agronomy

MDPI

*Article*

# Small-Sample Prediction and Uncertainty Assessment of Soil Organic Carbon Content in Cropland of the Liaohe Plain Based on the TabPFN Model

**Yong Yang [1], Yanzhi Zhao [2,3], Shuang Gang [2,3,4,*], Xianmin Chang [5] and Nicola Cannon [5]**

1 School of Environment, Shenyang University, Shenyang 110044, China; yangyong24083010@syu.edu.cn
2 International Research Center for Black Soil Carbon and Sustainable Land Management, Shenyang University, Shenyang 110044, China; zhaoyanzhi0509@163.com
3 Liaoning China-UK International Joint Laboratory for Soil Health and Restoration Technology Innovation and Application, Shenyang 110044, China
4 Shenyang Key Laboratory of Low-Carbon Simulation and Big Data, Shenyang 110044, China
5 School of Agricultural Science and Practice, Royal Agricultural University, Cirencester GL7 6JS, Gloucestershire, UK; xianmin.chang@rau.ac.uk (X.C.); nicola.cannon@rau.ac.uk (N.C.)
* Correspondence: gang_shuang@163.com

## Abstract

Soil organic carbon (SOC) is a key indicator of cropland quality, soil fertility, and the carbon sequestration potential of agroecosystems. Accurate characterization of its spatial distribution is essential for black soil conservation and regional soil carbon management. However, regional-scale SOC prediction is often constrained by limited field observations, which can reduce model generalizability and predictive reliability. In this study, we developed a limited-sample SOC prediction framework for the Liaohe Plain using 310 surface (0–20 cm) soil samples collected in 2025 and multi-source environmental covariates, including climate, vegetation, soil spectral, and terrain variables. The framework used the Tabular Prior-Data Fitted Network (TabPFN), whose performance was compared with that of Random Forest, Support Vector Machine, CatBoost, K-Nearest Neighbors, and XGBoost. Model performance was evaluated using 100 repetitions of random 80:20 holdout validation and repeated five-fold spatial cross-validation based on spatially constrained clustering, while sampling-induced relative uncertainty was quantified using 100 repeated random sampling and model-fitting runs. Under random holdout validation, TabPFN showed competitive predictive performance, with mean $R^2$ and RMSE values of 0.608 ± 0.038 and 4.026 ± 0.184 g kg$^{-1}$, respectively. Repeated spatial cross-validation yielded more conservative performance estimates, with mean $R^2$ and RMSE values of 0.535 ± 0.072 and 4.33 ± 0.31 g kg$^{-1}$, respectively, indicating that random splitting may overestimate model performance when sampling sites are spatially clustered. Spatial prediction showed that cropland SOC ranged from 4.63 to 27.04 g kg$^{-1}$, with generally lower values in the west and higher values in the northeast. Areas with high sampling-induced relative uncertainty were mainly concentrated in the northern, northeastern, and marginal regions. These findings provide a methodological basis for SOC mapping, supplementary sampling optimization, and regional soil carbon management under limited-sample conditions, although the temporal robustness of the results requires confirmation using independent data from additional years.

**Keywords:** soil organic carbon; TabPFN; small-sample prediction; uncertainty assessment; digital soil mapping



Academic Editor: Carlo Leifert

Received: 21 July 2026
Revised: 12 August 2026
Accepted: 13 August 2026
Published: 14 August 2026

## 1. Introduction

Soil organic carbon (SOC) is a key indicator of soil fertility, cropland quality, and the carbon sequestration capacity of agroecosystems. It plays an essential role in maintaining soil structure, enhancing nutrient supply, regulating regional carbon cycling, and supporting sustainable agricultural development [1–3]. Accurate characterization of the spatial distribution of cropland SOC is important for identifying areas prone to soil quality degradation and regions with high carbon sequestration potential. It also provides a scientific basis for black soil conservation, cropland quality improvement, and regional soil carbon management [4,5]. The Liaohe Plain is an important component of the black soil region in Northeast China, where cropland is highly concentrated, and agricultural activities are intensively managed. Under the combined influence of climatic conditions, topographic background, soil types, and long-term cultivation practices, cropland SOC in this region exhibits pronounced spatial heterogeneity [6]. Therefore, spatial prediction of cropland SOC in the Liaohe Plain is of great significance for revealing regional SOC differentiation patterns and supporting black soil conservation and farmland carbon management.

Traditional SOC mapping has mainly relied on field sampling, laboratory measurements, and spatial interpolation methods, such as ordinary kriging, inverse distance weighting, and spline interpolation [7–9]. These methods are effective for describing the spatial continuity of SOC. However, their predictive performance is highly dependent on sample size, the spatial distribution of sampling sites, and the strength of spatial autocorrelation. When measured samples are limited or unevenly distributed, conventional interpolation methods often fail to fully capture the complex nonlinear relationships between SOC and environmental factors, including climate, vegetation, soil spectral characteristics, and topography [10–12]. In recent years, digital soil mapping approaches integrating remote sensing, geographic information systems, and machine learning have become important tools for SOC spatial prediction [13–15]. Models such as random forest, support vector machine, CatBoost, K-nearest neighbors, and XGBoost have been widely applied to soil property prediction and SOC spatial mapping [16–19]. Nevertheless, regional-scale cropland SOC prediction still faces two main challenges. First, field-measured soil samples are usually costly and labor-intensive to obtain. Under limited-sample conditions, traditional machine learning models may be affected by insufficient sample representativeness, sensitivity to parameter settings, and perturbations in training samples, which can reduce their generalization ability and prediction stability [20,21]. Second, most existing studies have focused primarily on point prediction accuracy, whereas relatively limited attention has been paid to the spatial reliability of prediction results and pixel-scale uncertainty. This restricts the practical application of SOC mapping products in sampling-site optimization, risk identification, and soil carbon management [22,23]. Therefore, developing SOC prediction models suitable for limited-sample conditions and incorporating uncertainty assessment to characterize the spatial reliability of predictions are essential for improving the credibility of regional cropland SOC mapping.

TabPFN, namely the Tabular Prior-Data Fitted Network, is a pretrained foundation model designed for tabular prediction tasks. It learns general modeling priors from a large number of synthetic tabular tasks and enables rapid adaptation to new tasks through in-context learning [24,25]. Compared with traditional machine learning models, TabPFN depends less on complex hyperparameter tuning and large-scale training datasets, showing considerable potential for small- and medium-sized tabular data modeling. SOC prediction can be regarded as a typical tabular regression problem, in which the input variables consist of multi-source environmental covariates, such as climate factors, vegetation indices, soil spectral indices, and topographic variables, while the output variable is SOC content. Therefore, introducing TabPFN into cropland SOC spatial prediction under limited-sample conditions may help evaluate the applicability of tabular foundation models in digital soil mapping and provide a new modeling perspective for regional SOC prediction.

Based on this rationale, this study focuses on cropland in the Liaohe Plain. Using 310 surface soil samples collected in 2025 and multi-source environmental covariates, we developed a limited-sample SOC prediction framework based on the TabPFN model and compared its performance with RF, SVM, CatBoost, KNN, and XGBoost. In addition, model stability was evaluated using experiments with different training-sample proportions, and SOC spatial prediction and sampling-induced relative uncertainty assessment were conducted based on repeated random sampling. The main objectives of this study were to: (1) construct a multi-source environmental covariate system suitable for cropland SOC spatial prediction in the Liaohe Plain; (2) evaluate the applicability and stability of TabPFN under repeated validation and different training-sample conditions; and (3) reveal the spatial distribution pattern of cropland SOC and its sampling-induced relative uncertainty in the Liaohe Plain in 2025. The results provide methodological support for cropland SOC spatial mapping, sampling-site optimization, and regional soil carbon management under limited-sample conditions.

## 2. Materials and Methods

### *2.1. Study Area*

The Liaohe Plain (119.21–124.81° E, 39.92–43.48° N) is located in the southern part of Northeast China. It is an important component of the black soil region of Northeast China, as well as a major grain-producing area and typical agroecosystem in Liaoning Province. The study area covers approximately $8.23 \times 10^4$ km$^2$. The terrain is generally flat and characterized by extensive fluvial and alluvial deposition. Cropland resources are widely distributed and intensively utilized across the region (Figure 1).

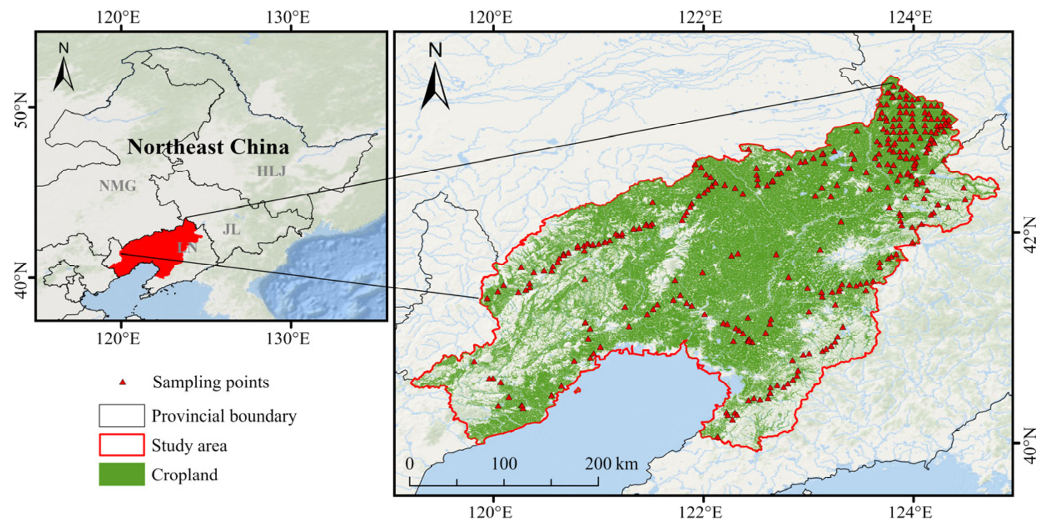


**Figure 1.** Location of the Liaohe Plain study area.

The Liaohe Plain has a temperate monsoon climate, with cold, dry winters and warm, rainy summers. The mean annual temperature ranges from approximately 4.3 to 11.6 °C, and annual precipitation ranges from approximately 386 to 1156 mm. The region has abundant hydrological resources, and its major rivers include the Liaohe River, Daling River, Hun River, and Taizi River. Under the combined influence of climate, parent material, topography, and long-term cultivation, cropland soils in this region are diverse, with the major soil types including brown soil, black soil, cinnamon soil, meadow soil, aeolian sandy soil, and paddy soil. Cropland is intensively managed, and the dominant crop types include maize, rice, and soybean.

