Journal of Integrative Agriculture 2026, 25(8): 3453–3468

Available online at ScienceDirect

Journal of Integrative Agriculture



RESEARCH ARTICLE

A zoning-based machine learning framework for accurate soil organic matter prediction across Mollisol and non-Mollisol regions

Xue Li[1, 2], Bo Jiang[2], Depiao Kong[1], Deqiang Zang[2], Ya Chen[2], Changkun Wang[3], Huanjun Liu[1], Chong Luo[1#]

[1] *State Key Laboratory of Black Soils Conservation and Utilization, Northeast Institute of Geography and Agroecology, Chinese Academy of Sciences, Changchun 130102, China*

[2] *School of Public Administration and Law, Northeast Agricultural University, Harbin 150030, China*

[3] *Nanjing Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008, China*

Highlights

- Propose a new soil organic matter (SOM) mapping framework that integrates remote sensing zoning, feature selection, and the Random Forest (RF) algorithm.
- High-precision classification of Mollisol and non-Mollisol based on Landsat-8 multi-temporal remote sensing images and environmental covariates.
- The optimal feature combinations for SOM mapping differ between Mollisol and non-Mollisol areas.
- The mean SOM value in the Mollisol region is slightly higher, while the spatial variability of SOM value is stronger in the non-Mollisol region.

Abstract

Soil organic matter (SOM) is a core indicator of soil fertility and ecosystem function. However, in regions where Mollisol and non-Mollisol coexist, high-precision spatial mapping faces significant challenges due to pronounced terrain heterogeneity and redundancy in high-dimensional covariates. This study proposes a "remote sensing zoning-feature selection optimization-random forest (RSZ-FSO-RF)" framework. By integrating Landsat-8 multi-temporal imagery from 2014–2023 with topographic and climatic factors, and leveraging the Google Earth Engine (GEE) platform, it achieves high-precision remote sensing zoning of Mollisol and non-Mollisol areas (overall accuracy: 92.13%, Kappa coefficient: 0.70). Subsequently, local Random Forest (RF) regression models were established within each zone for SOM prediction, with predictive variables optimized using recursive feature elimination (RFE). Results demonstrate that compared to FAO-zone-based modeling, the RSZ-FSO-RF framework significantly enhances prediction accuracy (R^2=0.619, RMSE=6.849 g kg^{-1}). And further feature optimization continued to enhance model performance (R^2=0.627, RMSE=6.781 g kg^{-1}). Notably, optimal predictor combinations varied significantly across zones, with SOM spatial variability generally higher in non-Mollisol areas than in Mollisol regions. By organically integrating remote sensing zoning with feature selection, this framework effectively mitigates covariate redundancy while accounting for local heterogeneity, significantly enhancing the accuracy and stability of high-resolution SOM mapping. Furthermore, this study provides scientific basis and decision support for soil resource management and sustainable agricultural development under complex topographic conditions.

Keywords: soil organic matter, remote sensing zoning, feature selection algorithm, random forest algorithm, prediction

1. Introduction

Northeast China is the main distribution area of Mollisol in China, and with its superior soil fertility conditions, it has long assumed an important food production function, and has developed into a key national commercial grain production base, which is of great strategic significance for guaranteeing China's food security (Yang *et al.* 2020; Mustafa *et al.* 2021). Soil organic matter (SOM) , as one of

Received 15 July 2025; Received in revised form 12 October 2025; Accepted 17 November 2025; Available online 14 January 2026
Xue Li, Mobile: +86-18846763278, E-mail: S231202038@neau.edu.cn; #Correspondence Chong Luo, Mobile: +86-13304807096, E-mail: luochong@iga.ac.cn

© 2026 CAAS. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/). Peer review under responsibility of Editorial Board of *Journal of Integrative Agriculture*.
doi: 10.1016/j.jia.2026.01.016

the core substances for the maintenance of Mollisol fertility, is mainly concentrated in the soil surface layer and has high accumulation characteristics (Liu *et al.* 2012). SOM is not only an important carbon reservoir in terrestrial ecosystems and has a key role in regulating the stability of the global carbon cycle (Huang *et al.* 2022; Broeg *et al.* 2024), but also directly affects the supply of nutrients, structural stability, and ecological functions of soils, and is an important indicator for evaluating the quality of arable soils and crop production potential (Peng *et al.* 2024). However, over time, the SOM value of soils decreases due to erosion, tillage and intensive farming. To address this challenge, there is a need for continuous monitoring of arable land and the production of soil maps with high spatial resolution, which play a key role in the proper management of soil quality and the sustainability of agriculture.

Remote sensing technology has become a key tool for obtaining the spatial distribution of SOM in large-scale regions (Zhou *et al.* 2016). Compared with traditional spatial interpolation methods, remote sensing prediction shows significant advantages in improving prediction accuracy and revealing spatial heterogeneity of soil properties (Ladoni *et al.* 2010; Mulder *et al.* 2011). It has been shown that SOM value is closely related to soil reflectance in a specific band range (Shonk *et al.* 1991; Henderson *et al.* 1992; Conforti *et al.* 2015). Currently widely used satellites (e.g., Landsat 5/7/8, Sentinel-2, MODIS) and airborne multispectral sensors cover the key band range of SOM response, which provides a theoretical basis for SOM mapping using multispectral images (Liu *et al.* 2009; Bao *et al.* 2017). However, in large-scale regions, single-day images are susceptible to interference from clouds, shadows, and differences in agricultural management (Luo *et al.* 2022a). In recent years, thanks to the relative stability of soil background and sensor performance, multi-temporal synthetic images of the bare soil period have been widely used in SOM remote sensing mapping, which effectively improves the accuracy of remote sensing predictions of soil properties (Dou *et al.* 2019; Wang X *et al.* 2022). Environmental covariates are key factors affecting soil properties, among which climatic and topographic variables are often used for soil property modeling due to their availability and accuracy (John *et al.* 2021). Numerous studies have shown that the introduction of multiple topographic parameters (e.g., elevation, slope) and climatic factors (e.g., temperature, precipitation) can significantly improve the accuracy of soil property prediction (Mahmoudzadeh *et al.* 2020). Combining these environmental covariates with measured soil property data can help construct more robust and accurate predictive models (Minasny and McBratney 2016). However, the complexity of the soil formation process and its high degree of interaction with environmental factors, which results in significant heterogeneity in the spatial distribution of SOM, still pose a serious challenge to the predictive ability of models (Luo *et al.* 2022a; Meng *et al.* 2025).

Optimized modeling strategies are widely recognized as an important approach to improving the accuracy of soil property predictions (Zhang Y *et al.* 2023). Traditional digital soil mapping (DSM) typically relies on a global modeling framework, which integrates all samples to construct a single model (Liang *et al.* 2024). Although this approach has demonstrated effectiveness in numerous studies, it often overlooks the inherent spatial heterogeneity of soils (Franklin and Mills 2009). For instance, in regions where Mollisol and non-Mollisol are intermingled, variations in erosion intensity, agricultural disturbance, and topographic undulation result in significant disparities in SOM value. The global model typically assume uniform soil properties across different areas, thereby compromising the accuracy of predictions (Hengl *et al.* 2017). In contrast, the zonal modeling strategy based on soil type, i.e., modeling separately within different soil type regions, is more consistent with the essential characteristics of soil spatial heterogeneity. This method can not only more accurately capture the differences in soil properties within each sub-area, but also effectively circumvent the prediction bias triggered by inter-area differences, thus improving the accuracy of SOM mapping (Sothe *et al.* 2022; Zhang *et al.* 2022). Building upon enhanced partition modeling predictive capabilities, further improvements rely on effectively identifying key environmental factors driving spatial variation within the SOM space (Chen S *et al.* 2022; van der Westhuizen *et al.* 2023). In DSM, the widespread application of machine learning — particularly the Random Forest (RF) algorithm — has made feature selection a critical step in constructing robust, high-precision predictive models.

Characteristic variable screening can effectively identify the optimal predictor combinations to improve the accuracy of SOM prediction (Rosero-Vlasova *et al.* 2017; Shi *et al.* 2021). Since multicollinearity and redundant information often exist between remote sensing and environmental covariates, direct inclusion of all variables in modeling may trigger model overfitting and reduce its robustness (Zhang X *et al.* 2023). It has been shown that using appropriate feature selection algorithms for variable optimization helps to eliminate redundant features and improve model performance (Keskin *et al.* 2019). In recent years, some scholars have further coupled feature selection algorithms with the modeling process in depth, aiming to synergistically optimize the model structure and input variable combinations to achieve higher prediction accuracy. For example, Vohland *et al.* (2014) significantly enhanced the predictive performance and robustness of the model by combining the CARS and PLSR models under visible-near-infrared (Vis-NIR) and mid-infrared (MIR) spectral data; He *et al.* (2023) combined multiple features such as principal component analysis (PCA), least absolute shrinkage and selection operator (Lasso), and SCARS-PCA to extract key variables from spectral and profile information, significantly improving the stability and accuracy of predictions; Wang H *et al.* (2021) integrated Lasso and correlation coefficient to screen key bands based on the Sentinel-1/2 data, and the constructed RF model realized high-precision SOC prediction (R^2=0.921) under multiple land use types. All these studies show that the reasonable use of feature selection algorithms to remove redundant variables can effectively improve the SOM modeling accuracy. The widespread application of methods like recursive feature elimination (RFE),

Boruta, and Lasso further demonstrates the value of feature selection in SOM mapping (Kebonye *et al.* 2023; Alshehri *et al.* 2024). Although feature selection algorithms demonstrate potential in extracting key variables, few studies have systematically coupled feature optimization with soil-type-based partitioned modeling to simultaneously address the dual challenges of spatial heterogeneity and variable redundancy in SOM mapping.

To address this research gap, this study proposes an innovative predictive framework applicable to complex landscapes where Mollisol and non-Mollisol areas are intermingled. Specifically: (1) integrating Landsat-8 multi-temporal remote sensing images and environmental covariates to construct a regional modeling system integrating remote sensing zoning and RF algorithms; (2) quantitatively evaluating the role of remote sensing zoning in improving the modeling accuracy and comparing it with the local modeling strategy of Food and Agriculture Organization of the United Nations (FAO) zoning; (3) incorporating the RFE algorithm into the modeling framework to optimize covariate selection within each region, thereby further enhancing the predictive accuracy of SOM; (4) exploring the spatial distribution and variation patterns of SOM in different soil types, so as to provide methodological support and scientific basis for the precise management and sustainable use of complex soil landscapes.

