https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

Open Access Earth System Science Data Discussions

A 1 km soil organic carbon density dataset with depth of 20cm and 100cm from 1985 to 2020 in China

Yi Dong[1,2&], Xinting Wang[1,2&], Wei Su[1,2*]

[1] College of Land Science and Technology, China Agricultural University, Beijing 100083, China

[2] Key Laboratory of Remote Sensing for Agri-Hazards, Ministry of Agriculture and Rural Affairs, Beijing 100083, China

Correspondence to: Wei Su (suwei@cau.edu.cn)

&These authors contributed equally to this work and should be considered co-first authors.

Abstract: Soil organic carbon (SOC) is an important component of the worldwide carbon cycle as a vital indicator of soil quality and ecosystem health, with significant implications for agricultural production and climate change adaptation and mitigation strategies. Although there are some studies on mapping the spatial distribution of soil organic carbon density (SOCD), the long-time series SOCD products in China are still lacking. Therefore, this study proposed a new algorithm with climatic zoning, aiming to improve the accuracy of predicting SOC densities with depths of 0-20 cm and 0-100 cm from 1985 to 2020. The data sources used in this study include Landsat archives, topographic data, meteorological data, and measured SOCD data. The innovation lies in the zoning models by climate regions using a random forest ensemble learning approach for SOCD estimation in China. The predicted results show that our zoning model outperformed the global model without climate zoning in predicting SOCD with R^2=0.55 and RMSE=2.19 for 0-20 cm SOCD estimation and R^2=0.52 and RMSE=6.50 for 0-100 cm. Comparably, the SOCD estimation using the global model is with R^2=0.46 and RMSE=2.36 for 0-20 cm SOCD estimation and R^2=0.44 and RMSE=8.09 for 0-100 cm. Moreover, our 0-20 cm SOCD predictions align well with independent samples (R^2=0.69, RMSE=2.01) and are further validated with Xu's dataset (R^2=0.63, RMSE=1.82). Furthermore, the comparisons with the published SOC content products including HWSD, SoilGrids250m, and GSOCmap have also shown good consistency, too. Comparably, our predicted SOCD is the best fit with SoilGrids250m products with R^2=0.72 and RMSE=1.35. Comparisons of model predictions to independent datasets from the 1980s, 2000s, and 2010s in China reveal substantial connections and a trend of increasing forecast accuracy over time. The predicted SOCD is available via the Figshare (https://doi.org/10.6084/m9.figshare.27290310.v1) (Dong et al., 2024).

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

1 Introduction

Soils are important because they enable the movement of carbon, energy, and water (Chaney et al., 2019; Crow et al., 2012). The foundation of soil fertility lies in soil carbon, a significant component of terrestrial carbon storage. SOC accounts for more than half of total soil carbon and is an essential component of the soil carbon cycle, which has a major impact on soil fertility and agricultural productivity (Baldock, 2007; Chen et al., 2022). A combination of natural and human forces is placing significant strain on the global SOC reservoir. SOC content estimation has become a hot spot in global climate change due to its close relationship with climate change. The sustainability of agricultural production is threatened worldwide by soil degradation and the loss of intimate relationships. As a largely agricultural country, the distribution and changes of SOC content in China have significant impacts on the global carbon balance and play an important role in balancing global carbon emissions and sequestration (Xu et al., 2018). However, China's complicated geography and varied climate have resulted in significant regional differences in SOC, leading to the difficulties of SOC content estimation in China.

There has been more and more interest in global, national, and regional SOC content estimation in the last few years(Padarian et al., 2022). In-depth studies to estimate subsurface SOC content estimation, particularly at a regional scale, remain challenging due to the difficulty of data collection, the lack of long-term observations, and the depth dependency of soil carbon sequestration (Padarian et al., 2022). The advancement of digital soil mapping technology opens up new paths for estimating SOC content in large-scale and long-term series (Li et al., 2024). The use of machine learning techniques for digital soil modeling is a common concept in DSM. Compared to traditional mapping methods such as geo-statistics, expert knowledge, and individual representation, machine learning techniques provide a new paradigm for estimating SOC content in large-scale and long-term series. To produce continental-scale SOC-weighted mean maps, Odgers et al. (Odgers et al., 2012) used an equal-area spline function for soil databases, while Mulder et al.(Mulder et al., 2016) used a machine learning model with a three-dimensional distribution to estimate SOC content in eastern France. These studies provide evidence for a comprehensive and accurate understanding of soil properties and their spatial variation. Despite these advances, most digital soil mapping studies have focused on a specific period and the long-term dynamics of SOCD mapping have not yet been developed. Emadi et al. (Emadi et al., 2020) predicted the SOCD in northern Iran using a sample of 1879 measurements, and Nabiollahi et al. (Nabiollahi et al., 2019) used a random forest (RF) model to predict the SOCD at 137 sites in Marivan, Kurdistan Province,

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

Iran. However, these studies only focus on local zones. In China, researchers have paid considerable attention to the sequestration potential of SOC storage, but most studies have focused on specific experimental areas or ecosystem types. Fang et al. (Fang et al., 2007) estimated the carbon sink of terrestrial vegetation in China. Furthermore, these studies often lack attention to long-term trends and dynamics, resulting in insufficient data sets to fully understand climate change and the impact of human activities on SOCD. At the national level, there is relatively little study on the potential for organic carbon storage across different ecosystem types (O'Rourke et al., 2015). The scarcity and unevenness of SOC data in China, as well as the lack of effective estimation methods, all contribute to the uncertainty of SOC prediction. In addition, the diverse and complex topography in China, as well as the lack of measured SOCD data, have increased the difficulty of SOC content estimation. Previous studies often used the data from inventories of relevant resources to make rough calculations of carbon sinks (Pan et al., 2004). Unfortunately, the spatial continuity and variability of SOC, the spatial differentiation of organic carbon sequestration potential, and the influence of environmental factors have not been considered in previous studies. Especially in western China, there is almost no measured SOC data (Liu et al., 2022), which poses a challenge for understanding terrestrial ecosystems and soil carbon sinks in China. Given these challenges, it is urgent to carry out SOCD mapping and analyze the temporal and spatial changes of SOCD in China.

To produce robust and accurate long-term SOCD products in China, we explore the RF models with climate zoning to predict SOCD in China from 1985 to 2020 and improve the study of SOCD maps for the 0-20 cm and 0-100 cm soil layers in China. The Landsat TM/ETM+/OLI images, topography, meteorology, and soil properties data are used for SOCD mapping in this study. The main contributions of this study can be summarized as follows.

(1) A nationwide, long-term soil organic carbon density dataset from 1985 to 2020 with depths of 20cm and 100cm in China is provided in this study.

(2) The machine learning RF models zoned by climate zones in China are developed for SOCD estimation, and the spatial-temporal variability of soil carbon is considered in our SOCD estimation.

(3) The proposed framework provides a comprehensive understanding of SOCD estimation including spectral indices of satellite remote sensing images, digital elevation model (DEM) and its topographic derivatives, meteorological features, and soil properties. The technique offers the potential for SOCD mapping with sufficiently measured SOC content data.

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

2 Study area and data sources

2.1 Study area

The study area, which extends throughout China, is characterized by complex and diverse terrains including mountains, plateaus, basins, plains, and deserts (Yuan et al., 2023). In addition, China has a large latitude difference from 4°N to 53°N and a large longitude difference from 73°E to 135°E. Therefore, there are obvious differences in precipitation and temperature in the study area, which bring significantly different accumulation processes and spatial patterns of soil carbon (Zheng et al., 2023). In addition, there are various soil types, including red soil, brown soil, brown soil, black soil, and chestnut calcium soil, which have obvious spatial characteristics in the study area (Shangguan et al., 2014)(Shangguan et al., 2014). For these reasons, we developed four different RF models for SOCD estimation for four temperature zones from south to north in China including humid area, semi-humid area, semi-arid area and arid area.

2.2 Data sources

(1) SOC content data

There were 8203 measured SOC content samples in the 1980s, 2000s, and 2010s in China collected for model building and validation of SOCD estimation. The SOC content and soil mass weight data of the 1980s were collected from the profile database of the Second National Soil Census (1980-1996) (http://www.geodata.cn). The SOCD data of the 2000s was collected from the China Terrestrial Ecosystem Carbon Density Dataset (2000-2014) (http://www.cnern.org.cn/). The SOC content data of the 2010s was collected from the Soil Attribute Data of the China Soil System Record (2010s) (https://www.resdc.cn/), which was measured in the China Soil System Survey Collection and China Soil System Journal Compilation Project. To validate the SOCD estimation results in this study, two independent SOCD data were used, including the measured SOC content data in the Heihe River basin and the measured SOCD data from Song et al. (Song et al., 2016). The SOC content data of the Heihe River basin were collected from the spatio-temporal Tripolar Environmental Big Data Platform (https://poles.tpdc.ac.cn/zh-hans/). The measured SOCD data from Xu et al. (Xu et al., 2018) focuses on SOC densities and soil carbon storage with a depth of 0-20 cm in various terrestrial ecosystems in China. The data was measured in field campaigns between 2004 and 2014, as well as some unpublished field measurements.

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

(2) Landsat archives

The time-series archived Landsat 4, 5, 7, and 8 TM/ETM+/OLI images spanning from 1985 to 2020 (Yu et al., 2023) are used for SOCD estimation, which are retrieved from the GEE cloud computing platform (Liu et al., 2024). Preprocessing of Landsat images, including radiometric calibration, atmospheric correction, geometric correction, cloud identification, and spectral index calculating are carried out on the GEE cloud computing platform. Random sampling and statistical regression analysis are performed to determine the calibration coefficients for each band spectral reflectance. Principal major axis regression models are used to normalize the reflectance data for different sensors. Radiometric correction coefficients of different Landsat sensors are calculated (Fig. 1). The spatially overlapping images are combined into one image using the aggregation function, and the combined image dataset is subjected to stitching operations to produce spatially coherent images. A variety of spectral indices were calculated using Landsat images after processing. Spectral indices normalized difference vegetation index (NDVI), enhanced vegetation index (EVI), soil adjusted vegetation index (SAVI), and land surface water index (LSWI) were calculated using Landsat images. The formulae for these spectral indices are as follows:

$$NDVI = \frac{NIR - Red}{NIR + Red} \tag{1}$$

$$EVI = 2.5 * \frac{NIR - Red}{NIR + 6 * Red - 7.5 * Blue + 1} \tag{2}$$

$$SAVI = \frac{NIR - Red}{(NIR + Red + 0.5) * 1.5} \tag{3}$$

$$LSWI = \frac{NIR - SWIR1}{NIR + SWIR1} \tag{4}$$

Where, NIR stands for the near-infrared band, Red for the red band, Blue for the blue band, and SWIR1 for the short-wave infrared band 1.

This study used the land cover dataset newly released by Wuhan University (Yang and Huang, 2021). This is the first China Land Cover Annual Data Set (CLCD) derived from Landsat on the GEE platform.

(3) DEM and its topographic derivatives

Terrain is an important factor affecting the formation of soil organic matter. The DEM data is used for SOCD estimation, which is downloaded from the Resource and Environment Science Data Platform of the Chinese Academy of Sciences

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

(https://www.resdc.cn) with a spatial resolution of 500 m. Topographic data and its topographic derivatives are extracted from the DEM data. There are eight terrain derivatives, including slope, aspect, elevation, profile curvature (Pr_c), topographic wetness index (TWI), analytical hill-shading (AH), channel network base level (CNBL), and channel network distance (CND), which are calculated using SAGA GIS version 8.0.1 (https://saga-gis.org/) (Zhang et al., 2023). The spatial resolution of all raster data was uniformly adjusted to 1000m using resampling techniques to achieve spatial consistency between different datasets.

(4) Meteorological data

The meteorological features including temperature (Tem), and precipitation (Pre) measured in 2,400 Chinese meteorological stations are used to quantify the effects of meteorological fluctuations. All meteorological data are downloaded from the China Meteorological Data Network (http://data.cma.cn/). For spatial consistency, the meteorological data is defined and projected into WGS 84 coordinates. All meteorological point data are interpolated into grid data with 1000 m spatial resolution using the ANUSPLIN program (Padarian et al., 2022).