### *2.2. Data Acquisition and Processing*

#### 2.2.1. Soil Data Acquisition and Processing

In April 2025, all 310 surface soil samples were collected from cropland across the Liaohe Plain. A stratified random sampling strategy was adopted to improve the representation of the major environmental conditions within the study area. Cropland distribution, site accessibility, topographic conditions, and soil types were considered in the sampling design. This approach has been widely applied in regional soil property prediction studies [26].

The soil sampling and SOC measurement procedures were as follows. Soil samples were collected from 30 m × 30 m grids, with each sampling site represented by a composite sample formed by thoroughly mixing five subsamples collected within the grid. The measured SOC content was therefore considered representative of the corresponding grid. The central coordinates of each grid were recorded using a portable GPS device (GPS G350, UniStrong, Beijing, China). To prevent cross-contamination, all samples were placed in individually labeled bags and transported to the laboratory. Upon arrival, the samples were air-dried, gently ground, and passed through a 2 mm sieve to remove coarse fragments and plant residues. The processed samples were then stored in sealed, labeled bags under cool, dry conditions at room temperature until analysis. SOC was determined from a 5 g subsample using the potassium dichromate oxidation–external heating method. Laboratory analyses were conducted from May to June 2025. Although this method is widely used, some analytical uncertainty may arise because of variations in oxidation efficiency.

#### 2.2.2. Remote Sensing Data Acquisition and Processing

Remote sensing data were obtained from the Landsat 8 OLI/TIRS and Landsat 9 OLI-2/TIRS-2 Collection 2 Level-2 surface reflectance and land surface temperature products, accessed through the Google Earth Engine (GEE) cloud computing platform (https://code.earthengine.google.com/, accessed on 11 August 2026). April–May 2025 was selected as the image acquisition period because it was close to the field sampling date and generally corresponded to the pre-sowing or early growth stages of major crops in the Liaohe Plain, when vegetation cover was relatively low [27]. This temporal window was therefore used to reduce vegetation interference with soil spectral signals while maintaining adequate spatial coverage. Images were filtered according to the study area, acquisition period, data quality, and cloud cover, with scenes exceeding 30% cloud cover excluded. A total of 87 Landsat 8/9 Collection 2 Level-2 images were retained. After application of the official scale factors and QA_PIXEL masking, observations from the two sensors were combined without additional cross-sensor harmonization.

All image preprocessing procedures were conducted on the GEE platform. The main steps included: (1) applying the official Collection 2 Level-2 scale factors to the optical and thermal bands to obtain surface reflectance and land surface temperature, respectively; (2) masking clouds, cloud shadows, cirrus, and snow or ice using the QA_PIXEL band; and (3) generating pixel-wise median composites for each band from all valid observations acquired during April–May 2025. This approach reduced the influence of residual cloud contamination, atmospheric noise, and short-term surface anomalies. The April–May LST composite was used as an auxiliary indicator of surface thermal conditions near the sampling period, rather than as a proxy for long-term thermal controls on SOC.

The resulting April–May composite was generated at a spatial resolution of 30 m. All Landsat-derived covariates, including LST and spectral indices, together with the topographic variables, were aligned to a common projection, spatial extent, pixel size, and grid origin. Continuous climate covariates were resampled from 1 km to 30 m using bilinear interpolation, whereas the categorical cropland mask was resampled using nearest-neighbor interpolation. Covariate values were extracted at the central coordinates of each 30 m $\times$ 30 m soil sampling grid and used as model inputs.

### 2.2.3. Acquisition and Calculation of Environmental Covariates

According to the soil–environment relationship theory, similar environmental conditions tend to produce similar soil properties [28]. Therefore, SOC content at a specific geographical location can be expressed as a function of the environmental covariates at that location [8]. In this study, 17 environmental covariates related to climate, vegetation, soil spectral characteristics, and topography were initially considered as candidate predictors for SOC prediction (Table 1). The corresponding sample-level covariate values are provided in Table S2. Multicollinearity assessment and feature selection were subsequently conducted before model development. The relationship between SOC content and environmental covariates can be expressed as follows:

$$SOC_i = g(F_i) + \varepsilon_i \tag{1}$$

where $SOC_i$ represents the soil organic carbon content at location $i$, $F_i$ denotes the vector of environmental covariates at location $i$, $g(\cdot)$ represents the nonlinear mapping relationship established by the prediction model, and $\varepsilon_i$ is the residual error.

**Table 1.** Environmental covariates initially considered as candidate predictors for SOC prediction.

| Category | Variable | Description | Spatial Resolution |
|---|---|---|---|
| Climatic factors | MAT | Mean annual temperature (°C) | 1 km (resampled to 30 m) |
| | MAP | Mean annual precipitation (mm) | |
| | EVP | Annual evaporation (mm) | |
| | MSR | Mean annual solar radiation (MJ $m^{-2}$) | |
| Vegetation indices | NDVI | Normalized difference vegetation index | 30 m |
| | EVI | Enhanced vegetation index | |
| Soil spectral indices | BI | Brightness index | 30 m |
| | CI | Coloration index | |
| | SI | Salinity index | |
| | LST | Land surface temperature | |
| | NDMI | Normalized difference moisture index | |
| Topographic factors | ELE | Elevation (m) | 30 m |
| | SLO | Slope (°) | |
| | ASP | Aspect (°) | |
| | PLC | Plan curvature | |
| | PRC | Profile curvature | |
| | TWI | Topographic wetness index | |

(1) Climatic Factors

Climatic data were obtained from station-based observations provided by the China Meteorological Data Service Centre (CMA, http://data.cma.cn/). Daily observations from 36 meteorological stations within and around the Liaohe Plain were spatially interpolated using ANUSPLIN, with elevation included as an auxiliary variable to account for terrain-related climatic gradients [29]. The interpolated daily surfaces of air temperature, precipitation, solar radiation, and evaporation were subsequently aggregated to generate the corresponding annual climatic covariates for 2025.

The climate surfaces were generated at a native spatial resolution of 1 km and resampled to the 30 m modeling grid solely for spatial alignment with the other covariates. They should therefore be interpreted as regional-scale climate surfaces rather than native 30 m observations.

(2) Vegetation Indices

The red band (Band 4, 0.64–0.67 µm), near-infrared band (Band 5, 0.85–0.88 µm), and blue band (Band 2, 0.45–0.51 µm) from Landsat 8/9 surface reflectance products were used to calculate the normalized difference vegetation index (NDVI; Equation (2)) and enhanced vegetation index (EVI; Equation (3)). NDVI reflects differences in surface vegetation cover, whereas EVI can partly reduce the saturation effect commonly associated with NDVI and improve vegetation sensitivity in areas with relatively high biomass [30].

$$NDVI = \frac{R_{NIR} - R_{Red}}{R_{NIR} + R_{Red}} \tag{2}$$

$$EVI = 2.5 \times \frac{R_{NIR} - R_{Red}}{(R_{NIR} + 6 \times R_{Red} - 7.5 \times R_{Blue} + 1)} \tag{3}$$

where $R_{NIR}$, $R_{Red}$, and $R_{Blue}$ represent the near-infrared, red, and blue bands of Landsat imagery, respectively.

(3) Soil Spectral Indices

The blue band (Band 2, 0.45–0.51 µm), red band (Band 4, 0.64–0.67 µm), near-infrared band (Band 5, 0.85–0.88 µm), shortwave infrared 1 band (Band 6, 1.57–1.65 µm), shortwave infrared 2 band (Band 7, 2.11–2.29 µm), and land surface temperature band (ST_B10, 10.6–11.2 µm) were selected to calculate soil spectral indices. Specifically, the brightness index (BI; Equation (4)), coloration index (CI; Equation (5)), salinity index (SI; Equation (6)), normalized difference moisture index (NDMI; Equation (7)), and land surface temperature (LST; Equation (8)) were derived. These indices were used to characterize soil brightness, soil color characteristics, salinity-related spectral response, surface moisture conditions, and the surface thermal environment, respectively [31,32].

$$BI = \sqrt{{R_{Red}}^2 + {R_{NIR}}^2} \tag{4}$$

$$CI = \frac{R_{SWIR1}}{R_{SWIR2}} \tag{5}$$

$$SI = \sqrt{R_{Blue} \times R_{Red}} \tag{6}$$

$$NDMI = \frac{R_{NIR} - R_{SWIR1}}{R_{NIR} + R_{SWIR1}} \tag{7}$$

$$LST = ST\_B10 \times 0.00341802 + 149.0 \tag{8}$$

where $R_{Red}$, $R_{NIR}$, $R_{Blue}$, $R_{SWIR1}$, and $R_{SWIR2}$ represent the surface reflectance of the red, near-infrared, blue, shortwave infrared 1, and shortwave infrared 2 bands, respectively. $ST\_B10$ denotes the original pixel value of the land surface temperature band, and LST was obtained in Kelvin after applying the scale factor conversion.

(4) Topographic Factors

Topographic factors were derived from the SRTM 30 m digital elevation model (DEM; http://srtm.csi.cgiar.org/). Based on the DEM, elevation (ELE), slope (SLO), aspect (ASP), plan curvature (PLC), profile curvature (PRC), and topographic wetness index (TWI) were extracted to characterize terrain relief, slope gradient and direction, surface curvature, and potential water accumulation conditions in the study area. Among these variables, TWI reflects the spatial tendency of surface moisture accumulation and is generally associated with soil erosion, material transport, and organic matter accumulation processes [33].