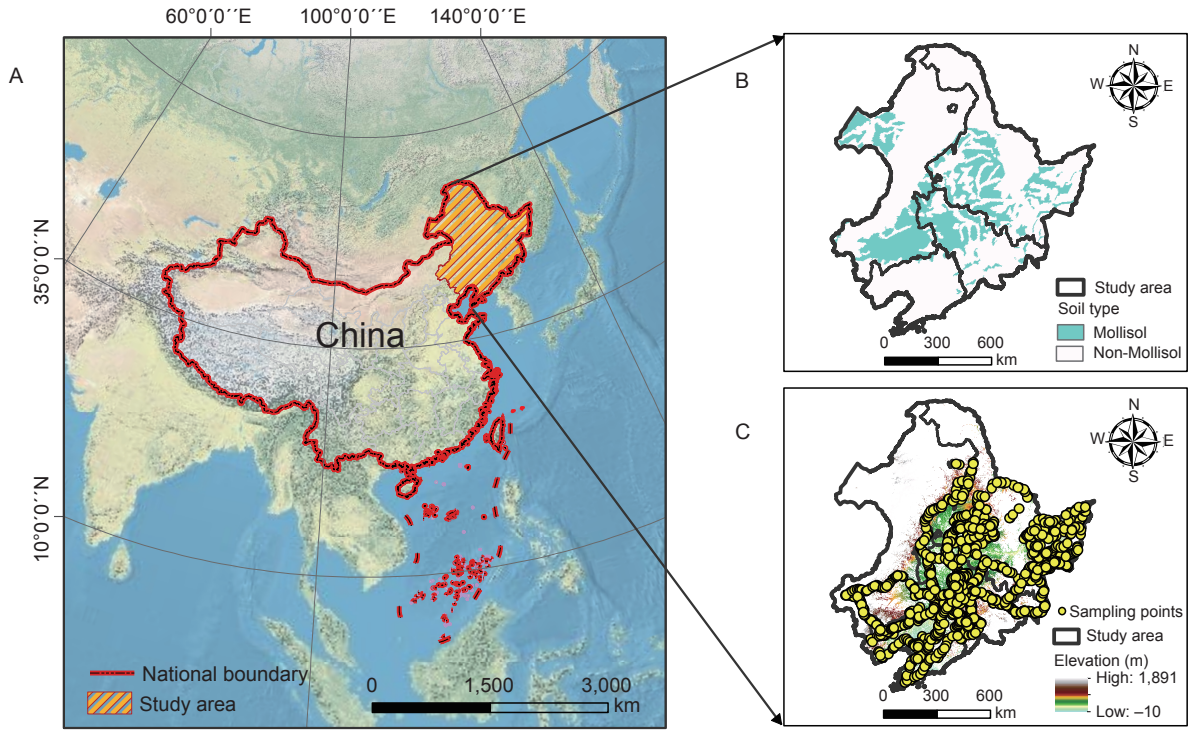
2. Materials and methods

2.1. Overview of the study area

This study area is located in Northeast China (38.72°N–53.56°N, 115.52°E–135.09°E), covering three provinces of Heilongjiang, Jilin, Liaoning and eastern Inner Mongolia, with a total area of about 788,000 km² (Luo *et al.* 2022b). The region mainly consists of Songnen Plain, Sanjiang Plain and Liaohe Plain, which is a typical temperate region with concentrated Mollisol distribution (Wang H *et al.* 2022), and is also the third largest Mollisol area in the world (Wang H *et al.* 2022; Wang *et al.* 2024). With a coverage area of 44,520 km², Mollisol represents the most characteristic soil resource in this region (Wang *et al.* 2024). The geomorphology is dominated by a mixture of low mountains, hills and plains, with an elevation range of –10 to 1,894 m, and slopes mainly distributed between 0°–5° (Wang and Fan 2019). The regional climate is characterized by a mesothermal semi-humid to humid climate with four distinct seasons, average annual temperatures ranging from –4 to 11°C, and annual precipitation of 207–1,026 mm, with a significant spatial and temporal distribution. Corn, soybeans and rice are the main crops, with the cultivated area accounting for more than 90% of the total arable land. April–May is the typical bare soil window period, with sparse vegetation cover and low cloud cover, which provides a favorable time for obtaining high-quality remote sensing images (Luo *et al.* 2024) (Fig. 1).

2.2. Workflow of SOM mapping

The overall technical flow of this study is shown in Fig. 2. First, Landsat-8 remote sensing images and major environmental covariates were collected and preprocessed, while regional zoning results based on the FAO traditional



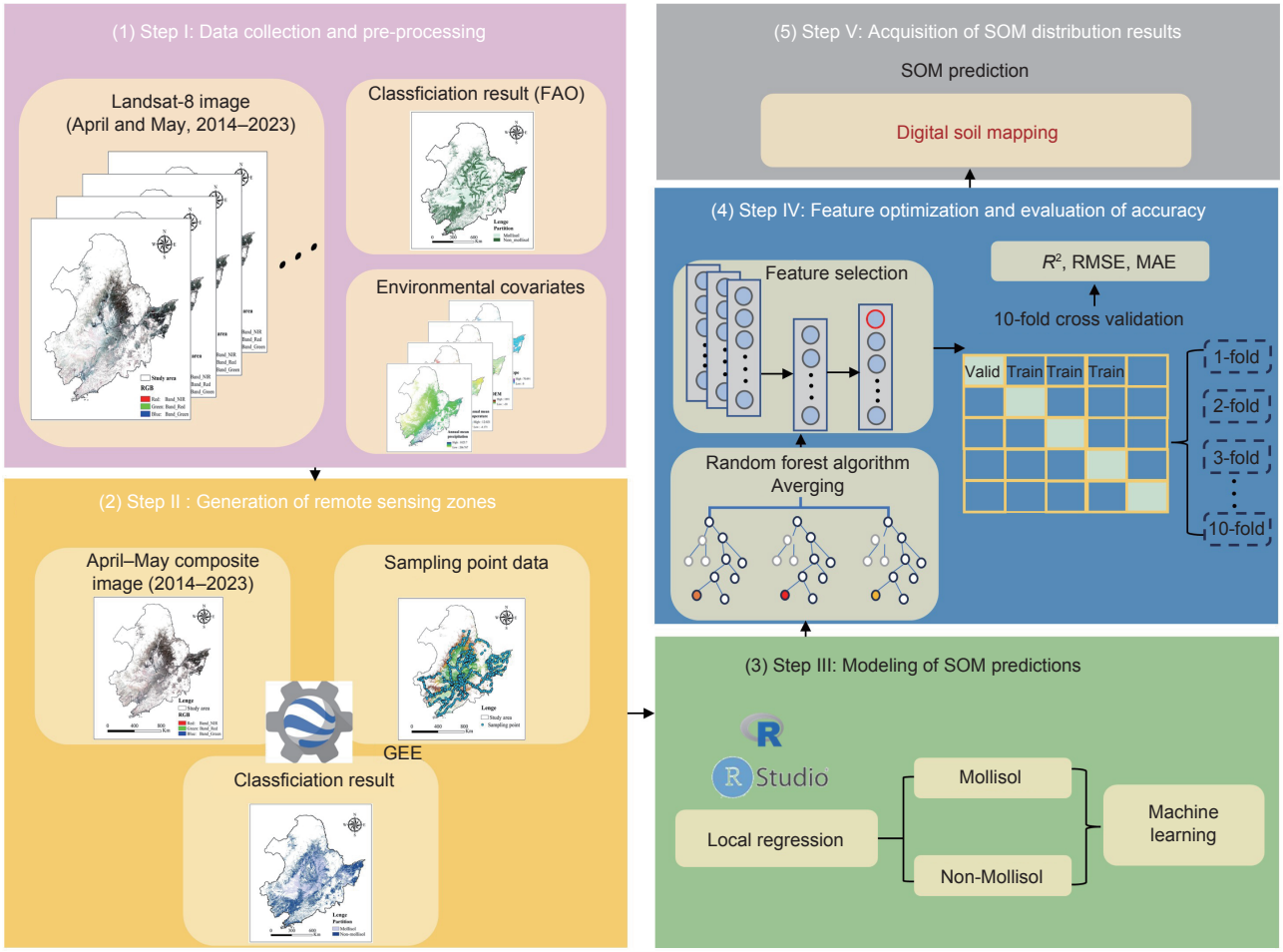
审图号：GS 京 (2026) 0965 号

Fig. 1 Location of the study area and soil sampling sites. A, geographic location of the northeast region. B, distribution of Mollisol and non-Mollisol based on FAO classification. C, soil sampling sites and elevation.

classification system were obtained. Subsequently, based on the multi-temporal Landsat-8 images, environmental covariates and field sampling data from April to May, 2014–2023, remote sensing zoning results were constructed with the help of Google Earth Engine (GEE) platform. Then, the SOM prediction model was constructed using the RF algorithm by combining the local modeling strategies, and the feature selection algorithm was introduced to optimize the variable combinations. The model performance is evaluated by ten-fold cross-validation. Finally, the optimal model was applied to generate the spatial distribution map of SOM value in the study area.

2.3. Data and preprocessing

Soil data collection and preprocessing This study obtained the spatial distribution of Mollisol and non-Mollisol zones in Northeast China based on the FAO/UNESCO Soil Map of the World and the subsequent World Reference Base for Soil Resources (WRB) Classification System. The system follows globally consistent standards and classifies soils according to morphology, pedogenic processes, and physicochemical properties. Based on this distribution, a stratified random sampling approach was applied to collect 1,603 surface (0–20 cm) soil samples in April and May (bare soil period) from 2023–2024 (Table 1), with their spatial distribution shown in Fig. 1. After the samples were naturally air-dried and passed through a 2 mm sieve, the soil organic carbon (SOC) content was determined by the potassium dichromate oxidation method, and the SOM value was calculated by a conversion factor of 1.724. Out of these



审图号 : GS 京 (2026) 0965 号

Fig. 2 The workflow of this study. GEE, Google Earth Engine; SOM, soil organic matter.

Table 1 Descriptive statistics of soil organic matter value in the study area

Sample type[1)]	*n*	Max. (g kg^{-1})	Min. (g kg^{-1})	Mean (g kg^{-1})	SD (g kg^{-1})	Skewness	Kurtosis	CV (%)
TS	1,603	127.576	2.138	32.481	19.123	1.429	2.785	58.875
P(M)S	577	97.078	2.138	32.111	16.561	0.823	0.525	51.573
P(N)S	1,026	127.576	3.827	32.690	20.428	1.577	3.053	62.492

[1)] TS, total sample; P(M)S, samples in Mollisol areas; P(N)S, samples in non-Mollisol areas.

samples, 577 samples were taken from the Mollisol area and 1,026 samples were taken from the non-Mollisol area. The range of SOM value in the study area was 2.138–127.576 g kg^{-1} with a mean value of 32.481 g kg^{-1} and a coefficient of variation as high as 58.875% indicating significant spatial heterogeneity. The mean value of SOM in the Mollisol area (32.111 g kg^{-1}) was slightly lower than that in the non-Mollisol area (32.690 g kg^{-1}), but its standard deviation was smaller, 16.561 g kg^{-1} in the Mollisol area, compared with 20.428 g kg^{-1} in the non-mollisol area, which indicated that SOM distribution was relatively homogeneous in the Mollisol area, whereas the variability in the non-Mollisol area was stronger.

Acquisition and preprocessing of Landsat-8 remote sensing imagery The Landsat-8 satellite carrying the operational land imager (OLI) was launched in 2013, with a revisit cycle of 16 days. During the period of 2014–2023, this study acquires all available Landsat-8 surface reflectance (SR) imagery of the Northeast China in April and May each year through the GEE platform. All available Landsat-8 SR images of the Northeast China are acquired through the GEE platform during the period 2014–2023. The data were obtained from the "LANDSAT/LC08/C02/T1_L2" dataset in the GEE platform, which was released by the U.S. Geological Survey (USGS), and the atmospheric correction processing was done using the land surface reflectance code (LaSRC) algorithm. In the image preprocessing process, cloud masking operations were performed using the "QA_PIXEL" and "QA_RADSAT" bands to remove clouds, shadows, and saturated pixels, resulting in a set of cloud-free Landsat 8 images covering the study area from 2014 to 2021. Then after atmospheric correction, six reflectance bands (B1-B6) were extracted from the remote sensing data, with a spatial resolution of 30 m (Table 2). To improve the SOM prediction accuracy, the median synthesis method was used to fuse the multi-temporal images from April to May to generate remote sensing images that stably represent the characteristics of the bare soil period, forming a multi-temporal Landsat-8 SR synthetic image for the years 2014–2023 that covers the northeastern part of China (You and Dong 2020; Wang X *et al.* 2022).

