property features through nonlinear activation functions (Chen et al., 2020; Qi et al., 2023). With the increasing dimensionality and temporal coverage of remote sensing data, deep learning methods, such as convolutional neural networks (CNNs), long short-term memory networks (LSTMs), and graph neural networks (GNNs) offer powerful feature extraction capabilities. These methods provide a novel technical pathway for characterizing SOC dynamics under the combined influence of human activities and environmental factors, and have demonstrated

significant advantages in estimating SOC content across regional-scale Mollisols (Meng et al., 2024; Meng et al., 2022; Meng et al., 2025; Tziolas et al., 2024; Wang et al., 2023c; Zhao and Efremova, 2023). However, deep learning models are reliant on large amounts of high-quality samples and require substantial computational resources, and thus face practical challenges in training costs, model transferability, and interpretability. These limitations constrain the widespread application of deep learning models in large-scale research (Ji et al., 2025). Compared to the highly complex deep learning methods, stacking models achieve multi-model fusion with a lower computational cost. By maintaining computational efficiency while retaining a certain level of interpretability, they have the potential to overcome the limitations of individual models in complex Mollisols systems, thereby demonstrating practical value in SOC estimation (Tan et al., 2021; Wu et al., 2023). Although previous studies have explored the use of stacked models for soil property mapping (Biney et al., 2022; Tan et al., 2021), investigations focusing on SOC estimation across global Mollisols based on long-term Landsat time series remain scarce. A key scientific challenge is how to leverage temporally consistent remote sensing data together with robust modeling strategies at large spatial scales to comprehensively capture the combined influences of natural environmental factors and human activities on SOC dynamics in Mollisols.

To address the above challenges, this study focuses on the long-term dynamics of SOC in global Mollisol cropland areas under intensive agricultural management and the multiple factors driving these changes. We developed a SOC content estimation framework incorporating a multi-modeling strategy based on the globally available Landsat multispectral data from 2000 to 2023. To compensate for the lack of temporal resolution in the digital mapping of large-scale SOC content, we generated new year-by-year SOC content distribution maps for the global Mollisol cropland areas from 2000 to 2023 with the spatial resolution of 30 m, to allow for a finer observation of the spatial and temporal variability patterns. Furthermore, the main environmental drivers affecting SOC changes were assessed based on the model input features, and the qualitative analyses of previous studies on the relationship between certain factors and SOC content were validated based on the estimation results. In addition, some typical Mollisol regions were analyzed to characterize the response of human activities to SOC variation, exploring how differences in the intensity of human activities regulate SOC dynamics across multiple spatial scales. These analyses provide a scientific basis for understanding the mechanisms of soil carbon evolution under anthropogenic influence and for informing sustainable land management strategies.

2. Materials and methods

The overall methodological framework of this study is illustrated in Fig. 1. First, multi-source soil sample data within the study area were integrated to construct a SOC sample database covering the topsoil of major global Mollisol croplands. Second, long-term consistent surface reflectance data were derived from Landsat multispectral imagery spanning 2000–2023 under bare-soil conditions, from which spectral features with ecological and soil-indicative significance were extracted. Together with auxiliary environmental variables, including topography and climate, a multidimensional input feature set was constructed.

On this basis, SOC estimation models were developed using feature selection and multi-model ensemble strategies, and model performance was evaluated through cross-validation. Finally, the trained models were applied to annual Landsat composite images to produce year-by-year SOC maps for global Mollisol croplands at a spatial resolution of 30 m, followed by analyses of spatiotemporal variations and their relationships with ecological background and anthropogenic disturbances.

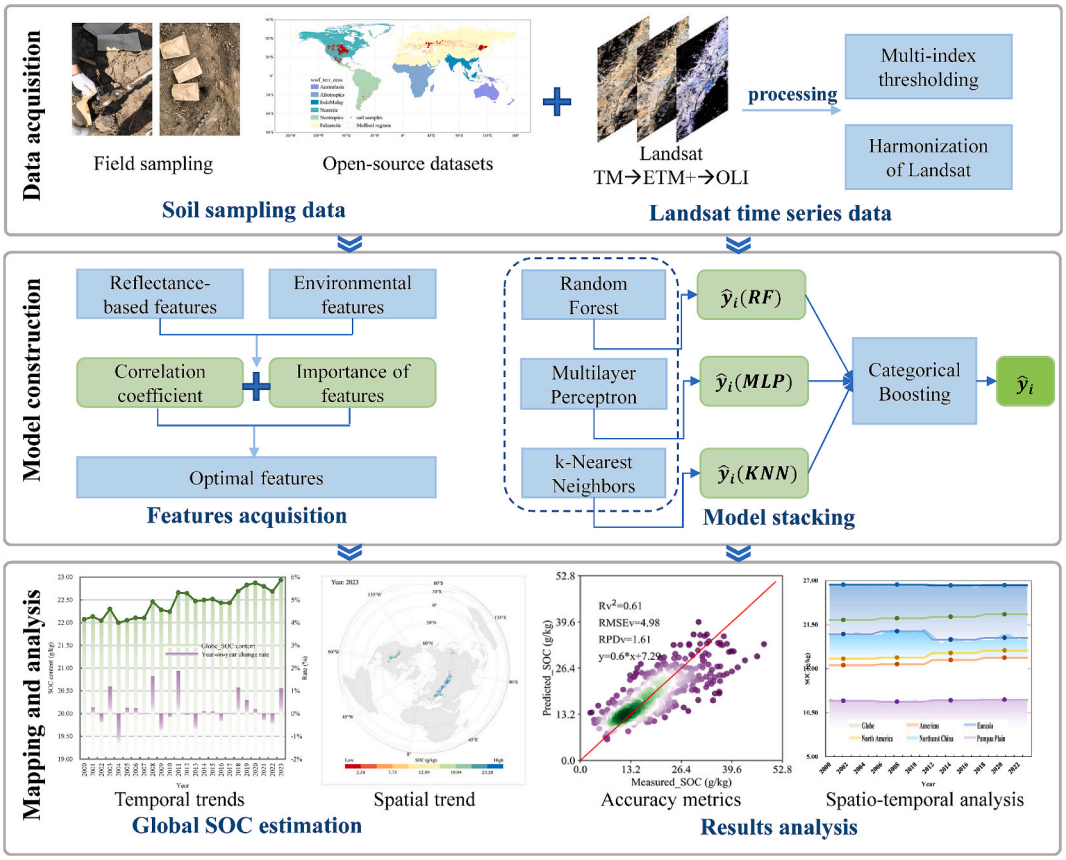

Fig. 1. Overall method flowchart.

2.1. Study areas

Mollisols are mainly distributed in Eastern Europe, Central Asia, and Eastern Asia, as well as in North and South America (Fig. 2), and are widely found in countries such as Russia, the United States, China, and Argentina. Globally, Mollisols cover a total area of about 725 Mha (Yuxin et al., 2024), of which approximately one-third is used for agriculture (Zanaga et al., 2022), forming an important basis for global food production. From a biogeographical perspective, although Mollisols occur across multiple climatic zones, they are mainly concentrated in the Nearctic, Palearctic, and Neotropics (Olson et al., 2001).

Mollisols are characterized by a thick humus-rich surface horizon, relatively low bulk density, well-developed aggregate structure, and high cation exchange capacity. Their topsoil layers are typically rich in organic matter and nutrients, conferring strong carbon sequestration capacity and high productive potential. Mollisols contribute approximately 9% to the estimated global total of 677 Gt of SOC stored within the upper 30 cm of the soil profile (Yuxin et al., 2024), highlighting their important role in the global terrestrial carbon pool.

In this study, the global Mollisol cropland areas were selected as the study area. Spatial constraints were defined using a global Mollisol distribution vector data provided by the National Earth System Science Data Center of the National Science and Technology Infrastructure of China (http://www.geodata.cn), in combination with the GLC_FCS30-1985-2020 land-use data (https://data.casearth.cn/). These datasets were jointly used to delineate the study area and to characterize the spatial distribution and long-term temporal dynamics of SOC content in the topsoil layer (0–30 cm).

There are four typical areas known worldwide for their high-quality crop production (Meng et al., 2024), namely, Northeast China (38°72′–53°56′N, 115°52′–135°09′E), the Russian Plain (43°10′–58°52′N, 22°26′–96°50′E), the Mississippi River Basin of north-central America (20°16′–54°50′N, 86°55′–124°16′W), and the Pampas Plain of South America (18°10′–39°36′S, 50°58′–66°16′W). Despite representing only a small proportion of the global soil, Mollisols provide a significant proportion of the cultivation of oilseeds, cereals, and tuber crops, which are essential to global food security and economic development (International Food Policy Research Institute, 2019).

2.2. Soil sample collection

To construct a representative dataset for model training and validation, soil samples were collected or integrated from multiple sources, covering the major global Mollisols regions. In areas with extensive Mollisols croplands, such as northeastern China and Russia, soil samples were obtained through field surveys or provided by collaborating institutions. In other regions, public datasets were used to supplement the temporal and spatial coverage of the data.

In northeastern China, a total of 639 topsoil samples (0–30 cm) were collected from Mollisol cropland areas during the bare-soil periods in 2005, 2011–2021, and 2023. The bare-soil period typically occurs in early spring (March–May) and late autumn (October–November) (Wang et al., 2022), minimizing vegetation interference (Meng et al., 2024) and ensuring reliable soil spectral signals (Meng et al., 2022). Sampling points were initially arranged based on a gridded layout across the study area and were adjusted on-site as needed to achieve a uniform spatial distribution while avoiding roads, buildings, and other anthropogenic disturbances. Adjacent sampling points were separated by at least 100 m. At each location, soil was collected using a five-point composite method, and information on geographic coordinates, crop residue type, and ground cover was recorded. The SOC content was then measured using the potassium dichromate heating method.

The 92 samples (0–30 cm) from Russia were provided by the V.V. Dokuchaev Soil Science Institute. Humus content was determined using the Tiurin method (FAO, 2021) (here regarded as soil organic matter content), and converted to SOC content by dividing by 1.724 (the Van Bemmelen coefficient).

Additionally, a total of 1,067 topsoil samples were obtained from the World Soil Information Service (WoSIS, December 2023 release, https://data.isric.org/), covering the United States, Canada, Argentina, and Mexico. The database provides quality-controlled and standardized soil property data and is widely used in global-scale soil mapping studies. To ensure compatibility with the study region and Mollisols, only samples with a sampling depth ≤30 cm and consistent soil type were retained. For locations with multiple records, the data were consolidated to derive representative topsoil SOC values. This screening and consolidation procedure ensured spatial representativeness, type consistency, and temporal applicability. It should be noted that although WoSIS contains samples from China and Russia, only those matching the study's spatial extent and soil type were included, whereas the majority were excluded due to mismatches in soil type or sampling period.

2.3. Landsat image acquisition

Since a single Landsat satellite cannot provide complete coverage of the global study area, we used multi-satellite data from the Landsat series for the modeling. For 2011 and earlier, Landsat 5 Thematic Mapper (TM) data were selected, Landsat 7 Enhanced Thematic Mapper Plus (ETM+) data were used for 2012, and Landsat 8 Operational Land Imager (OLI) data were applied for 2013 and after. Regarding Landsat 7

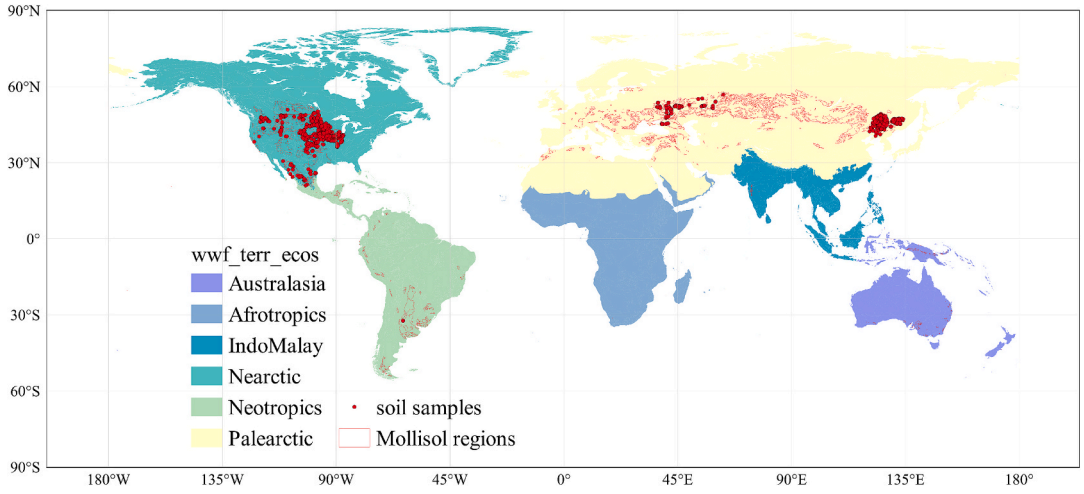

Fig. 2. Map of the global study area.