(5) Published soil database

There are four published soil databases used to validate the SOCD estimation results in this study. One is the Harmonized World Soil Database (HWSD v2.0), produced by the International Institute for Applied Systems in Vienna and the Food and Agriculture Organization of the United Nations. There are two soil properties including soil bulk weight and organic carbon content are used for SOCD estimation at depths of 0-20 cm, 20-40 cm, 40-60 cm, 60-80 cm, and 80-100 cm. The SoilGrids250m v2.0 dataset including the soil silt content, sand content, clay content, and organic carbon content data with the spatial resolution of 250 m are downloaded from FAQ SoilGrids (https://soilgrids.org/) for validation. For spatial consistency, this soil attribute datum is resampled to 1000 m. This soil product with five depth intervals (5, 15, 30, 60, and 100 cm) is used to calculate the soil silt content (Silt), sand content (Sand), clay content (Clay), and organic carbon at 0-20 cm and 0-100 cm (Zhang et al., 2023). Taking the clay content data as an example, the clay content with depths of 0-20 cm and 0-100 cm is calculated as follows:

$$CLY_{020} = \frac{CLY_{05}}{4} + \frac{CLY_{515}}{2} + \frac{CLY_{1530}}{4} \quad (5)$$

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

Open Access Earth System Science Data Discussions

$$CLY_{0100} = \frac{CLY_{05}}{20} + \frac{CLY_{515}}{10} + \frac{3}{20} * CLY_{1530} + \frac{3}{10} * CLY_{3060} + \frac{2}{5} * CLY_{60100} \quad (6)$$

Where, CLY_{05}, CLY_{515}, CLY_{1530}, CLY_{3060}, and CLY_{60100} are the clay content (g/kg) at depths of 0-5, 5-15, 15-30, 30-60, and 60-100 cm respectively.

The GSOCmap dataset (https://www.fao.org/), which is the first global SOC product led by FAO, is used for validation. GSOCmap is a 1-kilometer soil grid that covers depths ranging from 0 to 30 centimeters. The SOC Dynamics ML dataset in China is now available on the Dryad platform (https://datadryad.org/). Using machine learning, the dataset aims to capture the dynamics of SOC and its drivers in different soil horizons in China between the 1980s and 2010s (Li et al., 2022). The dataset contains valuable information such as SOC stocks, carbon fixation rates, and SOC content. The organic carbon density with the depth of 20 cm and 100 cm in the 1980s, 2000s, and 2010s in China is used. This study focuses on SOCD, which is different from SOC content. The conversion from SOC content to SOCD is presented in Section 3.1.

3 Methodology

3.1 Converting SOC to SOCD with normalized soil depth

The measured data in the 2000s is SOCD, while it is SOC content in the 1980s and 2010s. For consistency of these data, we convert the SOC content data to SOCD. The 1980s SOC content data were from the Second National Soil Census, and the 2010s SOC content data were from the Soil Attribute Data of China Soil System Record, which had several different soil depths. For the consistency of the measured data, we convert the soil data with different depths into the SOCD with the depth of 0-20 cm and 0-100 cm using the package "mpspline2" v.0.1.3 (Bishop et al., 1999). The default value of 0.1 was used for the spline smoothing parameter lambda. The SOCD (kg C/m^2) is calculated using bulk density (kg/m^3), and coarse fractions percentage (%) provided by SoilGrids 2.0 (Poggio et al., 2021; Zhang et al., 2023).

$$\text{SOCD} = \left\{\frac{(SOC \times BD \times SD)}{100}\right\} \times \left\{1 - \frac{CF}{100}\right\} \quad (7)$$

Where SOC is soil organic carbon content (%), SOCD is soil organic carbon density, BD is the soil bulk density, SD is the soil depth (cm), and CF is the coarse fractions proportion in a specific soil layer.

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

3.2 Feature optimization for RF modelling

Several factors affect the SOCD estimation, including soil properties, climate, topographical factors, vegetation cover, and land use type (Chen et al., 2020; Sun et al., 2019; Wu et al., 2022). To build more accurate SOCD estimation models, we optimize the features through the correlation between different features and the sensitive features should be used for building the SOCD estimation model. The correlation between two continuous variables can be determined statistically using the Pearson correlation coefficient method (Zhang et al., 2019). Its value ranges from -1 to 1, here +1 indicates a perfect positive correlation -1 indicates a perfect negative correlation, and 0 indicates no correlation between the analyzed features. The Pearson correlation coefficient is used to determine which variables are highly correlated with each other. The correlation would be determined using a 0.05 significance level. The features with significant correlation and the absolute value of correlation coefficients greater than 0.8 were removed (Jiarpakdee et al., 2020). The Pearson correlation coefficient between the target variable (i.e., SOCD) and environmental variables (i.e., topographic, meteorological, spectral features) was determined by calculating the relevance between the target variable and the environmental variables and eliminating the self-correlation on the diagonal (Fig. 2). Feature optimization is performed to remove insensitive variables, reduce model complexity, and improve model generalization ability. Twelve characteristic variables were selected from 19 environmental factors, including Clay, Sand, Temperature, Precipitation, NDVI, Elevation, Slope, Aspect, TWI, CND, AH, and CLCD.

3.3 Climate zoning in China

Climate zoning is carried out to quantify the differences in temperature and precipitation in China and improve the accuracy of SOCD estimation. According to Tang et al. (Tang et al., 2018), there are obvious differences in SOCD observed in different zones of China for the diverse and complex environmental factors under warm-temperate climate conditions with a mean precipitation (MAP) threshold of 400 mm and a mean annual temperature (MAT) threshold of 10 °C. To mitigate the interannual variability, the multi-annual average temperature and precipitation are used to classify the climatic differences in China into four subzones including humid areas (MAP $\geqslant$ 400 mm and MAT $\geqslant$ 10°C), semi-humid area (MAP $\geqslant$ 400 mm and MAT $\leqslant$ 10°C), semi-arid area (MAP $\leqslant$ 400 mm and MAT $\leqslant$ 10°C) and arid area (MAP $\leqslant$ 400 mm and MAT $\geqslant$ 10°C) (Fig. 3). Soil data and environmental variables are grouped in each subzone, and zonal SOCD estimation models are developed for each subzone with depths of 0-20 cm and 0-100 cm (Fig. 3).

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

3.4 SOCD estimation using zoned RF models

According to the four climate subzones, the SOCD estimation models are developed in arid, semi-arid, humid, and semi-arid zones with depths of 0-20 cm and 0-100 cm, respectively. Within each subzone, a corresponding random forest tree is used to generate and aggregate the output of numerous decision trees to increase forecast stability and accuracy (Wu et al., 2021). Each decision tree can help prevent overfitting to some extent because it is developed using samples and features from various random subsets of the input features (Sun et al., 2024). Nevertheless, merging or averaging the prediction results from each decision tree significantly improves the generalization ability of the model. The correlation between SOCD values and the optimized sensitive environmental covariates is used to build a separate random forest model in each subzone, and SOCD was estimated to use combined models of predictor variables. The inputs of the random forest model are defined by the machine learning library Scikit-Learn based on the division of the samples into four climate zones. Parameters include the number of trees, the percentage of randomly selected features, and the maximum depth of the tree. The param_dist dictionary defines the range of hyperparameters of the random forest model to be tuned. The goal of model tuning is to maximize the performance of the model in cross-validation.

The optimized random forest model is used to predict the SOCD in the entire study area by inputting the measured SOCD with spectral indices from soil properties, Landsat archives, topographic derivatives, and meteorological elements. Ten-fold cross-validation is used to evaluate the generalization ability of the model. These indicators, including the coefficient of determination (R^2) and the root mean square error (RMSE), are used to validate the performance of the random forest model for SOCD estimation. The trained model is saved using the joblib library. The estimations were then combined with a geographic coordinate system to create digital SOCD maps. The relationship between SOCD and optimized environmental variables could be explored.

4 Results and conclusions

4.1 Statistical analysis of sampling points

The descriptive statistical analysis of the measured SOCD is shown in Fig. 4. The SOCD value at 0-20 cm depth shows a range of 0.070 to 22.93 kg C/m² in the 1980s with an average SOCD of 4.12 kg C/m², showing a positive offset and sharp distribution

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

pattern. In the 2000s, average SOCD increased slightly to 4.30 kg C/m² and data variability increased with more measured samples. The number of samples reached its maximum in the 2010s, with the average density decreasing to 4.18 kg C/m², but the maximum reaching 26.58 kg C/m², suggesting that the skewness of the data distribution increased. For the SOCD value with a depth of 0-100 cm, the mean value in the 1980s was 13.23 kg C/m². In the 2000s, the average SOCD decreased significantly to 9.06 kg C/m² and the variability decreased. There is a significant increase in the maximum value of SOCD, with an average of 13.32 kg C/m² and a maximum of 132.92 kg C/m², with a more distorted data distribution and thicker tails (Fig. 4).

Figure 5 shows the geographical arrangement of SOCD data based on Whittaker biomes with depths of 20 cm and 100 cm in the 1980s, 2000s, and 2010s in China. The distribution of samples shows significant regional concentration and geographical variation, with most points concentrated in the northeastern plain, southwestern plateau, hilly zones, and southeastern coastal zones. There are fewer SOCD samples in northwestern China due to difficult human accessibility, lower vegetation cover, less human activity, and a dry environment. In terms of timing, there are fewer SOCD sample sites in the 1980s. The number of sampling sites increased in the 2000s, particularly in agriculturally developed and densely populated areas.

4.2 Model performance of SOCD estimation

To evaluate the model performance of SOCD estimation with depths of 0-20 cm and 0-100 cm, two indicators are used, including the coefficient of determination (R^2) and the root mean square error (RMSE). The variance of the dependent variable is calculated using the coefficient of determination and the discrepancy between model predictions and estimated results is assessed using the root mean square error (Fig. 6). The 0-20 cm SOCD prediction model has an accuracy of R^2=0.46 and RMSE=2.36 before zones and R^2=0.55 and RMSE=2.19 after zones, with R^2 increased 0.09 and RMSE decreased 0.17. The 0-100 cm SOCD prediction model has an accuracy of R^2=0.44 and RMSE=8.09 before zones and R^2=0.52 and RMSE=6.50 after zones, with R^2 increased by 0.08 and RMSE decreased by 1.59.

The R^2 values for the SOCD estimation model range from 0.43 to 0.59 at a depth of 0 to 20 cm, showing a moderate correlation between the estimated SOCD and measurements in different climate zones in China. The RMSE values for the SOCD estimation models range from 1.87 to 2.74. The R^2 values for the SOCD estimation model with a depth of 0-100 cm are between 0.50 and 0.54, showing a slightly worse correlation between the observed and estimated SOCD than that for the 0-20

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

cm soil layer. The RMSE values range from 3.17 to 8.57, indicating that the estimated SOCD in deeper soil layers is worse than that in the upper layer (Fig. 7). When performing SOCD estimation in different climate zones, the estimation accuracy of the semi-arid and semi-humid zones (R^2=0.57 and RMSE=1.94, R^2=0.59 and RMSE=2.74) exceeded that of the dry and humid zones (R^2=0.43 and RMSE=1.87, R^2=0.48 and RMSE=1.90) with a depth of 0–20 cm. This result can be explained by the fact that SOCD measurements are more evenly distributed in semi-arid and semi-humid zones, which reduces the influence of extreme values and allows a more accurate estimate. Estimation accuracy is improved in these zones because the environmental factors are more consistent with the measured SOCD used for model training. At a soil depth of 0-100 cm, the model performed slightly better in dry and semi-arid zones (R^2=0.54 and RMSE=3.17, R^2=0.53 and RMSE=5.03) than in wet and semi-humid regions (R^2=0.51 and RMSE=5.50, R^2=0.50 and RMSE=8.57). Soil moisture conditions in drylands are more extreme and change significantly, resulting in more sensitive and distinct responses of vegetation and soil microorganisms to water (Tietjen et al., 2010; Tripathi et al., 2024). This high sensitivity provides the model with a clearer prediction signal, which improves the model's prediction accuracy. In contrast, soil moisture in wet areas is influenced by multiple factors, including precipitation, temperature, and vegetation cover (Lozano-Parra et al., 2018), and these complex factors reduce the predictive power of the model. Our SOCD estimation models capture this feature of SOC accumulation successfully.