Acquisition and pre-processing of environmental covariates Based on existing studies, two types of environmental covariates were selected for this study: climatic variables and terrain variables (Table 3).

1) Climatic variables. Numerous studies have shown that climate plays an extremely important role in influencing the value of SOM, and among them, temperature and precipitation are the key variables most commonly used for SOM prediction (Fig. 3-A and B). Higher temperatures generally accelerate microbial activity and organic matter decomposition, thereby decreasing SOM (Weil and Magdoff 2004); whereas increased precipitation contributes to plant growth and residue inputs, promoting SOM accumulation (Gao *et al.* 2024). In addition, precipitation indirectly affects microbial activity and SOM decomposition rates by regulating soil moisture and aeration (Niu *et al.* 2024). Wet areas usually exhibit higher SOM value. In this study, monthly temperature and precipitation datasets for China from 2014 to 2023 were obtained from the National Tibetan Plateau Science Data Center (https://data.tpdc.ac.cn/). These datasets were processed in ArcGIS 10.8 to generate multi-year average data and resampled to a 30 m spatial resolution to derive the average precipitation index and multi-year average temperature index for the study area.

Table 2 Information on Landsat-8 SR data used in this study

Band name	Description	Wavelength (nm)	Resolution (m)
B1	Blue	450–520	30
B2	Green	520–600	30
B3	Red	630–690	30
B4	NIR	770–900	30
B5	SWIR1	1,550–1,750	30
B6	SWIR2	2,080–2,350	30

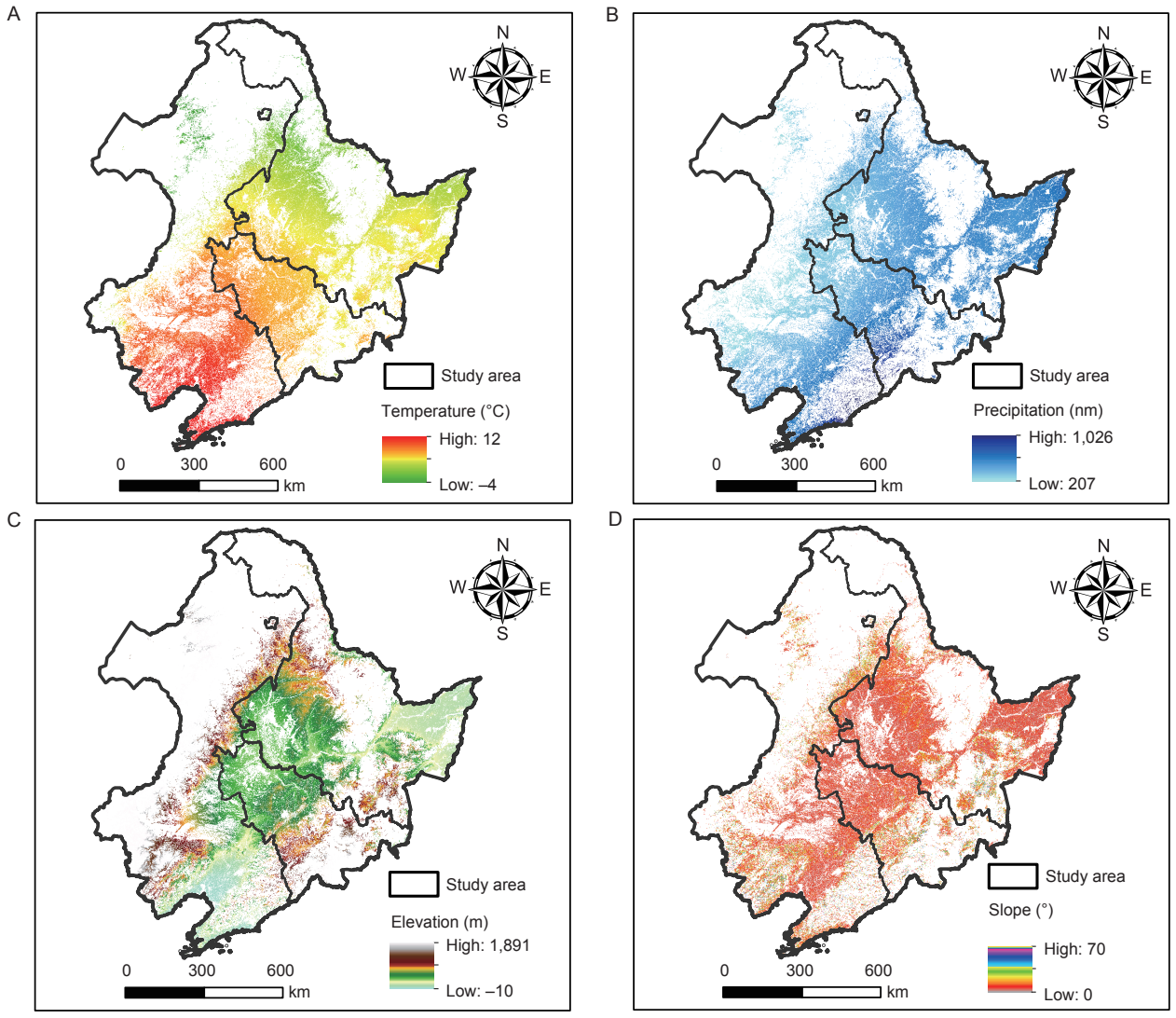
Table 3 Basic information on environmental covariates used in this study

Category	Environmental covariates	Abbreviation	Unit
Climate	Multi-year average temperature	PRE	°C
	Multi-year average precipitation	TEMP	nm
Topography	Slope	SLOPE	°
	Elevation	DEM	m

2) Topographic variables. Topographic conditions have significant effects on the spatial distribution of SOM value, with slope and elevation being the most commonly used topographic factors (Zhang *et al.* 2014; Niu *et al.* 2024) (Fig. 3-C and D). Gently sloping areas usually exhibit higher SOM value due to weaker erosion and long-term accumulation of surface organic matter (Zhu *et al.* 2019; Wang B *et al.* 2021). In contrast, higher elevations also favor SOM accumulation due to lower temperatures and shorter growing seasons, which slow down the decomposition rate of SOM (Xu *et al.* 2014; Gutiérrez-Girón *et al.* 2015; Dan *et al.* 2016). Therefore, slope and elevation are often selected as representative terrain variables to be included in SOM modeling. In this study, the "Elevation" dataset from the "NASADEM: NASA NASADEM Digital Elevation 30 m" data stored in GEE was used as the elevation input for the study area. In this study, the "Elevation" dataset from the "NASADEM: NASA NASADEM Digital Elevation 30 m" data stored in GEE was used as the input for the elevation of the study area, and the obtained 'Elevation' dataset was used to calculate "SLOPE" as the input for the slope of the study area.

2.4. Predictive modeling

Detection model for spatial distribution of Mollisol area RF is an integrated learning method that consists of multiple decision trees with good nonlinear modeling capability, which can significantly improve the prediction accuracy and effectively control the computational cost (Breiman *et al.* 2001; Belgiu and Drăguț 2016). In this study, a remote sensing detection model for identifying the spatial distribution of Mollisol areas was constructed based on the GEE platform. The model was trained using 577 Mollisol zone samples with 1,026 non-Mollisol zone samples, and 70% of these samples were randomly selected as the



审图号 : GS 京 (2026) 0965 号

Fig. 3 Environmental covariate data for the study area. A, multi-year average temperature data. B, multi-year average precipitation data. C, elevation data. D, slope data.

training set. Input features included six reflectance bands of Landsat-8 OLI (Blue, Green, Red, NIR, SWIR1 and SWIR2) and four major environmental covariates, totaling 10 variables. By optimizing the model parameters based on out-of-bag error (OOB), the number of regression trees (ntree) was finally determined to be 300, and the number of split variables (mtry) was 1/3 of the total number of variables. This parameter configuration effectively suppressed the overfitting, stabilized the OOB error rate, and strengthened the model's generalization ability and robustness while enhancing the model's prediction performance.

Regression mathods To better characterize the pronounced spatial heterogeneity between Mollisol and non-Mollisol regions and thus improve SOM prediction performance, this study developed two zoning-based local regression strategies.

1) Local regression model based on FAO zoning. According to the soil classification criteria of the Harmonized World Soil Database (HWSD) issued by FAO, Phaeozems and Chernozems were classified as Mollisol zones, while all other soil types were designated as non-Mollisol zones. In this study, separate sub-models were established for the Mollisol and non-Mollisol zones. Each sub-model was independently used to predict SOM within its corresponding zone, and the prediction results were subsequently integrated to produce the final local regression output.

2) Local regression model based on remote sensing zoning. With the support of the high-performance computing capability of GEE, Landsat-8 multi-band imagery and major environmental covariates were used to construct the basic datasets for zoning. All available soil samples were introduced as training data. The performance of different zoning schemes was evaluated using a confusion matrix and cross-validation (CV), and the optimal zoning scheme was selected accordingly. Subsequently, independent regression models were developed within each remote sensing zone to predict SOM values. Finally, the prediction results from all zones were integrated to generate a high-precision local

regression model based on remote sensing zoning.

2.5. Feature selection algorithms

It has been shown that redundant and noisy features may weaken model performance, whereas feature selection can effectively filter out the most representative combinations of independent variables and enhance SOM prediction accuracy (Visalakshi and Radha 2014; Chen Y *et al.* 2022). In this study, three typical feature selection algorithms, RFE, Boruta and Lasso, are used to evaluate their performance in SOM modeling. RFE is a model-based feature selection algorithm. It iteratively trains the model and removes the least important variables at each step until an optimal subset of features is obtained. This process retains only the variables that contribute most to the prediction performance (Chen *et al.* 2020); Boruta is a random-forest-based method for identifying all relevant predictors. The algorithm creates "shadow" (noise) features and evaluates each variable's importance by comparing its average contribution to model accuracy with that of the shadows. Variables that consistently fail to outperform their shadow counterparts are removed through iteration, producing a stable subset of fully relevant predictors (Ali *et al.* 2019); Lasso is an embedded feature selection algorithm based on linear regression. The algorithm improves the generalization ability of the model by adding an L1 regularization term (the sum of the absolute values of the regression coefficients) to the loss function. This regularization narrows the coefficients, automatically selects important features and eliminates irrelevant ones, enhancing the predictive performance of the model (Muthukrishnan and Rohini 2016). In this study, all three feature selection algorithms perform the optimal feature selection process in the *R* language environment.

2.6. Accuracy assessment

To avoid model overfitting during SOM spatial prediction, a 10-fold CV approach was employed for model validation. In this procedure, the dataset was randomly divided into 10 equal subsets. Each time, nine subsets were used for model training, and the remaining one was retained for validation. This process was repeated 10 times so that every subset served once as the validation data. Since the retained samples were not affected by the choice of model parameters, this dataset provided more accurate and precise estimates for the new data. To further evaluate the model results, we utilized the correlation coefficient (R^2) (eq. (1)), absolute root mean square error (RMSE) (eq. (2)), and mean absolute error (MAE) (eq. (3)) of the model predictions with the ground observations. The higher the R^2, the smaller the values of the RMSE and MAE, indicating the higher accuracy of the SOM prediction model (Zhang *et al.* 2019).

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

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

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\hat{y}_i\right| \quad (3)$$

where n represents the number of samples, y_i represents the observed value of sample i, $\hat{y}_i$ represents the predicted value of sample i, and $\bar{y}_i$ represents the corresponding mean of the observations.