ETM+ images, the raw data does exhibit striping problem. However, we retained the original data to ensure consistency across years and to avoid additional uncertainties introduced by gap-filling procedures.

In this study, six bands (Table 1) of the Landsat image data were downloaded from the United States Geological Survey (USGS) Earth Explorer website (https://earthexplorer.usgs.gov/) for the bare soil periods from 1985 to 2023, which were all obtained from the Landsat Collection 2 Level-2 Quality Assessment (QA) band. Moreover, these data are surface reflectance data that have been radiometrically calibrated and atmospherically corrected, and are an important and reliable data source for qualitative or quantitative studies of surface ecological and environmental elements.

According to the orbital coverage of the Worldwide Reference System (WRS-2), approximately 1,490 scenes are required each year to achieve complete coverage of the study area. Therefore, suitable images for SOC estimation must be selected from a large archive of satellite data. To obtain optimal bare-soil observations for annual SOC estimation, we adopted a pixel-based best bare-soil observation selection approach rather than using annual or seasonal image composites. Specifically, based on the WRS-2 orbital coverage, all available Landsat images covering the study area for a given year were first collected to construct an initial annual image pool. Subsequently, for each pixel, all images acquired in that year were examined, and a combination of multiple spectral indices (Eqs. (1)–(4)) was applied to identify valid observations. These criteria ensured that the pixel reflectance was free from contamination by clouds, snow/ice, built-up areas, and vegetation, thereby retaining spectral information representative of "pure bare-soil" conditions as much as possible. On this basis, the single observation corresponding to the minimum value of Eq. (3) was selected as the effective annual reflectance for that pixel, in order to minimize the influence of residual vegetation and atmospheric effects. Finally, the best observations from all pixels were assembled to generate the annual bare-soil image for that year. The threshold values used in the above equations were determined through a multi-threshold segmentation analysis applied to 100,000 randomly selected cropland pixels uniformly distributed across the study area, and subsequently validated using 100 Landsat scenes from Northeast China and the central United States. A total of 96,498 images were browsed, and 35,760 Landsat TM/OLI images with six bands were selected for mapping of the year-by-year SOC content for 2000–2023. Reflectance spectra of the soil sample points were also acquired under the same rules.

The range of available thresholds for the normalized difference snow index (Riggs et al., 1994) (NDSI) was used to exclude ice and snow pixels, as follows:

$$NDSI = \frac{G - SWIR1}{G + SWIR1} \in (-\infty, 0) \quad (1)$$

The cloud index (Zhai et al., 2018) (CI) enabled us to remove cloud-covered pixels as much as possible, with the threshold expression as follows:

$$CI = \frac{B + G + R + NIR + SWIR1 + SWIR2}{6} \in (0, 0.20) \quad (2)$$

The threshold expression for the normalized difference vegetation index (Rouse et al., 1974) (NDVI) used in order to exclude vegetation pixels during the growing period can be written as follows:

$$NDVI = \frac{NIR - R}{NIR + R} \in (0, 0.25) \quad (3)$$

In addition, building pixels are also prone to be incorrectly extracted as bare soil pixels and needed to be removed, which was achieved using the normalized difference building index (Zha et al., 2003) (NDBI) threshold expression:

$$NDBI = \frac{SWIR1 - NIR}{SWIR1 + NIR} \in (0.10, +\infty) \quad (4)$$

Table 1
The specific band information for the Landsat series of satellites employed in this study.

Sensor type	Band name	Description	Wavelength (μm)
Landsat TM/ ETM+	B	Band 1 (blue) surface reflectance	0.45–0.52
	G	Band 2 (green) surface reflectance	0.52–0.60
	R	Band 3 (red) surface reflectance	0.63–0.69
	NIR	Band 4 (near infrared) surface reflectance	0.77–0.90
	SWIR1	Band 5 (shortwave infrared 1) surface reflectance	1.55–1.75
	SWIR2	Band 7 (shortwave infrared 2) surface reflectance	2.08–2.35
Landsat OLI	B	Band 2 (blue) surface reflectance	0.45–0.51
	G	Band 3 (green) surface reflectance	0.53–0.59
	R	Band 4 (red) surface reflectance	0.64–0.67
	NIR	Band 5 (near infrared) surface reflectance	0.85–0.88
	SWIR1	Band 6 (shortwave infrared 1) surface reflectance	1.57–1.65
	SWIR2	Band 7 (shortwave infrared 2) surface reflectance	2.11–2.29

2.4. Harmonization of landsat

The acquisition and matching of images has an important impact on the analysis of feature characteristics in time-series quantitative remote sensing studies, mainly due to the fact that different sensors have different conditions of their own, and it is complicated to obtain consistent temporal characteristics and spectral effects in different periods. This can result in bias and non-optimal results in the long time-series analysis of soil properties based on multi-sensor images (Mishra et al., 2016). To address this issue, the Landsat ETM+ sensor was used to determine the cross-calibration factor for the three Landsat sensors to calibrate both the TM and the ETM+ data to the OLI level. Since the acquired images were concentrated in March to May and October to November, single-month monthly mean images from the different sensors in 2011, 2012, and 2013 were used for the cross-calibration. The calibration coefficients were ultimately the most consistent over the image spectra at random points in May, and May was the preferred time period for mapping the large areas of Mollisols (Luo et al., 2022), which allowed us to obtain a large number of images with wide coverage.

2.5. Environmental data acquisition

To characterize the climatic and topographic context relevant to SOC modeling, a number of environmental features were used as explanatory variables (McBratney et al., 2003; Meng et al., 2024). We used the digital elevation model (DEM) with a 30-m spatial resolution provided by the USGS-led Shuttle Radar Topography Mission (SRTM) to obtain elevation values, as well as slope and aspect calculated from the elevation, as auxiliary topographic data. The global annual average surface temperature (GAAST) and precipitation rate (GAAPR) for each year were obtained from the National Oceanic and Atmospheric Administration: Climate Forecast System (NOAA CFS) NOAA/CFSV2/FOR6H dataset. The above datasets were accessed through the Geospatial Processing Service on the Google Earth Engine platform, and all the data were analyzed at the raw resolution.

2.6. Feature construction and selection

In this study, a feature library (Table S1) of three clusters of spectra, spectral indices, and auxiliary data was established, covering a variety of factors, such as vegetation, soil, and topography. The light gradient

boosting machine (LGBM) model (Ke et al., 2017) was chosen to implement the feature extraction, due to the large number of features, the presence of covariance, and the tendency for there to be missing values. The feature importance provided by the LGBM model when each type of feature individually predicted the SOC content was used to extract the final input features. The top two to top ten most important features of each category were selected to be combined into the set of input features for the different cases. In addition, phenomena such as jumps in the values of features in contiguous geographic space or complex operations leading to image noise can affect the mapping of soil organic matter content. This is more restrictive for the spectral index features, as such, and after selecting the input features based on the LGBM model, 100 images were randomly selected within the global Mollisol cropland areas to perform the inversion testing, and features that would lead to spatial inconsistencies in the images were eliminated. We retained as few and as general and effective features as possible, to facilitate the global regression mapping. A total of 14 features were finally selected, including longitude, latitude, elevation, GAAST, GAAPR, and spectral-based calculations (descriptive statistics for the features are shown in Table S2).

The feature formulas used to represent the spectral differences were as follows:

$$B2\_B3\_DI = R - G \tag{5}$$

$$B4\_B5\_DI = SWIR1 - NIR \tag{6}$$

$$B5\_B6\_DI = SWIR2 - SWIR1 \tag{7}$$

The formulas related to soil properties and parent material (Drury, 1987; Hunt Jr and Rock, 1989; Smith et al., 2005) can be written as:

$$Char\ soil\ index(CSI) = NIR/SWIR2 \tag{8}$$

$$Moisture\ stress\ index(MSI) = SWIR1/NIR \tag{9}$$

$$Iron\ oxide\ ratio\ index(IORI) = R/B \tag{10}$$

where the ferrous minerals index (FMI) (Segal, 1982) has the same formula with MSI.

Vegetation indices (Gitelson et al., 2002; Pinty and Verstraete, 1992; Scudiero et al., 2015) were also utilized:

$$Visible\ atmospherically\ resistant\ index(VARI) = \frac{G - R}{G + R - B} \tag{11}$$

$$Canopy\ response\ salinity\ index(CRSI) = \sqrt{\frac{NIR \times R - G \times B}{NIR \times R + G \times B}} \tag{12}$$

$$\begin{aligned} &Global\ environmental\ monitoring\ index(GEMI) \\ &= eta \times (1 - 0.25 \times eta) - \frac{R - 0.125}{1 - R},\ eta \\ &= \frac{2 \times (NIR^2 - R^2) + 1.5 \times NIR + 0.5 \times R}{NIR + R + 0.5} \end{aligned} \tag{13}$$

2.7. Prediction models

In this study, to identify the optimal combination strategy among the different model structures, we built ensemble model frameworks based on the typical machine learning models, and conducted extensive experiments to evaluate the effects of different model types, parameter configurations, and ensemble learning structures on the performance of SOC content estimation. Finally, we constructed a global SOC ensemble (GloSOC-Ensemble) model, which adopted a typical two-layer stacked ensemble structure, in which the first layer contained three base learners, namely, RF, MLP, and KNN. The second layer employed a CatBoost model as the meta learner to further optimize the final SOC content estimation based on the predictions of the first-layer model.

RF is an extension of the bagging (bootstrap aggregating) ensemble method, featuring strong resistance to overfitting and the capability to assess variable importance (Breiman, 2001). Complementarily, MLP represents a classic form of feed-forward neural network that leverages nonlinear activation functions to capture complex feature interactions (Rumelhart et al., 1986). In addition, KNN, as a distance-based learning algorithm, offers good interpretability and implementation simplicity by directly using labeled training instances for prediction (Cover and Hart, 1967). The base learners covered three different types of modeling paradigms, which can provide more adaptive learning capabilities under the conditions of differences in data distribution, feature dimensions, and sample density. As an efficient gradient boosting algorithm, CatBoost has the advantage of dealing with unbalanced data and high-dimensional features, and it is especially good at improving model generalization and mitigating overfitting (Prokhorenkova et al., 2018).

Therefore, for sample i, with a set of features $x_i \in \mathbb{R}^d$, the training dataset for the first layer can be denoted as:

$$\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{n}, x_i \in \mathbb{R}^d \tag{14}$$

where x_i is the set of features, y_i denotes the true SOC content of sample i, and n is the number of samples.

After input to the first-layer model, the predicted values were generated as follows, respectively:

$$\hat{y}_i^{(RF)} = f_{RF}(x_i) = \frac{1}{T}\sum_{t=1}^{T} f_t(x_i) \tag{15}$$

$$\hat{y}_i^{(MLP)} = f_{MLP}(x_i) = W_L\sigma(W_{L-1}\sigma(\cdots\sigma(W_1x_i + b_1)\cdots) + b_{L-1}) + b_L \tag{16}$$

$$\hat{y}_i^{(KNN)} = f_{KNN}(x_i) = \sum_{j \in N_k(x_i)} w_j y_j \tag{17}$$

$$w_j = \frac{1/d(x_i, x_j)}{\sum_{L \in N_k(x_i)} 1/d(x_i, x_l)}$$

where T is the total number of trees in the RF model, $f_t(*)$ is the prediction function for the regression tree t, and $\hat{y}_i^{(RF)}$ is the predicted value of SOC based on RF. L denotes the number of layers of the MLP network (without input layers); W_L and b_L are the weight matrix and bias vector of layer L, respectively; $\sigma(*)$ is the activation function; and $\hat{y}_i^{(MLP)}$ is the value of the SOC content predicted based on MLP. $N_k(x_i)$ denotes the set of k training samples that are closest to sample x_i, y_j denotes the true SOC values of the j neighboring points, $d(*,*)$ is the Manhattan distance, and the weight w_j is inversely proportional to the distance.