#### 2.2.4. Land Use Data

Land use data were obtained from the Resource and Environment Science and Data Center (RESDC) of the Chinese Academy of Sciences (https://www.resdc.cn), with a spatial resolution of 30 m. In this study, the 2025 land use dataset was selected to extract cropland within the study area and generate a cropland mask.

### *2.3. Methods*

#### 2.3.1. Multicollinearity Assessment and Feature Selection

Before model development, multicollinearity among the 17 candidate environmental covariates was assessed using pairwise Pearson correlation coefficients and variance inflation factors (VIFs). Absolute Pearson correlation coefficients of $|r| \geq 0.80$ and VIF values of $\geq$10 were used to identify strongly correlated or potentially redundant predictors. The variable with the highest VIF was removed at each step, and the VIF values were recalculated after each removal. This procedure was repeated until all retained covariates had VIF values below 10. The covariates retained after screening were subsequently used for model evaluation, interpretation, spatial prediction, and sampling-induced relative uncertainty assessment.

#### 2.3.2. TabPFN Model

TabPFN is a pretrained foundation model designed for tabular prediction tasks and represents an emerging deep-learning approach for tabular data [24]. Based on the Transformer architecture, TabPFN is pretrained on a large number of synthetic tabular tasks to learn general modeling priors across different data distributions and variable relationships. When applied to a new dataset, both the training samples and the samples to be predicted are jointly used as contextual inputs. Model fitting and prediction are then completed through in-context learning in a single forward pass. In this study, TabPFN's attention operated within the tabular feature space and did not explicitly model geographic distance, spatial autocorrelation, or semivariogram structure. Compared with traditional machine learning models, TabPFN depends less on complex hyperparameter tuning and is suitable for small- and medium-sized tabular prediction tasks, particularly under limited-sample conditions [25].

#### 2.3.3. Comparative Models

To assess the applicability of TabPFN for SOC prediction, random forest (RF), support vector machine (SVM), CatBoost, k-nearest neighbors regression (KNN), and XGBoost were selected as benchmark models. These algorithms represent several widely used learning paradigms in soil property prediction and digital soil mapping, including bagging-based tree ensembles, gradient-boosting decision trees, kernel-based regression, and local distance-based regression [34–36]. Collectively, they provide a representative basis for comparing nonlinear modeling approaches for SOC prediction.

RF is a decision tree-based ensemble learning method that reduces the instability of a single model by constructing multiple regression trees and averaging their prediction results [37]. SVM maps input variables into a high-dimensional feature space through kernel functions and establishes a regression function in that space, making it suitable for nonlinear regression problems with small samples [38]. CatBoost and XGBoost are both gradient boosting tree models that improve predictive performance by iteratively fitting residuals and have strong nonlinear representation capabilities [39,40]. KNN is a nonparametric regression method based on sample similarity. It predicts target values through the weighted averaging of neighboring samples and can characterize local relationships among samples [41].

#### 2.3.4. Model Optimization and Evaluation

Model performance was evaluated using 100 repeated random 80:20 holdout validation runs. In each repetition, the full dataset was randomly divided into a training set of 248 samples and a held-out test set of 62 samples. A different random seed was used for each repetition. Model performance was assessed on the held-out test set and summarized as the mean ± standard deviation of $R^2$, RMSE, MAE, and RPD across the 100 repetitions.

Within each repetition, the hyperparameters of RF, SVM, CatBoost, KNN, and XGBoost were optimized exclusively using the training data through randomized search coupled with 5-fold cross-validation. Up to 30 hyperparameter configurations were sampled from the predefined search spaces listed in Table 2, and the configuration yielding the lowest cross-validated RMSE was selected. The held-out test set was not used to fit preprocessing procedures, select hyperparameters, or train the models. For SVM and KNN, feature standardization was incorporated into the cross-validation pipeline to prevent information leakage. All models were evaluated using identical training–test partitions. TabPFN was applied using its pretrained inference configuration without task-specific hyperparameter optimization.

**Table 2.** Hyperparameter search spaces used for model optimization.

| Model | Hyperparameter Search Space |
|---|---|
| RF | n_estimators: 300, 500; max_depth: None, 10, 20; min_samples_split: 2, 5; min_samples_leaf: 1, 2; max_features: sqrt, 1.0 |
| SVM | C: 0.1, 1, 10, 100; gamma: scale, 0.01, 0.1, 1; epsilon: 0.05, 0.1, 0.2 |
| CatBoost | iterations: 300, 600; depth: 4, 6, 8; learning_rate: 0.03, 0.05, 0.1; l2_leaf_reg: 1, 3 |
| KNN | n_neighbors: 3, 5, 7, 9, 11, 15; weights: uniform, distance; p: 1, 2 |
| XGBoost | n_estimators: 300, 600; max_depth: 3, 5; learning_rate: 0.03, 0.05, 0.1; subsample: 0.8, 1.0; colsample_bytree: 0.8, 1.0 |
| TabPFN | Pretrained inference configuration without task-specific hyperparameter optimization |

Note: Parameters not listed in the table were kept at their default settings.

The hyperparameters listed in Table 2 control key aspects of model complexity, regularization, and learning behavior. For RF, n_estimators determines ensemble size, while tree depth, minimum sample requirements for node splitting and leaf formation, and feature-sampling parameters regulate tree complexity. For SVM, C, gamma, and epsilon control the regularization–error trade-off, kernel behavior, and the $\varepsilon$-insensitive region, respectively. For CatBoost and XGBoost, the number of trees, depth, and learning rate govern boosting complexity, with L2 regularization in CatBoost and subsampling in XGBoost helping to limit overfitting. For KNN, n_neighbors, weights, and p define neighborhood size, weighting, and the order of the Minkowski distance metric, respectively.

For the repeated random holdout results, all models were evaluated using identical data partitions, resulting in a paired comparison structure. For each evaluation metric, differences among the six models were assessed using the Friedman test. When a significant overall difference was detected, pairwise comparisons were conducted using the Wilcoxon signed-rank test with Holm correction for multiple comparisons. Statistical significance was determined at $p < 0.05$.

In addition, repeated five-fold spatial cross-validation based on size-constrained k-means clustering was conducted to assess model transferability to geographically separated areas. Sampling coordinates were transformed to the WGS 84/UTM zone 51N projected coordinate reference system (EPSG:32651). In each repetition, the 310 sites were divided into five geographically compact spatial folds containing 50–74 samples each, using only the projected coordinates. SOC observations, environmental covariates, and model-performance information were not used to construct the folds. Each fold was used once as the outer test set, while the remaining four folds were used for model training. Hyperparameters of RF, SVM, CatBoost, KNN, and XGBoost were optimized within the outer training data using four-fold spatial group cross-validation and the search spaces listed in Table 2. The outer test fold was excluded from preprocessing, hyperparameter selection, and model fitting. TabPFN was applied using its pretrained inference configuration. Identical spatial partitions were used for all models. The procedure was repeated using 20 spatial partitions generated with different random seeds. Within each repetition, predictions from the five outer test folds were combined to obtain one spatially out-of-fold prediction for each sample. $R^2$, RMSE, MAE, and RPD were calculated from the combined predictions, and model performance was reported as the mean $\pm$ standard deviation across the 20 repetitions. All validation analyses were conducted within the 2025 dataset and were designed to assess robustness to sample partitioning and spatial separation.

The coefficient of determination ($R^2$), root mean square error (RMSE), mean absolute error (MAE), and ratio of performance to deviation (RPD) were used to evaluate model performance. $R^2$ was used to characterize the ability of the model to explain variations in SOC content. RMSE and MAE were used to quantify the differences between predicted and observed values, whereas RPD was used to assess the overall reliability of model prediction. In general, higher $R^2$ and RPD values, together with lower RMSE and MAE values, indicate better predictive performance [42]. For RPD interpretation, values below 1.4 were considered to indicate poor predictive ability, values between 1.4 and 2.0 indicated moderate predictive ability, and values above 2.0 indicated good predictive ability. This criterion was used as a supplementary reference for assessing the practical reliability of SOC prediction [43]. The formulas for these evaluation metrics are as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (C_i - \widehat{C}_i)^2}{\sum_{i=1}^{n} (C_i - \overline{C})^2} \tag{9}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (C_i - \widehat{C}_i)^2} \tag{10}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |C_i - \widehat{C}_i| \tag{11}$$

$$RPD = \frac{SD}{RMSE} \tag{12}$$

where $C_i$ represents the observed SOC content of the $i$-th sample; $\widehat{C}_i$ represents the predicted SOC content of the $i$-th sample; $\overline{C}$ represents the mean observed SOC content; $n$ represents the number of samples; and $SD$ represents the standard deviation of the observed SOC content.

#### 2.3.5. Small-Sample Stability and Sampling-Induced Uncertainty Assessment

To assess model stability under reduced training-sample conditions, a paired repeated-sampling design was used. In each of 100 repetitions, the full dataset was randomly divided into a training-candidate set of 248 samples and a held-out test set of 62 samples. A single random ordering of the training-candidate samples was then used to construct four nested training subsets: S1, S2, S3, and S4 contained 99, 149, 198, and 248 samples, respectively, corresponding to approximately 40%, 60%, 80%, and 100% of the training-candidate set. Thus, S1⊂S2⊂S3⊂S4 within each repetition.

The same held-out test set and nested training subsets were used for TabPFN, RF, XGBoost, and CatBoost. These models were selected because they showed relatively competitive performance in the initial comparison. For RF, XGBoost, and CatBoost, hyperparameters were optimized exclusively within each scenario-specific training subset using the procedure described in Section 2.3.4, whereas TabPFN was applied using its pretrained inference configuration. The test set was used only for performance evaluation. The procedure was repeated 100 times using different random seeds. Model performance was evaluated on the common held-out test set and reported as the mean ± standard deviation across repetitions. The experimental scenarios are summarized in Table 3.