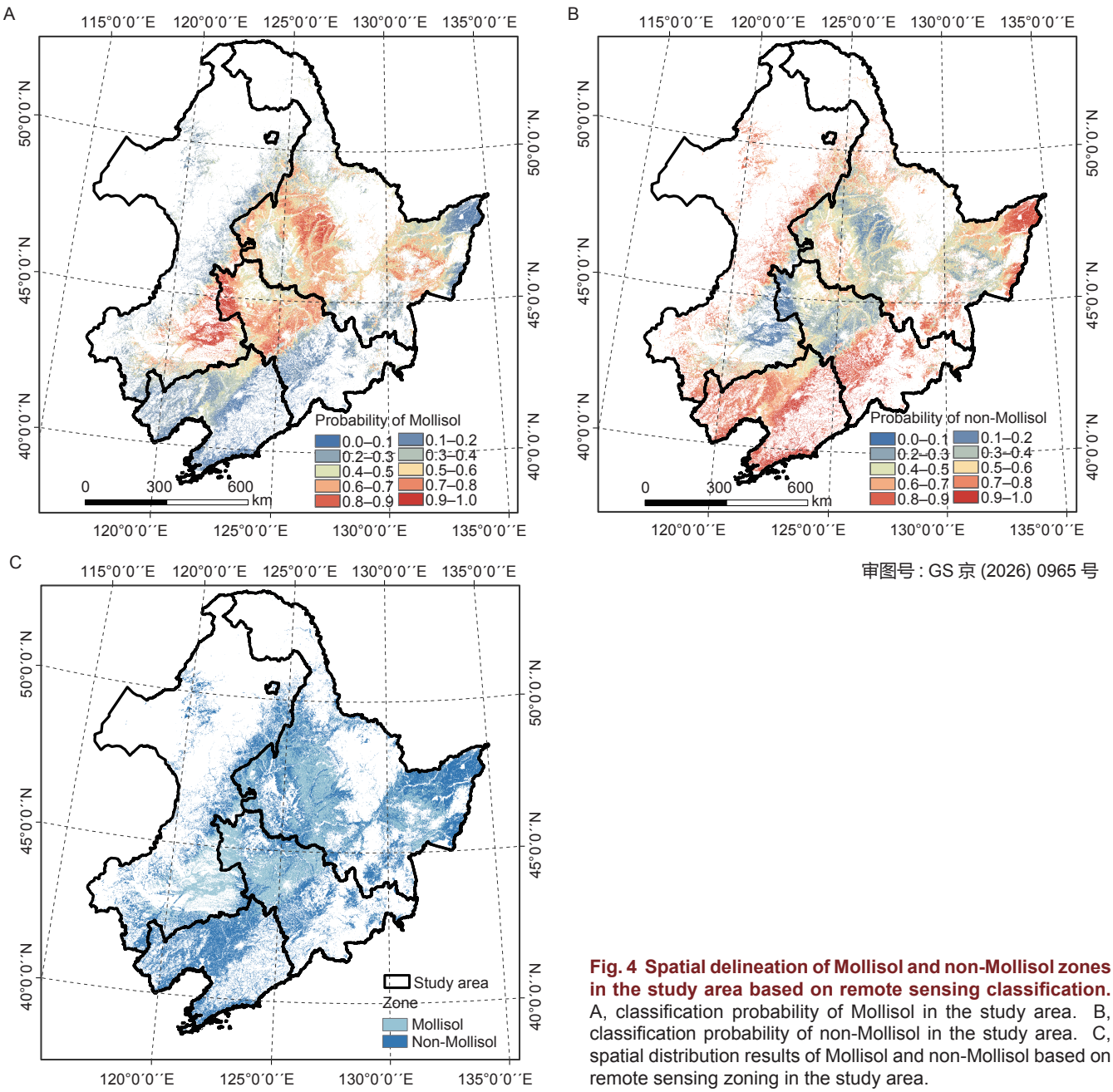
3. Results

3.1. Spatial distribution detection of Mollisol area

In this study, the multi-temporal remote sensing images (April–May) of Landsat-8 from 2014–2023 were called in the GEE platform, and a Mollisol area identification model based on the RF algorithm was constructed by taking its six spectral bands and four major environmental covariates as inputs (Fig. 4-A and B). The results show that the classification model integrating multi-temporal remote sensing data and environmental covariates has high accuracy, with an overall accuracy of 92.13% and a Kappa coefficient of 0.70, which indicates that the model has a good distinguishing ability in Mollisol zone identification. The detection results are shown in Fig. 4-C, and the Mollisol zone is mainly distributed in the southwest and central areas of the study area. Based on the classification results, the area of Mollisol zone in the study area is about 140.805 km², accounting for 34.007% of the total cultivated area; the area of non-Mollisol zone is 273.232 km², accounting for 65.993%. In order to further verify the classification effect and make comparisons, the results of Mollisol zoning delineated by FAO global Mollisol were introduced in this study as reference. According to the FAO data, the area of Mollisol zone in the study area is 154.071 km², accounting for 37.212% of the total cultivated area; the area of non-Mollisol zone is 259.966 km², accounting for 62.788% (Table 4).

3.2. Accuracy assessment of different local regression models based on remote sensing band inputs

Using the RF algorithm and Landsat-8 multi-temporal remote sensing imagery, two local regression models were constructed for SOM prediction, with results shown in Fig. 5 and Table 5. For the model based on FAO zoning, when spectral bands were used as predictors, its accuracy was R^2=0.519, RMSE=7.706 g kg^{-1}, and MAE=5.325. In contrast, the model based on remote sensing zoning demonstrated slightly higher prediction accuracy, achieving R^2=0.520, RMSE=7.697 g kg^{-1}, and MAE=5.302. From a zoning perspective, in the non-Mollisol region, the FAO-based model outperformed the remote sensing–based model, yielding an R^2 that was 0.014 higher and an RMSE that was 0.001 g kg^{-1} lower. In contrast, within the Mollisol region, the remote sensing-based model demonstrated substantially better performance, achieving an R^2 that was 0.029 higher than that of the FAO model and an RMSE that was 0.274 g kg^{-1} lower. This advantage in the Mollisol region outweighed the marginal advantage of the FAO-based model in the non-Mollisol area, resulting in higher overall prediction accuracy for the remote sensing-based model. The results indicate that remote sensing-based zoning demonstrates superior



审图号：GS 京 (2026) 0965 号

Fig. 4 Spatial delineation of Mollisol and non-Mollisol zones in the study area based on remote sensing classification. A, classification probability of Mollisol in the study area. B, classification probability of non-Mollisol in the study area. C, spatial distribution results of Mollisol and non-Mollisol based on remote sensing zoning in the study area.

Table 4 Area and proportion of each zone based on remote sensing zoning and FAO zoning

Zoning method	Zone	Area (km²)	Percentage (%)
Remote sensing zoning	Mollisol	140.805	34.007
	Non-Mollisol	273.232	65.993
FAO zoning	Mollisol	154.071	37.212
	Non-Mollisol	259.966	62.788

capability in identifying spatial heterogeneity of SOM in the transition zone between Mollisol and non-Mollisol areas.

3.3. Accuracy assessment of different local regression models incorporating environmental covariates

Fig. 6 illustrates the accuracy performance of local regression models based on FAO zoning and remote sensing zoning in SOM prediction after incorporating environmental covariates (topography and climate factors). Results indicate that adding topographic or climatic variables to the model substantially improves its predictive performance compared with using only remote sensing bands. When both types of environmental covariates are introduced simultaneously, the accuracy of both local regression models shows a consistent and robust upward trend. Specifically, when all remote sensing bands and environmental covariates were used as inputs, the remote sensing-based local regression model achieved the best performance: R^2=0.619, RMSE=6.849 g kg^{-1}, MAE=4.601 g kg^{-1}. Compared to the corresponding model based on FAO zoning, R^2 increased by 0.002, RMSE decreased by 0.006 g kg^{-1}, and MAE differed by 0.006 g kg^{-1}. Compared to the remote sensing zoning local model using only bands, R^2 increased by 0.099, while RMSE and MAE decreased by

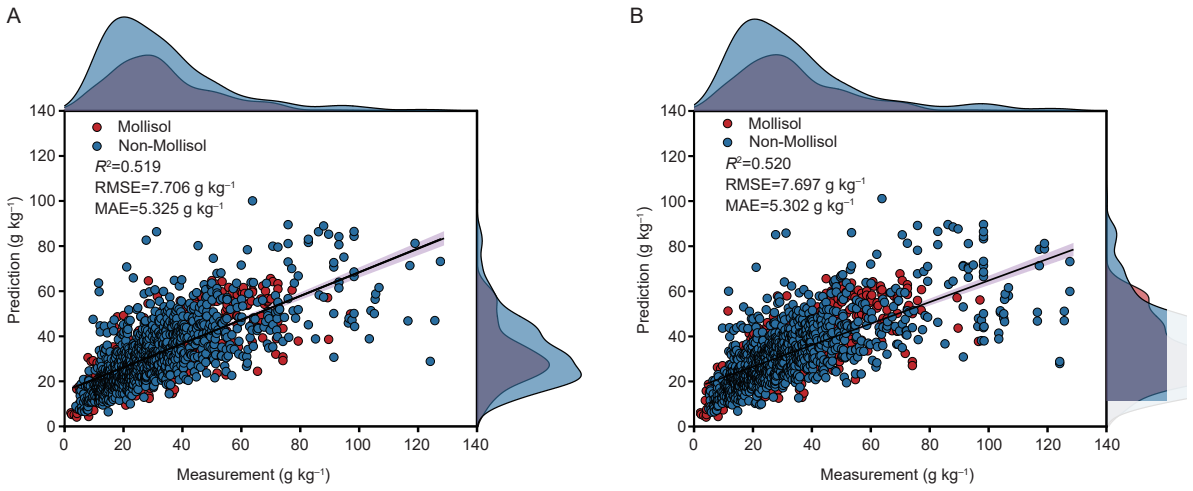


Fig. 5 Scatter plot of prediction accuracy using local regression with remote sensing bands only. A, local regression based on FAO zoning. B, local regression based on remote sensing zoning.

0.848 and 0.701 g kg^{-1}, respectively. Overall, the results demonstrate that the remote-sensing-based partitioned local regression framework maintains superior robustness and predictive accuracy compared to FAO-based local models. They also indicate that incorporating topographic and climatic covariates significantly enhances the spatial predictive capability of SOM.

3.4. Accuracy of different feature selection algorithms

In this study, three feature selection algorithms, RFE, Boruta and Lasso, were used to optimize the input variables of local regression models based on remote sensing zones (Table 6). The results show that both Boruta and Lasso algorithms select almost all input features in both zones with the same combination of selected variables, indicating that these two methods lack differentiation in variable selection. In contrast, the RFE algorithm demonstrated a stronger feature screening ability, and was able to extract more representative variable combinations based on environmental features in different regions. In the Mollisol region, the features selected by RFE included Temp, B4, B2, B3, B5, B1, and B6; in the non-Mollisol region, the selected variables were Temp, DEM, B1, B2, B4, Pre, B5, B6, and B3. The result indicates that the RFE algorithm demonstrates higher sensitivity and adaptability in capturing the relationships between regional environmental variables and SOM responses. This advantage enables it to more effectively reveal the heterogeneity mechanisms between Mollisol and non-Mollisol regions.

Feature selection plays an active role in improving the prediction performance of local regression models based on remote sensing zones. The results showed that both RFE and Boruta algorithms significantly improved the prediction accuracy of the model, with the RFE algorithm performing optimally (Fig. 7-A and B). Under RFE screening, the model performance showed a clear improvement, with R^2 increasing to 0.627 and RMSE decreasing to 6.781 g kg^{-1}. Compared with the unoptimized model, R^2 increased by 0.008 and RMSE decreased by 0.094 g kg^{-1}, demonstrating the stronger effectiveness of RFE in variable optimization. In contrast, the Lasso algorithm failed to significantly improve the model performance, and the prediction accuracy was still R^2=0.619, RMSE=6.849 g kg^{-1}, and MAE=4.601 g kg^{-1} (Fig. 7-C).