After five-fold cross-validation of the samples for each model in the first layer, a predicted set of stacked values was obtained:

$$S_i = \left[f_{RF}^{CV}(x_i), f_{MLP}^{CV}(x_i), f_{KNN}^{CV}(x_i)\right] \tag{18}$$

the training data for the input *meta*-learning model were then:

$$\mathscr{D}' = \{(S_i, y_i)\}_{i=1}^{n} \tag{19}$$

Finally, the new training set was learned to obtain the final predictions:

$$\hat{y}_i = g_{CatBoost}(S_i) \tag{20}$$

To further improve the model performance and avoid overfitting, all the sub-models were hyper-parameter optimized through the Optuna (3.3.0) framework (Prokhorenkova et al., 2018), which applies a Bayesian optimization strategy to automatically search for optimal hyperparameter combinations. The objective function was defined as the minimization of the mean squared error (MSE) on the validation set, ensuring that each model was optimized under a fair and comparable setting. Model construction and training were conducted in the Python

environment, primarily relying on libraries such as Vecstack (0.4.0) (Shcherbatyy, 2016), scikit-learn (1.0.2), and CatBoost (1.2).

2.8. Feature importance assessment

For each base learner in the first layer, the feature importance was calculated individually. The RF model utilized its built-in feature importance metric, while MLP and KNN utilized the SHapley Additive exPlanations (SHAP) algorithm to compute the mean absolute SHAP value as the feature importance (Lundberg and Lee, 2017). Subsequently, the second-layer CatBoost model evaluated the model contribution weights by a built-in algorithm. The feature importance of each base learner was scaled to the range [0,1], and the final feature importance values were calculated as follows:

$$Final\_Imp_j = \sum_{i=1}^{M} \left(w_i \times Imp_{i,j} \right) \tag{21}$$

where w_i is the weight for base learner i, $Imp_{i,j}$ denotes the normalized importance of feature j from learner i, and M is the number of base learners.

2.9. Estimation model and prediction accuracy

With the soil sampling points arranged in ascending order of true SOC content values, they were divided into two groups by a stratified sampling method, where 67% of the samples were used for the model calibration (N = 1204) and 33% of the samples were used for the validation set (N = 594). Five-fold cross validation was used to assess the model performance.

The coefficient of determination (R^2), root-mean-square error (RMSE), and residual prediction deviation (RPD) were also calculated to evaluate the model performance. In addition, the 1:1 line was used to measure how far the true SOC content values deviated from the predicted SOC content values. A detailed description of the equations is provided as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \tag{22}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \tag{23}$$

$$RPD = \frac{\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2 / n}}{RMSE} \tag{24}$$

where n is the number of samples, y_i is the true SOC content of sample i, $\hat{y}_i$ is the predicted SOC content of sample i, and $\bar{y}_i$ is the mean value of y_i. In general, a well-performing model usually exhibits high R^2 and low RMSE values (Williams, 1987). Furthermore, the RPD values were divided into five levels to interpret the model performance (Yuan et al., 2019): RPD < 1.4 (unacceptable), 1.4 ≤ RPD < 1.8 (fair), 1.8 ≤ RPD < 2.0 (good), 2.0 ≤ RPD < 2.5 (very good), and RPD ≥ 2.5 (excellent).

To quantify the overall uncertainty of the ensemble model, we adopted a five-fold cross-validation approach and calculated the SD of the RMSE across folds for the CatBoost model. The SD of the MSE was used to approximate the 95% confidence interval (CI) for the model performance using the following formula (Montgomery and Runger, 2019; Walpole et al., 1993):

$$CI_{95\%} = \mu + 1.96 \cdot \sigma \tag{25}$$

where μ denotes the mean RMSE, and σ the corresponding SD. This approach provides an interval-based estimation of model performance and serves as an initial approximation of predictive uncertainty.

To analyze the spatial and temporal volatility of the estimation results, we calculated the weighted SD using the mean, SD, and pixel count of the SOC values across the years and spatial blocks. For each year or spatial block, the weighted overall SD was calculated as follows:

$$\sigma_{total} = \sqrt{\frac{1}{N} \sum_{i=1}^{n} \left[\sigma_i^2 + (\mu_i - \bar{\mu})^2 \right] \cdot N_i} \tag{26}$$

where n is the number of sub-blocks or sub-years involved in the merger, μ_i is the mean value of component i, σ_i is the SD of component i, N_i denotes the number of pixels in component i, N is the total number of pixels, and $\bar{\mu}$ is the weighted average value.

In addition to the uncertainty associated with model performance stability and spatiotemporal variability, as reflected by fluctuations in cross-validation performance (Eq. (25)) and spatiotemporally weighted statistical features (Eq. (26)), this study further quantified the uncertainty of SOC spatiotemporal predictions using a bootstrap-based empirical prediction interval (PI) approach to characterize the predictive distribution of the test samples. This method does not rely on any specific distributional assumptions and has been widely applied in uncertainty assessment for digital soil mapping (Schmidinger and Heuvelink, 2023; Wadoux, 2019).

Specifically, the training samples were repeatedly resampled with replacement, and the model was retrained for each bootstrap replicate to generate an ensemble of predictions for the test samples across different model realizations. Based on the resulting predictive distribution, prediction intervals were constructed at a given confidence level (1−α) using quantiles, with the lower and upper bounds defined as the α/2 and (1−α/2) quantiles of the predictive distribution, respectively.

The prediction interval width (PIW) for an individual sample is defined as:

$$PIW_i = \hat{y}_i^{(1-\alpha/2)} - \hat{y}_i^{(\alpha/2)} \tag{27}$$

where $\hat{y}_i^{(1-\alpha/2)}$ and $\hat{y}_i^{(\alpha/2)}$ denote the upper and lower quantiles of the predictive distribution for the ith sample, respectively. In this study, a 90% empirical prediction interval was adopted (α = 0.10). Furthermore, the mean prediction interval width (MPIW) was calculated by averaging PIW across all samples:

$$MPIW = \frac{1}{N} \sum_{i=1}^{N} PIW_i \tag{28}$$

where N is the number of test samples. MPIW was used to characterize the overall level of model prediction uncertainty, while the distribution of PIW was employed to analyze the spatial heterogeneity of prediction uncertainty among individual samples.

2.10. Result normalization

Since Landsat images exhibit variations in effective coverage across different periods and regions, conducting direct temporal sequence statistics for the whole region may introduce spurious variations due to missing observations. To ensure spatial comparability among SOC estimates from different years, a spatial coverage consistency procedure was applied to the annual SOC prediction maps prior to the statistical analyses (the specific process is shown in Fig. S1). According to the WRS2 grid, there were 181 paths (Fig. S2(a)) and 80 rows (Fig. S2(b)) in the raw image data for every year. Considering the spatial continuity and aggregation of Landsat path numbers, we grouped adjacent orbital tracks by extracting the prefix of the WRSPR attribute (i.e., the higher-order digits of the path index), thereby simplifying the subsequent block-wise statistics and spatial consistency analysis. Combined with the GLC_FCS30-1985–2020 cropland mask (mosaicked into five regions), the annual images were ultimately divided into 28 spatial blocks (Fig. S2 (c)).

For each year, the number of valid SOC pixels within 28 spatial blocks and their proportion relative to the total number of valid pixels

were calculated, and coverage time series for each block were constructed for the period 2000–2023. The temporal stability of spatial coverage was evaluated using the coefficient of variation (CV), range, and relative range (RR) of the coverage ratio (Table S3). Spatial blocks exhibiting pronounced coverage fluctuations were identified as potential anomalous regions, and their coverage time series were corrected using linear interpolation (Shen et al., 2015) to mitigate the influence of interannual differences in spatial coverage. After this procedure, the effective spatial coverage used for statistical analyses remained stable at 42.5% ± 2.5% across years (Fig. S3). This essentially compensated for the pixel-coverage of the Mollisols in Eurasia for the years 2000–2005/2013 during the statistical analyses, and did not modify the SOC content estimation images in a practical sense.

3. Results

3.1. Soil sample characteristics

A total of 1798 sets of topsoil sample (Fig. 3) data with SOC contents ranging from 1.90 to 62.50 g/kg from Northeast China, Russia, the United States of America, Canada, Argentina, and Mexico were employed, with 827 samples from 1985 to 1999 and 971 samples from 2000 to 2023 (Fig. S4). The mean, standard deviation (SD), and coefficient of variation of the SOC content were 16.51 g/kg, 6.90 g/kg, and 41.79% for 1985–1999 and 20.15 g/kg, 8.35 g/kg, and 41.44% for 2000–2023, respectively.

3.2. Results of Landsat data harmonization

The results of Landsat data harmonization are listed in Table 2. All data were ultimately converted to OLI sensor reflectance values, and model training and validation were conducted using the converted data. The same reflectance transformation was applied to the images during image mapping.

3.3. Mapping and long-term trends of topsoil SOC content in the global Mollisol cropland areas

Annual maps of the topsoil SOC content in the global Mollisol cropland areas were produced using the GloSOC-Ensemble model (five-fold cross-validation, $R_c^2 = 0.71$, $RMSE_c = 4.25$ g/kg, $R_v^2 = 0.61$, $RMSE_v = 4.98$ g/kg, Fig. 4). Year-by-year results are provided in Fig. S5. The multi-year average map based on the valid data (see Section 2.10) is shown in Fig. 5(a). Since the beginning of the 21st century, the Mollisol cropland areas have largely maintained high soil fertility, with most areas exhibiting SOC contents above 18.04 g/kg and an average SOC content of 21.30 g/kg. In addition, the 45°N latitudinal zone passes through the world's three major grain-producing regions (Meng et al., 2024), i.e., the Russian Plain, Northeast China, and the Mississippi River Basin of north-central America, with regional mean SOC contents of 29.02 g/kg, 20.13 g/kg, and 17.72 g/kg, respectively. With reference to the terrestrial ecoregions of the world data (Olson et al., 2001), the three regions with the largest area and highest mean SOC content are the Palearctic, Nearctic, and Neotropical regions, which suggests that the relatively humid climate of the temperate zones provides favorable conditions for the development of Mollisols, while higher latitudes with lower temperatures allow for prolonged accumulation of humus.

The analysis of the interannual trend (Fig. 5(b)) shows that the topsoil SOC content in the global study area increased significantly between 2000 and 2023 (Mann-Kendall test, $p < 0.01$; Theil-Sen slope = 0.037 g/kg), which indicates that it was slowly accumulating overall. The average annual percentage change shows that the fluctuation of SOC content was generally less than ±2% from year to year, and the global average annual change rate was 0.15%, which further supports the stability and moderation of the growth trend. Overall, spatial and temporal continuity was maintained across the different spatio-temporal blocks, although variations remained among continental regions. SOC content in Eurasia was generally higher than in North America, and both were significantly higher than in South America. Dividing the period into four intervals of six years each (Fig. 5(c)), the SOC content increased

Table 2
The cross-calibration results for the TM, ETM+, and OLI data. ρ and RMSE represent the correlation coefficient and the root-mean-square error, respectively.

Band	Regression model between ETM+ and TM	ρ	RMSE	Regression model between OLI and ETM+	ρ	RMSE
B	ETM+ = 0.7092*TM + 0.0128	0.69	0.02	OLI = 0.6168*ETM+ + 0.0121	0.72	0.02
G	ETM+ = 0.7157*TM + 0.0162	0.71	0.02	OLI = 0.644*ETM+ + 0.0251	0.72	0.02
R	ETM+ = 0.8536*TM + 0.0071	0.81	0.02	OLI = 0.7871*ETM+ + 0.0114	0.84	0.02
NIR	ETM+ = 0.9009*TM + 0.0295	0.88	0.05	OLI = 0.855 *ETM+ + 0.0463	0.87	0.05
SWIR1	ETM+ = 0.8275*TM + 0.0347	0.81	0.04	OLI = 0.7767 *ETM+ + 0.0471	0.77	0.04
SWIR2	ETM+ = 0.8558*TM + 0.0207	0.83	0.04	OLI = 0.8091 *ETM+ + 0.0287	0.82	0.04

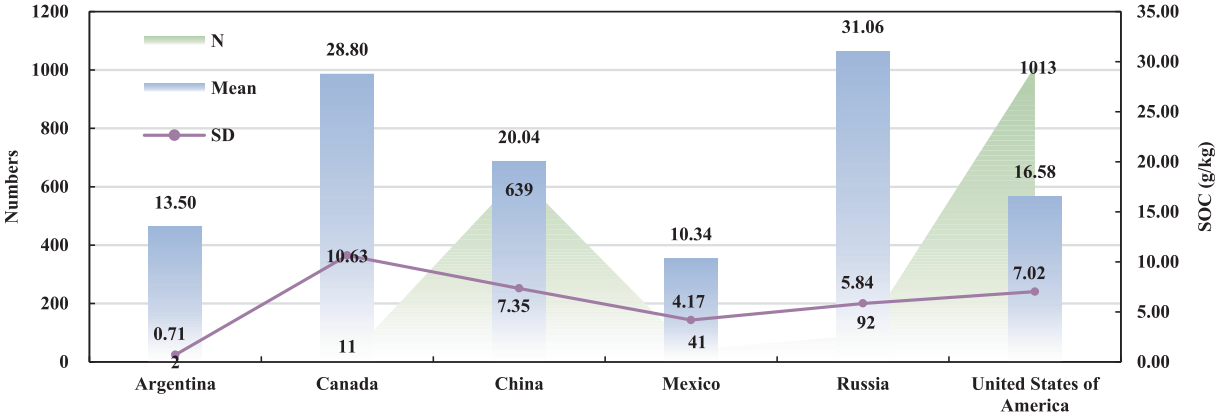

Fig. 3. Sample data from different countries. N denotes the number of samples, Mean denotes the mean value of the SOC content, and SD denotes the standard deviation of the SOC content.

