



OPEN Integrating transformer-based learning and Sentinel-2 bare soil composites for soil organic carbon mapping in the black soil region of Northeast China

Na Chen¹, Zhikang Wei¹, Xuancheng Jin², Nan Lin³, Fan Yang⁴, Ling Zhao¹ & Song Wu¹✉

Accurate assessment of soil organic carbon (SOC) is essential for sustainable cropland management and carbon sequestration monitoring. However, high-resolution SOC mapping remains challenging due to two persistent limitations: (1) the difficulty of extracting true bare-soil reflectance—especially when single-date imagery is used and spectral signals remain influenced by vegetation, residue, and soil moisture; and (2) reliance on models that require large training datasets and may underperform in typical small-sample soil survey settings. To address these challenges, we developed an approach that integrates multi-temporal Sentinel-2 bare-soil composites with a transformer-based foundation model—Tabular Prior-data Fitted Network (TabPFN)—for SOC prediction in the black soil region of Northeast China. Bare soil pixels were extracted using a Normalized Difference Vegetation Index threshold (0.1–0.4), and two compositing strategies—the 50th percentile (P50) and 90th percentile (P90)—were compared. We systematically evaluated three advanced algorithms: TabPFN, convolutional neural network (CNN), and Extreme Gradient Boosting (XGBoost). Results demonstrated that the TabPFN model coupled with P50 composites achieved the highest prediction accuracy ($R^2 = 0.78$, $RMSE = 1.90 \text{ g kg}^{-1}$), outperforming CNN and XGBoost by 4–6%. TabPFN's distinct advantage lies in its design as a prior-data fitted transformer, which enables robust generalization from limited samples ($N = 174$) without extensive hyperparameter tuning, effectively addressing the “small data” challenge pervasive in digital soil mapping. SHapley Additive exPlanations analysis indicated that shortwave infrared band (B12) and precipitation have the greatest effect on model output, indicating joint role of soil spectral response and climate variability. This is one of the first studies to apply the TabPFN architecture to SOC estimation, offering a novel, interpretable, and scalable workflow that bridges the gap between data scarcity and model complexity. The proposed framework provides a reliable tool for high-resolution SOC mapping in heterogeneous croplands, supporting precision agriculture and long-term carbon accounting initiatives.

Keywords Soil organic carbon, Sentinel-2, Digital soil mapping, Bare soil composite, TabPFN, SHAP

Soil organic carbon (SOC) plays an important role in soil fertility, agribusiness, and global carbon cycle. SOC can alter soil structure, water preservation, nutrient availability, and microbial activity, all of which influence crop yield and soil health^{1–3}. At large scale, SOC regulates atmospheric CO₂ levels, since soils are one of the planet's largest carbon reservoirs, holding more carbon than vegetation and atmosphere together⁴. As climate change intensifies and food security demands grow, a better understanding of SOC distribution across landscapes is needed^{5,6}. High-resolution SOC maps can guide soil management, support carbon sequestration initiatives, and inform land-use planning, particularly in regions with complex cropping environments⁷.

Digital Soil Mapping (DSM) has emerged as a key approach to estimate the spatial variability of soil properties using environmental covariates^{8–10}. Traditional field-based methods such as soil sampling and laboratory

¹College of Economics and Management, Jilin Agricultural University, Changchun 130118, China. ²College of Earth Sciences, Jilin University, Changchun 130061, China. ³Modern Industry College, Jilin Jianzhu University, Changchun 130118, China. ⁴College of Resources and Environment, Jilin Agricultural University, Changchun 130118, China. ✉email: wusong@jlau.edu.cn

analysis are costly, labor-intensive, and often unable to capture fine-scale spatial heterogeneity over large areas¹¹. DSM overcomes these limitations by integrating field observations with spatially continuous predictors derived from remote sensing, climate, and terrain data¹². Sentinel-2 satellite imagery, with its high spatial resolution (10–20 m), frequent revisit time, and multiple spectral bands, has been widely adopted in DSM for SOC prediction¹³. Combining remote sensing variables with machine learning algorithms can make SOC mapping faster and more cost-effective compared to purely sample-based approaches¹⁴.

However, several challenges persist in the remote sensing-based prediction of SOC. First, vegetation and crop residue can mask the spectral signal of bare soil, which is essential for reliable topsoil SOC estimation^{15–17}. A common strategy to address this is applying a Normalized Difference Vegetation Index (NDVI) threshold to extract bare soil pixels, yet the optimal threshold often varies with local vegetation and environmental conditions^{18,19}. Second, due to cloud cover and complex surface conditions, multi-temporal compositing of satellite imagery is necessary to obtain stable and representative reflectance values. The 50th percentile (P50, i.e., the median) composite is the most commonly used method²⁰. P50 composite effectively suppresses noise and outliers while reflecting the most typical bare soil conditions, making it a relatively robust approach²¹. Castaldi et al.²² adopted the 90th percentile (P90) composite method to capture the driest bare soil state. More recent studies have employed different compositing strategies and conducted comparisons, with results varying across regions^{23,24}. Nevertheless, how these compositing choices affect the predictive accuracy of organic carbon models still requires further exploration.

Machine learning methods, including Random Forest, Cubist, Extreme Gradient Boosting (XGBoost) etc., have been widely applied in SOC mapping^{14,25,26}. Building on this foundation, more recent research has turned to deep learning methods. Recent studies have demonstrated the superiority of deep learning methods (e.g., Convolutional Neural Network, CNN) in SOC mapping, owing to their powerful feature extraction capabilities, which enable the automatic identification of complex spatial patterns and hierarchical relationships from multi-dimensional remote sensing data^{27–29}. Nonetheless, deep learning methods typically require large and diverse training datasets to avoid overfitting and ensure generalization, which is often impractical in large agricultural landscapes where soil sampling is constrained by cost and accessibility³⁰. This is particularly true in China's vast Northeast black soil region, where systematic sampling is difficult, resulting in limited training data. A promising alternative in data-scarce scenarios is the Tabular Prior-data Fitted Network (TabPFN), a transformer-based foundation model³¹. Unlike conventional deep learning models, TabPFN is tailored for small-sample settings, as it is pre-trained on millions of synthetic tabular tasks to learn transferable priors before fine-tuning. This enables strong predictive performance with minimal hyperparameter tuning, even when training data are limited. Moreover, its probabilistic formulation provides uncertainty-aware predictions, which is particularly valuable when working with sparse or heterogeneous environmental datasets. These characteristics make TabPFN well suited for SOC prediction under common real-world sampling constraints.

Despite progress in remote sensing-based SOC modeling, two key gaps remain: the effects of different bare-soil compositing strategies on SOC prediction are not well assessed, and the potential of transformer-based foundation models such as TabPFN under small-sample conditions has not yet been explored. To address these gaps, this study proposes a workflow that integrates Sentinel-2 multi-temporal bare-soil composites with a TabPFN model for SOC mapping in Northeast China and compares its performance against XGBoost and CNN. Specifically, we (i) evaluate the effect of different bare-soil compositing strategies (P50 composite vs. P90 composite) on model performance, (ii) assess the predictive capability and uncertainty characteristics of TabPFN in a data-limited SOC mapping context, and (iii) interpret the drivers of SOC prediction using SHapley Additive exPlanations (SHAP) analysis. This work provides one of the first empirical evaluations of integrating foundation models with bare-soil remote sensing for digital soil mapping, offering a scalable and interpretable framework suitable for data-scarce agricultural environments.

Results

Model performance comparison

The SOC content of the 174 surface soil samples exhibited considerable variability across the study area. SOC values ranged from 4.05 to 25.29 g kg⁻¹, with a mean of 14.20 g kg⁻¹ and a median of 14.33 g kg⁻¹. The standard deviation of 4.12 g kg⁻¹ reflects substantial heterogeneity in soil carbon levels. Additionally, the near-zero skewness (–0.08) suggests an approximately symmetric distribution, indicating that both low-carbon and high-carbon croplands are well represented. This variability provides a robust foundation for model calibration and validation under diverse soil conditions.