Importance analysis of optimized features for SOCD estimation is performed to better understand the contribution of various environmental variables to SOCD estimation. Variables such as temperature and precipitation had the greatest influence on the SOCD estimation at both the 0-20 cm and 0-100 cm depths (Fig. 8). The influence of temperature on SOCD is mainly in its effect on soil microbial activity and respiration (Huang et al., 2024). An increase in temperature can accelerate the decomposition of soil organic carbon, but it may also increase the rate of decomposition of plant residues, thereby increasing the amount of carbon returned to the soil (Ofiti et al., 2021). Precipitation directly affects soil water status, and suitable soil water content is conducive to SOC accumulation (Cheng et al., 2020). When the moisture content is low, the decomposition of SOC is accelerated, thus reducing SOCD. In addition to these basic variables, some other variables significantly influence the SOCD estimation. Elevation and NDVI are also significant variables for SOCD estimation at both the 0-20 cm and 0-100 cm depths. Elevation is associated with the vertical distribution of surface hydrothermal conditions, which affects the soil formation process and organic carbon distribution properties (Wang et al., 2023). The importance of NDVI at both depths

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

underscores the critical role of vegetation health and productivity in contributing to SOCD. Vegetation cover influences the amount of organic matter returned to the soil through litterfall, and the activity of roots affects soil moisture and nutrient cycling (Chen et al., 2023). All of these factors contribute to the accumulation and storage of SOC. Climatic factors, topographic conditions, vegetation coverage, and soil physicochemical properties are the important factors determining SOCD estimation.

4.3 Validation with independent sample points

The SOCD estimation result is validated against independent SOC content data measured in the Heihe River basin and the simulated SOCD data by Li et al. (Li et al., 2022). The Heihe River basin is a major ecological and agricultural zone in northwest China. There are special geographical and climatic characteristics for the soil carbon accumulation in the Heihe River basin, which are important for exploring soil quality in arid and semi-arid zones. Validation is carried out by comparing measured data in the Heihe River basin with the estimated SOCD in this study. The comparison results show that our estimated SOCD is highly consistent with the measured SOC data from the Heihe River basin (Fig. 9). The estimated SOCD and the measured SOCD have a significant correlation, which is shown by the R^2 value of 0.69, and the RMSE value of 2.01 for the estimated result with the depth of 0-20 cm. Additionally, the proposed model demonstrates superior accuracy compared to Li's dataset, which reported an R^2 of 0.60 and an RMSE of 2.27.

The data reported by Xu et al. (Xu et al., 2018) created data on carbon storage of terrestrial ecosystems in China with a depth of 0-20 cm, which are used as independent samples for validation to solve the above problems of the Heihe River basin, such as the small distribution range of sample sites for validation and the small number of samples. These soil samples are widely distributed across the southern Tibet Autonomous Region, Qinghai Province, and eastern Inner Mongolia Autonomous Region. This is very good evidence for validating the robustness, reliability, and generalizability of the SOCD estimation model in this study. The estimated SOCD results are compared with the measured SOC data in the field campaign of Xu et al. (Xu et al., 2018). In addition, the field data were compared with 0-20 cm organic carbon density maps generated by a machine learning analysis dataset of SOC dynamics and their drivers in China during 2000-2014. The results of the comparative analysis are encouraging and show high agreement between the estimated SOCD using our developed model and the measured SOC data. Specifically, the R^2 value is 0.63 and the RMSE value is 1.82, which further confirms the accuracy of our SOCD estimation

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

model (Fig. 10). Furthermore, the model outperforms Li's dataset, which yielded an R^2 of 0.54 and an RMSE of 2.04, underscoring the enhanced predictive accuracy of our approach.

4.4 Comparison with published SOCD products

The 1-km-resolution SOCD dataset of China is created in this study, which is compared with the published SOCD products including HWSD v2.0, SoilGrids 250m, and GSOCmap datasets to validate and confirm its accuracy and reliability. The comparison results shown in Fig. 11 show that our produced 1-km-resolution SOCD dataset is largely consistent with published SOCD products, with the highest fit to SoilGrids250m and an R^2 of 0.72, significantly better than that of 0.62 with GSOCmap dataset and that of 0.51 with HWSD dataset. The HWSD v2.0 dataset is jointly published by the Food and Agriculture Organization of the United Nations (FAO) and the International Institute for Applied Systems (IIAS) in Vienna, which provides soil data on a global scale. Unfortunately, its applicability and accuracy are limited in China. The correlation of our SOCD dataset with HWSD is reported with the R^2 value of 0.51 and the RMSE value of 1.93. The GSOCmap dataset is led by the FAO and is intended to cover various ecosystems around the world. This is the first global SOC map. The correlation of our SOCD dataset with GSOCmap is reported with the R^2 of 0.62 and the RMSE of 1.57. The SoilGrids250m dataset is created using ISRIC's digital soil mapping technology, which is a global soil dataset. The correlation of our SOCD dataset with SoilGrids250m is reported with the R^2 of 0.72 and the RMSE of 1.35. Models are more accurate and applicable than global soil databases in capturing SOCD changes in China. This study highlights the need to create and implement region-specific models that utilize current geographic and environmental data to provide a more precise tool for accurately estimating soil carbon reserves.

For time series estimation accuracy, the estimated SOCDs in China are compared with the SOC Dynamics ML dataset in China in the 1980s, 2000s, and 2010s (Fig. 12). The comparison results show that there are significant correlations between estimated SOCDs and measured data with RMSE of 1.80, 1.51 and 1.52 and R^2 of 0.65, 0.69 and 0.67 in the 1980s, 2000s and 2010, respectively. The performance improvement in the later period is mainly due to the increased sample points. With more sample data available, the model has captured the spatial heterogeneity of SOCD more accurately. These comparisons confirm the robustness of the SOCD estimation model in this study and its potential to provide accurate estimates of SOCD. The

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

improvement over time highlights the importance of integrating current data and advanced analytical methods into soil carbon studies.

4.5 Spatiotemporal changing of SOCD in China

The SOCD changes over time from the 1980s to the 2010s are validated in Fig. 13 compared with the published investigations. Fig. 13 reveals that our estimated SOCD results with depths of 0-20 cm (a) and 0-100 cm (b) are falling in the value range of the previous investigations of Ni (Ni, 2001), Wu et al. (Wu et al., 2003), Wang et al. (Wang et al., 2004), Xu et al. (Xu et al., 2018), Wang et al. (Wang et al., 2021), Li et al. (Li et al., 2022), Zhang et al. (Zhang et al., 2023). We can find that SOCD in China has slightly upward increasing from the 1980s to the 2010s in the 0-20 cm topsoil (Fig. 13a). This may be resulted from that the topsoil is more susceptible to the direct effects of soil management practices and environmental changes. However, the estimated SOCD shows an increasing from 1980s to 1990s and keeps stable from 1990s to 2020s in the 0-100 cm deep soil (Fig. 13b). Fig. 14 and Fig. 15 show the spatiotemporal distributions of the estimated SOCD at the 5-year interval from the 1980s to the 2010s. And the regions with high SOCD value in depth of 0-20 cm are in northeast and southwest China with red color in Fig. 14. Comparablely speaking, there are the largest area of high SOCD value labeled dark red color bar in period of 2010-2015 (Fig. 14f). From the perspective of longitude, the SOCD distribution shows different pattern, and it is homogeneous in high and low longitudes where the land cover is forest mostly. Conversely, the variance of SOCD is higher in mid-longitude regions where is with distinct land cover types. Similarly, the regions with high SOCD value in depth of 0-100 cm are in northeast and southwest China with green color in Fig. 15. And there are smaller variance of SOCD in high and low longitudes, and there are higher variance of SOCD in mid-longitude regions. There are many driving factors for the changing of SOCD in China. Targeted monitoring and management practices should be implemented for SOCD trends at different soil depths to maximize soil carbon sequestration and continuously improve soil quality.

5 Data availability

The 1 km soil organic carbon density dataset with depths of 20cm and 100cm from 1985 to 2020 in China is currently freely available at https://doi.org/10.6084/m9.figshare.27290310.v1 (Dong et al., 2024). The data can be imported into remote sensi ng processing software (e.g., ENVI), standard geographical information system software (e.g., ArcGIS).

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

6 Conclusions

In this study, a SOCD dataset with a resolution of 1-kilometer resolution and soil depths of 0-20 cm and 0-100 cm is created from 1985 to 2020 in China. The accuracy and validity of this dataset are validated by two independent metrics and data and four types of published global products. The conclusions are as follows.

(1) The delineation of climatic zones for SOCD estimation modeling has been proven useful for enhancing the precision of the models and effectively addressing the uneven distribution of measured SOC.

(2) Comparison with independently measured data in the Heihe River basin shows that the estimated SOCD agrees with the measured data with R^2=0.69 and RMSE=2.01. The accuracy of the SOCD estimation model is corroborated by validation against the dataset of Xu with R^2=0.63 and RMSE=1.82.

(3) Compared to published global products including HWSD, SoilGrids250m, and GSOCmap, the estimated SOCD in this study was consistent and accurate. Comparison with the SoilGrids250m dataset shows the superiority of zoning RF models in capturing variations in SOCD in China with R^2=0.72 and RMSE=1.35.

(4) The model demonstrated excellent correlation with time series datasets and increased accuracy over time by comparing with the independently measured data from the 1980s, 2000s, and 2010s. This highlights the importance of integrating current data and advanced analytical methods into soil carbon studies. In addition, time series analyses showed the change of SOCD in China over time and at different soil depths, which can be influenced by many reasons such as agricultural management practices, land-use changes, and climate change.

Despite the impressive results of this study, more soil data are required to validate and improve the SOC estimation model. Future studies will focus on the effects of different land management strategies on SOC change as well as the development of more refined models for estimating soil organic carbon. Furthermore, given the uncertainties in existing global SOC estimates, we urge that future research focus on standardized soil sampling, cross-dataset comparisons, more validation, and global collaboration to improve the accuracy of SOC estimates.

Author contributions

Yi Dong and Xinting Wang designed the research, performed the analysis, and wrote the paper; Wei Su revised the manuscript.

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

Open Access Earth System Science Data Discussions

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgments

This study was supported by the National Natural Science Foundation of China under the project (No. 42171331 & No. 42471402), the 2115 Talent Development Program of China Agricultural University.

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

References

Baldock, J. A.: Composition and cycling of organic carbon in soil, in: Nutrient cycling in terrestrial ecosystems, edited, Springer, 1-35, 2007.

Bishop, T., Mcbratney, A. B., and Laslett, G. M.: Modelling soil attribute depth functions with equal-area quadratic smoothing splines, Geoderma, 91, 27-45, https://doi.org/10.1016/S0016-7061(99)00003-8, 1999.

Chaney, N. W., Minasny, B., Herman, J. D., Nauman, T. W., Brungard, C. W., Morgan, C. L., Mcbratney, A. B., Wood, E. F., and Yimam, Y.: Polaris soil properties: 30-m probabilistic maps of soil properties over the contiguous united states, Water Resour. Res., 55, 2916-2938, 2019.

Chen, J., Biswas, A., Su, H., Cao, J., Hong, S., Wang, H., and Dong, X.: Quantifying changes in soil organic carbon density from 1982 to 2020 in chinese grasslands using a random forest model, Front. Plant Sci., 14, 1076902, 2023.

Chen, S., Arrouays, D., Mulder, V. L., Poggio, L., Minasny, B., Roudier, P., Libohova, Z., Lagacherie, P., Shi, Z., and Hannam, J.: Digital mapping of globalsoilmap soil properties at a broad scale: a review, Geoderma, 409, 115567, 2022.

Chen, Y., Li, Y., Wang, X., Wang, J., Gong, X., Niu, Y., and Liu, J.: Estimating soil organic carbon density in northern china's agro-pastoral ecotone using vis-nir spectroscopy, J. Soils Sediments, 20, 3698-3711, 2020.

Cheng, R. R., Chen, Q. W., Zhang, J. G., Shi, W. Y., Li, G., and Du, S.: Soil moisture variations in response to precipitation in different vegetation types: a multi-year study in the loess hilly region in china, Ecohydrology, 13, e2196, 2020.

Crow, W. T., Berg, A. A., Cosh, M. H., Loew, A., Mohanty, B. P., Panciera, R., de Rosnay, P., Ryu, D., and Walker, J. P.: Upscaling sparse ground-based soil moisture observations for the validation of coarse-resolution satellite soil moisture products, Rev. Geophys., 50, 2012.

Dong, Y., Wang, X., and Su, W.: A 1 km soil organic carbon density dataset with depth of 20cm and 100cm from 1985 to 2020 in china, edited, https://doi.org/10.6084/m9.figshare.27290310.v1, 2024.

Emadi, M., Taghizadeh-Mehrjardi, R., Cherati, A., Danesh, M., Mosavi, A., and Scholten, T.: Predicting and mapping of soil organic carbon using machine learning algorithms in northern iran, Remote Sens., 12, https://doi.org/10.3390/rs12142234, 2020.