**Table 3.** Experimental scenarios for the paired small-sample stability analysis.

| Scenario | Proportion of Training-Candidate Set | Number of Training Samples | Number of Testing Samples | Experimental Purpose |
|---|---|---|---|---|
| S1 | 40% | 99 | 62 | Low training-sample condition |
| S2 | 60% | 149 | 62 | Moderate training-sample condition |
| S3 | 80% | 198 | 62 | Relatively sufficient training-sample condition |
| S4 | 100% | 248 | 62 | Full training-candidate condition |

Note: Within each of the 100 repetitions, S1–S4 were nested training subsets sharing the same 62-sample held-out test set and were applied identically to all four models.

Because S1–S4 shared the same held-out test set and were derived from the same training-candidate set within each repetition, their performance metrics had a paired structure. Differences among the four scenarios were assessed using the statistical procedure described in Section 2.3.4.

Model evaluation considered both point prediction accuracy and sampling-induced prediction variability. Point prediction accuracy was assessed using $R^2$, RMSE, MAE, and RPD, whereas sampling-induced variability was characterized from the empirical distribution of predictions across repeated random sampling and model-fitting runs. This measure reflects the sensitivity of SOC predictions to changes in training-sample composition.

In the spatial prediction stage, the 100 TabPFN models generated under the S4 scenario, each trained using all 248 samples in the corresponding training-candidate set, were applied to every cropland pixel in the study area. This yielded 100 SOC predictions for each pixel. The mean of these predictions was used as the final SOC estimate, while the 5th and 95th percentiles were used to calculate the sampling-induced relative uncertainty index (UNC). The UNC was calculated as follows:

$$UNC_j = \frac{Q_{0.95}\left(\widehat{C}_j\right) - Q_{0.05}\left(\widehat{C}_j\right)}{\overline{C}_j} \tag{13}$$

where $UNC_j$ represents the sampling-induced relative uncertainty of the $j$-th pixel; $Q_{0.95}\left(\widehat{C}_j\right)$ and $Q_{0.05}\left(\widehat{C}_j\right)$ represent the 95th and 5th percentiles of the 100 SOC prediction results for this pixel, respectively; and $\overline{C}_j$ represents the mean predicted SOC content of this pixel. A larger $UNC_j$ value indicates a wider empirical percentile range relative to the mean prediction and, therefore, greater sensitivity of the pixel-level SOC prediction to training-sample perturbations.

2.3.6. Environmental Similarity and Extrapolation Assessment

Environmental similarity and potential extrapolation risk were assessed using the 14 covariates retained after feature selection. All covariates were standardized using the mean and standard deviation of the 310 sampling sites, and environmental dissimilarity for each cropland pixel was defined as its minimum Euclidean distance to the sampled sites in the standardized covariate space. Pixels outside the sampled range of at least one covariate were identified as univariate extrapolations. The sample-supported environmental domain was defined using the 95th percentile of the leave-one-out nearest-neighbor distances among the sampling sites. Associations between environmental dissimilarity and sampling-induced relative uncertainty were evaluated using Spearman's rank correlation, comparisons of median UNC inside and outside the environmental domain, and the proportion of high-UNC pixels, defined as the upper 10%, located outside the domain.

## 3. Results

### *3.1. Dataset Characterization and Covariate Screening*

The SOC dataset was first characterized using descriptive statistics. As shown in Table 4, SOC content in the 310 cropland soil samples ranged from 1.24 to 38.62 g $kg^{-1}$, with a mean of 12.92 g $kg^{-1}$ and a standard deviation of 6.76 g $kg^{-1}$. The coefficient of variation was 0.52, indicating moderate variability in cropland SOC content across the Liaohe Plain.

**Table 4.** Descriptive statistics of SOC content in the soil samples.

| Soil Depth (cm) | N | Mean | Maximum | Minimum | SD | CV |
|---|---|---|---|---|---|---|
| 0–20 | 310 | 12.92 | 38.62 | 1.24 | 6.76 | 0.52 |

Note: SD, standard deviation; CV, coefficient of variation. Mean, maximum, minimum, and SD are expressed in g $kg^{-1}$.

Multicollinearity was assessed for the 17 candidate environmental covariates listed in Table 1. NDVI and EVI were strongly correlated (r = 0.943) and exhibited high initial VIF values of 93.255 and 68.521, respectively. Because NDVI had the higher VIF and provided largely redundant information relative to EVI, it was removed first. After the VIFs were recalculated, SI had the highest remaining VIF (17.390) and was strongly correlated with BI (r = 0.839); therefore, SI was excluded. A subsequent recalculation showed that EVP still exceeded the VIF threshold, with a value of 10.930, and it was also removed. The final predictor set comprised 14 environmental covariates, with VIF values ranging from 1.159 to 3.079, indicating that no substantial multicollinearity remained among the retained predictors. Detailed VIF values and the sequential variable-screening process are provided in Table S1.

### *3.2. Model Performance Under Random Holdout Validation*

In this study, TabPFN was applied to predict cropland SOC across the Liaohe Plain and was compared with five commonly used machine-learning models: RF, SVM, CatBoost, KNN, and XGBoost. To reduce the influence of any single random partition, model performance was evaluated using 100 repetitions of random 80:20 holdout validation. The performance of each model is reported as the mean ± standard deviation across the 100 repetitions in Table 5.

**Table 5.** Model performance across 100 repetitions of random 80:20 holdout validation.

| Model | Training Samples | Testing Samples | MAE (g kg$^{-1}$) | RPD | RMSE (g kg$^{-1}$) | $R^2$ |
|---|---|---|---|---|---|---|
| TabPFN | 248 | 62 | 3.19 ± 0.15 | 1.62 ± 0.07 | 4.026 ± 0.184 | 0.608 ± 0.038 |
| RF | 248 | 62 | 3.13 ± 0.11 | 1.60 ± 0.06 | 4.041 ± 0.132 | 0.603 ± 0.026 |
| SVM | 248 | 62 | 3.58 ± 0.24 | 1.39 ± 0.09 | 4.620 ± 0.291 | 0.472 ± 0.064 |
| CatBoost | 248 | 62 | 3.10 ± 0.17 | 1.58 ± 0.08 | 4.126 ± 0.211 | 0.585 ± 0.041 |
| KNN | 248 | 62 | 3.52 ± 0.21 | 1.42 ± 0.08 | 4.530 ± 0.262 | 0.493 ± 0.058 |
| XGBoost | 248 | 62 | 3.18 ± 0.19 | 1.57 ± 0.10 | 4.129 ± 0.243 | 0.584 ± 0.048 |

Note: Values are reported as the mean ± standard deviation across 100 repeated random 80:20 holdout validation runs. In each run, 248 samples were used for model training, and 62 samples were held out for testing.

As shown in Table 5, TabPFN achieved mean $R^2$, RMSE, MAE, and RPD values of 0.608 ± 0.038, 4.026 ± 0.184 g kg$^{-1}$, 3.19 ± 0.15 g kg$^{-1}$, and 1.62 ± 0.07, respectively. Based on the adopted RPD criterion, its predictive ability was classified as moderate. RF yielded closely comparable results, with an $R^2$ of 0.603 ± 0.026 and an RMSE of 4.041 ± 0.132 g kg$^{-1}$. CatBoost and XGBoost also showed similar predictive performance, with $R^2$ values of 0.585 ± 0.041 and 0.584 ± 0.048, respectively. By contrast, SVM and KNN produced lower $R^2$ values of 0.472 ± 0.064 and 0.493 ± 0.058, respectively.

The Friedman test detected significant overall differences among the six models for the evaluated performance metrics ($p < 0.05$). Subsequent pairwise Wilcoxon signed-rank tests with Holm correction indicated that TabPFN significantly outperformed SVM and KNN for most metrics. However, after correction for multiple comparisons, the differences between TabPFN and RF, CatBoost, or XGBoost were generally not statistically significant. Overall, TabPFN showed competitive predictive performance among the evaluated models. Although it achieved the highest mean $R^2$ and RPD values and the lowest mean RMSE, the numerical differences relative to RF were small. CatBoost achieved the lowest MAE, whereas TabPFN and RF produced nearly identical $R^2$ and RMSE values. These findings suggest that TabPFN is a viable option for limited-sample SOC prediction, although it did not consistently outperform the tree-based benchmark models.

### *3.3. Repeated Five-Fold Spatial Cross-Validation*

To further evaluate model transferability to geographically separated areas, repeated five-fold spatial cross-validation based on spatially constrained clustering was conducted. The performance of the six models under this validation strategy is summarized in Table 6.

**Table 6.** Model performance under repeated five-fold spatial cross-validation based on spatially constrained clustering.

| Model | MAE (g kg$^{-1}$) | RPD | RMSE (g kg$^{-1}$) | $R^2$ |
|---|---|---|---|---|
| TabPFN | 3.32 ± 0.26 | 1.47 ± 0.09 | 4.33 ± 0.31 | 0.535 ± 0.072 |
| RF | 3.36 ± 0.24 | 1.45 ± 0.08 | 4.39 ± 0.29 | 0.520 ± 0.067 |
| SVM | 3.82 ± 0.33 | 1.30 ± 0.10 | 4.92 ± 0.41 | 0.395 ± 0.096 |
| CatBoost | 3.39 ± 0.28 | 1.43 ± 0.09 | 4.45 ± 0.34 | 0.508 ± 0.078 |
| KNN | 3.75 ± 0.31 | 1.33 ± 0.10 | 4.83 ± 0.39 | 0.418 ± 0.091 |
| XGBoost | 3.44 ± 0.30 | 1.41 ± 0.10 | 4.52 ± 0.37 | 0.492 ± 0.083 |

Note: Values are reported as the mean ± standard deviation across 20 repeated spatial cross-validation runs.