The distribution of R² and RMSE values across all bootstrap runs in Fig. 1 indicates that the performance across the three models is quite stable across the iterations, suggesting that the bootstrap strategy is not adversely affected by random sampling. A comparison of the mean performance shows that TabPFN has higher accuracy than CNN and XGBoost, respectively. Under P50 composite, TabPFN attained an average R² of 0.781 ± 0.043 and an RMSE of 1.902 ± 0.153, outperforming CNN (R² = 0.737 ± 0.042, RMSE = 2.093 ± 0.171) and XGBoost (R² = 0.715 ± 0.049, RMSE = 2.174 ± 0.166). For P90 composite, all models have a slightly lower accuracy but the same relative rank. This performance difference suggests that P90 composite may introduce additional spectral variability or noise, while the P50 composite provides more stable and representative reflectance values for SOC prediction. Likewise, the clear ranking of TabPFN > CNN > XGBoost implies that the models differ in their ability to handle uncertainty and learn non-linear relationships under these compositing strategies.

Scatter plots of observed and predicted SOC values (Fig. 2) show visual information compared to the three methods. Among all three methods, TabPFN produced the points that closest to 1:1 and tended to be a little more dispersed and biased across the full SOC range. CNN returned moderate results, while XGBoost was much more varied and spreadable. Such distributions suggest that TabPFN made more consistent and reliable

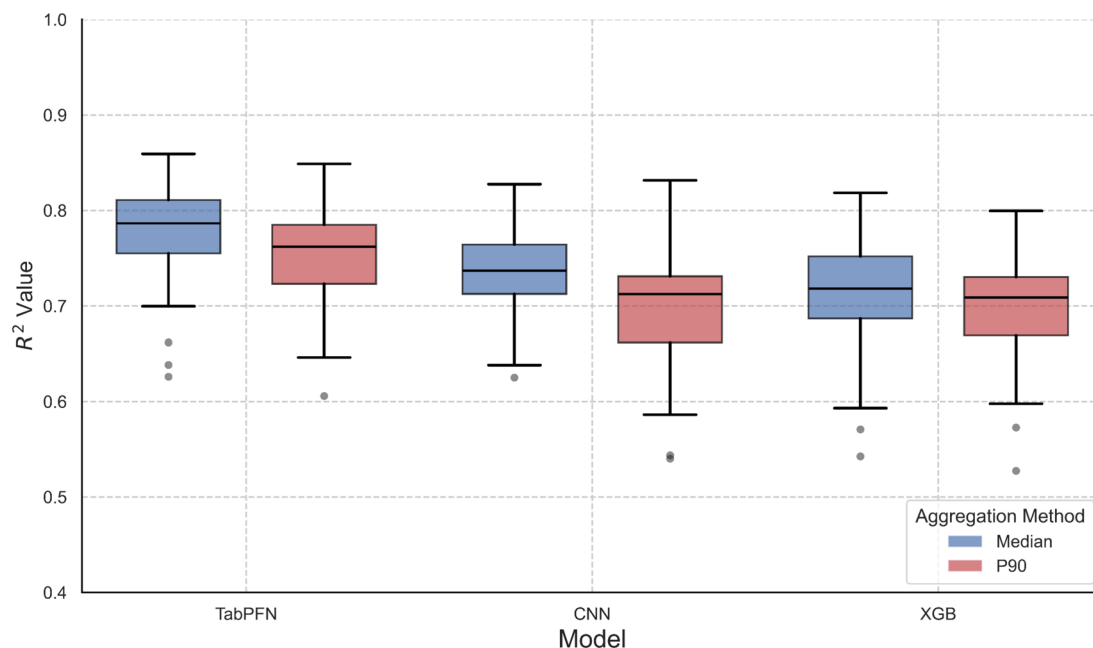


Fig. 1. Distribution of R^2 and RMSE across 100 bootstrap iterations for TabPFN, CNN, and XGBoost models using P50 and P90 composite.

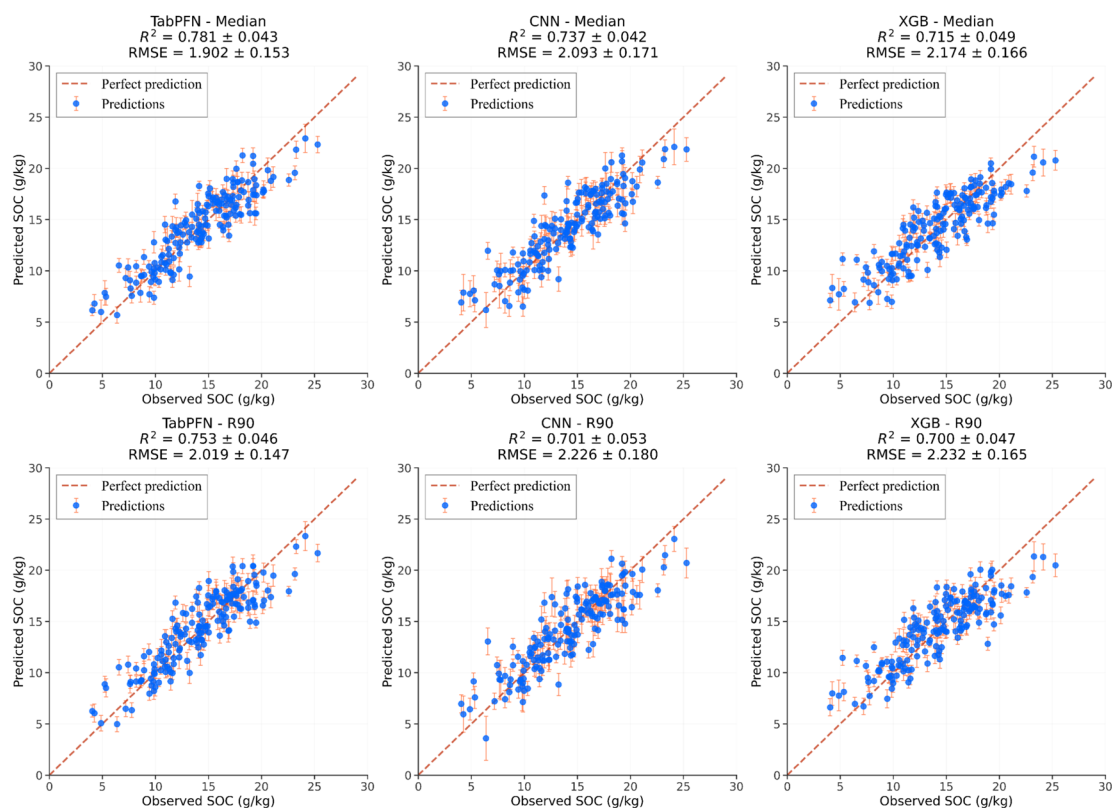


Fig. 2. Scatter plots of measured vs. predicted SOC values for each model, with error bars representing variability across 100 bootstrap iterations.

predictions, especially for lower and higher SOC concentrations, suggesting it may be an effective tool for SOC modeling in the black soil region.

Spatial prediction of SOC

With TabPFN, our best predictive model, we mapped SOC continuously across the study region at a spatial resolution of 10 m. The resulting distribution (Fig. 3) shows significant spatial variations in the surface SOC in the Black soil region, including a mean SOC content of 16.28 g kg^{-1} , which is consistent with earlier results from similar agro-ecological contexts⁹. The comparison between the local zoom-in map and the high-resolution Sentinel-2 synthetic image (using P50 composite method) indicates that the predicted SOC effectively captures the changes at the field scale and reflects the subtle differences in soil conditions of individual plots. In undulating terrain, organic carbon tends to accumulate in sedimentary areas, and this distribution pattern is more obvious in space. Overall, high-resolution organic carbon maps provide detailed baselines for assessing the spatial distribution of soil fertility and carbon storage in heterogeneous cropland landscapes. A resolution of 10 m enables differentiation at the field division level, which is crucial for precise soil management, targeted intervention, and continuous carbon monitoring efforts.