Fang, J., Guo, Z., Piao, S., and Chen, A.: Terrestrial vegetation carbon sinks in china, 1981–2000, Science in China Series D: Earth Sciences, 50, 1341-1350, 2007.

Huang, K., Ma, Z., Xia, P., Lin, T., Zhang, Z., Jiang, X., Wang, X., and Huang, X.: Spatial pattern and controlling factors of soil organic carbon density in a typical karst province, china, Soil and Tillage Research, 242, 106160, 2024.

Jiarpakdee, J., Tantithamthavorn, C., and Treude, C.: The impact of automated feature selection techniques on the interpretation of defect models, Empir. Softw. Eng., 25, 3590-3638, 2020.

Li, H., Wu, Y., Liu, S., Xiao, J., Zhao, W., Chen, J., Alexandrov, G., and Cao, Y.: Decipher soil organic carbon dynamics and driving forces across china using machine learning, Glob. Change Biol., 28, 3394-3410, 2022.

Li, T., Cui, L., Wu, Y., Mclaren, T. I., Xia, A., Pandey, R., Liu, H., Wang, W., Xu, Z., and Song, X.: Soil organic carbon estimation via remote sensing and machine learning techniques: global topic modeling and research trend exploration, Remote Sens., 16, 3168, 2024.

Liu, F., Wu, H., Zhao, Y., Li, D., Yang, J., Song, X., Shi, Z., Zhu, A., and Zhang, G.: Mapping high resolution national soil information grids of china, Sci. Bull., 67, 328-340, 2022.

Liu, Z., Chen, G., Wen, Q., Yi, L., and Zhao, J.: Extraction of rocky desertification information based on multi-feature combination optimization and random forest algorithm: a case study of zhaotong city in yunnan province, Science of Soil and Water Conservation, 22, 95-105, 2024.

Lozano-Parra, J., Pulido, M., Lozano-Fondón, C., and Schnabel, S.: How do soil moisture and vegetation covers influence soil temperature in drylands of mediterranean regions? Water, 10, 1747, 2018.

Mulder, V. L., Lacoste, M., Richer-De-Forges, A. C., Martin, M. P., and Arrouays, D.: National versus global modelling the 3d distribution of soil organic carbon in mainland france, Geoderma, 263, 16-34, 2016.

Nabiollahi, K., Eskandari, S., Taghizadeh-Mehrjardi, R., Kerry, R., and Triantafilis, J.: Assessing soil organic carbon stocks under land-use change scenarios using random forest models, Carbon Manag., 10, 63-77, 2019.

Ni, J.: Carbon storage in terrestrial ecosystems of china: estimates at different spatial resolutions and their responses to climate change, Clim. Change, 49, 339-358, 2001.

Odgers, N. P., Libohova, Z., and Thompson, J. A.: Equal-area spline functions applied to a legacy soil database to create weighted-means maps of soil organic carbon at a continental scale, Geoderma, 189, 153-163, https://doi.org/10.1016/j.geoderma.2012.05.026, 2012.

Ofiti, N. O., Zosso, C. U., Soong, J. L., Solly, E. F., Torn, M. S., Wiesenberg, G. L., and Schmidt, M. W.: Warming promotes loss of subsoil carbon through accelerated degradation of plant-derived organic matter, Soil Biology and Biochemistry, 156, 108185, 2021.

O'Rourke, S. M., Angers, D. A., Holden, N. M., and Mcbratney, A. B.: Soil organic carbon across scales, Glob. Change Biol., 21, 3561-3574, 2015.

Padarian, J., Minasny, B., Mcbratney, A., and Smith, P.: Soil carbon sequestration potential in global croplands, Peerj, 10, e13740, 2022.

Padarian, J., Stockmann, U., Minasny, B., and Mcbratney, A. B.: Monitoring changes in global soil organic carbon stocks from space, Remote Sens. Environ., 281, 113260, 2022.

Pan, Y., Luo, T., Birdsey, R., Hom, J., and Melillo, J.: New estimates of carbon storage and sequestration in china's forests: effects of age–class and method on inventory-based carbon estimation, Clim. Change, 67, 211-236, 2004.

Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., and Rossiter, D.: Soilgrids 2.0: producing soil information for the globe with quantified spatial uncertainty, Soil, 7, 217-240, https://doi.org/10.5194/soil-7-217-2021, 2021.

Shangguan, W., Gong, P., Liang, L., Dai, Y., and Zhang, K.: Soil diversity as affected by land use in china: consequences for

soil protection, Scientific World Journal, https://doi.org/10.1155/2014/913852, 2014.

Song, X., Brus, D. J., Liu, F., Li, D., Zhao, Y., Yang, J., and Zhang, G.: Mapping soil organic carbon content by geographically weighted regression: a case study in the heihe river basin, china, Geoderma, 261, 11-22, 2016.

Sun, B., Wang, Y., Li, Z., Gao, W., Wu, J., Li, C., Song, Z., and Gao, Z.: Estimating soil organic carbon density in the otindag sandy land, inner mongolia, china, for modelling spatiotemporal variations and evaluating the influences of human activities, Catena, 179, 85-97, 2019.

Sun, Z., Wang, G., Li, P., Wang, H., Zhang, M., and Liang, X.: An improved random forest based on the classification accuracy and correlation measurement of decision trees, Expert Syst. Appl., 237, 121549, 2024.

Tang, X., Zhao, X., Bai, Y., Tang, Z., Wang, W., Zhao, Y., Wan, H., Xie, Z., Shi, X., and Wu, B.: Carbon pools in china's terrestrial ecosystems: new estimates based on an intensive field survey, Proceedings of the National Academy of Sciences, 115, 4021-4026, 2018.

Tietjen, B., Jeltsch, F., Zehe, E., Classen, N., Groengroeft, A., Schiffers, K., and Oldeland, J.: Effects of climate change on the coupled dynamics of water and vegetation in drylands, Ecohydrology: Ecosystems, Land and Water Process Interactions, Ecohydrogeomorphology, 3, 226-237, 2010.

Tripathi, I. M., Mahto, S. S., Kushwaha, A. P., Kumar, R., Tiwari, A. D., Sahu, B. K., Jain, V., and Mohapatra, P. K.: Dominance of soil moisture over aridity in explaining vegetation greenness across global drylands, Sci. Total Environ., 917, 170482, 2024.

Wang, S., Huang, M., Shao, X., Mickler, R. A., Li, K., and Ji, J.: Vertical distribution of soil organic carbon in china, Environ. Manage., 33, S200-S209, 2004.

Wang, S., Xu, L., Zhuang, Q., and He, N.: Investigating the spatio-temporal variability of soil organic carbon stocks in different ecosystems of china, Sci. Total Environ., 758, 143644, 2021.

Wang, Y., Liu, X., Lv, M., and Zhang, Z.: Mechanisms and influencing factors of hydrothermal processes in active layer soils on the qinghai-tibet plateau under freeze–thaw action, Catena, 220, 106694, 2023.

Wu, H., Guo, Z., and Peng, C.: Land use induced changes of organic carbon storage in soils of china, Glob. Change Biol., 9, 305-315, 2003.

Wu, J. Y., Lin, Y., Lin, K., Hu, Y. H., and Kong, G. L.: [Predicting prolonged length of intensive care unit stay via machine learning]., Beijing Da Xue Xue Bao. Yi Xue Ban = Journal of Peking University. Health Sciences, 53, 1163-1170, 2021.

Wu, Z., Liu, Y., Li, G., Han, Y., Li, X., and Chen, Y.: Influences of environmental variables and their interactions on chinese farmland soil organic carbon density and its dynamics, Land, 11, 208, 2022.

Xu, L., Yu, G., He, N., Wang, Q., Gao, Y., Wen, D., Li, S., Niu, S., and Ge, J.: Carbon storage in china's terrestrial ecosystems: a synthesis, Sci Rep, 8, 2806, 2018.

Xu, L., Yu, G., and He, N.: Changes of soil organic carbon storage in chinese terrestrial ecosystems from the 1980s to the 2010s, Acta Geographica Sinica, 73, 2150-2167, 2018.

Yang, J. and Huang, X.: 30 m annual land cover and its dynamics in china from 1990 to 2019, Earth System Science Data

Discussions, 2021, 1-29, 2021.

Yu, Z., Ma, R., Xu, J., Wang, Z., and Hu, M.: A dataset of the tsi of hulun lake in summer,1986-2020, China Scientific Data, 8, 21-25, 2023.

Yuan, J., Chen, W., and Zeng, J.: Spatio-temporal differentiation of cropland use change and its impact on cropland npp in china, Journal of Natural Resources, 38, 3135-3149, 2023.

Zhang, A., Dong, Z., and Kang, X.: Feature selection algorithms of airborne lidar combined with hyperspectral images based on xgboost, Chinese Journal of Lasers, 46, 2019.

Zhang, Z., Ding, J., Zhu, C., Wang, J., Ge, X., Li, X., Han, L., Chen, X., and Wang, J.: Historical and future variation of soil organic carbon in china, Geoderma, 436, https://doi.org/10.1016/j.geoderma.2023.116557, 2023.

Zhang, Z., Ding, J., Zhu, C., Wang, J., Li, X., Ge, X., Han, L., Chen, X., and Wang, J.: Exploring the inter-decadal variability of soil organic carbon in china, Catena, 230, https://doi.org/10.1016/j.catena.2023.107242, 2023.

Zheng, Z., Zhang, K., Shi, J., and Zhang, M.: Analysis of gnss water vapor detection accuracy and temporal sequence characteristics in different climate types in china, Science of Surveying and Mapping, 48, 68-77, 2023.

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

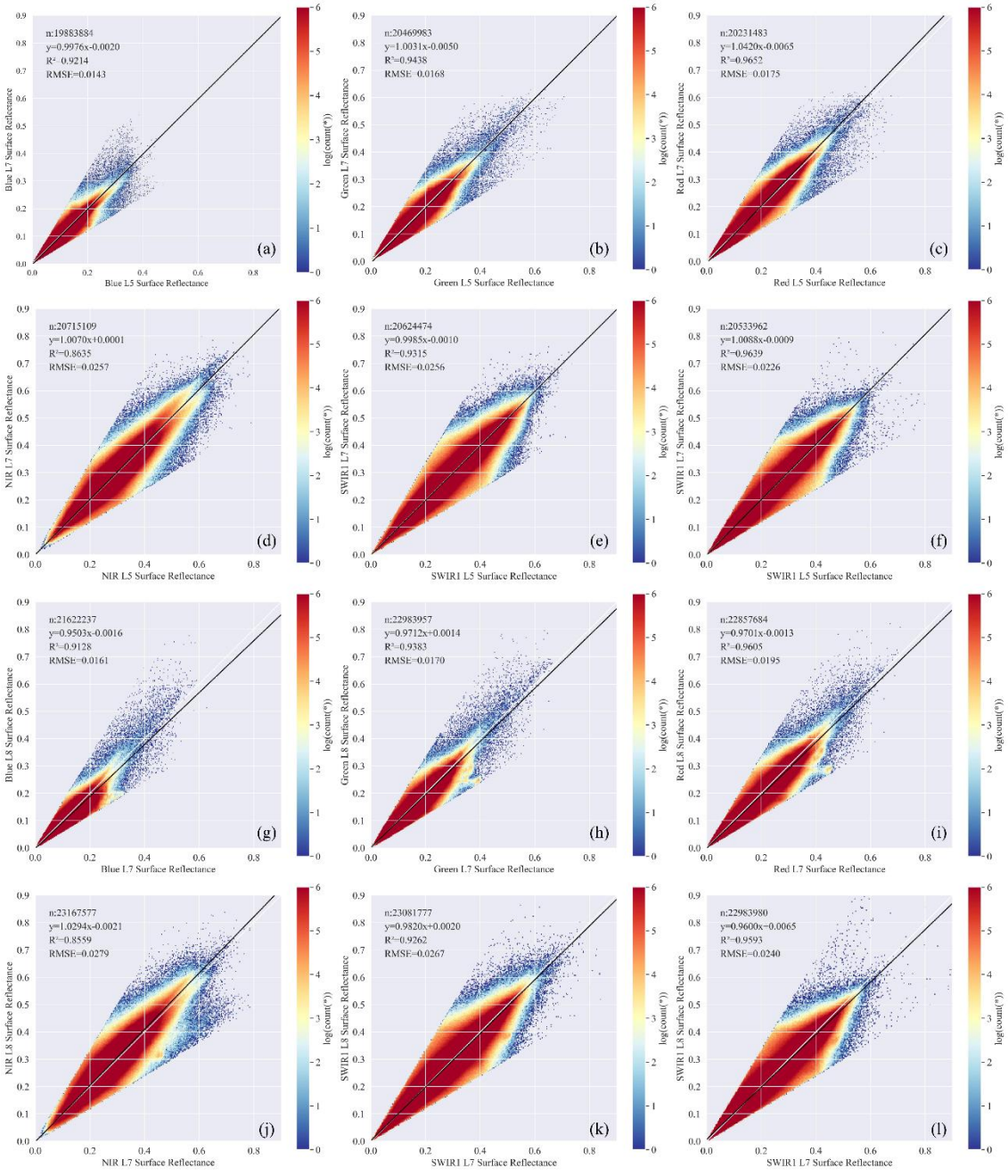


Figure 1. Radiometric normalization coefficients between Landsat 5 TM、Landsat 7 ETM+ (a-f) and Landsat 7 ETM+、Landsat 8 OLI (e-j) sensors for different bands including blue, green, red, NIR, SWIR1, and SWIR2. The radiometric normalization coefficients for each sensor are represented by fitted lines and correlation coefficients, indicating the correlation between the reference of different sensors, and characterizing the spectral response of the sensors in the different wavelength bands.