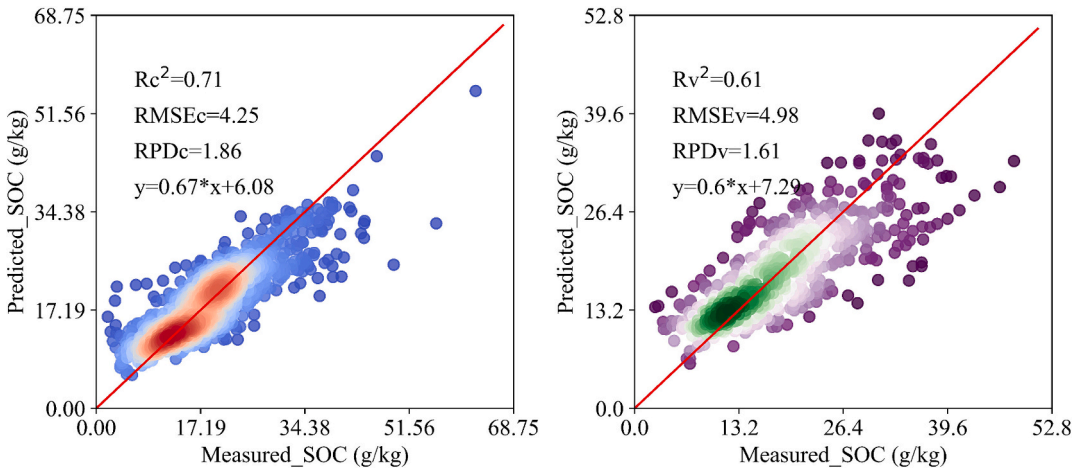

Fig. 4. Predicted and measured SOC content of the GloSOC-Ensemble model: (a) training samples based on cross-validation and (b) independent testing samples.

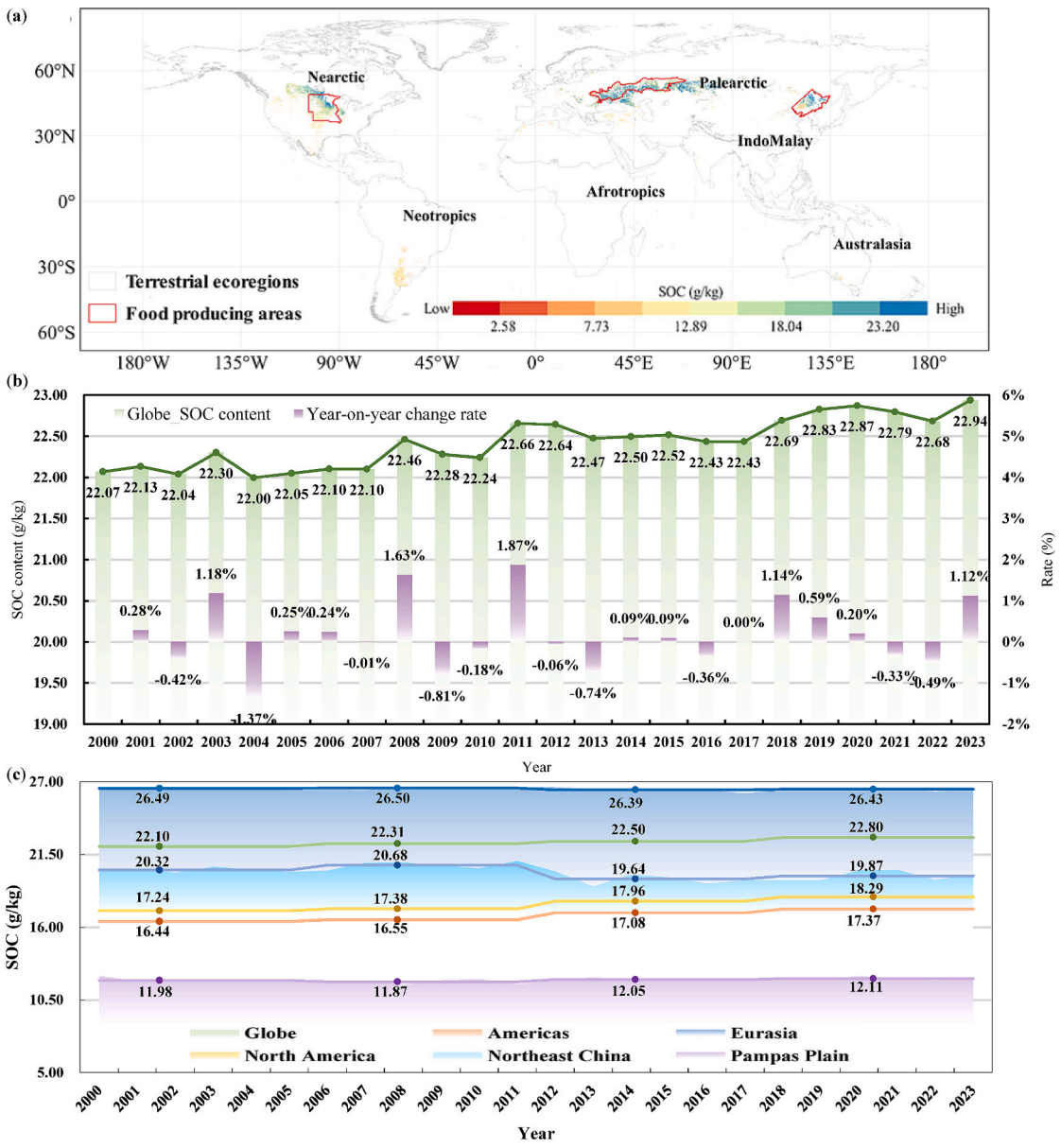

Fig. 5. Temporal and spatial trends of the SOC content in the topsoil of the global Mollisol cropland areas. (a) Spatial representation and (b) average annual percentage change of the mean SOC content for the full period. (c) Variations of the SOC content in the different regions.

from 22.10 g/kg in the early 2000 s to 22.80 g/kg in recent years, with a maximum stage-specific increase of 0.30 g/kg. The spatio-temporal trends of SOC content varied across regions. North America experienced a notable increase, whereas Eurasia exhibited an alternating pattern of gains and losses, with a pronounced decrease in Northeast China. Furthermore, the increase observed in the Americas outweighed the decrease in Eurasia, resulting in an overall upward trend.

3.4. Assessment of the predictive feature importance and environmental correlations

The correlation analysis results (Fig. 6(a)) show that latitude, the canopy response salinity index (CRSI), and the global environmental monitoring index (GEMI) exhibit moderate correlations with SOC content ($|\rho| > 0.4$, $p < 0.001$), representing the strongest correlations among all the features. The feature importance assessment further reveals considerable variation in the importance of individual features to SOC estimation. Notably, spatial location information accounts for more than 50% of the total importance, highlighting the strong geographic dependency of SOC distribution. Elevation, as a primary topographic factor, contributes 18.71%, indicating the potential influence of terrain on SOC accumulation (Gibson et al., 2021; Shen et al., 2015). The spectral difference indices contribute approximately 10.56% in total, reflecting their capacity to characterize soil and vegetation surface conditions through remote sensing. Vegetation indices contribute 8.42%, indicating an indirect modulation effect via vegetation cover and primary productivity. Soil properties and parent material related variables contribute 7.42%, suggesting that pedogenic characteristics partially influence the spatial pattern of carbon stocks. Although climatic factors contribute relatively little (4.80%), they still constitute an important environmental background for SOC spatial variability.

In this study, to preliminarily investigate the dynamic responses of SOC, we incorporated variables such as net primary productivity (NPP), evapotranspiration (ET), and potential evapotranspiration (PET) for the temporal analyses (the MOD17A3HGF and MOD16A2GF products were obtained from https://lpdaac.usgs.gov/products/). The results show (Fig. 6(b)) that there is a significant moderate positive correlation between NPP and SOC content ($\rho = 0.46$, $p < 0.05$), with some consistency between the two in terms of multi-year trends. This suggests that possible positive feedbacks exist between the increase of NPP and the accumulation of SOC. Regarding ET and PET, which are closely related to irrigated agriculture, the former is positively correlated with SOC content in the time series ((Fig. 6(c)), $\rho = 0.60$, $p < 0.01$), while the latter is significantly negatively correlated with SOC content ((Fig. 6(d)), $\rho = -0.55$, $p < 0.01$). Further regression analyses showed that the average annual variation in SOC content explains about 30% of the variation in ET/PET, implying that it affects vegetation water use efficiency to some extent (Huang et al., 2021; Yang et al., 2014).

3.5. Regional associations between SOC content variations and major human activities

In addition to natural environmental factors, human activities have been suggested to be associated with regional-scale SOC dynamics (Beillouin et al., 2023). Based on the predicted SOC results described above, we further examine, from a results-oriented perspective, the correspondence between SOC changes and major human activity indicators in selected representative regions.

In Northeast China, maize and soybean are the most widely grown crops, and crop rotation between the two is the predominant

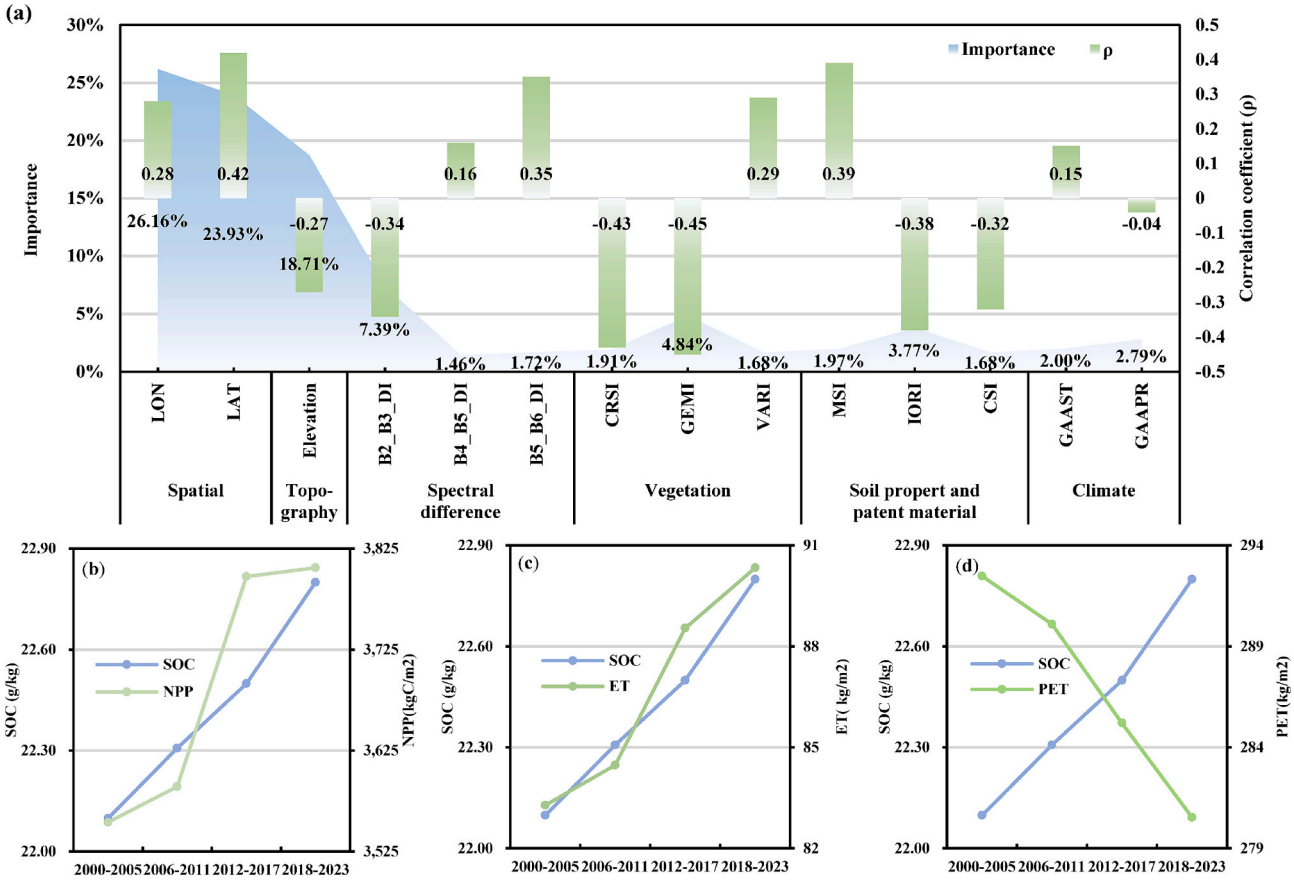

Fig. 6. Importance and temporal variation of the different features. (a) The features employed for SOC content estimation, including their correlation with SOC content and their importance in the regression model. (b)–(d) The relationship with mean SOC content at the different time periods.

agricultural practice. By analyzing the annual statistics (https://data.stats.gov.cn/easyquery.htm?cn=E0103) from the region (Fig. 7(a)), it was found that the sown area of maize is significantly and negatively correlated with the SOC content of the surface Mollisols throughout Northeast China ($|\rho| > 0.53$, $p < 0.01$), showing that the expansion of maize planting may be accompanied by a decreasing trend of SOC content. In contrast, the soybean planting area shows a significant positive correlation with SOC content only in Jilin province ($|\rho| = 0.59$, $p < 0.01$). The temporal trend analysis further indicates that fluctuations in soybean planting area and SOC content exhibit complementary patterns in multiple years. These regionally and temporally differentiated correlations reflect the potential links between cropping patterns and SOC dynamics.