Model interpretation

To better understand how different predictors affect the SOC estimates, we performed SHAP interpretation on best-performing TabPFN model. The 10 variables with the largest mean absolute SHAP values were B12, mean annual precipitation (PRE), mean annual temperature (TMP), clay index (CI), B11, Potential Evapotranspiration (PET), Digital Elevation Model (DEM), B2, B7, and B8 (Fig. 4a). B12 had the most influence on the model output, which indicates that it is sensitive to soil organic matter. Climatic variables (PRE, TMP, PET) also had a prominent role, as well as DEM. These results suggest that spectral information and environmental conditions contributed to the variation of SOC in the region.

The SHAP beeswarm plot (Fig. 4b) also helps clarify how each predictor has affected the SOC estimation. B12 gave the largest range of SHAP (roughly -6 to 4), which implies that it could be quite different across samples. In general, darker surfaces (low B12 reflectance) correlate with higher SOC predictions, and lighter ones (high reflectance) with lower. PRE was on the opposite side: more rain contributed to SOC, most likely through enhanced organic matter input and reduced decomposition. TMP had negative effects at high values, which is normal for SOC loss at warmer temperatures. Other predictors like CI, B11 and PET also contributed moderately, reflecting their relationship to soil texture, moisture, and microtopographic variation.

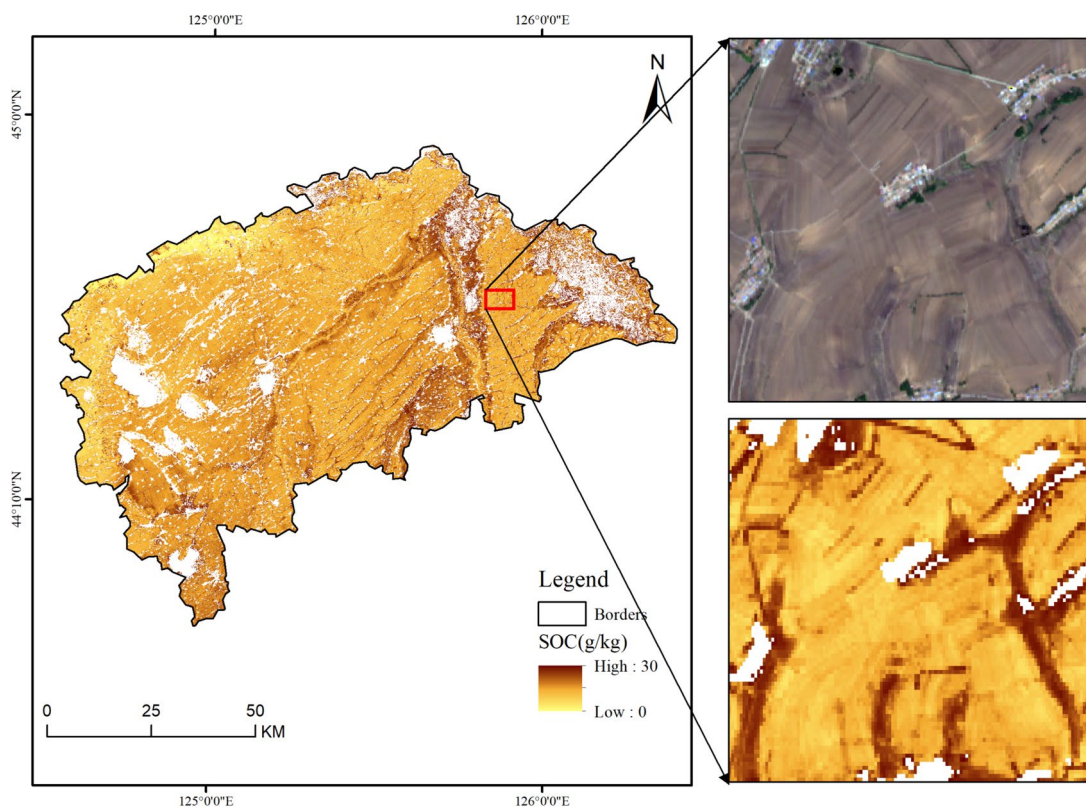


Fig. 3. Spatial distribution of predicted SOC content across cropland areas in the study region at 10 m resolution, with local zoom-in views highlighting field-scale variability. The map was created using QGIS Desktop 3.40.4 (<https://qgis.org>). The background imagery uses Sentinel-2 data courtesy of the European Space Agency (ESA).

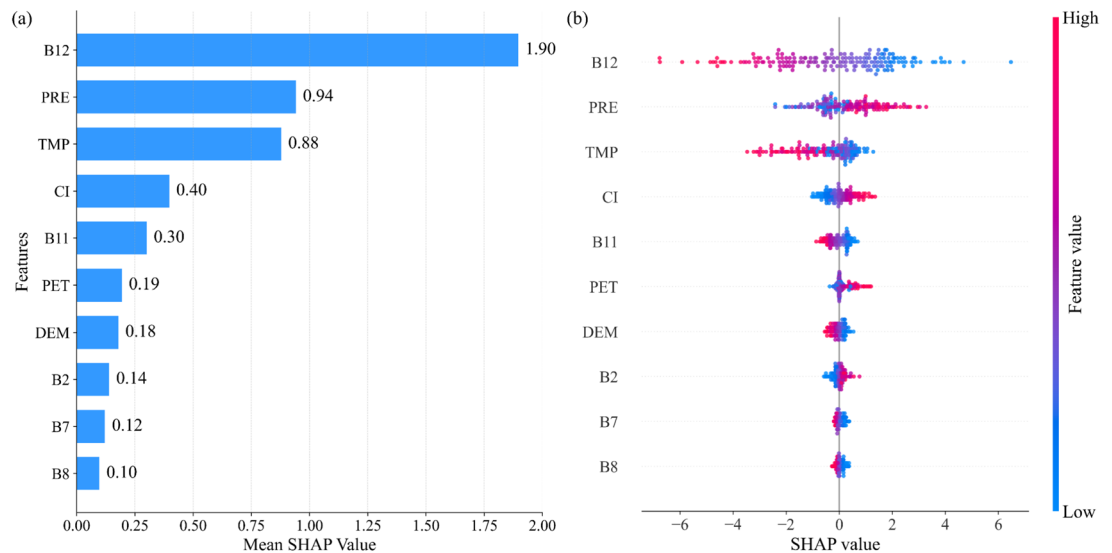


Fig. 4. SHAP analysis for the TabPFN model: (a) variable importance ranking; (b) beeswarm plot of feature effects.

Discussion

Effect of image compositing strategies on SOC prediction

Our findings demonstrate that the P50 composite consistently outperformed the P90 composite across all models for SOC prediction in the black soil region. In all three models the inputs from P50 composite generally achieve slightly higher R^2 values and lower RMSE, meaning that this approach provides representative surface reflectance. The result can be attributed to the distinct spectral characteristics and noise profiles captured by each compositing strategy. The P50 composite, representing the central tendency of reflectance over time, effectively filters out transient anomalies caused by residual moisture, subtle vegetation greening, or atmospheric perturbations²³. In the context of Northeast China's croplands, where the pre-planting window for bare soil observation is relatively short and often characterized by variable surface conditions, P50 composite provides more temporally consistent, radiometrically stable images as well as more reliable reflectance soil studies³².

In contrast, the P90 composite strategy, designed to select drier and presumably “cleaner” soil pixels, may inadvertently introduce bias. In our study area, higher reflectance values (targeted by P90 composite) are not solely indicative of dry soil; they can also correspond to surface crusting, residual crop residues with higher cellulose reflectance, or increased sensor noise under certain sun-target-sensor geometries²². This is consistent with the findings of Hong, et al.²³, who suggested that the superiority of the P50 composite over P90 composite can be explained by stronger spectral correlations and the P90 composite method may amplify non-soil signals in complex agroecosystems. Consequently, while P90 composite aims to reduce moisture interference, it may compromise the spectral integrity related to SOC by incorporating pixel values influenced by confounding surface factors, thereby reducing model generalizability and accuracy. Our study underscores that in intensively managed, heterogeneous croplands, robustness against outliers (achieved by P50 composite) is more critical for reliable SOC prediction than targeting an assumed optimal soil state.