In summary, RFE can not only effectively screen out the key features reflecting the regional heterogeneity response, but also further improve the predictive ability of the model, showing greater adaptability and stability in SOM spatial distribution modeling, which is suitable for being an important part of a high-performance modeling scheme.

3.5. SOM mapping results

Based on the 2014–2023 Landsat-8 multi-temporal remote sensing images, the spatial distribution of SOM in the study area was mapped using the local regression method of remote sensing zoning and combining the optimal features screened by the RFE algorithm (Fig. 8-A). Overall, the SOM value showed a spatial pattern of decreasing from northeast to southwest. Analyzed from a zonal perspective, the average SOM value in the Mollisol zone was 32.700 g kg^{-1}, which was slightly higher than that in the non-Mollisol zone at 32.614 g kg^{-1}, and the

Table 5 Evaluation of soil organic matter prediction accuracy using local regression models based on FAO zones and remote sensing zones with remote sensing bands as input variables

Model[1)]	Input Features	Zone	R^2	RMSE (g kg^{-1})	MAE (g kg^{-1})
F	Bands	Mollisol	0.567	6.612	4.567
		Non-Mollisol	0.498	8.389	5.715
		–	0.519	7.706	5.325
R	Bands	Mollisol	0.596	6.338	4.564
		Non-Mollisol	0.484	8.399	5.812
		–	0.520	7.697	5.302

[1)] F, the local regression model based on FAO zoning; R, the local regression model based on remote sensing zoning.

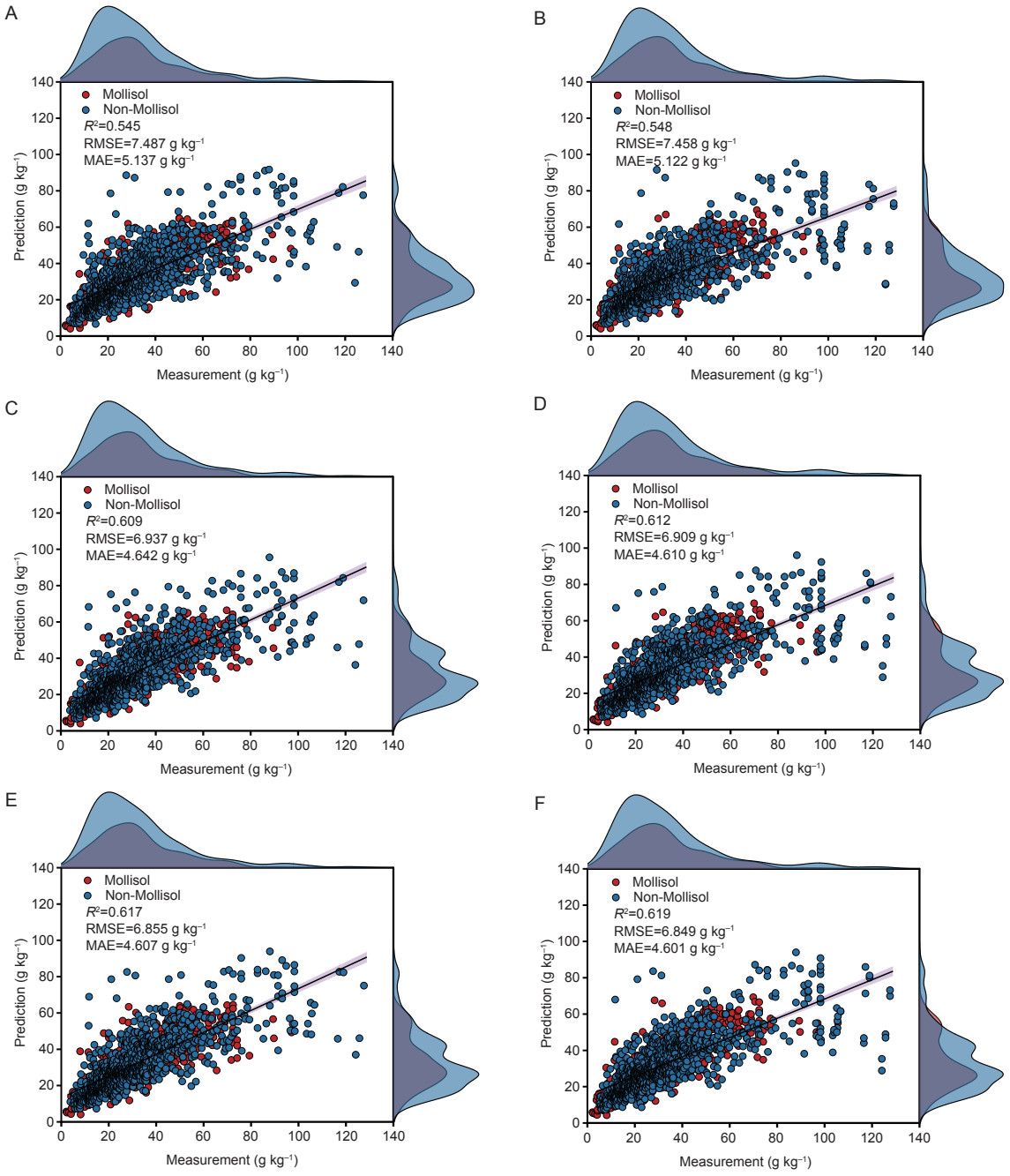


Fig. 6 Scatter plot of prediction accuracy for local regression models based on FAO zoning and remote sensing-based zoning, incorporating different combinations of environmental covariates. A and B, local regression analyses conducted using remote sensing bands and terrain data. C and D, local regression analyses conducted using remote sensing bands and climate data. E and F, local regression analyses conducted using remote sensing bands, terrain data, and climate data.

difference between the two was relatively small. However, a significant difference was observed in the spatial variability of SOM between the two zones. The standard deviation in the non-Mollisol zone reached 15.456 g kg^{-1}, which was 1.344 g kg^{-1} higher than that in the Mollisol zone (14.122 g kg^{-1}), indicating stronger spatial heterogeneity of SOM distribution in the non-Mollisol area (Fig. 8-B and C). This variability may stem from the more complex topographic relief, diverse land cover types, and more pronounced climatic and environmental gradient changes in the non-Mollisol zone (Zhou *et al.* 2019). This result further highlights the significant value of the synergistic application of remote sensing zoning strategies and feature selection methods in enhancing the spatial modeling accuracy of SOM and revealing the intra-regional heterogeneity.

4. Discussion

4.1. Spatial differences between FAO zoning and remote sensing zoning

When carrying out SOM spatial prediction at the regional

scale, the accurate delineation of the Mollisol and non-Mollisol zones not only determines the scientific validity of the modeling stratification strategy, but also directly affects the cognitive accuracy and quantitative explanation ability of the soil evolution process. In this study, we systematically compared the differences in spatial expression between two typical Mollisol zone delineation methods. One is a zoning method based on the FAO Global Mollisol Database, which is built on a wide-area soil classification system that integrates low-resolution soil survey data, theory of soil formation and expert knowledge judgment (Jones and Thornton 2015; Schad *et al.* 2015). The zoning method has strong global consistency and cross-regional comparability, and is suitable for macro-scale soil resource management and policy development. However, the FAO-based zoning method shows evident limitations when applied at the regional scale due to its low spatial resolution and static classification framework. Consequently, it fails to adequately represent the complex spatial structure of the transitional zones between Mollisol and non-Mollisol. This limitation leads to blurred boundaries, omission of localized patches, and an excessive generalization of topography-related heterogeneity (Fig. 9-A and B).

The other method is based on Landsat-8 multi-temporal remote sensing images and combines regional environmental covariates (e.g., topography, climate, etc.) to construct remote sensing zones. Compared with the FAO method, the remote sensing zoning has significant advantages in terms of the richness of data sources, timeliness and spatial resolution, and it can dynamically reflect the coupling relationship between surface reflectance features and soil-forming environment. With the integrated modeling of RF algorithm, this method can effectively identify the spatial distribution of Mollisol patches, especially in areas with significant spatial

Table 6 Selection of optimal features in each zone within remote sensing zones under different feature selection algorithms

Feature selection algorithmm[1)]	Optimal features[2)]
RFE+RF (R_M)	Temp, B4, B2, B3, B5, B1, B6
RFE+RF (R_N)	Temp, Elev, B1, B2, B4, Pre, B5, B6, B3
Boruta+RF (R_M)	B1, B2, B3, B4, B5, B6, Elev, Temp, Pre
Boruta+RF (R_N)	B1, B2, B3, B4, B5, B6, Elev, Temp, Pre
Lasso+RF (R_M)	B1, B2, B3, B4, B5, B6, Elev, Slope, Temp, Pre
Lasso+RF (R_N)	B1, B2, B3, B4, B5, B6, Elev, Slope, Temp, Pre

[1)] RFE, the recursive feature elimination algorithm; RF, the random forest algorithm; R_M, the Mollisol zone in the remote sensing-based zoning; R_N, the non-Mollisol zone in the remote sensing-based zoning.
[2)] Temp, the multi-year average temperature; B4, the NIR in Landsat-8; B2, the Green in Landsat-8; B3, the Red in Landsat-8; B5, the SWIR1 in Landsat-8; B1, the Blue in Landsat-8; B6, the SWIR2 in Landsat-8; Elev, the elevation; Pre, the multi-year average precipitation.