The number of land protection policies also exhibits clear regional differences worldwide. Among the 73 countries with Mollisols, the number of policy documents related to land and soil protection (from https://www.fao.org/faolex/en; Fig. 7(b)) is significantly higher in Eurasia than in the Americas and Oceania, with Russia, China, and Canada having the largest number of such policies. In general, the spatial distribution of the number of policies corresponds to the topsoil SOC contents of the Mollisol cropland areas of Eurasia and North America.

However, abrupt human disturbances have the potential to cause significant fluctuations in SOC content. In Ukraine, for instance, the average SOC content in the Mollisol cropland areas decreased significantly following the large-scale conflict that began in 2022. Compared with 2021, although the main conflict area recovered partially in 2023, the SOC content in the secondary impact area decreased by 0.92 g/kg in two years (Fig. 7(c) and (d)), revealing that social conflicts can cause short-term disturbances in the soil system and generate heterogeneous responses.

Therefore, human activities, such as agricultural restructuring, institutional security, and social stability, can affect the spatial and temporal distribution and dynamic characteristics of SOC content at different scales.

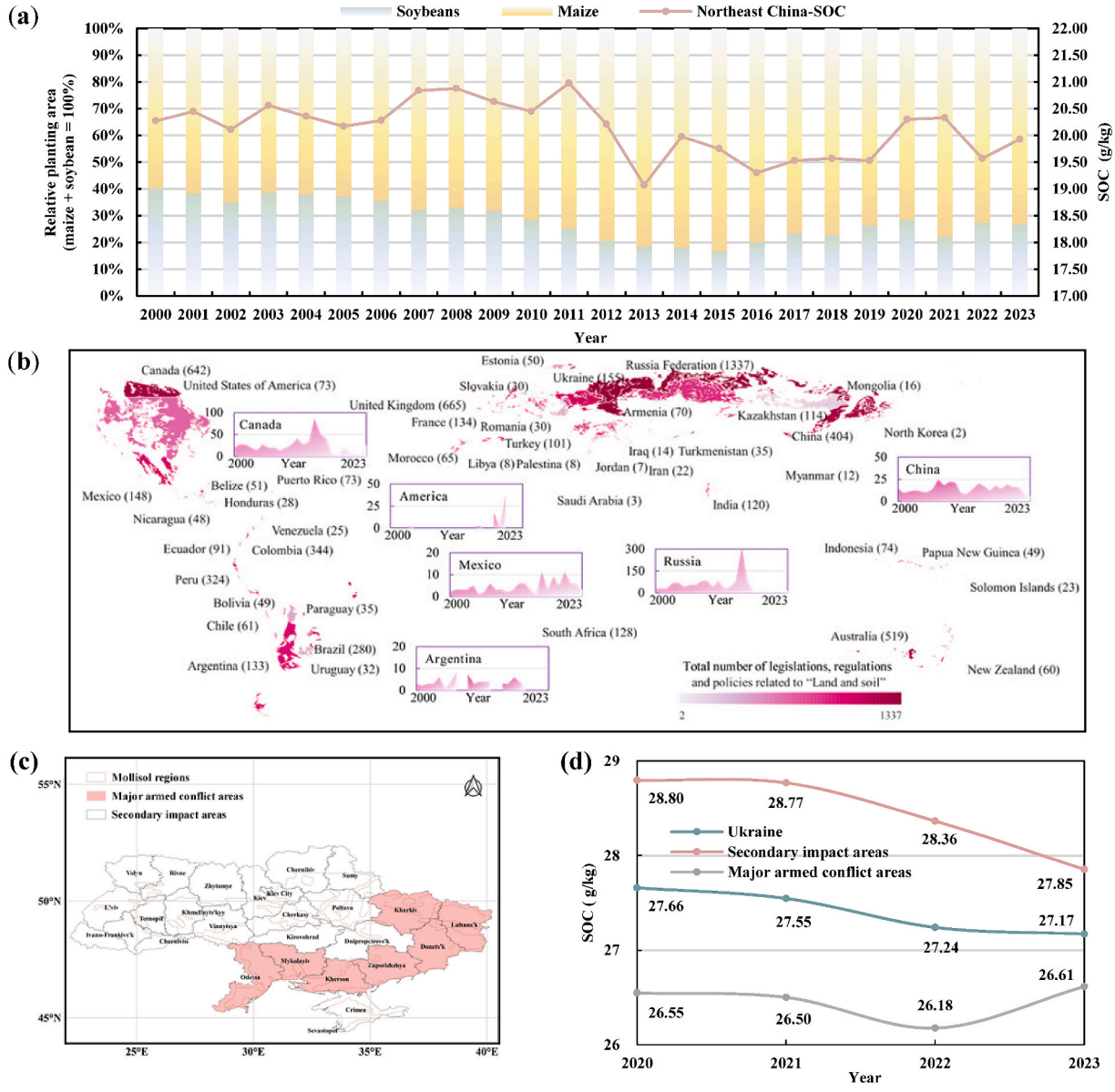

Fig. 7. Associations between human activities and SOC content variations. (a) Sowing areas of soybean and maize in Northeast China. (b) Global map showing the total number of legislations, regulations, and policies related to "land and soil" for each country in the FAOLEX database. (c) Map of Ukraine, showing the areas affected by the armed conflict. (d) Variation of the topsoil SOC content in the Mollisol cropland areas of Ukraine.

4. Discussion

4.1. Spatial and temporal characteristics of regional SOC trends

According to the year-by-year spatial distribution maps (Fig. S5) and statistical results (Fig. 5), it can be observed that the SOC contents in the different regions have shown significant differences between 2000 and 2023. In general, the spatial distribution pattern of SOC is highly consistent with the relative distribution pattern of the existing data products (e.g., GSOC v1.5 (https://data.apps.fao.org/), SoilGrids 2.0 (https://soilgrids.org/) (Poggio et al., 2021), and the SOC estimation results based on a geographic knowledge dataset and probability hybrid model (GEKD-PHM) (Meng et al., 2025)). The spatial distribution shows a gradient of "Europe > Northeast China > North America > South America". This consistency reflects the robustness of multi-source modeling in spatial pattern identification. Nevertheless, the range of SOC values varies between products. The results of this study have lower values over the whole global study area, compared to the other products (SoilGrids 2.0 (0–116.00 g/kg), GSOC v1.5 (0–99.70 g/kg), GEKD-PHM (0–82.07 g/kg), this study (0–55.51 g/kg)), which may stem from the range of SOC values in the training samples, the type of input variables used for the estimation, the modeling method, the time horizon, the scale sensitivity, and the range of pixels used for the statistical analysis. However, the temporal trend of SOC content shows some differentiation when compared to the spatial pattern. Remote sensing estimation studies (Meng et al., 2024; Meng et al., 2025) based on the reflectance of multi-year synthetic remote sensing data at longer time scales (e.g., the 1980 s to the present) have generally reported a continuous downward trend in SOC content in the Mollisols globally. The findings of this study based on high temporal resolution remote sensing data show that this trend is not generalized over the period of 2000–2023.

Specifically, Northeast China still showed a certain degree of decreasing trend (−0.34 g/kg) during the whole period. The most significant changes in SOC were observed in each region during 2006–2017 (Fig. 5(c)). Although high-intensity farming activities in Northeast China, as a major grain-producing region in China, have made important contributions to national food security, long-term irrational farming, over-tillage, and soil erosion have also accelerated the loss of SOC (Jiang et al., 2024a), which is consistent with the results of the existing studies (Meng et al., 2024; Wang et al., 2023b). Since 2017 (Ministry of Agriculture of the People's Republic of China, 2017), the Chinese government has implemented land protection policies every year (Fig. 7(b)), including the introduction of a protection law for Mollisols (Standing Committee of the National People's Congress of China, 2022), which aims to improve soil quality through measures such as erosion control, increased application of organic fertilizers, and returning straw to the soil. This may have had a positive effect on the SOC content in Northeast China, which did rebound somewhat during 2018–2023. In contrast, North America showed a more consistent and significant SOC growth trend over the study period (+1.28 g/kg). Previous studies based on counterfactual scenarios have indicated that SOC stocks increased annually in US cropland areas from 1995 to 2015 (Ogle et al., 2023). This can be mainly attributed to the continued promotion of efficient management practices such as no-till agriculture, rotational grazing, and cover crops (Joshi et al., 2023; Sangotayo et al., 2023). These differences may reflect the complex coupling effects of the different regions in terms of agricultural management policies, land-use change, and climate change response.

4.2. Ecological and anthropogenic coupling of regional SOC dynamics

In this study, to better interpret the possible background factors for SOC changes in the different regions, we attempted to combine ecological and anthropogenic disturbances in a preliminary analysis. Although we did not explicitly model ecological process variables in this study, the temporal changes in SOC (Fig. 6(b)–(d)) are closely related to the increase in NPP (Kida et al., 2017) and decrease in PET (Mishra et al., 2022), which indirectly supports the SOC accumulation mechanism proposed by the previous studies. This provides a reference for the subsequent exploration of the relationship between carbon stocks and ecosystem functions. In addition, the results of the feature importance analysis showed that the remote sensing indices related to vegetation coverage, moisture status, and soil mineral composition made a high contribution to the spatial estimation of SOC (the proportion of related features was 15.84%), which reflects the key role of regional ecological base conditions in soil carbon distribution (Wang et al., 2024; Xiao et al., 2024). Based on the model estimation performance, the spectral indices selected in this study exhibited notable ecological relevance and application potential. Spectral indices such as the char soil index (CSI), moisture stress index (MSI), and iron oxide ratio index (IORI) reflect the surface soil characteristics related to carbonaceous materials, moisture stress, and iron oxide distribution, while the visible atmospherically resistant index (VARI), canopy response salinity index (CRSI), and global environmental monitoring index (GEMI) correspond to vegetation density, salinity stress, and overall ecosystem status, respectively. These indices not only capture surface environmental changes effectively, but also help overcome the limitations of the conventional vegetation indices such as NDVI in soil property estimation (Liu et al., 2021; Sims and Gamon, 2002). Compared to the traditional indices, these structural or composite indices show greater sensitivity to ecological changes such as land degradation, crop structure adjustment, and moisture disturbance (Lei et al., 2024; Sun et al., 2021). This enhances their capacity to reflect the spatial heterogeneity of SOC distribution and highlights their potential utility in large-scale soil carbon monitoring.

In terms of anthropogenic influences, regional shifts in cropping structure and changes in socio-political stability can partly explain the spatial variability of SOC dynamics. For example, in Northeast China, the reduction in soybean cultivation and the expansion of maize led to a decline in the root decomposition and biological nitrogen fixation associated with shallow-rooted soybean (Thibodeau and Jaworski, 1975). The increased reliance on nitrogen fertilizers, combined with the prolonged nutrient depletion by deep-rooted maize, may have accelerated the mineralization and loss of topsoil SOC (Hong et al., 2023). In addition, the case of Ukraine (Fig. 7(c)–(d)) illustrates that modern armed conflict does not necessarily halt national food production and export (FAO, 2022b). In the major conflict zones, SOC levels showed signs of recovery in 2023, possibly due to farmland abandonment and subsequent natural vegetation regrowth (Shi et al., 2024; Zhao et al., 2023). In contrast, intensified agricultural activities in the secondary affected areas may have increased the cultivation pressure, thereby further depleting topsoil SOC content (FAO, 2022b).