Model performance comparison and underlying causes

The comparative analysis revealed a clear performance hierarchy: TabPFN > CNN > XGBoost. This hierarchy elucidates the strengths and limitations of different algorithmic paradigms for the specific task of SOC mapping from tabularized remote sensing data.

The superior performance of TabPFN stems from its unique integration of prior knowledge and attention-based learning. As a transformer-based model pre-trained on a broad spectrum of synthetic tabular data, TabPFN adds prior distributions over functions to approximate Bayesian inference³¹. This allows it to make strong inferences even with limited training samples ($n = 174$), effectively mitigating overfitting—a common challenge in digital soil mapping with sparse point data. Furthermore, its self-attention mechanism excels at identifying and weighting complex, non-linear interactions between predictors, such as the synergistic effect between precipitation and temperature revealed by our SHAP analysis. This capability to model high-order interactions without explicit feature engineering provides a significant advantage in capturing the multifaceted controls on SOC distribution.

The CNN model, while powerful for spatial feature extraction from image patches, showed suboptimal performance here. This is primarily because our input data structure—a feature vector per pixel—does not fully leverage CNN's strength in learning local spatial contexts and textures from neighboring pixels³³. When spectral and environmental data are processed as independent pixel-wise observations, the convolutional layers' capacity to abstract spatial patterns is underutilized, rendering it similar to a standard feedforward network but with higher parametric complexity prone to overfitting on small datasets. XGBoost, a robust and widely used

ensemble method, served as a strong baseline. Its slightly lower accuracy compared to TabPFN may be attributed to its inherent structure: while excellent at capturing non-linearities through boosted trees, it primarily models additive effects and may not as efficiently capture the complex, multiplicative interactions between climatic and spectral variables that are characteristic of soil forming processes³⁴.

Together, these results suggest that transformer-based models could be a valuable alternative to traditional machine learning and deep learning in tabular remote sensing. Furthermore, their scalability and interpretability make them particularly suitable for soil property mapping and environmental prediction.

Mechanistic insights from SHAP interpretation

The SHAP dependence analysis transcended mere feature ranking, offering a mechanistic window into the drivers of SOC variability that aligns with and extends established pedological understanding (Fig. 5). The Shortwave-Infrared (SWIR) band (B12) was the most influential predictor, which is consistent with previous studies and highlights the direct spectral response of soil to organic matter^{9,35}. Organic compounds, particularly humic substances, exhibit strong absorption features in the SWIR region due to overtones and combinations of C–H, N–H, and O–H bonds³⁶. Our observed negative relationship (higher B12 reflectance → lower SOC) is a classic spectral signature of soils, where darker, carbon-rich soils absorb more radiation across the spectrum^{36,37}. This confirms that despite compositing, Sentinel-2's SWIR bands retain critical information on SOC content, validating their central role in bare soil spectroscopy.

PRE ranked second in importance, showing a strong positive association with SOC, consistent with its recognized role in sustaining biomass production and contributing organic matter to the soil³⁸. Interestingly, the SHAP dependence plots revealed a positive interaction between PRE and TMP, suggesting that warmer conditions do not necessarily accelerate SOC loss when moisture is abundant. This interaction reflects the buffering effect of precipitation on temperature-driven decomposition, a pattern also reported in temperate agroecosystems where carbon accumulation is governed by the balance between organic matter input and decay processes³⁹.

The negative SHAP response of TMP aligns with the expectation that higher temperatures accelerate microbial activity and organic matter decomposition, thereby reducing SOC storage⁴⁰. At the same time, its interaction with PRE highlights the context-dependent nature of climatic effects: under wetter conditions, enhanced plant

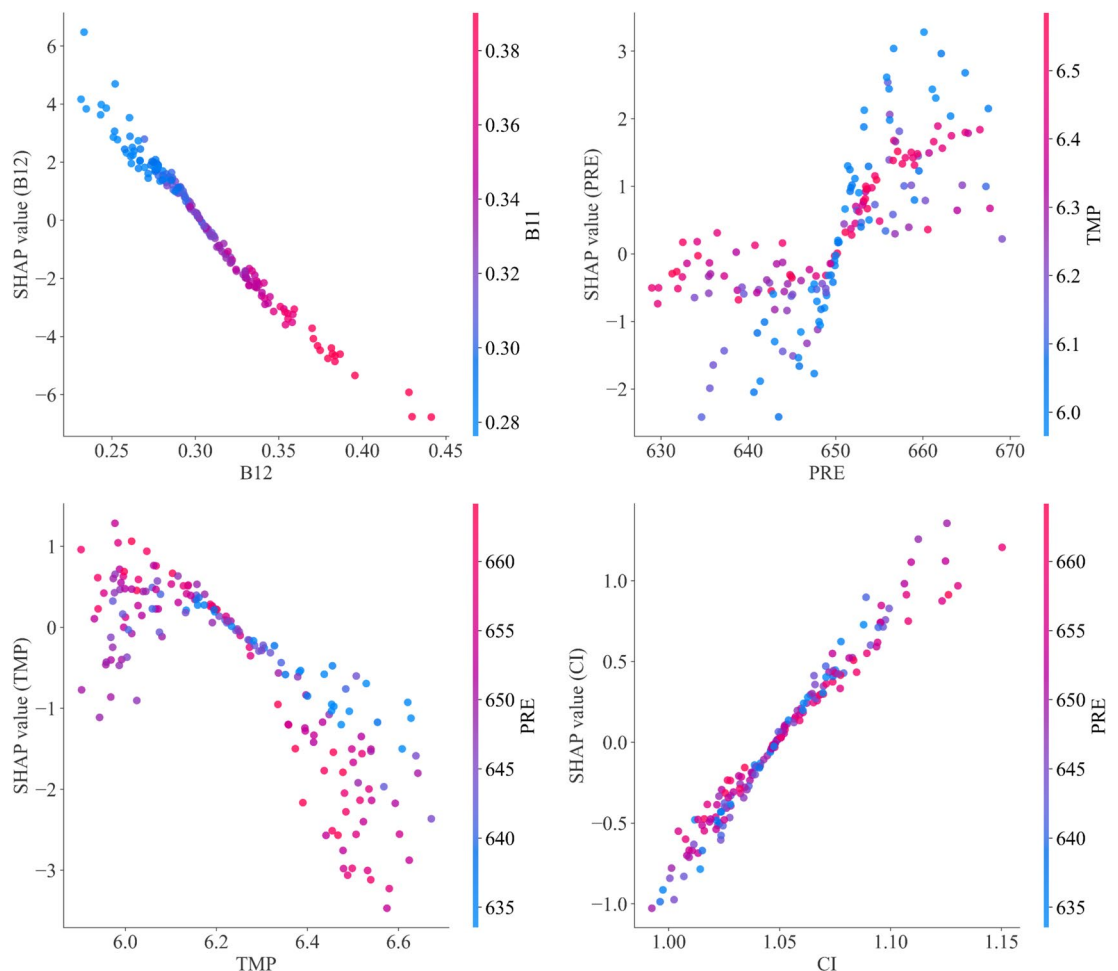


Fig. 5. SHAP dependence plots for the top 4 predictors of SOC.

productivity can partially compensate for the increased decomposition caused by higher temperatures⁴¹. The CI exhibited a consistently positive influence on SOC predictions, reflecting the established role of clay minerals in protecting organic matter through mineral associations and aggregate stabilization⁴². Clay particles provide vast surface areas for organo-mineral complexation, physically protecting organic molecules from microbial access and enzymatic breakdown. The fact that CI's importance was largely independent of climatic variables in our analysis suggests that this stabilization mechanism operates as a fundamental, geographically persistent control on SOC levels, providing a buffer against climatic fluctuations⁴³.

Collectively, these results indicate that both spectral variables (B12, CI) and environmental factors (PRE, TMP) provide complementary, mechanistically meaningful information for explaining SOC variability. The TabPFN model's ability to capture complex, nonlinear, and interactive relationships underscores its suitability for cropland systems characterized by limited data and heterogeneous environmental conditions. At the same time, the influence of spectral predictors such as B12 on transient surface states—like soil moisture and crop residue cover—points to the potential benefit of incorporating direct indicators of surface conditions in future modeling efforts to enhance generalizability.