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

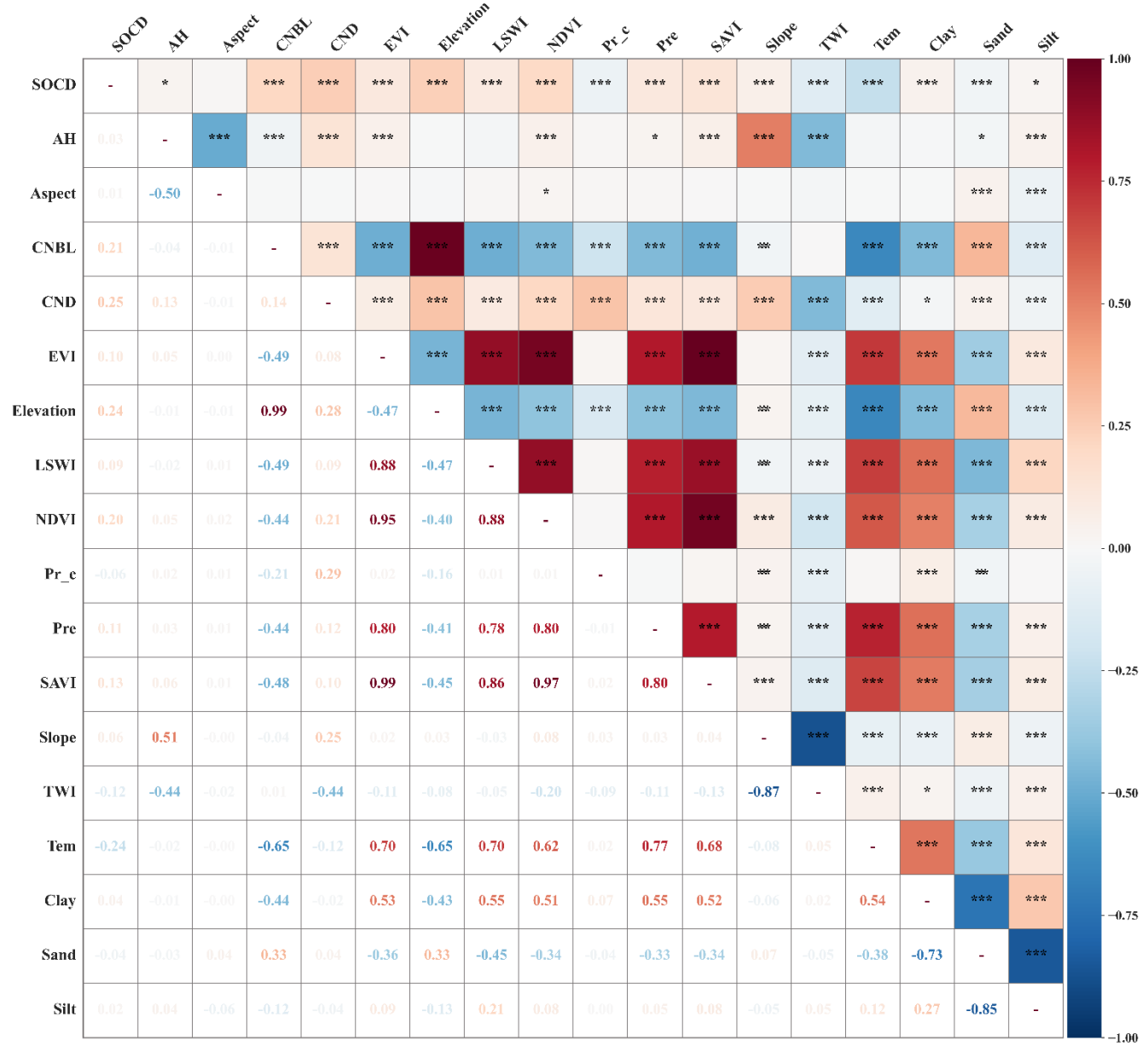


Figure 2. Correlation of optimized features for SOCD estimation.

Note: Correlations were significant at * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.01$. The number of asterisks indicates the statistical significance of the correlation. Warm hues like red typically suggest a positive correlation, whereas cold hues like blue typically indicate a negative connection. Color shades also indicate the magnitude of the correlation coefficient.

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

Open Access Earth System Science Data Discussions

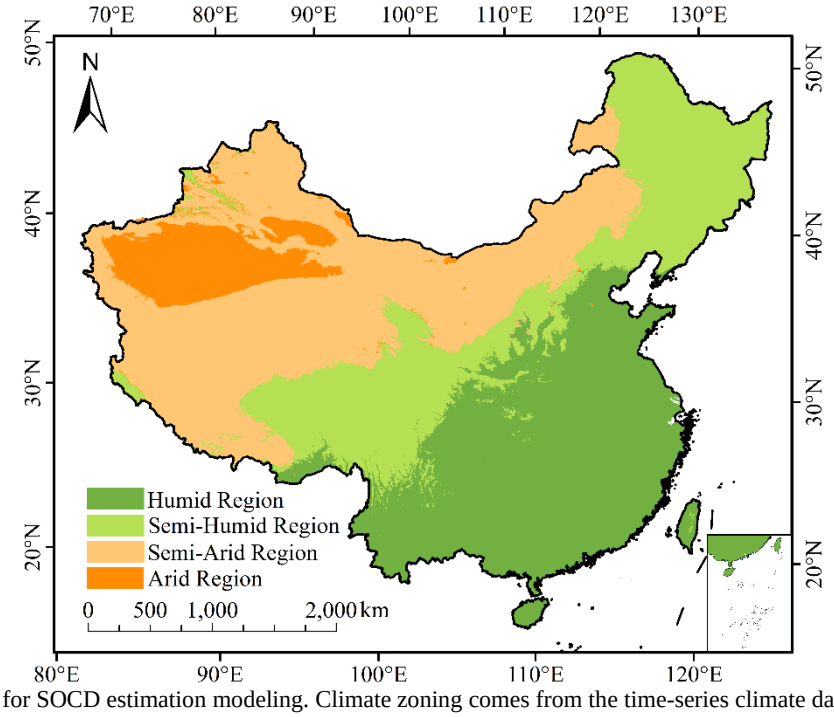


Figure 3. Climatic zones for SOCD estimation modeling. Climate zoning comes from the time-series climate data including temperature and precipitation. According to the difference in climate zones, it can be divided into humid, semi-humid, arid, and semi-arid zones.

490

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

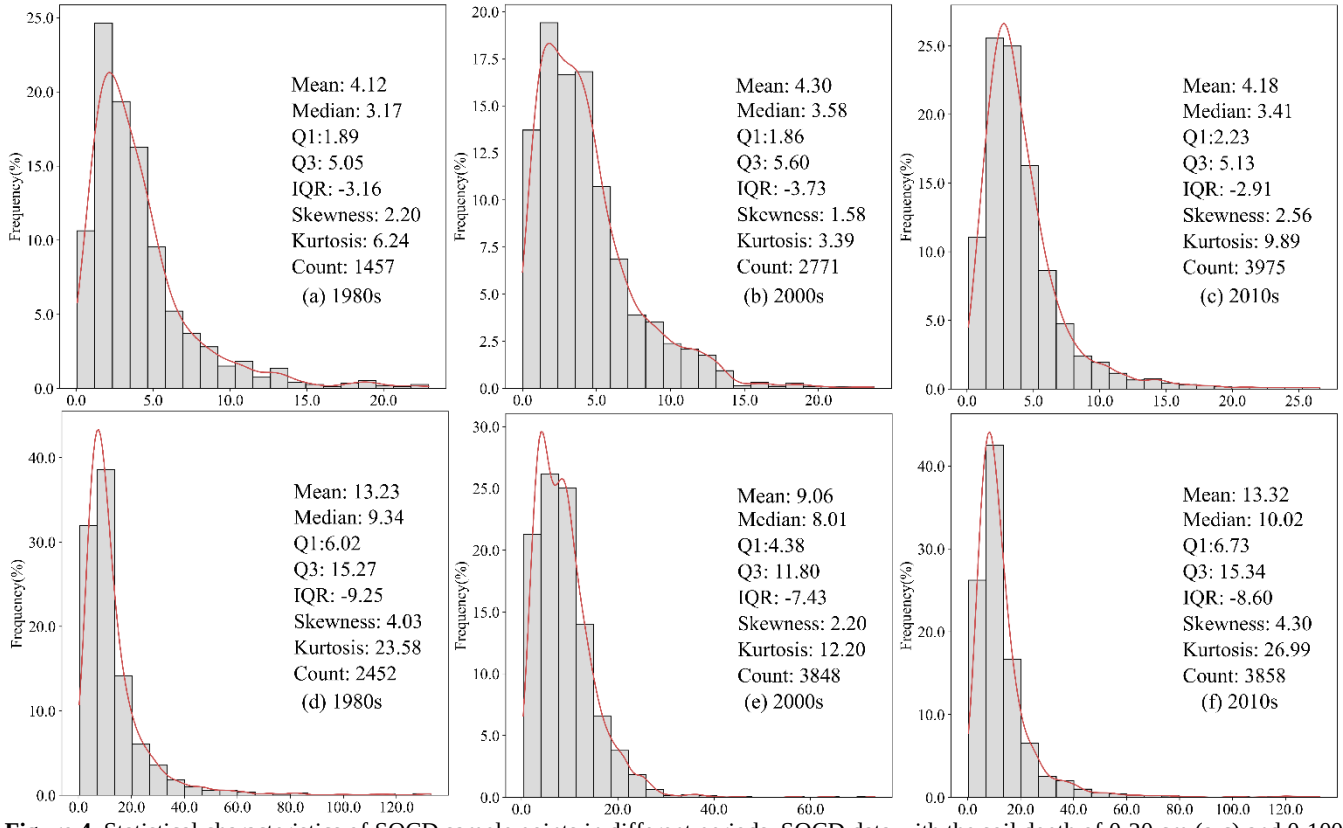


Figure 4. Statistical characteristics of SOCD sample points in different periods. SOCD data with the soil depth of 0-20 cm (a-c) and 0-100 cm (d-f) in China during the 1980s, 2000s, and 2010s are evaluated comprehensively.