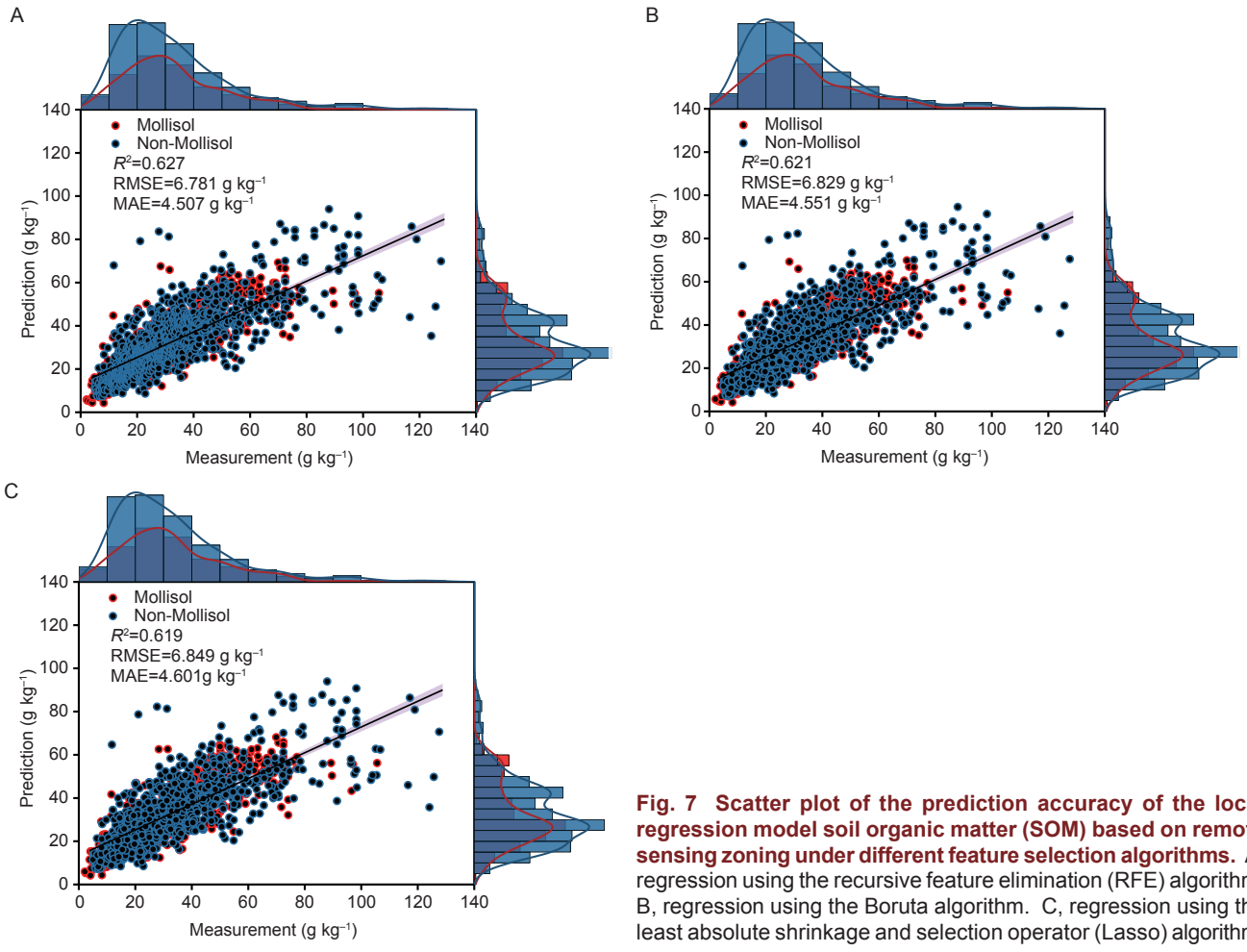
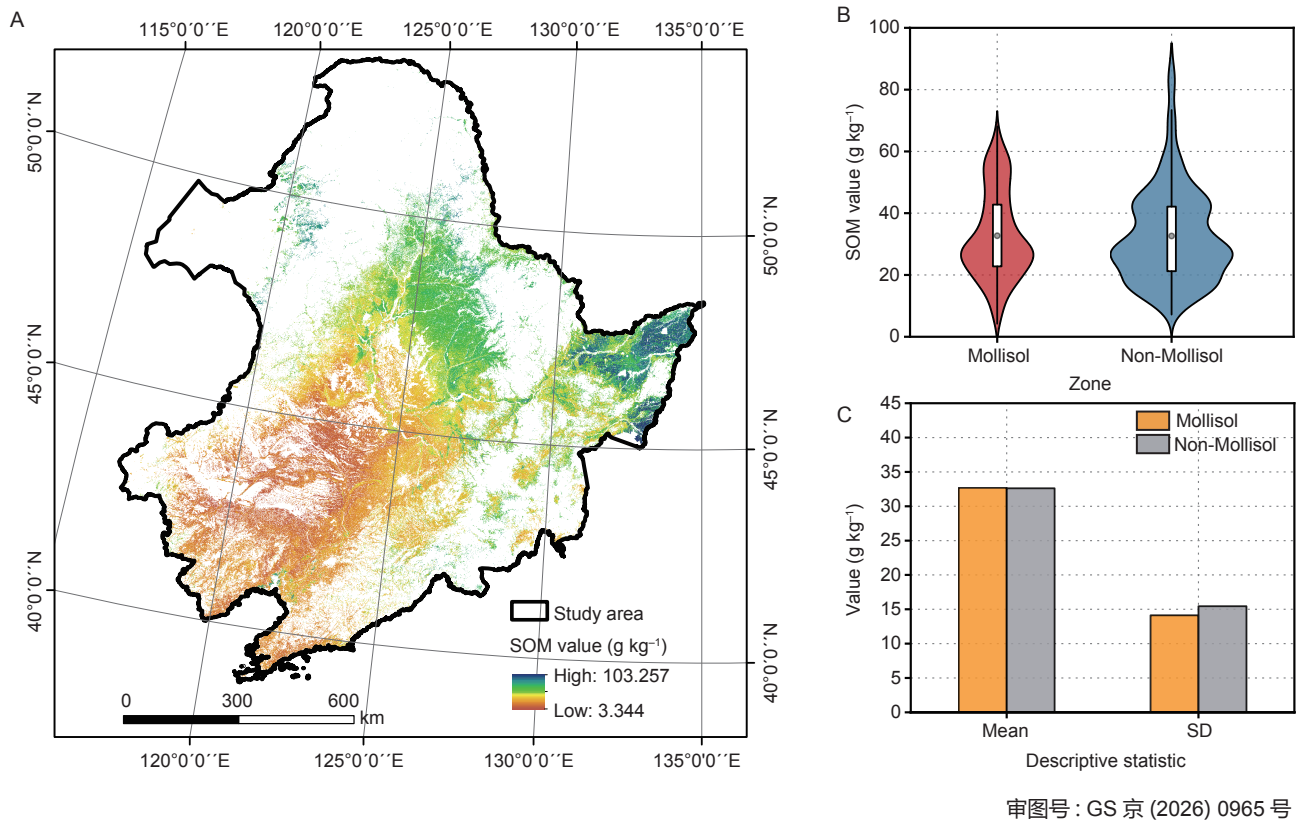


Fig. 7 Scatter plot of the prediction accuracy of the local regression model soil organic matter (SOM) based on remote sensing zoning under different feature selection algorithms. A, regression using the recursive feature elimination (RFE) algorithm. B, regression using the Boruta algorithm. C, regression using the least absolute shrinkage and selection operator (Lasso) algorithm.



审图号 : GS 京 (2026) 0965 号

Fig. 8 Spatial distribution map of soil organic matter (SOM) value in the study area. A, SOM inversion results based on remote sensing zoning strategy. B, concentration of SOM value in Mollisol and non-Mollisol areas. C, mean and standard deviation of SOM value in Mollisol and non-Mollisol areas.

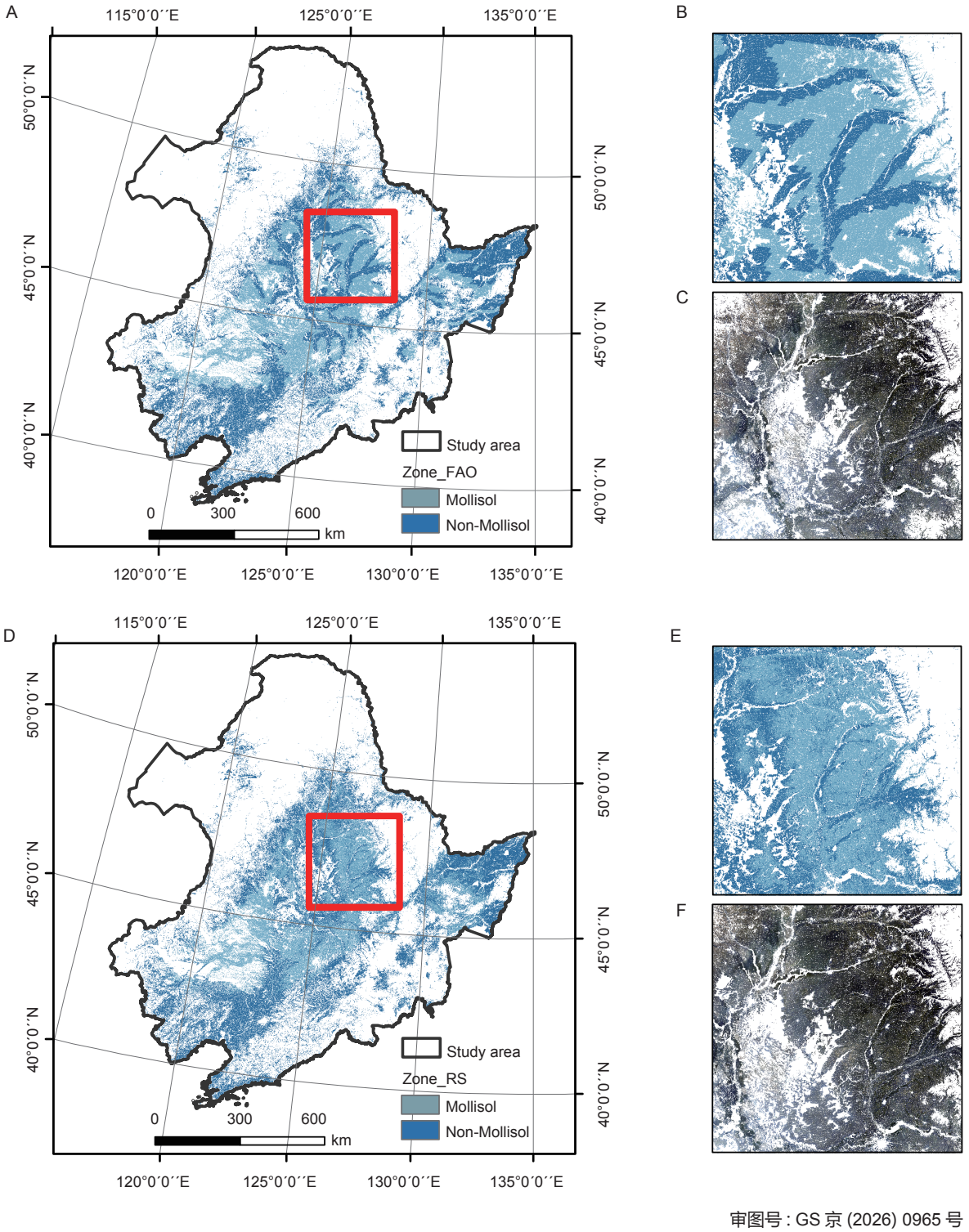
heterogeneity, such as marginal, fragmented and degraded areas, and shows higher identification sensitivity and discrimination ability (Fig. 9-D and E). This zoning strategy, which is closely matched with regional ecological processes, effectively improves the accuracy of portraying the spatial structural complexity of the Mollisol area, and enhances the adaptability and generalization ability of the model in the context of heterogeneity.

In summary, FAO zoning is suitable for macro-assessment of soil resources at global or national scales, while remote sensing zoning is more suitable for regional-scale SOM prediction and fine modeling tasks. It has significant advantages in capturing the dynamic evolution characteristics of Mollisol and supporting regional adaptive modeling, which provides a key support for the in-depth understanding of the spatial evolution mechanism of Mollisol and the enhancement of SOM mapping accuracy.

4.2. Improving SOM prediction accuracy through remote sensing-based zoning modeling and feature optimization

Reasonable spatial zoning and feature optimization strategies play a key role in enhancing the spatial prediction accuracy of SOM (Gomes *et al.* 2019; Guo *et al.* 2019). To systematically evaluate the effectiveness of zoning modeling, this study constructed two models for comparison. Results indicate that when all variables are included as inputs, the model based on remote sensing zoning outperforms the local regression strategy using FAO zoning (Table 7). This improvement stems from the significant enhancement in the model's response to spatial variations in SOM, achieved through the coupling of remote sensing bands with environmental covariates (Grimm *et al.* 2008; Guo *et al.* 2015; Hounkpatin *et al.* 2021). Further comparisons revealed that prediction accuracies were generally higher in the Mollisol region than in the non-Mollisol region, reflecting the significant non-homogeneity of the SOM distribution within the region and re-emphasizing the importance of employing local models to accommodate spatial variability.