Methodological implications, limitations, and future perspectives

This study introduces and validates an integrated framework that advances SOC mapping in cultivated landscapes by making three key methodological contributions. First, it establishes the practical advantage of P50 composite over high-percentile methods for generating spectrally stable bare soil inputs in the heterogeneous black soil region, a finding critical for operational DSM. Second, it demonstrates, for the first time in SOC mapping, the superior efficacy of the transformer-based TabPFN model under data-scarce conditions, outperforming established CNN and XGBoost approaches and offering a promising new paradigm for tabular remote sensing data analysis. Third, by coupling this high-performing model with SHAP analysis, the framework moves beyond prediction to provide mechanistically interpretable outputs. It explicitly quantifies the dominant role of the SWIR band (B12) and reveals the critical interactive effect between precipitation and temperature on SOC—a nuanced insight into climate controls that refines understanding of carbon vulnerability in temperate croplands. This combination of robust compositing, data-efficient learning, and process-aware interpretation offers a practical and insightful template for local-scale cropland SOC assessment.

We acknowledge several limitations that contextualize our findings. The spatial extent and sample size of our dataset, while adequate for this regional proof-of-concept, may constrain the direct transferability of the trained model to other agroecological zones with different soil-forming factors. Although the NDVI-based bare soil extraction effectively minimized green vegetation interference, it did not differentiate bare soil from non-photosynthetic vegetation (crop residues), which can spectrally confound SOC estimation. Furthermore, the static nature of some environmental covariates (e.g., long-term climate averages) fails to capture intra- and inter-annual dynamics that influence SOC. Most importantly, while SHAP enhances post-hoc interpretability, the TabPFN model remains purely data-driven. Its predictions are not constrained by established soil biogeochemical principles, potentially limiting its physical plausibility and reliability when extrapolating to conditions not represented in the training data.

Future research should build on this framework along three interconnected fronts. To improve transferability, efforts must focus on assembling larger, multi-region soil-spectral databases and developing domain adaptation or meta-learning techniques that allow models like TabPFN to generalize across diverse agricultural systems. Regarding input data, next-generation bare soil composites should integrate multi-sensor data (e.g., Sentinel-1 SAR for moisture and residue detection) and employ more sophisticated unmixing or classification algorithms to isolate the purest soil signal. The most transformative direction lies in developing hybrid mechanistic-machine learning models. Embedding physical constraints (e.g., decomposition kinetics) or process-model outputs into the architecture or loss function of predictive algorithms will be crucial. This shift from purely data-driven to knowledge-informed learning is essential for generating SOC maps that are not only accurate but also truly consistent with pedological theory, thereby enabling trustworthy projections under novel climate and management scenarios⁴⁴.

Conclusions

This study demonstrates that integrating multi-temporal Sentinel-2 bare-soil composites with the TabPFN transformer model provides an effective approach for SOC mapping under limited sampling conditions. Among the tested methods, the TabPFN model using the P50 composite achieved the highest accuracy, indicating that P50 composite better captures representative soil reflectance for SOC prediction. The SHAP analysis further revealed that SWIR (B12) and precipitation were the dominant predictors, highlighting the combined influence of soil spectral properties and climatic controls on SOC variability. Overall, the proposed workflow provides a practical pathway for scalable SOC mapping in agricultural regions where dense soil sampling is not feasible. Future work should evaluate the transferability of this approach across broader environmental gradients and integrate additional sensing modalities to further enhance robustness and generalization.

Materials and methods

Study area

The area is Nong'an County and Dehui City, in Changchun City, Jilin Province (124°31'–126°24'E, 43°55'–44°55'N)⁴⁵. Nong'an County lies on the eastern border of the plain and covers about 5400 km². It is a landscape that ranges from rolling grasslands in the west to wide plains in the east and is connected by a major river such as Songhua and Yitong Rivers. Dehui City is situated in the central–northern part of Jilin Province, adjacent to Changchun, with a total area of about 3,322 km². Both Nong'an and Dehui belong to the Songnen Plain, where

the landscape is slightly steep in the southwest and in the northeast. The soil is very fertile, forming part of the Northeast Black Soil Zone. Black and brown soils cover the area and are rich in organic matter, which is well used for growing⁴⁶. The area is single cropping with maize being the main food crop. The maize season runs from April to October, during which the land is prepared from April–May, producing large soil exposures, which provides good conditions to monitor SOC⁴⁷.

Soil sampling and laboratory analysis

174 surface soil samples (0–20 cm) were taken in 2022 from croplands throughout the area (Fig. 6). Sampling sites are selected for a wide variety of cultivated soil conditions corresponding to the regional heterogeneity. Five subsamples from each location are taken within a 10 × 10 m region and mixed. The geographic point of the center is observed with Global Positioning System (GPS). The locations of the centers of interest are recorded. The soil samples in the laboratory are air-dried, gently ground, and filtered through a 2 mm sieve before the analysis. The SOC content is determined with the dichromate oxidation (Walkley–Black) method⁴⁸ and the measured SOC concentrations (g kg⁻¹) are used for model calibration and independent validation.

Sentinel-2 imagery and preprocessing

Sentinel-2 Level-2A imagery from April 15 to May 31 of each year between 2021 and 2023 was obtained from the Google Earth Engine (GEE) platform. This period coincides with the pre-planting stage, during which most cropland soils in the study area remain bare, providing favorable conditions for the retrieval of soil reflectance data⁹. Images with cloud cover less than 20% were selected, and additional cloud masking was performed using the S2Cloudless algorithm to further eliminate cloud and cloud-shadow contamination.

To minimize the influence of vegetation, a NDVI threshold of 0.1–0.4 was applied to extract pixels corresponding to bare or sparsely vegetated soil^{49,50}. To generate representative multi-year bare-soil reflectance composites, two compositing strategies were applied: P50 composite, which provides robust reflectance estimates with strong stability and low sensitivity to outliers; and P90 composite, which reduces the influence of moisture²². Ten Sentinel-2 spectral bands (B2–B8, B8A, B11, and B12) were used as predictors. The bands with coarser spatial resolution (B5, B6, B7, B8A, B11, and B12) were resampled to 10 m using the nearest-neighbor method to ensure spatial consistency across all bands.

Environmental covariates

In addition to Sentinel-2 spectral reflectance, a suite of environmental variables known to influence SOC distribution were incorporated as auxiliary predictors (Table 1). These variables encompass climatic, topographic, and spectral indices derived from remote sensing data. Climatic variables, including TMP, PET, and PRE, were