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

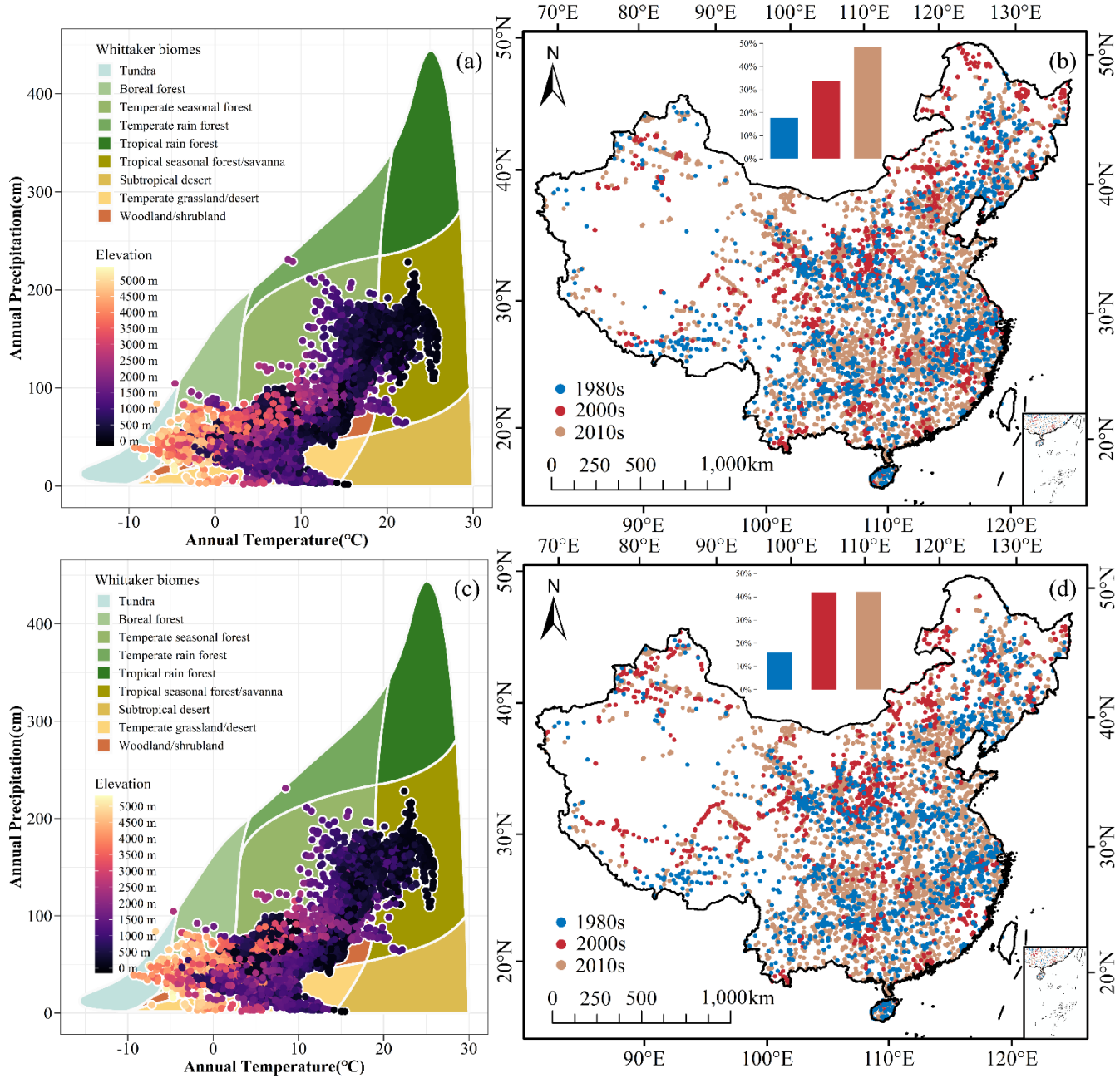


Figure 5. Spatial distribution of SOC sample points. (a-d) Distribution of the SOCD sampling sites with data used in this study for (b) top 20 cm soil and (d) top 100 cm soil. The distribution of site-level training data is based on Whittaker biomes for (a) top 20 cm soil and (c) top 100 cm soil.

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

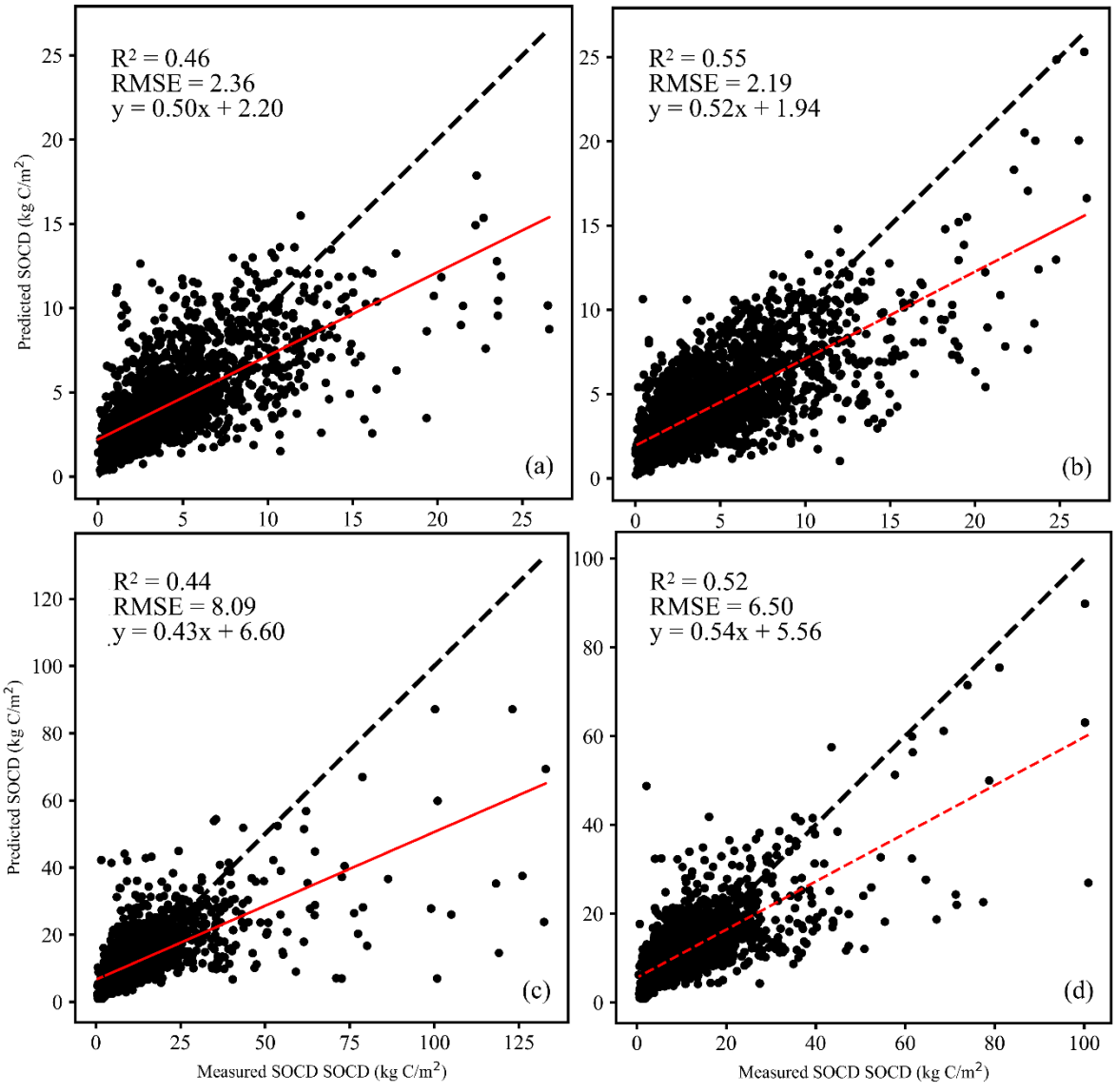


Figure 6. The model performance of global and zoning models with the depth of 0-20 cm and 0-100 cm. The SOCD prediction model of 0-20 cm and 0-100 cm soil depth is evaluated strictly by using a variety of statistical indicators, corresponding to four evaluation results, 0-20 cm global model (a), 0-20 cm regional model (b), 0-100 cm global model (c), and 0-100 cm regional model (d).

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

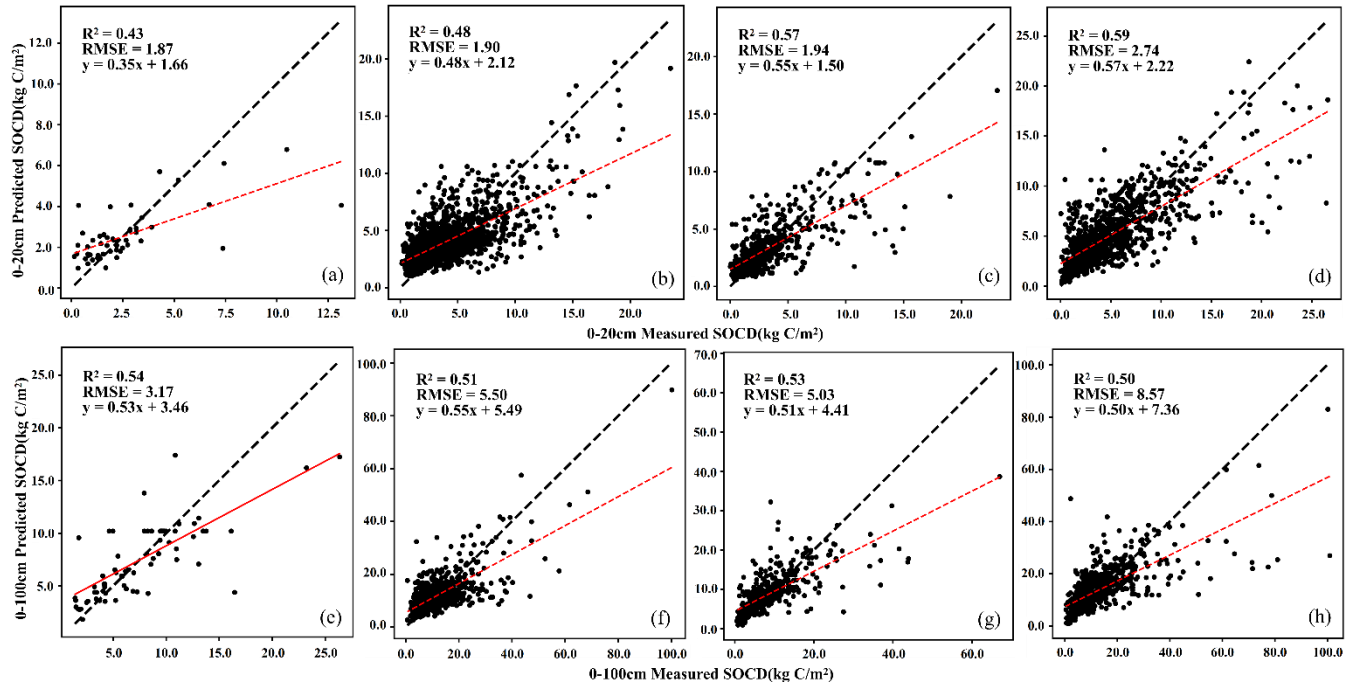


Figure 7. The model performance of different zoning models with the depth of 0-20 cm and 0-100 cm. Panels (a) and (e) depict the model performance for arid regions, where water scarcity is a predominant factor affecting SOCD. Panels (b) and (f) illustrate results for humid regions characterized by high moisture availability. Panels (c) and (g) showcase semi-arid regions, where the balance between precipitation and evaporation influences SOCD patterns. Finally, panels (d) and (h) display model accuracy in semi-humid regions, which exhibit intermediate conditions between arid and humid environments.

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

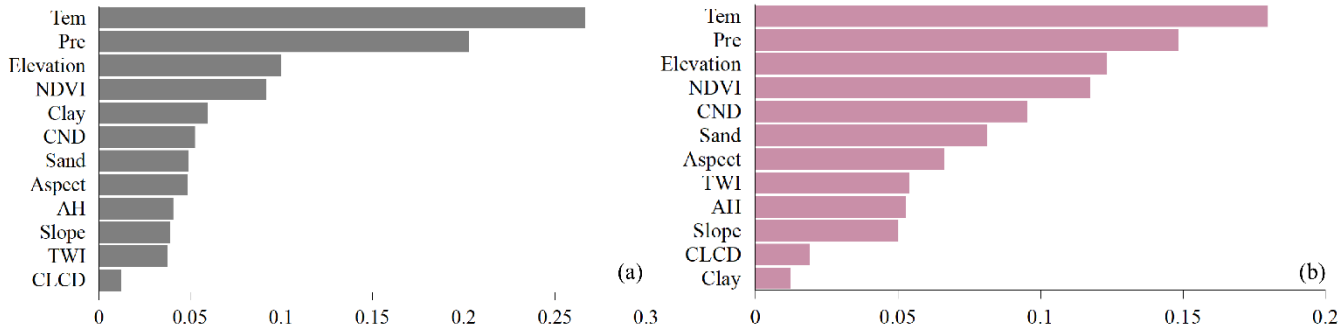


Figure 8. Importance ranking of features for SOCD estimation with the depth of 0-20 cm and 0-100 cm. It reports the contribution of different environmental variables to the SOCD estimation with different soil depths, including feature importance ranking for 0-20 cm depth (a) and feature importance ranking for 0-100 cm depth (b).