To address the variable redundancy problem, three algorithms, RFE, Boruta and Lasso, are introduced in this study for evaluation, and the results show that the RFE algorithm is the most superior in terms of variable screening effect and prediction accuracy enhancement. The optimized feature combinations using RFE significantly improved the prediction performance of the models in each zone (Table 8). It is worth noting that there are differences in the optimal variable combinations screened in different subregions, reflecting the regional variation characteristics of the main control factors of SOM. Specifically, in the Mollisol zone, the surface SOM value is high and the exposure period is long, and the multi-temporal remote sensing band is more sensitive to the soil spectra, which can effectively capture the spatial and temporal heterogeneity of SOM, and thus the remote sensing band is the dominant predictor in this zone; whereas, the non-



审图号：GS 京 (2026) 0965 号

Fig. 9 Comparison of the spatial distribution results of Mollisol and non-Mollisol under two zoning methods. A, distribution results based on FAO zoning. B, local map based on FAO zoning results. C, local map of Landsat-8 remote sensing imagery in the study area. D, distribution results based on remote sensing (RS) imagery. E, local map based on remote sensing zoning results. F, local map of Landsat-8 remote sensing imagery in the study area.

Mollisol area is characterized by large topographic relief and pronounced climatic gradients, SOM in this region is jointly regulated by both terrain and climate. Variables such as slope, elevation, and surface temperature effectively represent the influence of ecological conditions on SOM accumulation (Guo *et al.* 2013; Sun *et al.* 2015). Consequently, the contribution of topographic and climatic variables to SOM prediction is more significant in this region.

In summary, this study demonstrates that the synergistic strategy of remote sensing zoning and feature optimization can significantly enhance the spatial prediction accuracy

Table 7 Total soil organic matter (SOM) prediction accuracy of the local regression model and SOM prediction accuracy of each zone

Model[1)]	R^2	RMSE (g kg^{-1})	MAE (g kg^{-1})
F	0.617	6.855	4.607
R	0.619	6.849	4.601
F_M	0.639	5.762	4.076
F_N	0.608	7.414	4.905
R_M	0.669	5.742	4.074
R_N	0.597	7.429	4.993

[1)] F, the local regression model based on FAO zoning; R, the local regression model based on remote sensing zoning; F_M, the Mollisol zone in the FAO-based zoning; F_N, the non-Mollisol zone in the FAO-based zoning; R_M, the Mollisol zone in the remote sensing-based zoning; R_N, the non-Mollisol zone in the remote sensing-based zoning.

Table 8 Soil organic matter (SOM) prediction accuracy of the optimal zoning strategy under the RFE algorithm

Model[1)]	Optimal features[2)]	R^2	RMSE (g kg^{-1})	MAE (g kg^{-1})
R_M	Temp, B4, B2, B3, B5, B1, B6	0.675	5.609	4.063
R_N	Temp, Elev, B1, B2, B4, Pre, B5, B6, B3	0.601	7.384	4.992
R	–	0.627	6.781	4.507

[1)] R_M, the Mollisol zone in the remote sensing-based zoning; R_N, the non-Mollisol zone in the remote sensing-based zoning; R, the local regression model based on remote sensing zoning.
[2)] Temp, the multi-year average temperature; B4, the NIR in Landsat-8; B2, the Green in Landsat-8; B3, the Red in Landsat-8; B5, the SWIR1 in Landsat-8; B1, the Blue in Landsat-8; B6, the SWIR2 in Landsat-8; Elev, the elevation; Pre, the multi-year average precipitation.

of SOM, which is superior to the ability to characterize the spatial heterogeneity of the traditional FAO delineation. This strategy enhances the regional adaptability of the model, which in turn is applicable to the fine modeling of soil-forming-environment-driven Mollisol and non-Mollisol intertwined zones.

4.3. Limitations and future prospects

In this study, we developed a spatial prediction framework for SOM based on optimized Mollisol and non-Mollisol zoning (RSZ-FSO-RF). By integrating remote sensing-based regionalization, feature selection optimization, and the RF algorithm, this framework substantially improved SOM mapping accuracy in the typical interlaced zones of Mollisol and non-Mollisol areas in Northeast China, demonstrating strong regional adaptability and modeling efficiency. However, the following limitations still exist: (1) The spatial distribution of the samples is not yet balanced, and the sampling density of some ecological transition zones and highly heterogeneous regions is low, which limits the generalization ability of the local regression model in such regions and may trigger prediction errors; (2) this study is mainly based on the RF algorithm for modeling, and has not yet systematically compared the modeling advantages and applicability of other machine learning methods under diverse ecological backgrounds and spatial heterogeneity; (3) the geographic zoning strategy is only based on soil type classification, and the effects of different zoning schemes (e.g., regular grids, administrative boundaries, ecological zoning, etc.) on modeling performance have not been fully explored (Brungard *et al.* 2021; Du *et al.* 2024). There is a lack of systematic comparative studies on modeling unit scale and boundary definition of multiple zoning approaches. In view of the above limitations, future research should focus on the following aspects: firstly, optimize the sample deployment strategy, especially to strengthen the sampling density and environmental gradient coverage of ecological transition zones and high heterogeneity areas, and construct a more representative and robust training sample set through the introduction of spatially optimized sampling techniques; secondly, incorporate a variety of advanced machine learning algorithms (e.g., neural networks, deep learning, etc.), systematically evaluate their modeling performance in different ecological contexts, and identify the most regionally-adaptive modeling strategies and hyper-parameter configurations (Ou *et al.* 2021; Odebiri *et al.* 2022); finally, it is necessary to construct and compare various spatial zoning schemes, clarify the applicability of different zoning rules under complex geomorphological and environmental conditions, and deeply explore the influence of scale effect and boundary sensitivity of modeling units on the prediction results, so as to construct a more stable and optimal modeling zoning system applicable to specific regional characteristics.

5. Conclusion

This study developed an integrated RSZ-FSO-RF framework combining remote sensing zoning, feature selection optimization, and random forest modeling techniques to enhance spatial prediction capabilities for SOM in landscapes with intermingled Mollisol and non-Mollisol. Local models stratified by remote-sensing-derived zones consistently outperformed those based on FAO zoning in both accuracy and stability, and the incorporation of a targeted feature-selection strategy yielded further gains in predictive performance. Distinct spatial patterns and controlling factors were observed between Mollisol and non-Mollisol areas: Mollisol zones exhibited higher and more spatially uniform SOM, whereas non-Mollisol zones showed lower mean SOM but greater spatial variability, reflecting more complex topographic-climatic coupling and soil-forming processes. These results validate the necessity of zone-specific modeling and highlight the central roles of remote-sensing zoning and feature selection in constructing regional adaptive models. The RSZ-FSO-RF framework thus provides a theoretical basis and practical pathway for high-precision SOM mapping in complex ecological regions. Future work should test the framework across varied soil types and climate regimes and pursue long-term, mechanism-driven monitoring of soil carbon stocks by integrating higher spatio-temporal remote sensing and multi-source environmental data.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (42401460) and the National Key R&D Program of China (2021YFD1500100). Sincere thanks to anonymous reviewers and members of the editorial team, for the comments and contributions.

Declaration of competing interest

The authors declare that they have no conflict of interest.

Declaration of generative AI and AI-assisted technologies in the writing process

The authors declare that they did not use AI in the preparation and writing of this manuscript.

References

Ali Z, Hussain I, Faisal M, Elashkar E E, Gani S, Shehzad M A. 2019. Selection of appropriate time scale with Boruta algorithm for regional drought monitoring using multi-scaler drought index. *Tellus* (A: Dynamic Meteorology and Oceanography), **71**, 1604057.

Alshehri F, Shahfahad, Rahman A. 2024. Optimizing machine learning models with bayesian techniques for prediction of groundwater quality index in Southwest Saudi Arabia. *Earth Systems and Environment*, **8**, 1417–1436.

Bao N, Wu L, Ye B, Yang K, Zhou W. 2017. Assessing soil organic matter of reclaimed soil from a large surface coal mine using a field spectroradiometer in laboratory. *Geoderma*, **288**, 47–55.

Belgiu M, Drăguț L. 2016. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, **114**, 24–31.

Breiman L. 2001. Random forests. *Machine Learning*, **45**, 5–32.

Broeg T, Don A, Gocht A, Scholten T, Taghizadeh-Mehrjardi R, Erasmi S. 2024. Using local ensemble models and Landsat bare soil composites for large-scale soil organic carbon maps in cropland. *Geoderma*, **444**, 116850.

Brungard C, Nauman T, Duniway M, Veblen K, Nehring K, White D, Salley S, Anchang J. 2021. Regional ensemble modeling reduces uncertainty for digital soil mapping. *Geoderma*, **397**, 114998.

Chen R C, Dewi C, Huang S W, Caraka R E. 2020. Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, **7**, 52.

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

Chen Y, Ma L, Yu D, Zhang H, Feng K, Wang X, Song J. 2022. Comparison of feature selection methods for mapping soil organic matter in subtropical restored forests. *Ecological Indicators*, **135**, 108545.

Conforti M, Castrignanò A, Robustelli G, Scarciglia F, Stelluti M, Buttafuoco G. 2015. Laboratory-based Vis–NIR spectroscopy and partial least square regression with spatially correlated errors for predicting spatial variation of soil organic matter content. *Catena*, **124**, 60–67.

Dan W, Nianpeng H, Qing W, Yuliang L, Qiufeng W, Zhiwei X, Jianxing Z. 2016. Effects of temperature and moisture on soil organic matter decomposition along elevation gradients on the Changbai Mountains, Northeast China. *Pedosphere*, **26**, 399–407.

Dou X, Wang X, Liu H, Zhang X, Meng L, Pan Y, Yu Z, Cui Y. 2019. Prediction of soil organic matter using multi-temporal satellite images in the Songnen Plain, China. *Geoderma*, **356**, 113896.

Du R, Xiang Y, Chen J, Lu X, Wu Y, He Y, Xiang R, Zhang Z, Chen Y. 2024. Potential of solar-induced chlorophyll fluorescence (SIF) to access long-term dynamics of soil salinity using OCO-2 satellite data and machine learning method. *Geoderma*, **444**, 116855.

Franklin R B, Mills A L. 2009. Importance of spatially structured environmental heterogeneity in controlling microbial community composition at small spatial scales in an agricultural field. *Soil Biology and Biochemistry*, **41**, 1833–1840.

Gao Y, Huang D, Zhang Y, McLaughlin N, Zhang Y, Wang Y, Chen X, Zhang S, Lu Y, Liang A. 2024. Precipitation increment reinforced warming-induced increases in soil mineral-associated and particulate organic matter under agricultural ecosystem. *Applied Soil Ecology*, **196**, 105301.

Gomes L C, Faria R M, de Souza E, Veloso G V, Schaefer C E G, Fernandes Filho E I. 2019. Modelling and mapping soil organic carbon stocks in Brazil. *Geoderma*, **340**, 337–350.

Grimm R, Behrens T, Märker M, Elsenbeer H. 2008. Soil organic carbon concentrations and stocks on Barro Colorado Island-Digital soil mapping using Random Forests analysis. *Geoderma*, **146**, 102–113.

Guo P T, Li M F, Luo W, Tang Q F, Liu Z W, Lin Z M. 2015. Digital mapping of soil organic matter for rubber plantation at regional scale: An application of random forest plus residuals kriging approach. *Geoderma*, **237**, 49–59.

Guo P T, Wu W, Sheng Q K, Li M F, Liu H B, Wang Z Y. 2013. Prediction of soil organic matter using artificial neural network and topographic indicators in hilly areas. *Nutrient Cycling in Agroecosystems*, **95**, 333–344.

Guo Z, Adhikari K, Chellasamy M, Greve M B, Owens P R, Greve M H. 2019. Selection of terrain attributes and its scale dependency on soil organic carbon prediction. *Geoderma*, **340**, 303–312.

Gutiérrez-Girón A, Díaz-Pinés E, Rubio A, Gavilán R G. 2015. Both altitude and vegetation affect temperature sensitivity of soil organic matter decomposition in Mediterranean high mountain soils. *Geoderma*, **237**, 1–8.