4.3. Uncertainties and limitations of the SOC estimation framework

In this study, the topsoil SOC content of the global Mollisol cropland areas was estimated based on the GloSOC-Ensemble model. While the overall prediction performance was good, inherent uncertainties in both the model and data warrant careful consideration. Through five-fold cross-validation, the mean RMSE of the model training phase yielded 5.21 g/kg, with an SD of 0.42 g/kg, and the 95% CI was estimated to be approximately ±0.82 g/kg. However, cross-validation metrics primarily reflect the stability of model performance and do not adequately capture the uncertainty of individual predictions across temporal and spatial dimensions. Although the model achieved relatively high predictive accuracy on the test dataset, the bootstrap-based empirical prediction intervals indicate that SOC predictions are still associated with considerable uncertainty. The MPIW value suggests that the average width of the prediction intervals at the overall scale is 5.52 g/kg. In contrast, the distribution of PIW (Fig. 8) exhibits a pronounced right-skewed pattern, with most samples characterized by relatively narrow prediction intervals, indicating stable predictive performance across the dominant

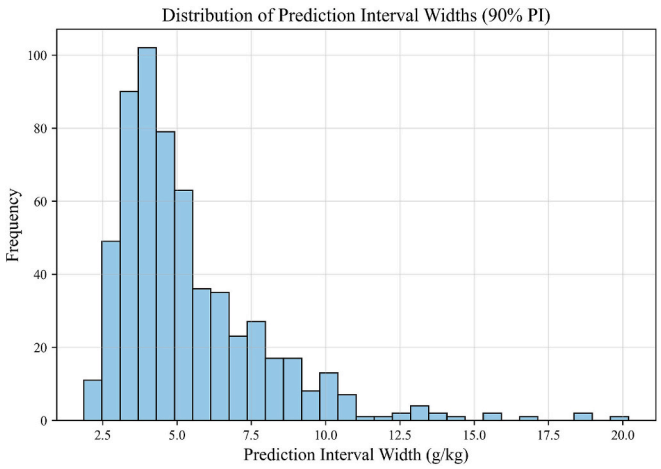

Fig. 8. The distribution of PIW.

sample space. However, a small number of samples show substantially larger PIW values, reflecting elevated prediction uncertainty under extreme or complex environmental conditions. These results demonstrate that, in addition to uncertainties arising from input data and environmental covariates, the predictive model itself constitutes a non-negligible source of uncertainty in SOC spatiotemporal mapping (Heuvelink et al., 2021; Szatmári et al., 2024). Therefore, in practical applications, SOC spatiotemporal variations should be interpreted by jointly considering both prediction results and their associated uncertainty information, in order to avoid over-interpretation of predictions in localized areas.

The spatial coverage of the remote sensing data was also not complete, especially in some areas of Eurasia between 2000 and 2005, which affected the temporal continuity and led to uncertainty (Fig. 9(a)) (Shen et al., 2016). Therefore, different data were used for the statistical analyses in this study: i) only valid data were used; and ii) all the data were used, where the data pending calibration were adjusted by the interpolation method. Despite being effective in mitigating the problem of incompleteness and outliers in different spatio-temporal image data, the compensation accuracy was still limited in high-latitude and complex terrain regions (Fig. 9(b)). Further minimization of the inaccuracy due to the differences in image coverage will be needed in the future, in order to improve the reliability of the results (Zhu and Woodcock, 2014). In addition, the Landsat 7 ETM+ data used for 2012 are affected by localized pixel gaps caused by the failure of the scan line corrector (SLC) onboard Landsat 7 ETM+ (hereafter referred to as the SLC-off issue), which represents a potential source of data-related uncertainty in SOC mapping. However, in the global-scale analyses conducted in this study, SOC estimates are primarily derived from regional statistics based on a large number of valid pixels, and both the spatial patterns and long-term temporal trends show strong robustness to a limited proportion of missing pixels. For future studies aiming at higher-resolution, pixel-level SOC mapping, gap-filling or multi-source remote sensing data fusion approaches could be considered to further reduce this type of uncertainty.

The analysis of global-scale SOC content variations also faces the challenge of differences in data resolution and time span across regions, leading to unequal model confidence among regions (Jungkunst et al., 2022). Global models can provide a unified analysis framework for large-scale regions, but can ignore local features, resulting in a lower accuracy (Chen et al., 2022). In contrast, region-based models can better capture local features and improve the accuracy, with the disadvantage that they can cause inconsistency in cross-region analysis (Liu et al., 2019; Padarian et al., 2017). Therefore, prospective studies should consider appropriately combining regional models under a global analysis framework, to ensure the accuracy while enhancing the overall consistency. In addition, the input features of this study included nine Landsat-derived spectral indices (see Section 2.6), which had the

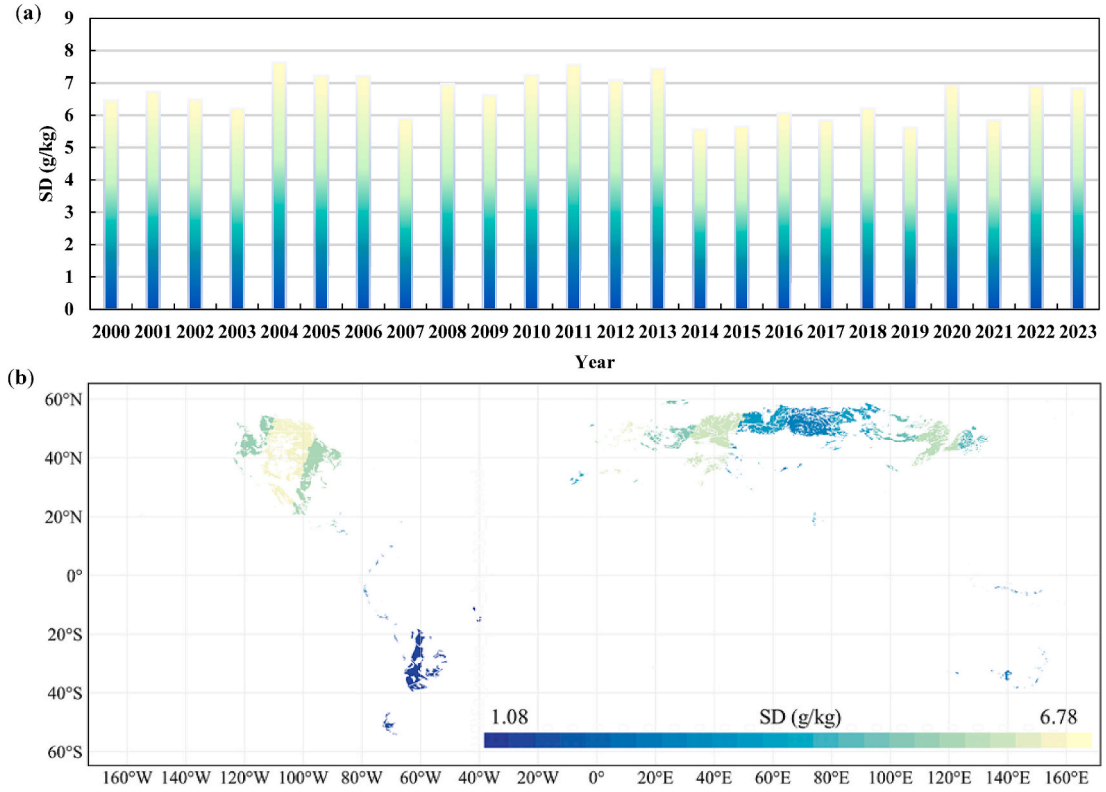

Fig. 9. Temporal and spatial patterns of uncertainty (standard deviation) in the estimated SOC content based on the GloSOC-Ensemble model. (a) Interannual variability of SOC uncertainty summarized at the regional scale; (b) spatial distribution of SOC uncertainty across the Mollisol cropland regions.

advantage of temporal coverage but did not adequately account for the effects of factors such as soil moisture and texture on reflectance (Dharumarajan et al., 2024). The nonlinear and nonstationary relationship between SOC and spectral response may be simplified, affecting the predictive stability (Gomez et al., 2022). Finally, the spatial distribution of the ground sampling was uneven, mainly concentrated in the United States and China, and the temporal distribution of sample collection was also unbalanced, limiting the model generalization ability (Zhou et al., 2023). In the future, more public datasets should be assembled to optimize the structure of the spatial and temporal distribution of the samples and improve the regional applicability and representativeness. In summary, insufficient data coverage, limited feature selection, and inadequate representative sampling still limit the accuracy and reliability of SOC content estimation.

5. Conclusion

In this study, the annual spatial distribution of topsoil SOC content in global Mollisol cropland areas was mapped based on ensemble learning and Landsat satellite imagery from 2000 to 2023, providing a long-term and spatially explicit assessment of SOC dynamics in major Mollisol croplands. The results indicated an overall increase in SOC levels across global Mollisol croplands during the study period, with a net rise of 3.17%. However, this global tendency masked pronounced regional heterogeneity. Mollisol croplands in the Americas exhibited a relatively consistent increase in SOC, exceeding 5.00% over the past two decades, whereas Eurasian regions showed more variable trajectories with localized fluctuations and occasional declines. These contrasting patterns highlight the spatial complexity of SOC dynamics in major black soil agricultural systems.

Further analyses revealed the close associations between SOC dynamics and environmental factors such as vegetation productivity and soil–water conditions. The integration of agro-environmental indicators, including—such as NPP, ET, and vegetation indices—provided additional insights into the coupling between ecosystem functioning and cropping systems, which jointly regulate the spatial and temporal patterns of SOC in croplands. Beyond environmental control factors, our findings also demonstrated that human interventions—such as cultivation practices, agricultural policies adjustments, and socio-political conditions—may exert significant influence over the dynamics of SOC. This underlines the dual sensitivity of soil carbon dynamics to both ecological and socio-economic factors.

Overall, these insights will enhance our understanding of soil carbon dynamics and underscore the importance of region-specific soil management strategies for maintaining and enhancing soil carbon stocks. The findings provide valuable scientific evidence for precision agricultural management, soil conservation planning, and the formulation of policies promoting sustainable land use and soil carbon sequestration. Future investigations should incorporate multi-source observational data, refine spatio-temporal modelling methods, and integrate socio-economic drivers to more accurately quantify the underlying mechanisms of SOC dynamics. These efforts will support more effective soil carbon management under evolving environmental and land-use conditions.

CRediT authorship contribution statement

Yayu Yang: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Linya Zhao:** Investigation, Data curation, Conceptualization. **Renjie Ji:** Visualization, Conceptualization. **Huimin Dai:** Validation, Resources, Data curation. **Xue Wang:** Writing – review & editing, Methodology. **Chao Niu:** Writing – review & editing, Methodology, Formal analysis. **Kun Tan:** Writing – review & editing, Supervision, Resources, Methodology, Funding acquisition, Conceptualization.

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.

Acknowledgements

We would like to thank the United States Geological Survey for providing the global Landsat satellite images, and Google Earth for providing the data processing resources. We would also like to thank Prof. Igor Yu. Savin (V.V. Dokuchaev Soil Science Institute, Moscow, Russia) for providing the Mollisol soil samples for Russia. This work was supported in part by the Yangtze River Delta Science and Technology Innovation Community Joint Research (Basic Research) Project (No. 2024CSJZN1300), the Shanghai Municipal Education Commission Science and Technology Project (2024AI02002), the National Natural Science Foundation of China (No. 42171335), and the National Civil Aerospace Project of China (No. D040102).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.geoderma.2026.117789.

Data availability

Data will be made available on request.