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

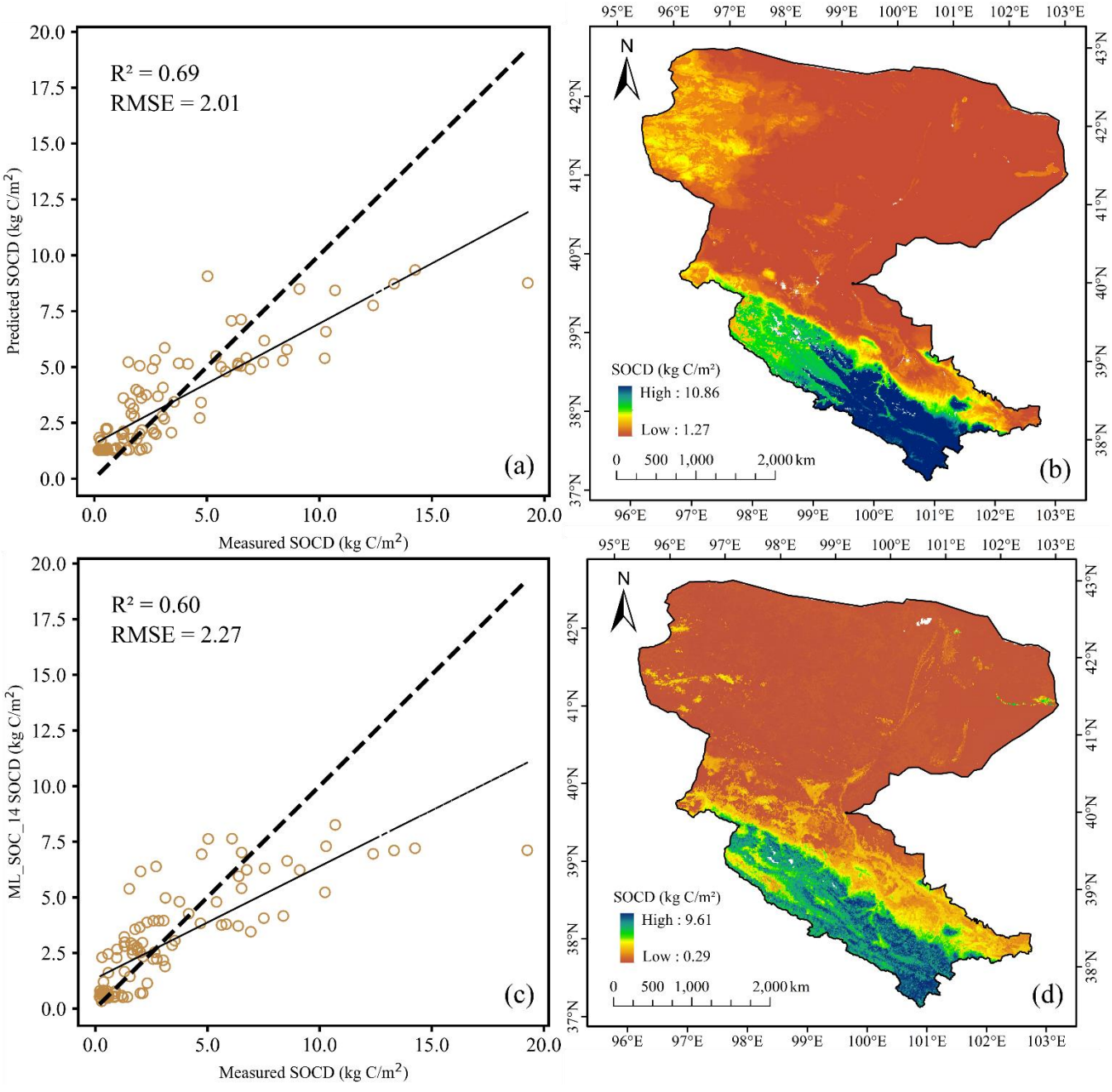


Figure 9. Comparison with the independently measured data in the Heihe River basin. (a) and (c) are the correlations of the estimated SOCD and SOC Dynamics ML dataset with the measured SOC for the soil with a depth of 0-20 cm, respectively. (b) and (d) are the spatial distribution of the estimated SOCD and SOC Dynamics ML dataset for the soil with a depth of 0-20 cm in this study.

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

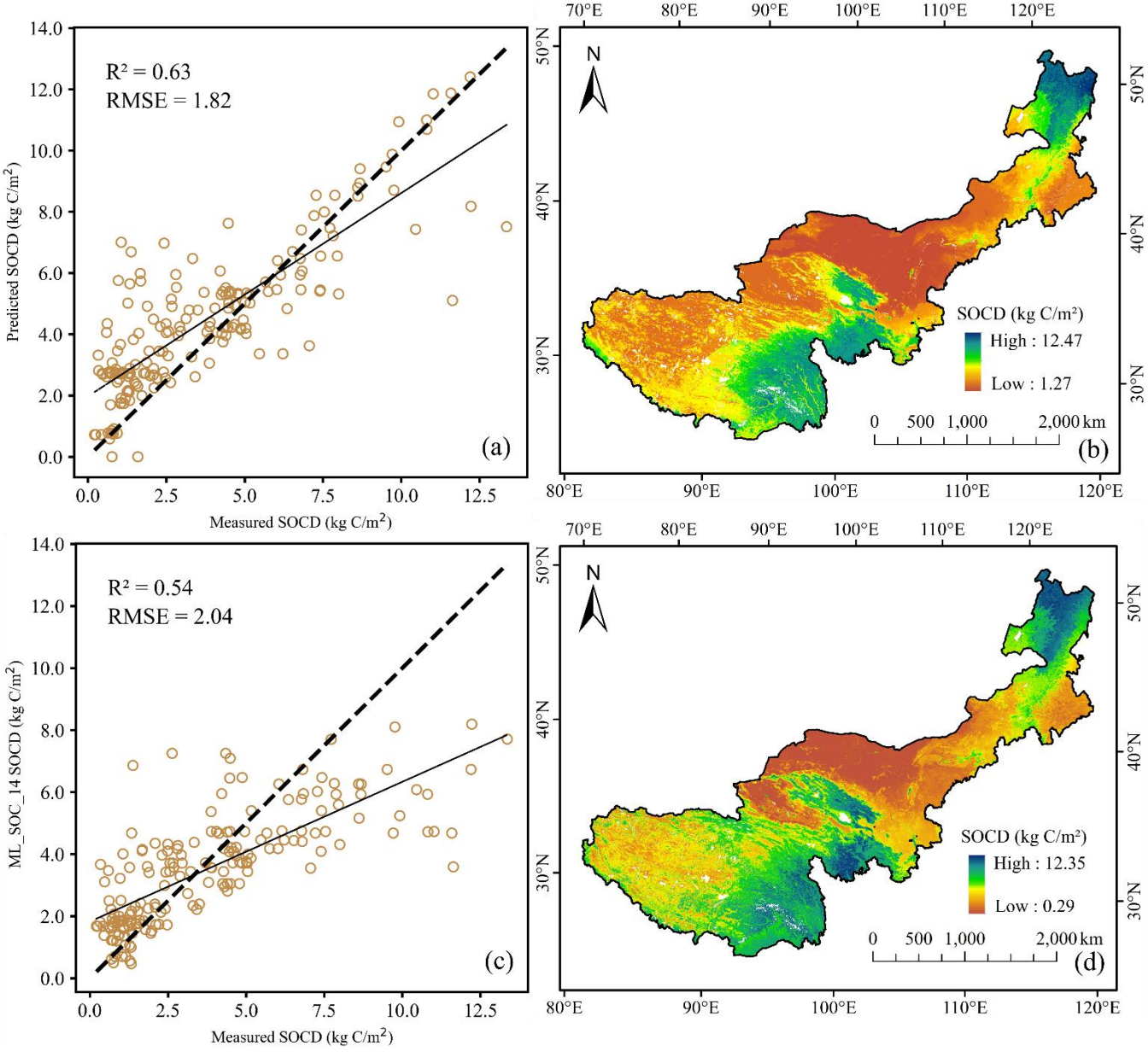


Figure 10. Comparison with the independently measured data of Xu. (a) and (c) are the correlations of the estimated SOCD and SOC Dynamics ML dataset with the published SOCD product of Xu for the soil with a depth of 0-20 cm, respectively. (b) and (d) are the spatial distribution of the estimated SOCD and SOC Dynamics ML dataset for the soil with a depth of 0-20 cm in this study.

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

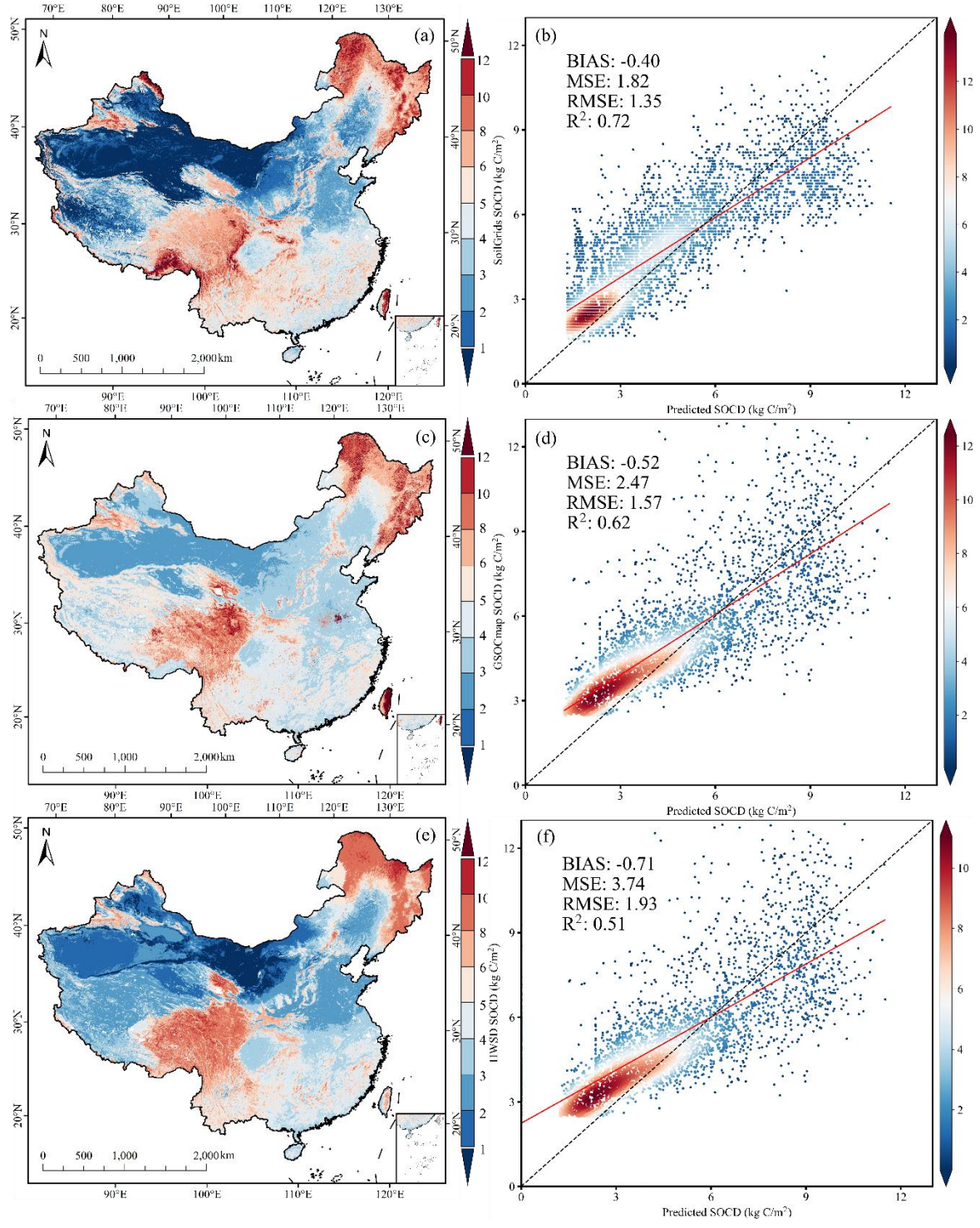


Figure 11. Comparison with three published global products. Our estimated SOCD is compared with the SoilGrids250m (a & b), GSOCmap (c & d), and HWSD v2.0 (e & f) datasets.