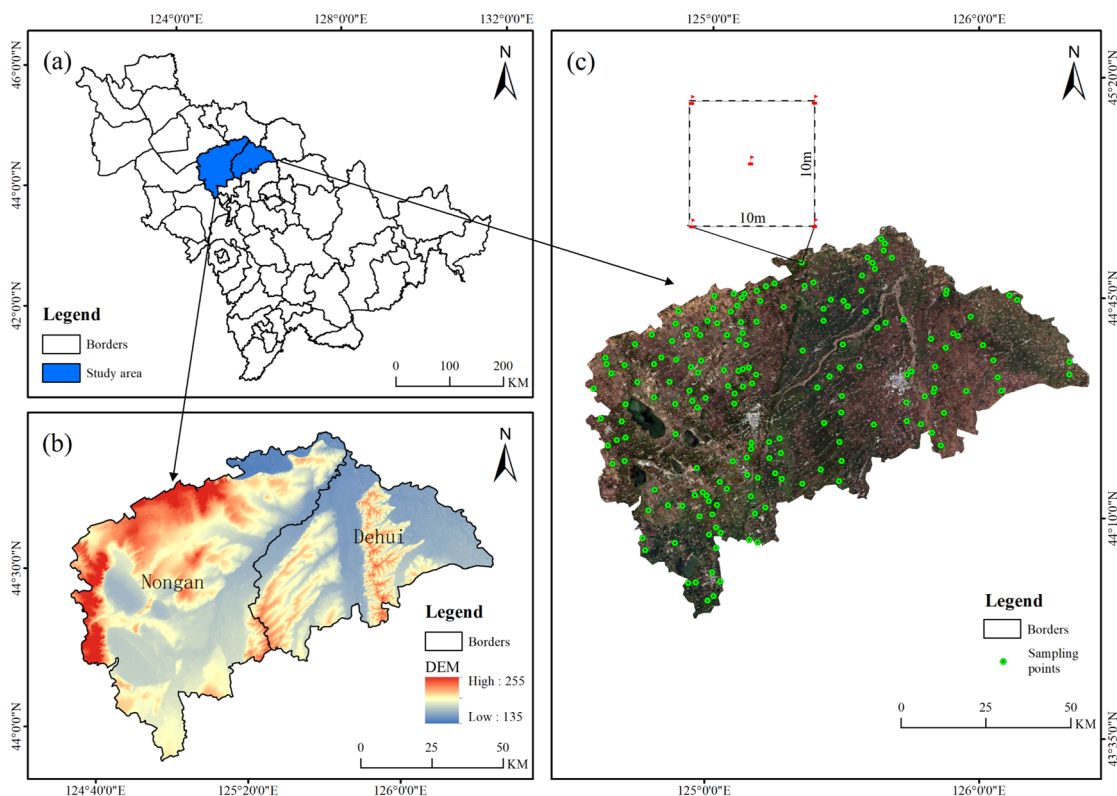


Fig. 6. Location of the study area and distribution of soil sampling points. The map was generated using QGIS Desktop 3.40.4 (<https://qgis.org>). The background satellite imagery uses Sentinel-2 data courtesy of the European Space Agency (ESA).

Category	Abbreviation	Source / Formula	Spatial resolution	Description
Climatic	TMP	National Earth System Science Data Center	1 km (resampled to 10 m)	Represents long-term thermal conditions
	PET	Same as above	1 km (resampled to 10 m)	Indicates atmospheric water demand
	PRE	Same as above	1 km (resampled to 10 m)	Reflects long-term moisture availability
Topographic	DEM	NASADEM	30 m (resampled to 10 m)	Absolute terrain height
	TPI	Derived from DEM	30 m (resampled to 10 m)	Relative elevation position
Spectral	SOCI	$B2 / (B3 \times B4)$	10 m	Sensitive to SOC-related spectral absorption ⁵⁴
	BI	$\sqrt{((B3^2 + B4^2)/2)}$	10 m	Indicates overall soil brightness ⁵⁵
	RECI	$(B5 + B6)/2$	10 m	Captures red-edge reflectance response of organic matter ⁵⁶
	CI	$B11 / B12$	10 m	Related to clay mineral absorption features ²⁹

Table 1. List of environmental covariates used in SOC modeling.

obtained from the National Earth System Science Data Center (National Science & Technology Infrastructure of China; <http://www.geodata.cn>). Topographic variables such as elevation and the topographic position index (TPI) were derived from the NASADEM dataset. Additionally, several spectral indices sensitive to soil properties were computed, including the soil organic carbon index (SOCI), brightness index (BI), Red-Edge carbon index (RECI), and CI. All covariates were resampled to a 10 m spatial resolution and spatially aligned with Sentinel-2 composites to ensure consistency among predictors.

SOC modeling

SOC prediction was conducted using three modeling approaches: TabPFN, CNN, and XGBoost (Fig. 7). TabPFN is a transformer-based probabilistic model specifically designed for tabular data. It integrates prior knowledge into a Bayesian framework and employs self-attention mechanisms to capture nonlinear and high-order dependencies among predictors without explicit hyperparameter tuning³¹. CNN is a deep learning model capable of learning spatial-spectral representations from multi-band raster data, effectively capturing the local spatial context and spectral variability related to SOC⁵¹. XGBoost is a widely used gradient boosting algorithm that builds an ensemble of decision trees to iteratively minimize prediction errors and has demonstrated high predictive performance in soil property estimation⁵².

To ensure the robustness of the evaluation, a bootstrap sampling strategy was adopted. For each model, perform 100 bootstrap iterations. In each iteration, a training set is generated through random sampling and replacement, while the remaining out-of-bag samples are used for validation. The final model performance metrics, including the R^2 and the RMSE, are calculated as the average of all 100 iterations. The model input includes bare soil reflectance composites from Sentinel-2 images. Each composite is combined with a set of environmental covariates to form two different pre-diction datasets for model comparison. The model performance metrics are defined as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (x_1 - x_2)^2}{n}} \quad (1)$$

$$R^2 = \frac{\sum_{i=1}^n (x_1 - \bar{x})^2}{\sum_{i=1}^n (x_2 - \bar{x})^2} \# \quad (2)$$

x_1 represents the actual measurement value of SOC while x_2 represents the predicted value of SOC; \bar{x} represents the average of SOC measured values; n represents the number of sampling points.

Model Interpretation and mapping

SHAP was used to interpret the model and quantify the relative contributions of each predictor to the SOC prediction of the best-performing model⁵³. For spatial mapping, 100 trained models from bootstrap iterations were respectively applied to the prediction dataset to generate separate SOC maps for the entire research area. Then, pixel averaging is performed on these maps to generate the final SOC prediction map with a spatial resolution of 10 m, representing the average predicted SOC value and effectively reducing the uncertainty caused by model variability. The final map was then visualized to analyze the spatial levels and distributions of SOC in the black soil region.