References

Abdulraheem, M.I., Zhang, W., Li, S., Moshayedi, A.J., Farooque, A.A., Hu, J., 2023. Advancement of remote sensing for soil measurements and applications: a comprehensive review. Sustainability 15 (21), 15444.
Bartholomeus, H., Schaepman, M.E., Kooistra, L., Stevens, A., Hoogmoed, W., Spaargaren, O., 2008. Spectral reflectance based indices for soil organic carbon quantification. Geoderma 145 (1–2), 28–36.
Beillouin, D., Corbeels, M., Demenois, J., Berre, D., Boyer, A., Fallot, A., Feder, F., Cardinael, R., 2023. A global meta-analysis of soil organic carbon in the Anthropocene. Nat. Commun. 14 (1), 3700.
Biney, J.K.M., Vašát, R., Bell, S.M., Kebonye, N.M., Klement, A., John, K., Borůvka, L., 2022. Prediction of topsoil organic carbon content with Sentinel-2 imagery and spectroscopic measurements under different conditions using an ensemble model approach with multiple pre-treatment combinations. Soil Tillage Res. 220, 105379.
Breiman, L., 2001. Random forests. Mach. Learn. 5–32.
Burke, M., Driscoll, A., Lobell, D.B., Ermon, S., 2021. Using satellite imagery to understand and promote sustainable development. Science 371 (6535), eabe8628.
Chen, L., Ding, Y., Pirasteh, S., Hu, H., Zhu, Q., Ge, X., Zeng, H., Yu, H., Shang, Q., Song, Y., 2022. Meta-learning an intermediate representation for few-shot prediction of landslide susceptibility in large areas. Int. J. Appl. Earth Obs. Geoinf. 110, 102807.
Chen, S., Xu, D., Li, S., Ji, W., Yang, M., Zhou, Y., Hu, B., Xu, H., Shi, Z., 2020. Monitoring soil organic carbon in alpine soils using in situ vis-NIR spectroscopy and a multilayer perceptron. Land Degrad. Dev. 31 (8), 1026–1038.
Chinilin, A., Savin, I.Y., 2023. Combining machine learning and environmental covariates for mapping of organic carbon in soils of Russia. Egypt. J. Remote Sens. Space Sci. 26 (3), 666–675.
Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory 13 (1), 21–27.
Cui, Z., Chen, S., Hu, B., Wang, N., Zhai, J., Peng, J., Bai, Z., 2025. High-accuracy mapping of soil organic carbon by mining sentinel-1/2 radar and optical time-series data with super ensemble model. Remote Sens. (Basel) 17 (4), 678.
Dharumarajan, S., Lalitha, M., Kalaiselvi, B., Kaliraj, S., Adhikari, K., Vasundhara, R., Niranjana, K., Hegde, R., Pradeep, C., Hittanagi, P., 2024. Remote sensing of soils: Spectral signatures and spectral indices, Remote Sensing of Soils. Elsevier, pp. 13-23.
Drury, S.A., 1987. Image interpretation in geology.
Dvorakova, K., Heiden, U., Pepers, K., Staats, G., van Os, G., van Wesemael, B., 2023. Improving soil organic carbon predictions from a Sentinel–2 soil composite by assessing surface conditions and uncertainties. Geoderma 429, 116128.
FAO, 2021. Standard Operating Procedure for Soil Organic Carbon: Tyurin Spectrophotometric Method. FAO Rome, Italy.
FAO, 2022a. Global Soil Organic Carbon Sequestration Potential Map—GSOCseq v. 1.1, FAO Rome, Italy.
FAO, 2022b. Impact of the Ukraine-Russia conflict on global food security and related matters under the mandate of the Food and Agriculture Organization of the United Nations (FAO).

Gibson, A., Hancock, G., Bretreger, D., Cox, T., Hughes, J., Kunkel, V., 2021. Assessing digital elevation model resolution for soil organic carbon prediction. Geoderma 398, 115106.
Gitelson, A.A., Kaufman, Y.J., Stark, R., Rundquist, D., 2002. Novel algorithms for remote estimation of vegetation fraction. Remote Sens. Environ. 80 (1), 76–87.
Gomez, C., Lagacherie, P., Coulouma, G., 2008. Continuum removal versus PLSR method for clay and calcium carbonate content estimation from laboratory and airborne hyperspectral measurements. Geoderma 148 (2), 141–148.
Gomez, C., Vaudour, E., Féret, J.-B., De Boissieu, F., Dharumarajan, S., 2022. Topsoil clay content mapping in croplands from Sentinel-2 data: Influence of atmospheric correction methods across a season time series. Geoderma 423, 115959.
Hengl, T., Mendes de Jesus, J., Heuvelink, G.B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS One 12 (2), e0169748.
Heuvelink, G.B., Angelini, M.E., Poggio, L., Bai, Z., Batjes, N.H., van den Bosch, R., Bossio, D., Estella, S., Lehmann, J., Olmedo, G.F., 2021. Machine learning in space and time for modelling soil organic carbon change. Eur. J. Soil Sci. 72 (4), 1607–1623.
Hong, S., Ding, J., Kan, F., Xu, H., Chen, S., Yao, Y., Piao, S., 2023. Asymmetry of carbon sequestrations by plant and soil after forestation regulated by soil nitrogen. Nat. Commun. 14 (1), 3196.
Huang, X., Wang, H., Zhang, M., Horn, R., Ren, T., 2021. Soil water retention dynamics in a Mollisol during a maize growing season under contrasting tillage systems. Soil Tillage Res. 209, 104953.
Hunt Jr, E.R., Rock, B.N., 1989. Detection of changes in leaf water content using near- and middle-infrared reflectances. Remote Sens. Environ. 30 (1), 43–54.
International Food Policy Research Institute, 2019. Global spatially-disaggregated crop production statistics data for 2010 version 2.0. Harvard Library Cambridge, MA.
Jenny, H., 1994. Factors of Soil Formation: A System of Quantitative Pedology. Courier Corporation.
Ji, R., Tan, K., Wang, X., Tang, S., Sun, J., Niu, C., Pan, C., 2025. PatchOut: A novel patch-free approach based on a transformer-CNN hybrid framework for fine-grained land-cover classification on large-scale airborne hyperspectral images. International Journal of Applied Earth Observation and Geoinformation 138, 104457.
Jiang, M., Jia, Z., Wen, Y., Xu, H., Wang, H., Zeng, Y., Li, L., Cui, M., Li, H., Zhang, J., 2024a. Safeguarding the "black soil granary": innovations in soil conservation and sustainable agriculture. Bull. Chin. Acad. Sci. 38, 2024018.
Jiang, R., Sui, Y., Zhang, X., Lin, N., Zheng, X., Li, B., Zhang, L., Li, X., Yu, H., 2024b. Estimation of soil organic carbon by combining hyperspectral and radar remote sensing to reduce coupling effects of soil surface moisture and roughness. Geoderma 444, 116874.
Jin, X., Du, J., Liu, H., Wang, Z., Song, K., 2016. Remote estimation of soil organic matter content in the Sanjiang Plain, Northest China: the optimal band algorithm versus the GRA-ANN model. Agric. For. Meteorol. 218, 250–260.
Joshi, D.R., Sieverding, H.L., Xu, H., Kwon, H., Wang, M., Clay, S.A., Johnson, J.M., Thapa, R., Westhoff, S., Clay, D.E., 2023. A global meta-analysis of cover crop response on soil carbon storage within a corn production system. Agron. J. 115 (4), 1543–1556.
Jungkunst, H.F., Göpel, J., Horvath, T., Ott, S., Brunn, M., 2022. Global soil organic carbon–climate interactions: why scales matter. Wiley Interdiscip. Rev. Clim. Chang. 13 (4), e780.
Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y., 2017. Lightgbm: a highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 30.
Kida, M., Tomotsune, M., Iimura, Y., Kinjo, K., Ohtsuka, T., Fujitake, N., 2017. High salinity leads to accumulation of soil organic carbon in mangrove soil. Chemosphere 177, 51–55.
Lei, J., Zeng, C., Zhang, L., Wang, X., Ma, C., Zhou, T., Laffitte, B., Luo, K., Yang, Z., Tang, X., 2024. Prediction of soil organic carbon stock combining Sentinel-1 and Sentinel-2 images in the Zoige Plateau, the northeastern Qinghai-Tibet Plateau. Ecol. Process. 13 (1), 32.
Li, Y., Li, J., Tan, J., Ma, T., Yan, X., Chen, Z., Li, K., 2025. Fine resolution mapping of forest soil organic carbon based on feature selection and machine learning algorithm. Remote Sens. (Basel) 17 (12), 2000.
Liu, J., Maeda, E.E., Wang, D., Heiskanen, J., 2021. Sensitivity of spectral indices on burned area detection using Landsat time series in savannas of southern Burkina Faso. Remote Sens. (Basel) 13 (13), 2492.
Liu, S., Shen, H., Chen, S., Zhao, X., Biswas, A., Jia, X., Shi, Z., Fang, J., 2019. Estimating forest soil organic carbon content using vis-NIR spectroscopy: implications for large-scale soil carbon spectroscopic assessment. Geoderma 348, 37–44.
Liu, X., Lee Burras, C., Kravchenko, Y.S., Duran, A., Huffman, T., Morras, H., Studdert, G., Zhang, X., Cruse, R.M., Yuan, X., 2012. Overview of Mollisols in the world: distribution, land use and management. Can. J. Soil Sci. 92 (3), 383–402.
Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 30.
Luo, C., Zhang, X., Meng, X., Zhu, H., Ni, C., Chen, M., Liu, H., 2022. Regional mapping of soil organic matter content using multitemporal synthetic Landsat 8 images in Google Earth Engine. Catena 209, 105842.
Mansuy, N., Thiffault, E., Paré, D., Bernier, P., Guindon, L., Villemaire, P., Poirier, V., Beaudoin, A., 2014. Digital mapping of soil properties in Canadian managed forests at 250 m of resolution using the k-nearest neighbor method. Geoderma 235, 59–73.
McBratney, A.B., Santos, M.M., Minasny, B., 2003. On digital soil mapping. Geoderma 117 (1–2), 3–52.
Meersmans, J., De Ridder, F., Canters, F., De Baets, S., Van Molle, M., 2008. A multiple regression approach to assess the spatial distribution of Soil Organic Carbon (SOC) at the regional scale (Flanders, Belgium). Geoderma 143 (1–2), 1–13.
Meng, X., Bao, Y., Luo, C., Zhang, X., Liu, H., 2024. SOC content of global Mollisols at a 30 m spatial resolution from 1984 to 2021 generated by the novel ML-CNN prediction model. Remote Sens. Environ. 300, 113911.
Meng, X., Bao, Y., Wang, Y., Zhang, X., Liu, H., 2022. An advanced soil organic carbon content prediction model via fused temporal-spatial-spectral (TSS) information based on machine learning and deep learning algorithms. Remote Sens. Environ. 280, 113166.
Meng, X., Bao, Y., Zhang, X., Luo, C., Liu, H., 2025. A long-term global Mollisols SOC content prediction framework: Integrating prior knowledge, geographical partitioning, and deep learning models with spatio-temporal validation. Remote Sens. Environ. 318, 114592.
Ministry of Agriculture of the People's Republic of China, 2017. Northeast Black Soil Conservation Plan (2017–2030).
Misebo, A.M., Hawryło, P., Szostak, M., Pietrzykowski, M., 2024. Spatial estimation of soil organic carbon, total nitrogen, and soil water storage in reclaimed post-mining site based on remote sensing data. Ecol. Ind. 166, 112228.
Mishra, N., Helder, D., Barsi, J., Markham, B., 2016. Continuous calibration improvement in solar reflective bands: Landsat 5 through Landsat 8. Remote Sens. Environ. 185, 7–15.
Mishra, U., Yeo, K., Adhikari, K., Riley, W.J., Hoffman, F.M., Hudson, C., Gautam, S., 2022. Empirical relationships between environmental factors and soil organic carbon produce comparable prediction accuracy to machine learning. Soil Sci. Soc. Am. J. 86 (6), 1611–1624.
Montgomery, D.C., Runger, G.C., 2019. Applied Statistics and Probability for Engineers. John Wiley & Sons.
Moura-Bueno, J.M., Dalmolin, R.S.D., Horst-Heinen, T.Z., Ten Caten, A., Vasques, G.M., Dotto, A.C., Grunwald, S., 2020. When does stratification of a subtropical soil spectral library improve predictions of soil organic carbon content? Sci. Total Environ. 737, 139895.
Ogle, S.M., Breidt, F.J., Del Grosso, S., Gurung, R., Marx, E., Spencer, S., Williams, S., Manning, D., 2023. Counterfactual scenarios reveal historical impact of cropland management on soil organic carbon stocks in the United States. Sci. Rep. 13 (1), 14564.
Olson, D., Dinerstein, E., Wikramanayake, E., Burgess, N., Powell, G., Underwood, E., d'Amico, J., Itoua, I., Strand, H., Morrison, J., 2001. Terrestrial Ecoregions of the World: A New Map of Life on Earth (PDF, 1.1 M) BioScience 51: 933-938. Online_ Linkage.
Padarian, J., Minasny, B., McBratney, A., 2017. Chile and the Chilean soil grid: a contribution to GlobalSoilMap. Geoderma Reg. 9, 17–28.
Padarian, J., Stockmann, U., Minasny, B., McBratney, A., 2022. Monitoring changes in global soil organic carbon stocks from space. Remote Sens. Environ. 281, 113260.
Pinty, B., Verstraete, M., 1992. GEMI: a non-linear index to monitor global vegetation from satellites. Vegetatio 101, 15–20.
Poggio, L., De Sousa, L.M., Batjes, N.H., Heuvelink, G.B., Kempen, B., Ribeiro, E., Rossiter, D., 2021. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. Soil 7 (1), 217–240.
Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., Gulin, A., 2018. CatBoost: unbiased boosting with categorical features. Adv. Neural Inf. Process. Syst. 31.
Qi, Q., Yue, X., Duo, X., Xu, Z., Li, Z., 2023. Spatial prediction of soil organic carbon in coal mining subsidence areas based on RBF neural network. Int. J. Coal Sci. Technol. 10 (1), 30.
Riggs, G.A., Hall, D.K., Salomonson, V.V., 1994. A snow index for the Landsat thematic mapper and moderate resolution imaging spectroradiometer, Proceedings of IGARSS'94-1994 IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 1942-1944.
Rouse, J.W., Haas, R.H., Schell, J.A., Deering, D.W., 1974. Monitoring vegetation systems in the great plains with ERTS. NASA Spec. Publ. 351 (1), 309.
Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning representations by back-propagating errors. Nature 323 (6088), 533–536.
Sangotayo, A.O., Chellappa, J., Sekaran, U., Bansal, S., Angmo, P., Jasa, P., Kumar, S., Iqbal, J., 2023. Long-term conservation and conventional tillage systems impact physical and biochemical soil health indicators in a corn–soybean rotation. Soil Sci. Soc. Am. J. 87 (5), 1056–1071.
Santillano Cázares, J., Roque Díaz, L.G., Núñez Ramírez, F., Grijalva Contreras, R.L., Robles Contreras, F., Macías Duarte, R., Escobosa García, I., Cárdenas Salazar, V., 2019. Soil fertility affects the growth, nutrition and yield of cotton cultivated in two irrigation systems and different nitrogen rates. Terra Latinoamericana 37 (1), 7–14.
Schmidinger, J., Heuvelink, G.B., 2023. Validation of uncertainty predictions in digital soil mapping. Geoderma 437, 116585.
Scudiero, E., Skaggs, T.H., Corwin, D.L., 2015. Regional-scale soil salinity assessment using Landsat ETM+ canopy reflectance. Remote Sens. Environ. 169, 335–343.
Segal, D., 1982. Theoretical basis for differentiation of ferric-iron bearing minerals, using Landsat MSS data, Proceedings of symposium for remote sensing of environment, 2nd Thematic Conference on Remote Sensing for Exploratory Geology, Fort Worth, TX, pp. 951.
Shcherbatyy, I., 2016. Vecstack: Stacking made easy.
Shen, H., Huang, L., Zhang, L., Wu, P., Zeng, C., 2016. Long-term and fine-scale satellite monitoring of the urban heat island effect by the fusion of multi-temporal and multi-sensor remote sensed data: a 26-year case study of the city of Wuhan in China. Remote Sens. Environ. 172, 109–125.
Shen, H., Li, X., Cheng, Q., Zeng, C., Yang, G., Li, H., Zhang, L., 2015. Missing information reconstruction of remote sensing data: a technical review. IEEE Geosci. Remote Sens. Mag. 3 (3), 61–85.