As shown in Table 6, repeated spatial cross-validation produced more conservative estimates of model performance than random holdout validation. Under repeated spatial cross-validation, TabPFN achieved an average $R^2$ of 0.535 ± 0.072 and an RMSE of

$4.33 \pm 0.31$ g kg$^{-1}$. Its performance was comparable to that of RF, CatBoost, and XGBoost, which achieved $R^2$ values of $0.520 \pm 0.067$, $0.508 \pm 0.078$, and $0.492 \pm 0.083$, respectively. In contrast, SVM and KNN showed relatively lower spatial generalization performance, with $R^2$ values of $0.395 \pm 0.096$ and $0.418 \pm 0.091$, respectively.

Compared with random holdout validation, the lower predictive accuracy obtained from repeated spatial cross-validation indicates that random splitting may overestimate model performance when sampling points are spatially clustered. Nevertheless, the spatial validation results suggest that TabPFN and the tree-based models retained moderate spatial generalization ability under limited-sample conditions.

### *3.4. Small-Sample Stability Analysis*

To further assess the sensitivity of model performance to training-sample size, four training-sample proportion scenarios, namely 40%, 60%, 80%, and 100%, were examined. The TabPFN, RF, XGBoost, and CatBoost models were evaluated through 100 repeated experiments under each scenario, and the results are presented in Table 7.

**Table 7.** Model performance under different training-sample proportion scenarios.

| Scenario | Model | MAE (g kg$^{-1}$) | RPD | RMSE (g kg$^{-1}$) | $R^2$ |
|---|---|---|---|---|---|
| S1 | TabPFN | $3.41 \pm 0.29$ | $1.49 \pm 0.12$ | $4.365 \pm 0.401$ | $0.531 \pm 0.088$ |
| | RF | $3.58 \pm 0.22$ | $1.42 \pm 0.10$ | $4.698 \pm 0.291$ | $0.425 \pm 0.064$ |
| | XGBoost | $3.52 \pm 0.29$ | $1.43 \pm 0.11$ | $4.655 \pm 0.381$ | $0.443 \pm 0.086$ |
| | CatBoost | $3.45 \pm 0.27$ | $1.47 \pm 0.09$ | $4.547 \pm 0.346$ | $0.477 \pm 0.074$ |
| S2 | TabPFN | $3.29 \pm 0.21$ | $1.54 \pm 0.10$ | $4.236 \pm 0.289$ | $0.560 \pm 0.058$ |
| | RF | $3.41 \pm 0.16$ | $1.48 \pm 0.08$ | $4.421 \pm 0.191$ | $0.507 \pm 0.037$ |
| | XGBoost | $3.33 \pm 0.22$ | $1.50 \pm 0.10$ | $4.362 \pm 0.286$ | $0.524 \pm 0.057$ |
| | CatBoost | $3.28 \pm 0.23$ | $1.51 \pm 0.09$ | $4.319 \pm 0.309$ | $0.536 \pm 0.061$ |
| S3 | TabPFN | $3.16 \pm 0.15$ | $1.60 \pm 0.07$ | $4.094 \pm 0.207$ | $0.593 \pm 0.039$ |
| | RF | $3.29 \pm 0.13$ | $1.53 \pm 0.07$ | $4.196 \pm 0.176$ | $0.566 \pm 0.032$ |
| | XGBoost | $3.24 \pm 0.18$ | $1.52 \pm 0.08$ | $4.224 \pm 0.219$ | $0.557 \pm 0.041$ |
| | CatBoost | $3.20 \pm 0.20$ | $1.57 \pm 0.08$ | $4.121 \pm 0.236$ | $0.585 \pm 0.044$ |
| S4 | TabPFN | $3.19 \pm 0.15$ | $1.62 \pm 0.07$ | $4.026 \pm 0.184$ | $0.608 \pm 0.038$ |
| | RF | $3.13 \pm 0.11$ | $1.60 \pm 0.06$ | $4.041 \pm 0.132$ | $0.603 \pm 0.026$ |
| | XGBoost | $3.18 \pm 0.19$ | $1.57 \pm 0.10$ | $4.129 \pm 0.243$ | $0.584 \pm 0.048$ |
| | CatBoost | $3.10 \pm 0.17$ | $1.58 \pm 0.08$ | $4.126 \pm 0.211$ | $0.585 \pm 0.041$ |

Note: Values are reported as mean ± standard deviation across 100 repeated experiments. In each run, 20% of the full dataset was used as the testing set, and different proportions of the remaining training candidate samples were used for model training. S1, S2, S3, and S4 represent 40%, 60%, 80%, and 100% of the training candidate samples, respectively.

As shown in Table 7, model performance generally improved as the proportion of training samples increased from 40% to 100%. This trend was reflected mainly by increasing $R^2$ and RPD values and decreasing RMSE values, although minor non-monotonic changes were observed for MAE. For TabPFN, mean $R^2$ increased from $0.531 \pm 0.088$ under S1 to $0.608 \pm 0.038$ under S4, while RMSE decreased from $4.365 \pm 0.401$ to $4.026 \pm 0.184$ g kg$^{-1}$. Over the same scenarios, RPD increased from $1.49 \pm 0.12$ to $1.62 \pm 0.07$. Similar overall improvements were observed for RF, XGBoost, and CatBoost, indicating that both training-sample size and sample composition influenced SOC prediction performance under limited-sample conditions.

The Friedman test showed that training-sample proportion had a significant effect on model performance for TabPFN, RF, XGBoost, and CatBoost for most evaluation metrics ($p < 0.05$). Post hoc Wilcoxon signed-rank tests with Holm correction indicated that the 100% training-candidate-sample scenario generally performed significantly better than the 40% scenario, whereas differences between adjacent scenarios were not always statistically significant. These results indicate that increasing the training-sample proportion generally improved SOC prediction performance, although the magnitude and statistical significance of this improvement varied among models and evaluation metrics.

Under S1, TabPFN achieved the highest mean $R^2$ and RPD values and the lowest mean RMSE and MAE among the four models, with an $R^2$ of $0.531 \pm 0.088$ and an RMSE of $4.365 \pm 0.401$ g kg$^{-1}$. Its numerical advantage was therefore most evident under the most restricted training-sample condition. Under S4, TabPFN and RF showed closely comparable performance, with $R^2$ values of $0.608 \pm 0.038$ and $0.603 \pm 0.026$ and RMSE values of $4.026 \pm 0.184$ and $4.041 \pm 0.132$ g kg$^{-1}$, respectively. CatBoost achieved the lowest MAE, whereas TabPFN obtained the highest $R^2$ and RPD and the lowest RMSE. Overall, TabPFN was particularly competitive under reduced training-sample conditions, but its advantage over the tree-based models was not consistent across all metrics or scenarios.

### 3.5. Model Interpretability Analysis

To further reveal the effects of different environmental covariates on SOC prediction, SHAP was used to interpret the TabPFN model. The overall contributions of environmental covariates were summarized according to four groups of factors: climate, vegetation, soil spectral indices, and topography. The results are shown in Figures 2 and 3.

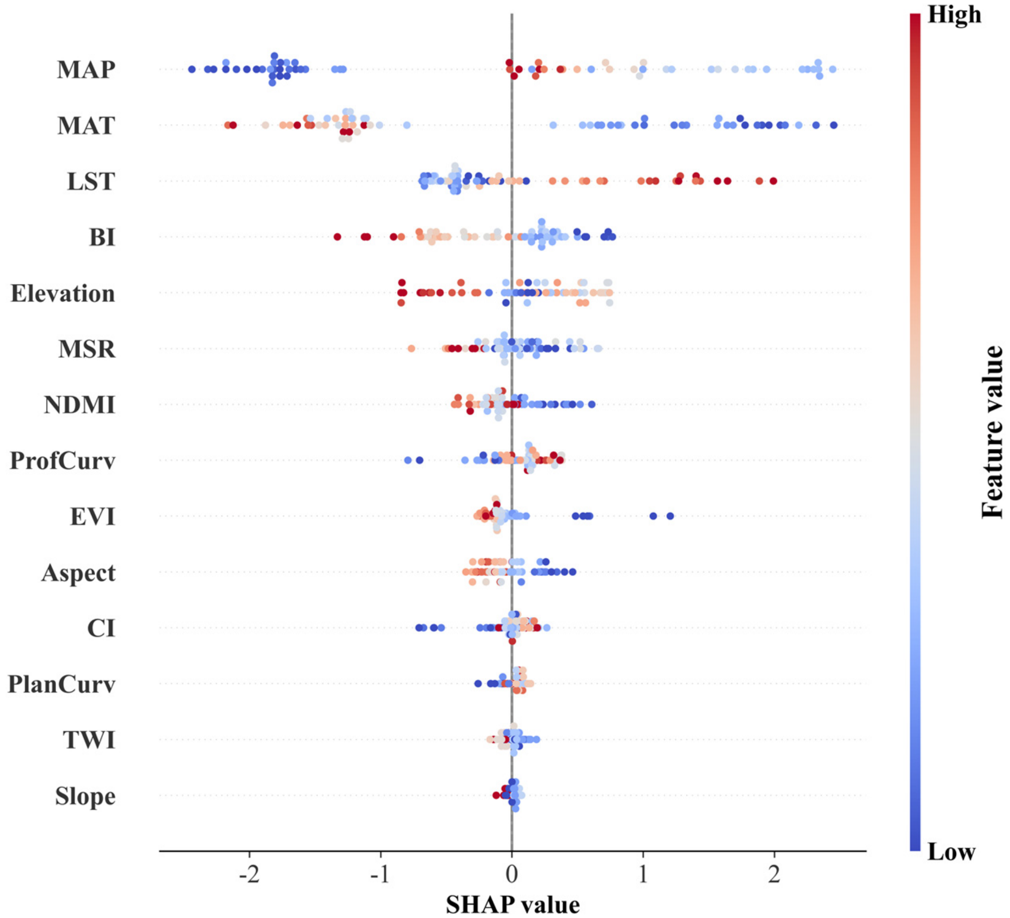


**Figure 2.** SHAP summary plot of environmental covariates based on the TabPFN model.

As shown in Figure 2, the magnitude and direction of the SHAP values varied substantially among the retained environmental covariates. MAP, MAT, LST, and BI exhibited relatively wide SHAP value distributions, indicating that these variables had comparatively strong influences on SOC prediction. Higher LST values were generally associated with positive SHAP contributions, whereas higher MAT and BI values tended to produce negative contributions. MAP showed a broader and partly nonlinear response, with both positive and negative SHAP values across its observed range. Elevation, MSR, and NDMI showed intermediate effects, whereas the SHAP values of PlanCurv, TWI, and Slope were mainly concentrated near zero, indicating relatively limited marginal contributions to the model predictions.