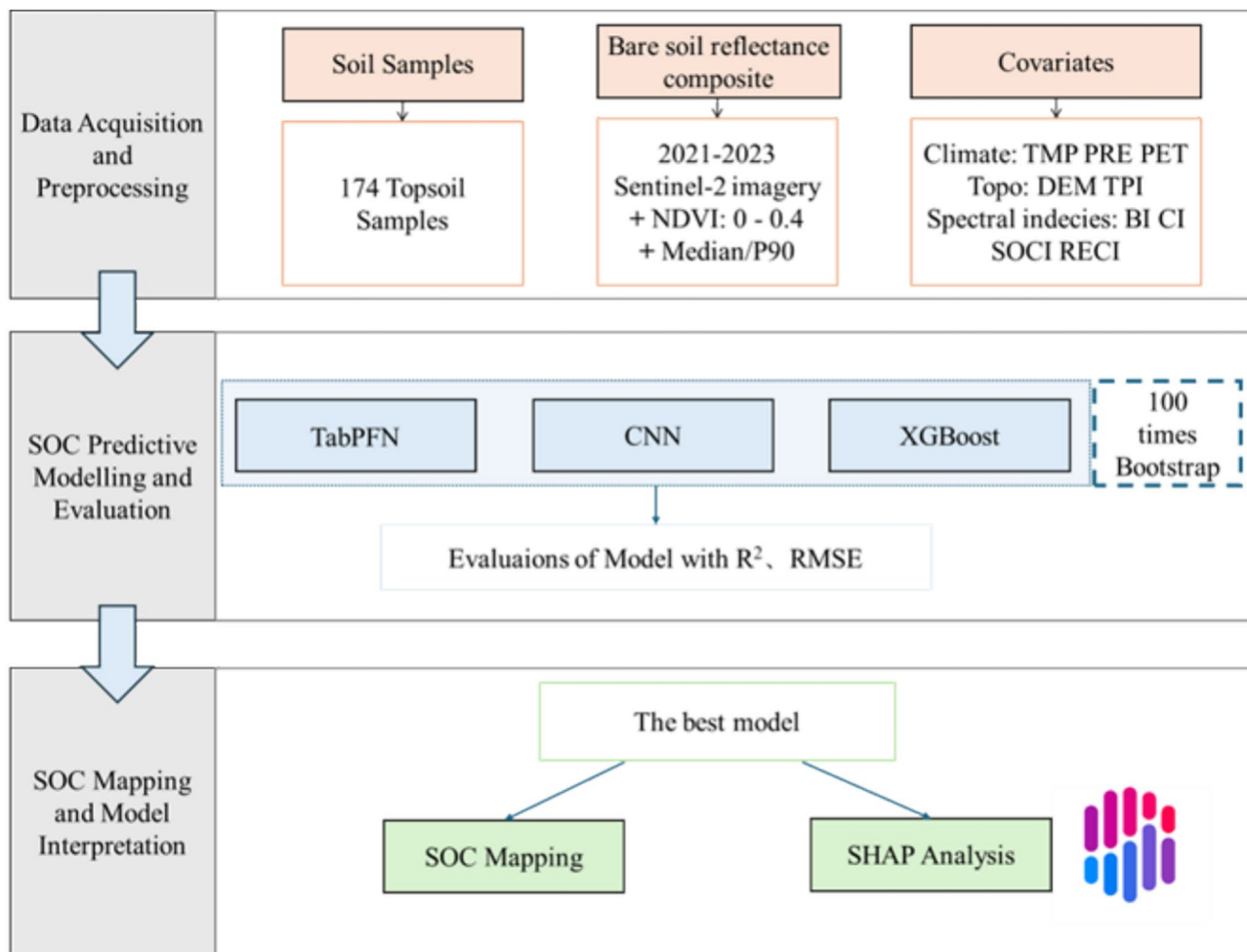


Fig. 7. Workflow of this study.

Data availability

The datasets analyzed during the current study are not publicly available due to existing agreements and data-use restrictions but are available from the corresponding author on reasonable request.

Received: 14 November 2025; Accepted: 22 December 2025

Published online: 05 January 2026

References

- Lal, R. Soil carbon sequestration impacts on global climate change and food security. *Science* **304**, 1623–1627. <https://doi.org/10.1126/science.1097396> (2004).
- Leifeld, J. & Menichetti, L. The underappreciated potential of peatlands in global climate change mitigation strategies. *Nat. Commun.* **9**, 1071. <https://doi.org/10.1038/s41467-018-03406-6> (2018).
- Six, J., Elliott, E. T. & Paustian, K. Soil Structure and Soil Organic Matter II. A Normalized Stability Index and the effect of mineralogy. *Soil Sci. Soc. Am. J.* **64**, 1042–1049. <https://doi.org/10.2136/sssaj2000.6431042x> (2000).
- Minasny, B., McBratney, A. B., Malone, B. P. & Wheeler, I. In *Advances in Agronomy* Vol. 118 (ed Donald L. Sparks) 1–47 (Academic Press, 2013).
- Qi, L., Ma, J., Sun, Q. & Shi, P. Mapping soil organic carbon sequestration potential in croplands using a combined proximal and remote sensing approach. *Soil Tillage Res.* **254**, 106733. <https://doi.org/10.1016/j.still.2025.106733> (2025).
- Minasny, B. et al. Soil carbon 4 per mille. *Geoderma* **292**, 59–86. <https://doi.org/10.1016/j.geoderma.2017.01.002> (2017).
- Chen, S. et al. Digital mapping of GlobalSoilMap soil properties at a broad scale: A review. *Geoderma* **409**, 115567. <https://doi.org/10.1016/j.geoderma.2021.115567> (2022).
- Zhou, F. et al. Integrating historical crop rotation changes into soil organic matter mapping in the Cropland of Southeastern China. *Earth's Future* **13**, e2025EF006117. <https://doi.org/10.1029/2025EF006117> (2025).
- Shi, P., Six, J., Sila, A., Vanlauwe, B. & Van Oost, K. Towards spatially continuous mapping of soil organic carbon in croplands using multitemporal Sentinel-2 remote sensing. *ISPRS J. Photogramm. Remote. Sens.* **193**, 187–199. <https://doi.org/10.1016/j.isprsjprs.2022.09.013> (2022).
- McBratney, A. B., Mendonça Santos, M. L. & Minasny, B. On digital soil mapping. *Geoderma* **117**, 3–52. [https://doi.org/10.1016/S0016-7061\(03\)00223-4](https://doi.org/10.1016/S0016-7061(03)00223-4) (2003).
- Poppiel, R. R., Paiva, A. F. S. & Demattê, J. A. M. Bridging the gap between soil spectroscopy and traditional laboratory: Insights for routine implementation. *Geoderma* **425**, 116029. <https://doi.org/10.1016/j.geoderma.2022.116029> (2022).