Shi, J., Deng, L., Wu, J., Bai, E., Chen, J., Shangguan, Z., Kuzyakov, Y., 2024. Soil organic carbon increases with decreasing microbial carbon use efficiency during vegetation restoration. Glob. Chang. Biol. 30 (12), e17616.
Sims, D.A., Gamon, J.A., 2002. Relationships between leaf pigment content and spectral reflectance across a wide range of species, leaf structures and developmental stages. Remote Sens. Environ. 81 (2–3), 337–354.
Smith, A.M., Wooster, M.J., Drake, N.A., Dipotso, F.M., Falkowski, M.J., Hudak, A.T., 2005. Testing the potential of multi-spectral remote sensing for retrospectively estimating fire severity in African Savannahs. Remote Sens. Environ. 97 (1), 92–115.
Smith, D., 2014. Soil Survey Staff: Keys to Soil Taxonomy. Natural Resources Conservation Service, Washington.
Standing Committee of the National People's Congress of China, 2022. Black Soil Protection Law of the People's Republic of China, National People's Congress (NPC).
Stockmann, U., Adams, M.A., Crawford, J.W., Field, D.J., Henakaarchchi, N., Jenkins, M., Minasny, B., McBratney, A.B., Courcelles, V.d.R., Singh, K., 2013. The knowns, known unknowns and unknowns of sequestration of soil organic carbon. Agric. Ecosyst. Environ. 164, 80–99.
Sun, H., Liu, H., Ma, Y., Xia, Q., 2021. Optical remote sensing indexes of soil moisture: evaluation and improvement based on aircraft experiment observations. Remote Sens. (Basel) 13 (22), 4638.
Szatmári, G., Pásztor, L., Takács, K., Mészáros, J., Benő, A., Laborczi, A., 2024. Space-time modelling of soil organic carbon stock change at multiple scales: case study from Hungary. Geoderma 451, 117067.
Tan, K., Ma, W., Chen, L., Wang, H., Du, Q., Du, P., Yan, B., Liu, R., Li, H., 2021. Estimating the distribution trend of soil heavy metals in mining area from HyMap airborne hyperspectral imagery based on ensemble learning. J. Hazard. Mater 401, 123288.
Thibodeau, P.S., Jaworski, E.G., 1975. Patterns of nitrogen utilization in the soybean. Planta 127, 133–147.
Tziolas, N., Tsakiridis, N., Heiden, U., van Wesemael, B., 2024. Soil organic carbon mapping utilizing convolutional neural networks and Earth observation data, a case study in Bavaria state Germany. Geoderma 444, 116867.
Urbina-Salazar, D., Vaudour, E., Richer-de-Forges, A.C., Chen, S., Martelet, G., Baghdadi, N., Arrouays, D., 2023. Sentinel-2 and Sentinel-1 bare soil temporal mosaics of 6-year periods for soil organic carbon content mapping in Central France. Remote Sens. (Basel) 15 (9), 2410.
Wadoux, A.-M.-C., 2019. Using deep learning for multivariate mapping of soil with quantified uncertainty. Geoderma 351, 59–70.
Walpole, R.E., Myers, R.H., Myers, S.L., Ye, K., 1993. Probability and Statistics for Engineers and Scientists. Macmillan New York, p. 5.
Wang, J., Zhen, J., Hu, W., Chen, S., Lizaga, I., Zeraatpisheh, M., Yang, X., 2023a. Remote sensing of soil degradation: progress and perspective. Int. Soil Water Conserv. Res. 11 (3), 429–454.
Wang, S., Xu, L., Zhuang, Q., He, N., 2021. Investigating the spatio-temporal variability of soil organic carbon stocks in different ecosystems of China. Sci. Total Environ. 758, 143644.
Wang, X., Li, S., Wang, L., Zheng, M., Wang, Z., Song, K., 2023b. Effects of cropland reclamation on soil organic carbon in China's black soil region over the past 35 years. Glob. Chang. Biol. 29 (18), 5460–5477.
Wang, X., Wang, L., Li, S., Wang, Z., Zheng, M., Song, K., 2022. Remote estimates of soil organic carbon using multi-temporal synthetic images and the probability hybrid model. Geoderma 425, 116066.
Wang, Y., Chen, S., Hong, Y., Hu, B., Peng, J., Shi, Z., 2023c. A comparison of multiple deep learning methods for predicting soil organic carbon in Southern Xinjiang, China. Comput. Electron. Agric. 212, 108067.
Wang, Z., Wu, W., Liu, H., 2024. Spatial estimation of soil organic carbon content utilizing PlanetScope, Sentinel-2, and Sentinel-1 data. Remote Sens. (Basel) 16 (17), 3268.
Williams, P., 1987. Interpretation of statistical evaluation of NIR analysis. Variables affecting near-infrared reflectance spectroscopic analysis. In: Near-Infrared Technology in the Agricultural and Food Industries, pp. 146–148.
Wu, M., Dou, S., Lin, N., Jiang, R., Zhu, B., 2023. Estimation and mapping of soil organic matter content using a stacking ensemble learning model based on hyperspectral images. Remote Sens. (Basel) 15 (19), 4713.
Wu, F., Tan, K., Wang, X., Ding, J., Liu, Z., 2023a. A novel semi-empirical soil multi-factor radiative transfer model for soil organic matter estimation based on hyperspectral imagery. Geoderma 437, 116605.
Xiao, X., He, Q., Ma, S., Liu, J., Sun, W., Lin, Y., Yi, R., 2024. Environmental variables improve the accuracy of remote sensing estimation of soil organic carbon content. Sci. Rep. 14 (1), 18964.
Yang, F., Zhang, G.-L., Yang, J.-L., Li, D.-C., Zhao, Y.-G., Liu, F., Yang, R.-M., Yang, F., 2014. Organic matter controls of soil water retention in an alpine grassland and its significance for hydrological processes. J. Hydrol. 519, 3086–3093.
Yang, R.-M., Huang, L.-M., Zhang, X., Zhu, C.-M., Xu, L., 2023. Mapping the distribution, trends, and drivers of soil organic carbon in China from 1982 to 2019. Geoderma 429, 116232.
Yuan, J., Gao, J., Yu, B., Yan, C., Ma, C., Xu, J., Liu, Y., 2024. Estimation of soil organic matter content based on spectral indices constructed by improved Hapke model. Geoderma 443, 116823.
Yuan, J., Wang, X., Yan, C.-X., Wang, S.-R., Ju, X.-P., Li, Y., 2019. Soil moisture retrieval model for remote sensing using reflected hyperspectral information. Remote Sens. (Basel) 11 (3), 366.
Yuxin, T., Angelini, M.E., Yigini, Y., Luotto, I., 2024. Global black soil distribution. Front. Agric. Sci. Eng. 11 (2).
Zanaga, D., Van De Kerchove, R., Daems, D., De Keersmaecker, W., Brockmann, C., Kirches, G., Wevers, J., Cartus, O., Santoro, M., Fritz, S., 2022. ESA WorldCover 10 m 2021 v200.
Zha, Y., Gao, J., Ni, S., 2003. Use of normalized difference built-up index in automatically mapping urban areas from TM imagery. Int. J. Remote Sens. 24 (3), 583–594.
Zhai, H., Zhang, H., Zhang, L., Li, P., 2018. Cloud/shadow detection based on spectral indices for multi/hyperspectral optical remote sensing imagery. ISPRS J. Photogramm. Remote Sens. 144, 235–253.
Zhang, T., Li, Y., Wang, M., 2024a. Remote sensing-based prediction of organic carbon in agricultural and natural soils influenced by salt and sand mining using machine learning. J. Environ. Manage. 352, 120107.
Zhang, W.-C., Wan, H.-S., Zhou, M.-H., Wu, W., Liu, H.-B., 2022. Soil total and organic carbon mapping and uncertainty analysis using machine learning techniques. Ecol. Ind. 143, 109420.
Zhang, W., Luo, C., Meng, X., Zang, D., Zhang, X., Liu, H., 2024b. Predicting regional soil organic matter content utilizing conventional satellites: assessing the influence of temporal, spatial, and spectral disparities. Catena 237, 107821.
Zhang, Y., Runting, R.K., Webb, E.L., Edwards, D.P., Carrasco, L.R., 2021. Coordinated intensification to reconcile the 'zero hunger' and 'life on land' sustainable development goals. J. Environ. Manage. 284, 112032.
Zhao, Q., Shi, P., Li, P., Li, Z., Min, Z., Sun, J., Cui, L., Niu, H., Zu, P., Cao, M., 2023. Effects of vegetation restoration on soil organic carbon in the Loess Plateau: a meta-analysis. Land Degrad. Dev. 34 (7), 2088–2097.
Zhao, W., Efremova, N., 2023. Soil organic carbon estimation from climate-related features with graph neural network. arXiv preprint arXiv:2311.15979.
Zhou, T., Lv, W., Geng, Y., Xiao, S., Chen, J., Xu, X., Pan, J., Si, B., Lausch, A., 2023. National-scale spatial prediction of soil organic carbon and total nitrogen using long-term optical and microwave satellite observations in Google Earth Engine. Comput. Electron. Agric. 210, 107928.
Zhu, Z., Woodcock, C.E., 2014. Continuous change detection and classification of land cover using all available Landsat data. Remote Sens. Environ. 144, 152–171.