He S, Tan S, Shen L, Zhou Q. 2023. Soil organic matter estimation model integrating spectral and profile features. *Sensors*, **23**, 9868.

Henderson T, Baumgardner M, Franzmeier D, Stott D, Coster D. 1992. High dimensional reflectance analysis of soil organic matter. *Soil Science Society of America Journal*, **56**, 865–872.

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**, e0169748.

Hounkpatin K O, Stendahl J, Lundblad M, Karltun E. 2021. Predicting the spatial distribution of soil organic carbon stock in Swedish forests using a group of covariates and site-specific data. *Soil*, **7**, 377–398.

Huang H, Yang L, Zhang L, Pu Y, Yang C, Wu Q, Cai Y, Shen F, Zhou C. 2022. A review on digital mapping of soil carbon in cropland: Progress, challenge, and prospect. *Environmental Research Letters*, **17**, 123004.

John K, Afu S, Isong I, Aki E, Kebonye N, Ayito E, Chapman P, Eyong M, Penížek V. 2021. Mapping soil properties with soil-environmental covariates using geostatistics and multivariate statistics. *International Journal of Environmental Science and Technology*, **18**, 1–16.

Jones P G, Thornton P K. 2015. Representative soil profiles for the Harmonized World Soil Database at different spatial resolutions for agricultural modelling applications. *Agricultural Systems*, **139**, 93–99.

Kebonye N M, Agyeman P C, Biney J K. 2023. Optimized modelling of countrywide soil organic carbon levels via an interpretable decision tree. *Smart Agricultural Technology*, **3**, 100106.

Keskin H, Grunwald S, Harris W G. 2019. Digital mapping of soil carbon fractions with machine learning. *Geoderma*, **339**, 40–58.

Ladoni M, Bahrami H A, Alavipanah S K, Norouzi A A. 2010. Estimating soil organic carbon from soil reflectance: A review. *Precision Agriculture*, **11**, 82–99.

Liang H, Japaer G, Yu T, Zhang L, Chen B, Lin K, Ju T, Zhao Y, Pei T, Aizizi Y. 2024. Comparison of global and zonal modeling strategies-A case study of soil organic matter and C: N ratio mapping in Altay, Xinjiang, China. *Ecological Informatics*, **84**, 102882.

Liu H, Zhang Y, Zhang B. 2009. Novel hyperspectral reflectance models for estimating black-soil organic matter in Northeast China. *Environmental Monitoring and Assessment*, **154**, 147–154.

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. *Canadian Journal of Soil Science*, **92**, 383–402.

Luo C, Wang Y, Zhang X, Zhang W, Liu H. 2022a. Spatial prediction of soil organic matter content using multiyear synthetic images and partitioning algorithms. *Catena*, **211**, 106023.

Luo C, Zhang W, Meng X, Yu Y, Zhang X, Liu H. 2024. Mapping the soil organic matter content in Northeast China considering the

difference between dry lands and paddy fields. *Soil and Tillage Research*, **244**, 106270.
Luo C, Zhang X, Meng X, Zhu H, Ni C, Chen M, Liu H. 2022b. Regional mapping of soil organic matter content using multitemporal synthetic Landsat 8 images in Google Earth Engine. *Catena*, **209**, 105842.
Mahmoudzadeh H, Matinfar H R, Taghizadeh-Mehrjardi R, Kerry R. 2020. Spatial prediction of soil organic carbon using machine learning techniques in western Iran. *Geoderma Regional*, **21**, e00260.
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 Sensing of Environment*, **318**, 114592.
Minasny B, McBratney A B. 2016. Digital soil mapping: A brief history and some lessons. *Geoderma*, **264**, 301–311.
Mulder V, De Bruin S, Schaepman M E, Mayr T. 2011. The use of remote sensing in soil and terrain mapping — A review. *Geoderma*, **162**, 1–19.
Mustafa A, Hu X, Abrar M M, Shah S A A, Sun N, Saeed Q, Kamran M, Naveed M, Conde-Cid M, Gao H J, Zhu P, Xu M G. 2021. Long-term fertilization enhanced carbon mineralization and maize biomass through physical protection of organic carbon in fractions under continuous maize cropping. *Applied Soil Ecology*, **165**, 103971.
Muthukrishnan R, Rohini R. 2016. LASSO: A feature selection technique in predictive modeling for machine learning. In: *2016 IEEE International Conference on Advances in Computer Applications* (ICACA). Institute of Electrical and Electronics Engineers, India. pp. 18–20.
Niu X, Zhang S, Zhang C, Yan P, Wang H, Xu W, Song M, Aurangzeib M. 2024. Key factors influencing the spatial distribution of soil organic carbon and its fractions in Mollisols. *Catena*, **247**, 108522.
Odebiri O, Mutanga O, Odindi J. 2022. Deep learning-based national scale soil organic carbon mapping with Sentinel-3 data. *Geoderma*, **411**, 115695.
Ou D, Tan K, Lai J, Jia X, Wang X, Chen Y, Li J. 2021. Semi-supervised DNN regression on airborne hyperspectral imagery for improved spatial soil properties prediction. *Geoderma*, **385**, 114875.
Peng L, Wu X, Feng C, Gao L, Li Q, Xu J, Li B. 2024. Assessing the potential of multi-source remote sensing data for cropland soil organic matter mapping in hilly and mountainous areas. *Catena*, **245**, 108312.
Rosero-Vlasova O, Alves D B, Vlassova L, Perez-Cabello F, Lloveria R M. 2017. Modeling soil organic matter (SOM) from satellite data using VISNIR-SWIR spectroscopy and PLS regression with step-down variable selection algorithm: Case study of Campos Amazonicos National Park savanna enclave, Brazil. In: *Remote Sensing for Agriculture*, *Ecosystems*, *and Hydrology XIX*. International Society for Optics and Photonics, Ecuador. pp. 166–172.
Schad P. 2015. The international soil classification system WRB, 2014. In: *Novel Methods for Monitoring and Managing Land and Water Resources in Siberia*. Springer, Germany. pp. 563–571.
Shi Y, Zhao J, Song X, Qin Z, Wu L, Wang H, Tang J. 2021. Hyperspectral band selection and modeling of soil organic matter content in a forest using the Ranger algorithm. *PLoS ONE*, **16**, e0253385.
Shonk J, Gaultney L, Schulze D, Van Scoyoc G. 1991. Spectroscopic sensing of soil organic matter content. *Transactions of the ASAE*, **34**, 1978–1984.
Sothe C, Gonsamo A, Arabian J, Snider J. 2022. Large scale mapping of soil organic carbon concentration with 3D machine learning and satellite observations. *Geoderma*, **405**, 115402.
Sun W, Zhu H, Guo S. 2015. Soil organic carbon as a function of land use and topography on the Loess Plateau of China. *Ecological Engineering*, **83**, 249–257.
Visalakshi S, Radha V. 2014. A literature review of feature selection techniques and applications: Review of feature selection in data mining. In: *2014 IEEE International Conference on Computational Intelligence and Computing Research*. Institute of Electrical and Electronics Engineers, India. pp. 1–6.
Vohland M, Ludwig M, Thiele-Bruhn S, Ludwig B. 2014. Determination of soil properties with visible to near-and mid-infrared spectroscopy: Effects of spectral variable selection. *Geoderma*, **223**, 88–96.
Wang B, Liu D, Yang J, Zhu Z, Darboux F, Jiao J, An S. 2021. Effects of forest floor characteristics on soil labile carbon as varied by topography and vegetation type in the Chinese Loess Plateau. *Catena*, **196**, 104825.
Wang D, Fan H. 2019. Distribution characteristics of gullies with slope gradient in Northeast China. *Environmental Monitoring and Assessment*, **191**, 1–12.
Wang H, Yang S, Wang Y, Gu Z, Xiong S, Huang X, Sun M, Zhang S, Guo L, Cui J. 2022. Rates and causes of black soil erosion in Northeast China. *Catena*, **214**, 106250.
Wang H, Zhang X, Wu W, Liu H. 2021. Prediction of soil organic carbon under different land use types using Sentinel-1/-2 data in a small watershed. *Remote Sensing*, **13**, 1229.
Wang W, Deng X, Yue H. 2024. Black soil conservation will boost China's grain supply and reduce agricultural greenhouse gas emissions in the future. *Environmental Impact Assessment Review*, **106**, 107482.
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.
Weil R R, Magdoff F. 2004. Significance of soil organic matter to soil quality and health. In: *Soil Organic Matter in Sustainable Agriculture*. CRC Press, USA. pp. 1–43.
van der Westhuizen S, Heuvelink G B, Hofmeyr D P. 2023. Multivariate random forest for digital soil mapping. *Geoderma*, **431**, 116365.
Xu M, Li X, Cai X, Gai J, Li X, Christie P, Zhang J. 2014. Soil microbial community structure and activity along a montane elevational gradient on the Tibetan Plateau. *European Journal of Soil Biology*, **64**, 6–14.
Yang H, Zhang X, Xu M, Shao S, Wang X, Liu W, Wu D, Ma Y, Bao Y, Zhang X. 2020. Hyper-temporal remote sensing data in bare soil period and terrain attributes for digital soil mapping in the Black soil regions of China. *Catena*, **184**, 104259.
You N, Dong J. 2020. Examining earliest identifiable timing of crops using all available Sentinel 1/2 imagery and Google Earth Engine. *ISPRS Journal of Photogrammetry and Remote Sensing*, **161**, 109–123.
Zhang J, Okin G S, Zhou B. 2019. Assimilating optical satellite remote sensing images and field data to predict surface indicators in the Western US: Assessing error in satellite predictions based on large geographical datasets with the use of machine learning. *Remote Sensing of Environment*, **233**, 111382.
Zhang M, Shi W, Ren Y, Wang Z, Ge Y, Guo X, Mao D, Ma Y. 2022. Proportional allocation with soil depth improved mapping soil organic carbon stocks. *Soil and Tillage Research*, **224**, 105519.
Zhang S, Zhang X, Liu Z, Sun Y, Liu W, Dai L, Fu S. 2014. Spatial heterogeneity of soil organic matter and soil total nitrogen in a Mollisol watershed of Northeast China. *Environmental Earth Sciences*, **72**, 275–288.
Zhang X, Chen S, Xue J, Wang N, Xiao Y, Chen Q, Hong Y, Zhou Y, Teng H, Hu B. 2023. Improving model parsimony and accuracy by modified greedy feature selection in digital soil mapping. *Geoderma*, **432**, 116383.
Zhang Y, Wei L, Lu Q, Zhong Y, Yuan Z, Wang Z, Li Z, Yang Y. 2023. Mapping soil available copper content in the mine tailings pond with combined simulated annealing deep neural network and UAV hyperspectral images. *Environmental Pollution*, **320**, 120962.
Zhou Y, Biswas A, Ma Z, Lu Y, Chen Q, Shi Z. 2016. Revealing the scale-specific controls of soil organic matter at large scale in Northeast and North China Plain. *Geoderma*, **271**, 71–79.
Zhou Y, Hartemink A E, Shi Z, Liang Z, Lu Y. 2019. Land use and climate change effects on soil organic carbon in North and Northeast China. *Science of the Total Environment*, **647**, 1230–1238.
Zhu M, Feng Q, Qin Y, Cao J, Zhang M, Liu W, Deo R C, Zhang C, Li R, Li B. 2019. The role of topography in shaping the spatial patterns of soil organic carbon. *Catena*, **176**, 296–305.

Executive Editor-in-Chief Keke Yi
Managing Editor Lujuan Sun