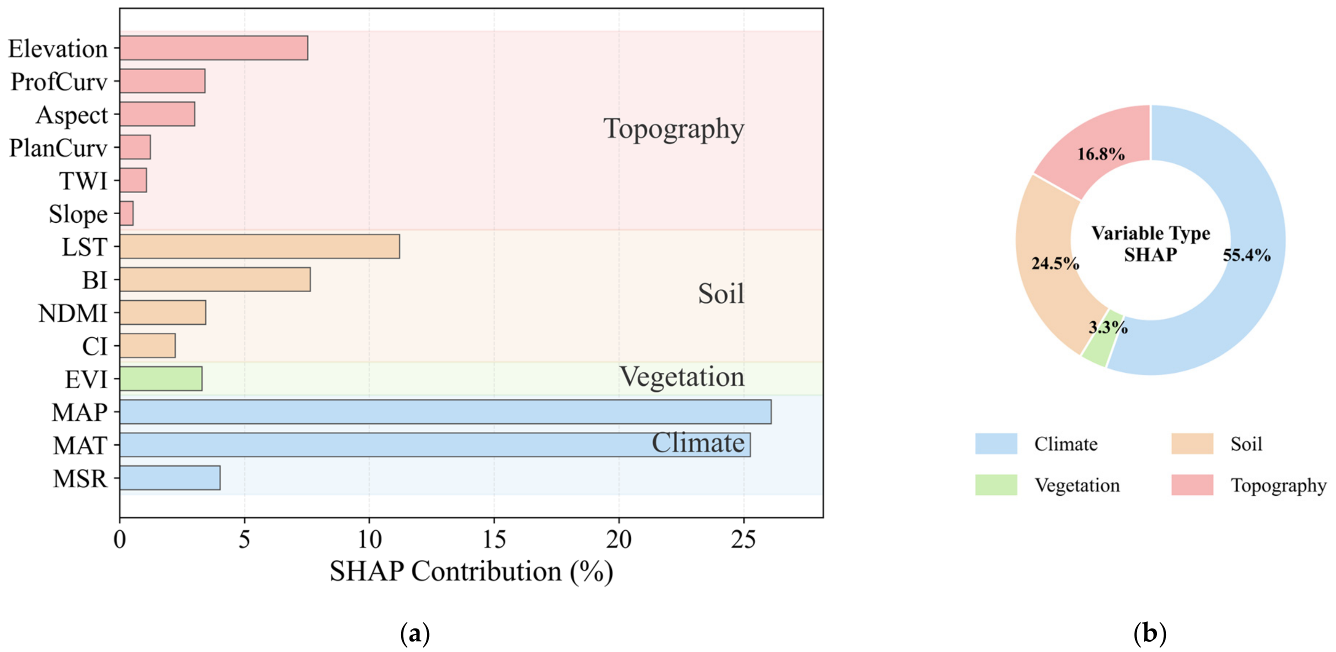


**Figure 3.** SHAP contribution analysis based on the TabPFN model. (**a**) Variable importance ranking; (**b**) contribution proportions of variable groups.

As shown in Figure 3, the retained environmental covariates differed substantially in their contributions to SOC prediction. Climatic variables made the largest overall contribution, accounting for 55.4% of the total SHAP importance, with MAP and MAT emerging as the dominant predictors and MSR contributing comparatively less. Soil spectral variables ranked second, accounting for 24.5%, primarily driven by LST and BI, whereas NDMI and CI showed smaller contributions. Topographic variables collectively accounted for 16.8%, with elevation contributing most strongly, followed by ProfCurv and aspect. Vegetation variables contributed only 3.3%, with EVI representing the sole retained vegetation index. Overall, climatic and soil spectral variables provided the principal predictive information for the TabPFN model, while topographic factors also made a meaningful contribution to cropland SOC prediction.

### *3.6. Spatial Prediction and Uncertainty Assessment*

Given its competitive performance under both repeated random holdout validation and repeated spatial cross-validation, TabPFN was used to predict cropland SOC across the Liaohe Plain in 2025. These predictions were subsequently used to generate the spatial distribution map of cropland SOC (Figure 4) and the corresponding sampling-induced relative uncertainty map (Figure 5).

As shown in Figure 4, cropland SOC content in the Liaohe Plain exhibited pronounced spatial variation, ranging from 4.63 to 27.04 g $kg^{-1}$. High SOC values were mainly distributed in the northeastern part of the study area, especially in northern Tieling and northeastern Fushun. In contrast, low SOC values were primarily observed in the western and northwestern parts of the study area, including parts of Fuxin, Jinzhou, Chaoyang, and

https://doi.org/10.3390/agronomy16161562

Huludao. Overall, the spatial pattern showed a gradual transition from low SOC values in the west to high SOC values in the northeast.

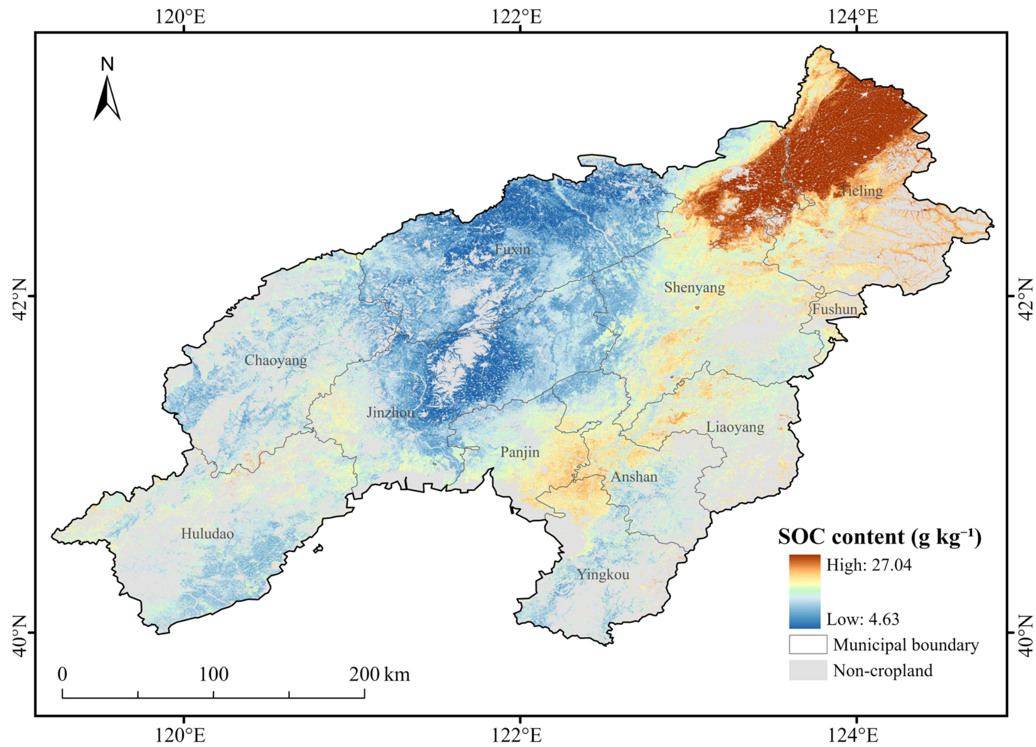


**Figure 4.** Spatial distribution of soil organic carbon content in cropland across the Liaohe Plain in 2025. Municipal boundaries are shown only within the study area.

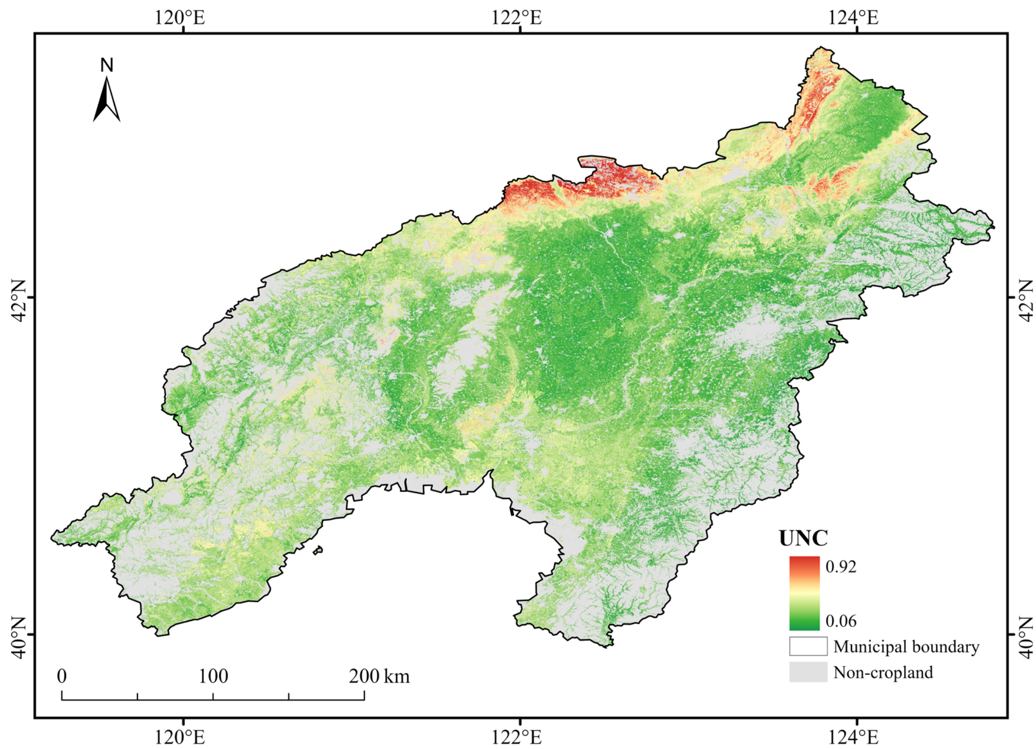


**Figure 5.** Spatial distribution of sampling-induced relative uncertainty in cropland SOC prediction across the Liaohe Plain. Higher UNC values indicate greater sensitivity of pixel-level SOC predictions to training-sample perturbations.