12. Castaldi, F., Chabrilat, S. & van Wesemael, B. Sampling strategies for soil property mapping using multispectral Sentinel-2 and Hyperspectral EnMAP Satellite Data. *Remote Sensing* **11**, 309 (2019).
13. Gholizadeh, A., Žižala, D., Saberioon, M. & Borůvka, L. Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging. *Remote Sens. Environ.* **218**, 89–103. <https://doi.org/10.1016/j.rse.2018.09.015> (2018).
14. Wadoux, A. M. J. C., Minasny, B. & McBratney, A. B. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth Sci. Rev.* **210**, 103359. <https://doi.org/10.1016/j.earscirev.2020.103359> (2020).
15. Dvorakova, K., Shi, P., Limbourg, Q. & van Wesemael, B. Soil Organic carbon mapping from remote sensing: The effect of crop residues. *Remote Sens.* <https://doi.org/10.3390/rs12121913> (2020).
16. Rogge, D. et al. Building an exposed soil composite processor (SCMaP) for mapping spatial and temporal characteristics of soils with Landsat imagery (1984–2014). *Remote Sens. Environ.* **205**, 1–17. <https://doi.org/10.1016/j.rse.2017.11.004> (2018).
17. Melo Dematte, J. A., Fongaro, C. T., Rizzo, R. & Safanelli, J. L. Geospatial Soil Sensing System (GEOS3): A powerful data mining procedure to retrieve soil spectral reflectance from satellite images. *Remote Sens. Environ.* **212**, 161–217. <https://doi.org/10.1016/j.rse.2018.04.047> (2018).
18. Castaldi, F. et al. Evaluating the capability of the Sentinel 2 data for soil organic carbon prediction in croplands. *ISPRS J. Photogramm. Remote Sens.* **147**, 267–282. <https://doi.org/10.1016/j.isprsjprs.2018.11.026> (2019).
19. Vaudour, E. et al. Temporal mosaicking approaches of Sentinel-2 images for extending topsoil organic carbon content mapping in croplands. *Int. J. Appl. Earth Obs. Geoinf.* **96**, 102277. <https://doi.org/10.1016/j.jag.2020.102277> (2021).
20. Xue, J. et al. National-scale mapping topsoil organic carbon of cropland in China using multitemporal Sentinel-2 images. *Geoderma* **456**, 117272. <https://doi.org/10.1016/j.geoderma.2025.117272> (2025).
21. Zhu, Y., Qi, L., Wu, Z. & Shi, P. Spectra-based predictive mapping of soil organic carbon in croplands: Single-date versus multitemporal bare soil compositing approaches. *Geoderma* **449**, 116987. <https://doi.org/10.1016/j.geoderma.2024.116987> (2024).
22. Castaldi, F. et al. Assessing the capability of Sentinel-2 time-series to estimate soil organic carbon and clay content at local scale in croplands. *ISPRS J. Photogramm. Remote Sens.* **199**, 40–60. <https://doi.org/10.1016/j.isprsjprs.2023.03.016> (2023).
23. Hong, Y. et al. Bridging the gap between laboratory VNIR-SWIR spectra and Landsat-8 bare soil composite image for soil organic carbon prediction. *Remote Sens. Environ.* **328**, 114874. <https://doi.org/10.1016/j.rse.2025.114874> (2025).
24. Zhang, M.-W. et al. Predicting spatial-temporal soil organic matter dynamics in a Mollisols region of the northern Songnen Plain, China, during 2009–2018 using a spectral-temporal feature set. *Geoderma* **461**, 117461. <https://doi.org/10.1016/j.geoderma.2025.117461> (2025).
25. Minasny, B. & McBratney, A. B. Digital soil mapping: A brief history and some lessons. *Geoderma* **264**, 301–311. <https://doi.org/10.1016/j.geoderma.2015.07.017> (2016).
26. Vaudour, E. et al. Satellite imagery to map topsoil organic carbon content over cultivated areas: An overview. *Remote Sens.* <https://doi.org/10.3390/rs14122917> (2022).
27. Padarian, J., Minasny, B. & McBratney, A. B. Using deep learning for digital soil mapping. *SOIL* **5**, 79–89. <https://doi.org/10.5194/soil-5-79-2019> (2019).
28. Meng, X., Bao, Y., Wang, Y., Zhang, X. & Liu, H. An advanced soil organic carbon content prediction model via fused temporal-spatial-spectral (TSS) information based on machine learning and deep learning algorithms. *Remote Sens. Environ.* <https://doi.org/10.1016/j.rse.2022.113166> (2022).
29. Meng, X., Bao, Y., Luo, C., Zhang, X. & Liu, H. SOC content of global Mollisols at a 30 m spatial resolution from 1984 to 2021 generated by the novel ML-CNN prediction model. *Remote Sens. Environ.* **300**, 113911. <https://doi.org/10.1016/j.rse.2023.113911> (2024).
30. Žižala, D. et al. Soil sampling design matters - Enhancing the efficiency of digital soil mapping at the field scale. *Geoderma Reg.* **39**, e00874. <https://doi.org/10.1016/j.geodrs.2024.e00874> (2024).
31. Hollmann, N. et al. Accurate predictions on small data with a tabular foundation model. *Nature* **637**, 319–326. <https://doi.org/10.1038/s41586-024-08328-6> (2025).
32. Heiden, U. et al. Soil reflectance composites-improved thresholding and performance evaluation. *Remote Sens.* <https://doi.org/10.3390/rs14184526> (2022).
33. Kalopesa, E. et al. Large-scale soil organic carbon estimation via a multisource data fusion approach. *Remote Sens.* **17**, 771 (2025).
34. Zhang, Y. et al. Estimation of coastal wetland soil organic carbon content in Western Bohai Bay Using Remote Sensing, Climate, and Topographic Data. *Remote Sens.* <https://doi.org/10.3390/rs15174241> (2023).
35. Song, J. et al. Mapping soil organic matter in cultivated land based on multi-year composite images on monthly time scales. *J. Integr. Agric.* **23**, 1393–1408. <https://doi.org/10.1016/j.jia.2023.09.017> (2024).
36. Ben-Dor, E., Inbar, Y. & Chen, Y. The reflectance spectra of organic matter in the visible near-infrared and short wave infrared region (400–2500 nm) during a controlled decomposition process. *Remote Sens. Environ.* **61**, 1–15. [https://doi.org/10.1016/S0034-4257\(96\)00120-4](https://doi.org/10.1016/S0034-4257(96)00120-4) (1997).
37. Angelopoulou, T., Tziolas, N., Balafoutis, A., Zalidis, G. & Bochtis, D. Remote sensing techniques for soil organic carbon estimation: A review. *Remote Sens.* **11**, 676 (2019).
38. Chen, X. et al. Effects of precipitation on soil organic carbon fractions in three subtropical forests in southern China. *J. of Plant Ecol.* **9**, 10–19. <https://doi.org/10.1093/jpe/rtv027> (2015).
39. Pallandt, M. et al. Modelling the effect of climate–substrate interactions on soil organic matter decomposition with the Jena Soil Model. *Biogeosciences* **22**, 1907–1928. <https://doi.org/10.5194/bg-22-1907-2025> (2025).
40. Wu, Y. et al. Mechanisms behind the soil organic carbon response to temperature elevations. *Agriculture* **15**, 1118 (2025).
41. Zhao, M. et al. Soil mineral-associated organic carbon and its relationship to clay minerals across grassland transects in China. *Appl. Sci.* **14**, 2061 (2024).
42. Xu, Z. & Tsang, D. C. W. Mineral-mediated stability of organic carbon in soil and relevant interaction mechanisms. *Eco-Environ. Health* **3**, 59–76. <https://doi.org/10.1016/j.eehl.2023.12.003> (2024).
43. Xue, B. et al. Effect of clay mineralogy and soil organic carbon in aggregates under straw incorporation. *Agronomy* **12**, 534 (2022).
44. Minasny, B. et al. Soil science-informed machine learning. *Geoderma* **452**, 117094. <https://doi.org/10.1016/j.geoderma.2024.117094> (2024).
45. Ge, H. et al. Enhancing yield prediction in maize breeding using UAV-derived RGB imagery: a novel classification-integrated regression approach. *Front. Plant Sci.* <https://doi.org/10.3389/fpls.2025.1511871> (2025).
46. Dong, Y. et al. A 30-m annual corn residue coverage dataset from 2013 to 2021 in Northeast China. *Sci. Data* **11**, 216. <https://doi.org/10.1038/s41597-024-02998-7> (2024).
47. Ma, J. & Shi, P. Remotely sensed inter-field variation in soil organic carbon content as influenced by the cumulative effect of conservation tillage in northeast China. *Soil Tillage Res.* **243**, 106170. <https://doi.org/10.1016/j.still.2024.106170> (2024).
48. Nelson, D. W. & Sommers, L. E. in *Methods of Soil Analysis* 539–579 (1982).
49. Broeg, T., Don, A., Wiesmeier, M., Scholten, T. & Erasmí, S. Spatiotemporal Monitoring of Cropland Soil Organic Carbon Changes From Space. *Glob. Change Biol.* **30**, e17608. <https://doi.org/10.1111/gcb.17608> (2024).
50. Broeg, T. et al. Using local ensemble models and Landsat bare soil composites for large-scale soil organic carbon maps in cropland. *Geoderma* **444**, 116850. <https://doi.org/10.1016/j.geoderma.2024.116850> (2024).
51. Zayani, H. et al. Using machine-learning algorithms to predict soil organic carbon content from combined remote sensing imagery and laboratory vis-NIR spectral datasets. *Remote Sens.* <https://doi.org/10.3390/rs15174264> (2023).

52. Wang, Q. et al. Incorporating agricultural practices in digital mapping improves prediction of cropland soil organic carbon content: The case of the Tuojiang River Basin. *J. Environ. Manage.* **330**, 117203. <https://doi.org/10.1016/j.jenvman.2022.117203> (2023).
53. Lundberg, S. M. & Lee, S.-I. in *Proceedings of the 31st International Conference on Neural Information Processing Systems* 4768–4777 (Curran Associates Inc., Long Beach, California, USA, 2017).
54. Thaler, E. A., Larsen, I. J. & Yu, Q. A New Index for remote sensing of soil organic carbon based solely on visible wavelengths. *Soil Sci. Soc. Am. J.* **83**, 1443–1450. <https://doi.org/10.2136/sssaj2018.09.0318> (2019).
55. Marques, M. J., Alvarez, A., Carral, P., Sastre, B. & Bienes, R. The use of remote sensing to detect the consequences of erosion in gypsiferous soils. *Int. Soil Water Conserv. Res.* **8**, 383–392. <https://doi.org/10.1016/j.iswcr.2020.10.001> (2020).
56. Castaldi, F., Chabrilat, S., Don, A. & van Wesemael, B. Soil organic carbon mapping using LUCAS topsoil database and sentinel-2 data: An approach to reduce soil moisture and crop residue effects. *Remote Sens.* **11**, 2121 (2019).

Author contributions

Conceptualization, Song Wu, Na Chen, and Zhikang Wei; methodology, Na Chen and Xuancheng Jin; software, Nan Lin; validation, Nan Lin; formal analysis, Zhikang Wei; resources, Ling Zhao and Na Chen; data curation, Xuancheng Jin; writing—original draft preparation, Song Wu, Zhikang Wei and Na Chen; writing—review and editing, Xuancheng Jin, Na Chen and Song Wu; visualization, Song Wu and Xuancheng Jin; supervision, Nan Lin; project administration, Song Wu; funding acquisition, Fan Yang. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the “Study on the Retrogressive Erosion Mechanism of Gully Heads with Different Parent Materials in Black Soil Regions of Low Mountains and Hills” project of Natural Science Foundation of Jilin Province, China (20250102200JC).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-33682-4>.

Correspondence and requests for materials should be addressed to S.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2026