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

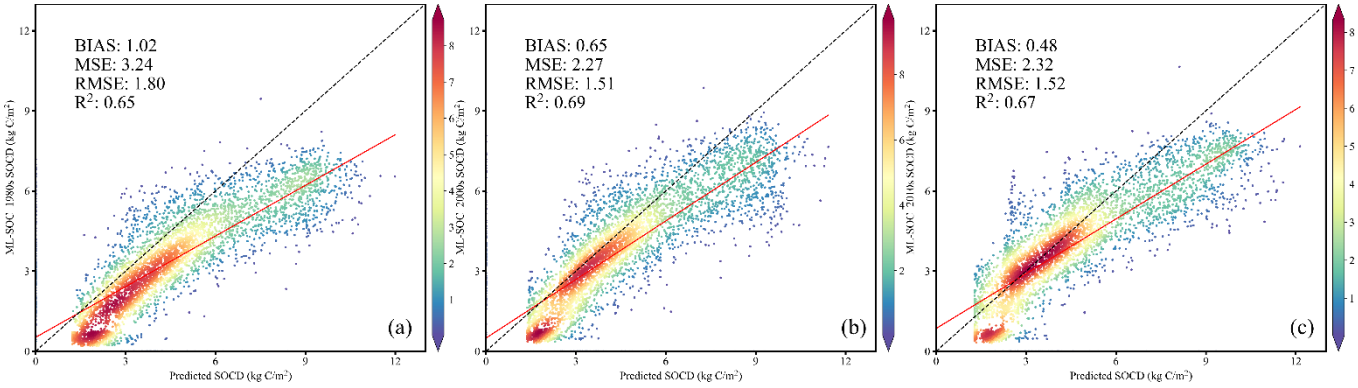
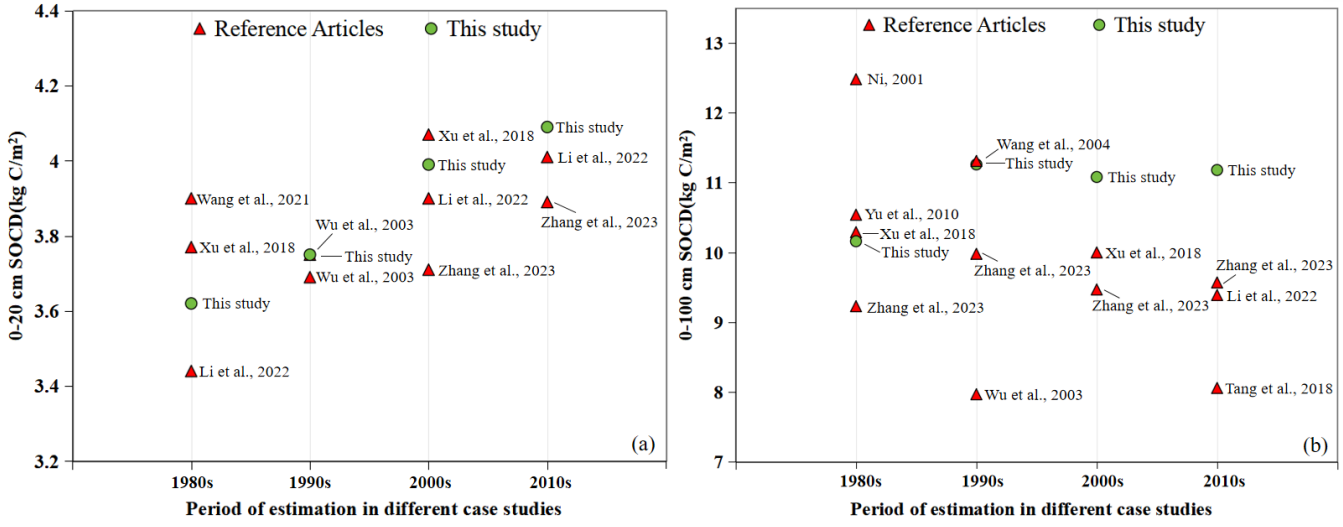


Figure 12. Comparison with the SOC Dynamics ML dataset with a depth of 0-20 cm in China in the 1980s (a), 2000s (b), and 2010s (c).

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025



535 **Figure 13.** Aggregated results of estimated SOCD with the depth of 0-20 cm (a) and 0-100 cm (b) in China from this study and previous investigations.

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

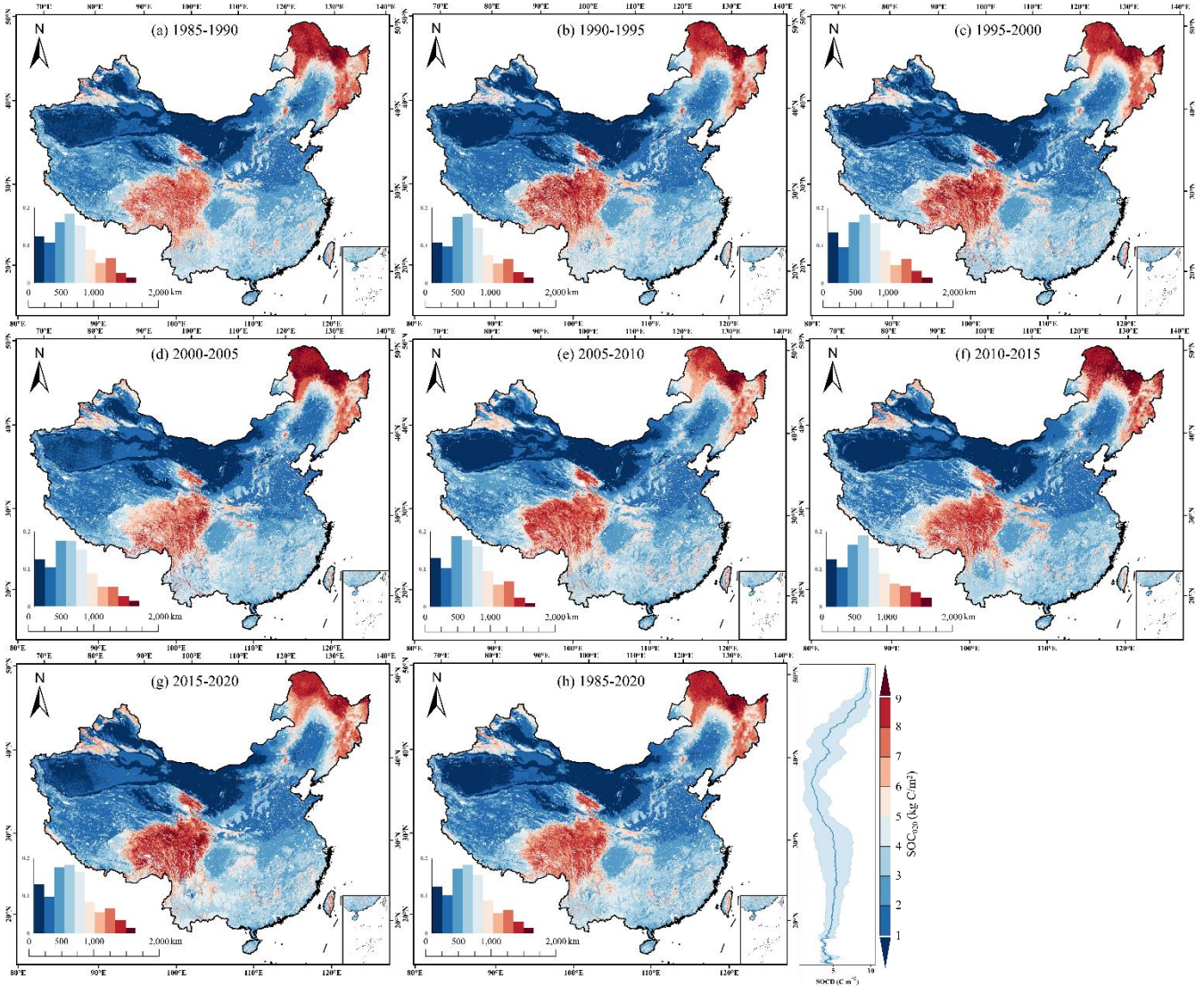


Figure 14. Spatial distribution of estimated SOCD at a depth of 0-20 cm in 1985-1990 (a), 1990-1995 (b), 1995-2000(c), 2000-2005 (d), 2005-2010 (e), 2010-2015 (f), 2015-2020 (g) and average from 1985 to 2020 (h). The lower left histograms in each panel show the area ratios for different SOCD levels.

https://doi.org/10.5194/essd-2024-588
Preprint. Discussion started: 4 February 2025

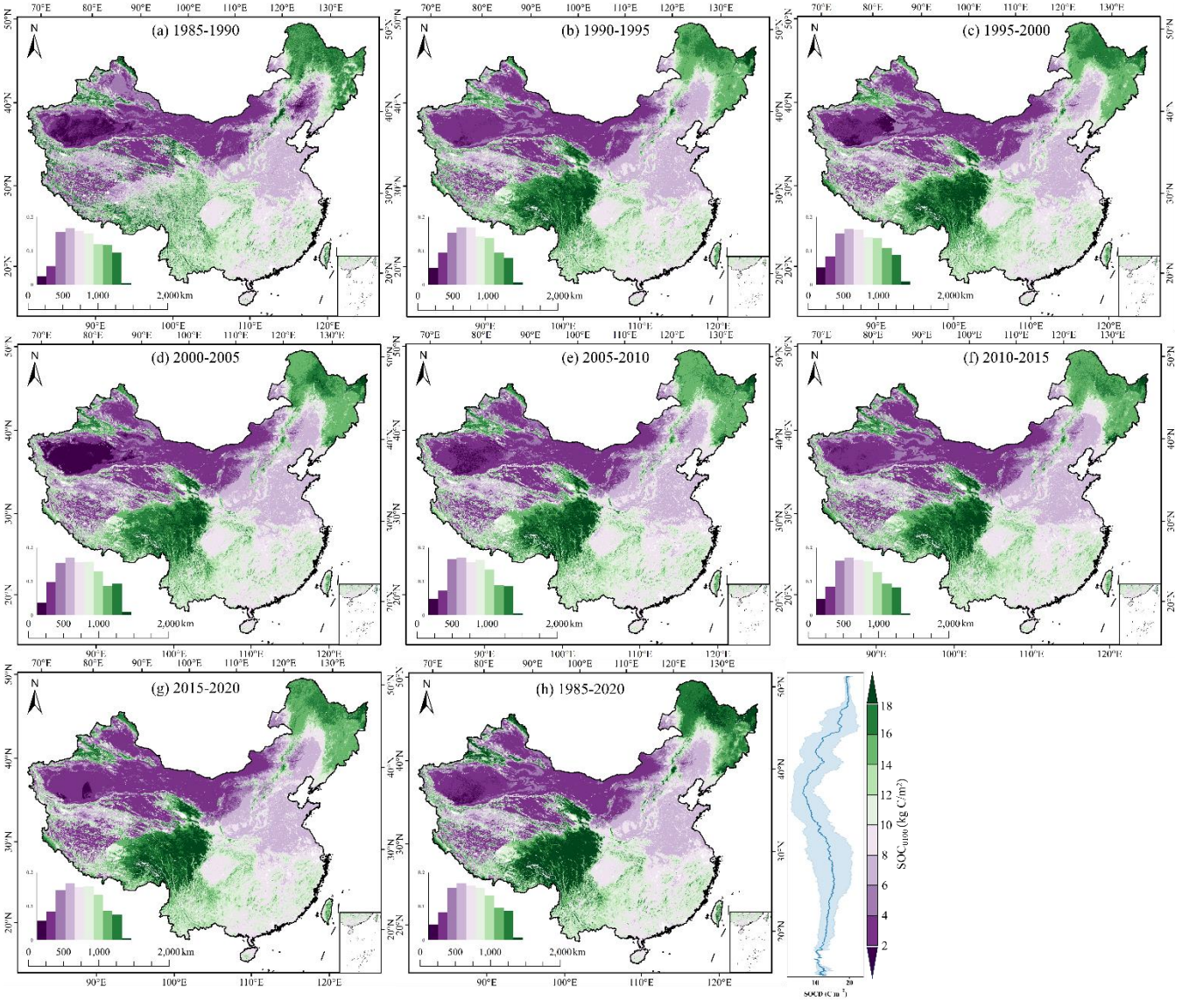


Figure 15. Spatial distribution of estimated SOCD at a depth of 0-100 cm in 1985-1990 (a), 1990-1995 (b), 1995-2000(c), 2000-2005 (d), 2005-2010 (e), 2010-2015 (f), and 2015-2020 (g) and average from 1985 to 2020 (h). The lower left histograms in each panel show the area ratios for different SOCD levels.