As shown in Figure 5, the spatial distribution of UNC exhibited clear spatial heterogeneity. High-UNC areas were mainly concentrated in the northern and northeastern parts of the study area, as well as in some marginal regions, where they displayed patchy and belt-like clustering patterns. Comparison with the SOC prediction map showed that some areas with relatively high predicted SOC values in the northeast also exhibited elevated sampling-induced relative uncertainty. This suggests that SOC predictions in these areas were more sensitive to training-sample perturbations, reflecting greater variability in the learned relationships between environmental covariates and SOC under limited-sample conditions.

The predicted SOC values ranged from 4.63 to 27.04 g $kg^{-1}$ and remained within the observed range of 1.24–38.62 g $kg^{-1}$, indicating that the final spatial predictions did not extend beyond the observed response range. However, the environmental-domain assessment showed that 4.8% of cropland pixels exceeded the sampled range of at least one covariate, while 14.1% were classified as environmentally dissimilar based on their multivariate covariate combinations. Environmental dissimilarity showed a positive association with UNC (Spearman's $\rho = 0.32$). The median UNC was 0.217 within the sample-supported environmental domain and 0.291 outside the domain. Among pixels within the upper 10% of UNC values, 37.5% occurred outside the sample-supported environmental domain. These results indicate that covariate shift was associated with some uncertainty hotspots, although it did not fully explain the observed UNC pattern.

## 4. Discussion

### 4.1. Applicability and Stability of the Model

This study evaluated the applicability and stability of TabPFN for cropland SOC prediction in the Liaohe Plain using repeated random holdout validation, repeated spatial cross-validation, and small-sample stability experiments (Tables 5–7). Across 100 repetitions of random 80:20 holdout validation, TabPFN achieved the highest mean $R^2$ and RPD values and the lowest mean RMSE, with values of $0.608 \pm 0.038$, $1.62 \pm 0.07$, and $4.026 \pm 0.184$ g $kg^{-1}$, respectively. However, its performance was closely comparable to that of RF, which achieved an $R^2$ of $0.603 \pm 0.026$ and an RMSE of $4.041 \pm 0.132$ g $kg^{-1}$. CatBoost produced the lowest MAE, while RF also achieved a slightly lower MAE than TabPFN. Moreover, the post hoc comparisons showed that the differences between TabPFN and RF, CatBoost, or XGBoost were generally not statistically significant after Holm correction. These results demonstrate that TabPFN provided competitive predictive performance under limited-sample conditions. However, its numerical advantage over the tree-based models was relatively small and was not uniformly reflected across all evaluation metrics.

Repeated spatial cross-validation produced more conservative estimates of model performance and provided a stricter assessment of transferability to geographically separated areas. Under this validation strategy, TabPFN achieved a mean $R^2$ of $0.535 \pm 0.072$ and an RMSE of $4.33 \pm 0.31$ g $kg^{-1}$. RF, CatBoost, and XGBoost showed broadly comparable, although slightly lower, spatial prediction performance. The reduction in accuracy relative to random holdout validation indicates that random splitting may yield optimistic performance estimates when spatially adjacent or environmentally similar samples occur in both the training and testing sets. Nevertheless, TabPFN and the tree-based models retained moderate predictive ability under spatially separated validation. This result emphasizes that model performance assessed solely through random splitting should be interpreted cautiously, particularly in regions characterized by uneven sampling density and substantial environmental heterogeneity. Taken together, the repeated random holdout and spatial cross-validation results characterize model robustness to sample partitioning and spatial separation within the 2025 dataset. Validation using independent observations from additional years would further assess the temporal stability of these comparative results.

The competitive performance of TabPFN may be associated with its pretrained tabular-learning framework. Through pretraining on a diverse collection of synthetic tabular tasks, TabPFN learns generalizable modeling priors that capture complex and nonlinear relationships among variables and transfers this knowledge to new datasets through in-context learning [44,45]. This characteristic reduces its dependence on extensive task-specific training and hyperparameter optimization and may be advantageous for SOC prediction, where field observations are often limited, but the relationships between SOC and environmental covariates are complex. However, the synthetic tasks used during TabPFN pretraining were not designed to reproduce the empirical spatial covariance structure of SOC in the Liaohe Plain. Its attention mechanism therefore operated in tabular feature space rather than explicitly representing geographic distance, spatial autocorrelation, or semivariogram structure. The spatial transferability of TabPFN must consequently be evaluated independently, as conducted through repeated spatial cross-validation in this study.

The comparable performance of RF, CatBoost, and XGBoost indicates that tree-based ensemble models also remained effective for modeling nonlinear relationships between SOC and environmental covariates [46]. Their performance, however, may depend on training-sample composition, environmental representativeness, and model configuration. By contrast, the lower performance of SVM and KNN may be associated with their greater sensitivity to feature-space structure, local sample configuration, and spatial heterogeneity, despite the use of standardized inputs [47]. These findings suggest that the pretrained architecture of TabPFN does not guarantee universal superiority over established machine-learning methods, but it provides a practical alternative for limited-sample tabular prediction without extensive task-specific optimization.

The small-sample stability experiments further demonstrated that training-sample size and composition influenced SOC prediction performance. As the proportion of training-candidate samples increased from S1 to S4, all four models generally showed higher $R^2$ and RPD values and lower RMSE values, while MAE also declined overall despite minor local fluctuations. For TabPFN, the mean $R^2$ increased from $0.531 \pm 0.088$ under S1 to $0.608 \pm 0.038$ under S4, whereas the RMSE decreased from $4.365 \pm 0.401$ to $4.026 \pm 0.184$ g kg$^{-1}$. Under S1, TabPFN achieved a higher $R^2$ and a lower RMSE than RF, XGBoost, and CatBoost, suggesting that its relative advantage was more evident when the available training data were substantially reduced. Under S4, however, TabPFN and RF showed nearly identical performance, while CatBoost achieved the lowest MAE. Thus, within the present experiments, TabPFN was particularly competitive under the most restricted training-sample condition, whereas its relative advantage over the tree-based models decreased as the number of training samples increased. This observation is consistent with the established suitability of TabPFN for small tabular datasets [24,25] and its successful application to other small-data regression problems [44,45], although the specific trend observed here may depend on the dataset and experimental design.

The improvement in TabPFN performance with increasing sample size also demonstrates that pretrained tabular models cannot replace high-quality and environmentally representative field observations. Insufficient sampling density or incomplete coverage of environmental gradients may still increase prediction error and restrict spatial generalization [48]. Future sampling should therefore prioritize environmentally heterogeneous areas, transition zones, sparsely sampled regions, and areas characterized by high sampling-induced relative uncertainty. Such improvements would enhance model transferability and strengthen the reliability of regional SOC mapping.

*4.2. Spatial Differentiation and Interpretation of Environmental Factors*

The SOC spatial prediction results showed that cropland SOC in the Liaohe Plain generally exhibited a spatial pattern transitioning from low values in the west to high values in the northeast (Figure 4), indicating pronounced spatial heterogeneity in SOC distribution across the study area. This pattern may be related to the combined effects of regional hydrothermal conditions, soil types, topographic background, and cropland utilization patterns. Previous studies have shown that climatic conditions, soil moisture, vegetation productivity, and agricultural management activities jointly affect SOC accumulation and decomposition processes [49], which is generally consistent with the findings of this study.

The higher SOC content in the northeastern part of the study area may be associated with more favorable moisture conditions and greater organic matter input. Suitable hydrothermal conditions are conducive to crop growth and can promote SOC accumulation by increasing crop residue and root returns [50]. In contrast, the lower SOC content in the western and northwestern regions may be related to insufficient precipitation, stronger evaporation, higher wind erosion risk, and weaker soil water retention capacity. These factors may limit biomass input and increase the risk of SOC loss [51]. Therefore, the regional hydrothermal gradient may be an important environmental background explaining the spatial differentiation of cropland SOC in the Liaohe Plain.

The SHAP analysis revealed clear differences in the contributions of the retained environmental covariates to TabPFN predictions (Figures 2 and 3). Climatic variables made the largest overall contribution, accounting for 55.4% of the total SHAP importance. MAP and MAT were the dominant predictors within this group, whereas MSR contributed comparatively less, indicating that regional hydrothermal conditions were the primary sources of predictive information for SOC estimation [52]. Soil spectral variables ranked second, accounting for 24.5% of the total contribution, mainly because of the importance of LST and BI. These variables may capture spatial differences in surface thermal conditions, soil brightness, moisture status, and related spectral characteristics during the low-vegetation image-acquisition period. Topographic variables collectively accounted for 16.8%, with elevation, ProfCurv, and aspect making the largest contributions within this group. Although the Liaohe Plain is generally characterized by low relief, local topographic variation may still influence SOC distribution by regulating water redistribution, erosion and deposition, and material transport [53]. Vegetation variables contributed only 3.3%, with EVI being the only vegetation index retained in the final predictor set. Its relatively limited contribution may be attributable to the April–May image-acquisition period, when crop cover was generally low and vegetation-related spatial differences were less pronounced.

*4.3. Sampling-Induced Relative Uncertainty Assessment*

The SOC spatial prediction map represents the spatial distribution of predicted SOC content but does not indicate the sensitivity of pixel-level predictions to changes in training-sample composition. In this study, a sampling-induced relative uncertainty index (UNC) was calculated from repeated random sampling results to identify areas with relatively high prediction variability (Figure 5). Areas with high UNC values were mainly distributed in the northern and northeastern parts of the study area and in some marginal regions, indicating that SOC predictions in these areas were more sensitive to training-sample perturbations. When considered together with the distribution of sampling sites, some high-UNC areas corresponded to regions with insufficient sample coverage or weak marginal representation, suggesting that sampling layout can affect model stability under limited-sample conditions.

High-UNC areas were associated not only with limited sample coverage and local environmental heterogeneity but also with environmental dissimilarity. Environmentally dissimilar pixels generally exhibited higher UNC values, and environmental dissimilarity was positively associated with UNC, suggesting that covariate shift may partly explain the greater prediction sensitivity observed in some northern, northeastern, and marginal regions. However, the incomplete spatial correspondence indicates that covariate shift was not the sole explanation and that variability in the learned environment–SOC relationships may also have contributed. In particular, some high-SOC areas in the northeastern part of the study area also showed relatively high UNC values. This indicates that although SOC content was predicted to be high in these areas, model predictions still varied across repeated sampling runs, reflecting sampling-induced variability in learning local environment–SOC relationships. Previous studies have shown that sparsely sampled areas, environmental-gradient transition zones, and marginal regions often exhibit higher uncertainty in soil property prediction [54], which is consistent with the results of this study. Combining the SOC prediction map with the sampling-induced relative uncertainty map can help distinguish predicted SOC levels from prediction reliability. This provides useful information for identifying priority areas for supplementary sampling, field verification, and dynamic monitoring, thereby supporting more targeted SOC mapping and regional soil carbon management [55].

### *4.4. Limitations and Future Research*

Several limitations should be acknowledged. Because the model comparison was based on soil observations and year-specific climate and remote-sensing covariates from 2025, the consistency of the comparative findings across years remains to be evaluated. The sampling-induced relative uncertainty assessment captures only prediction variability arising from changes in training-sample composition and does not represent total prediction uncertainty. Other unquantified sources include remote-sensing error, smooth interpolated climate surfaces with limited fine-scale representation, model structural uncertainty, and SOC measurement variability. In particular, variation in carbon recovery associated with the potassium dichromate oxidation–external heating method may have contributed to model residuals and may limit transferability to datasets generated using alternative analytical protocols. Nevertheless, all samples were analyzed using the same procedure, ensuring internal consistency within this study. Accordingly, the uncertainty map should be interpreted as an indicator of sampling-induced variability rather than as a complete uncertainty estimate. Because the empirical 5th–95th percentile range was not evaluated for coverage or calibration, it should not be regarded as a formal or calibrated prediction interval. In addition, although repeated spatial cross-validation was conducted, the limited number and uneven spatial distribution of sampling sites may still constrain spatial generalization, particularly in environmentally heterogeneous regions. The environmental-domain assessment should likewise be interpreted as a diagnostic of potential extrapolation risk rather than as a formal guarantee of model applicability.

Future studies should incorporate independent soil observations and temporally matched environmental covariates from additional years, quantify additional sources of uncertainty, improve sampling strategies in high-uncertainty and environmentally dissimilar regions, and integrate more detailed climate information, standardized SOC analytical protocols, and multi-year thermal indicators. Spatially adapted conformal prediction could be explored to construct calibrated prediction intervals, while Bayesian and quantile-regression approaches may provide complementary uncertainty estimates and further improve model robustness and transferability. SHAP interaction analysis should also be considered to examine the combined effects of environmental covariates.

https://doi.org/10.3390/agronomy16161562

## 5. Conclusions

This study focused on cropland in the Liaohe Plain and developed a TabPFN-based framework for SOC prediction using 310 surface soil samples and multi-source environmental covariates. The framework incorporated model comparison, repeated random holdout validation, repeated spatial cross-validation, small-sample stability analysis, spatial prediction, and sampling-induced relative uncertainty assessment. The main conclusions are as follows:

(1) TabPFN demonstrated competitive predictive performance under limited-sample conditions. Across 100 repetitions of random 80:20 holdout validation, it achieved mean $R^2$ and RMSE values of $0.608 \pm 0.038$ and $4.026 \pm 0.184$ g kg$^{-1}$, respectively. Its performance was closely comparable to that of RF and comparable to that of the other tree-based models. (2) Repeated spatial cross-validation yielded more conservative performance estimates. Under this validation strategy, TabPFN achieved mean $R^2$ and RMSE values of $0.535 \pm 0.072$ and $4.33 \pm 0.31$ g kg$^{-1}$, respectively. The reduction in predictive performance relative to random holdout validation indicates that random splitting may overestimate model performance when sampling sites are spatially clustered. (3) Model performance generally improved with increasing training-sample size. TabPFN was most competitive under S1, but under S4 its performance was comparable to RF, indicating no consistent advantage across metrics or scenarios. (4) Predicted cropland SOC content in the Liaohe Plain in 2025 ranged from 4.63 to 27.04 g kg$^{-1}$, with generally lower values in the west and higher values in the northeast. Areas with relatively high sampling-induced relative uncertainty were mainly concentrated in the northern, northeastern, and marginal regions, indicating greater sensitivity of pixel-level predictions to variation in training-sample composition.

Overall, the integration of TabPFN with repeated random holdout validation, repeated spatial cross-validation, and sampling-induced relative uncertainty assessment provides a feasible framework for spatial mapping of cropland SOC under limited-sample conditions. However, because the present findings are based on the 2025 dataset, their temporal generalizability should be confirmed using independent soil observations from additional years. Future research should increase sampling density, optimize sampling design, and quantify additional sources of uncertainty to improve the reliability of SOC spatial prediction.

**Supplementary Materials:** The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/agronomy16161562/s1. Table S1: Initial and final variance inflation factor values and sequential screening of the environmental covariates; Table S2: Sample-level values of the environmental covariates considered in this study.

**Author Contributions:** Conceptualization, Y.Y. and S.G.; methodology, Y.Y.; software, Y.Y.; validation, Y.Y. and Y.Z.; formal analysis, Y.Y.; investigation, Y.Y. and Y.Z.; resources, S.G. and X.C.; data curation, Y.Y.; writing—original draft preparation, Y.Y.; writing—review and editing, Y.Z., S.G., X.C., and N.C.; visualization, Y.Y.; supervision, S.G. and X.C.; project administration, Y.Z.; funding acquisition, Y.Z. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the Ministry of Science and Technology of the People's Republic of China through the National Key R&D Program of China (Grant No. 2025YFE0124700).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author due to restrictions.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References


1. Abrar, M.M.; Waqas, M.A.; Mehmood, K.; Fan, R.; Memon, M.S.; Khan, M.A.; Siddique, N.; Xu, M.; Du, J. Organic carbon sequestration in global croplands: Evidenced through a bibliometric approach. *Front. Environ. Sci.* **2025**, *13*, 1495991. [CrossRef]
2. Lal, R. Managing soils for negative feedback to climate change and positive impact on food and nutritional security. *Soil Sci. Plant Nutr.* **2020**, *66*, 1–9. [CrossRef]
3. Lal, R. Digging deeper: A holistic perspective of factors affecting soil organic carbon sequestration in agroecosystems. *Glob. Change Biol.* **2018**, *24*, 3285–3301. [CrossRef] [PubMed]
4. Meng, X.; Bao, Y.; Luo, C.; Zhang, X.; Liu, H. A new methodology for establishing an SOC content prediction model that is spatiotemporally transferable at multidecadal and intercontinental scales. *ISPRS J. Photogramm. Remote Sens.* **2024**, *218*, 531–550. [CrossRef]
5. Li, Y.; Zheng, S.; Wang, L.; Dai, X.; Zang, D.; Qi, B.; Meng, X.; Mei, X.; Luo, C.; Liu, H. Systematic identification of factors influencing the spatial distribution of soil organic matter in croplands within the black soil region of Northeastern China across multiple scales. *Catena* **2025**, *249*, 108633. [CrossRef]
6. Xu, Y.D.; Pei, J.B.; Li, S.Y.; Zou, H.T.; Wang, J.K.; Zhang, J.B. Main characteristics and protection and utilization countermeasures of different types of black soil areas in Northeast China. *Chin. J. Soil Sci.* **2023**, *54*, 495–504.
7. Hengl, T.; Heuvelink, G.B.M.; Stein, A. A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma* **2004**, *120*, 75–93. [CrossRef]
8. McBratney, A.B.; Mendonça Santos, M.L.; Minasny, B. On digital soil mapping. *Geoderma* **2003**, *117*, 3–52. [CrossRef]
9. Liu, X.; Wang, M.; Liu, Z.; Bao, Y.; Li, X.; Wang, F.; Ji, X. Improving spatial prediction of soil organic matter in typical black soil area of Northeast China using structural equation modeling integration framework. *Comput. Electron. Agric.* **2025**, *236*, 110404. [CrossRef]
10. Pouladi, N.; Gholizadeh, A.; Khosravi, V.; Borůvka, L. Digital mapping of soil organic carbon using remote sensing data: A systematic review. *Catena* **2023**, *232*, 107409. [CrossRef]
11. Radočaj, D.; Gašparović, M.; Jurišić, M. Open remote sensing data in digital soil organic carbon mapping: A review. *Agriculture* **2024**, *14*, 1005. [CrossRef]
12. Gopp, N.V.; Meshalkina, J.L.; Narykova, A.N.; Plotnikova, A.S.; Chernova, O.V. Mapping of soil organic carbon content and stocks at the regional and local levels: The analysis of modern methodological approaches. *For. Sci. Issues* **2023**, *6*, 1–46. [CrossRef]
13. Kumar, A.; Moharana, P.C.; Jena, R.K.; Malyan, S.K.; Sharma, G.K.; Fagodiya, R.K.; Shabnam, A.A.; Jigyasu, D.K.; Kumari, K.M.V.; Doss, S.G. Digital mapping of soil organic carbon using machine learning algorithms in the upper Brahmaputra valley of northeastern India. *Land* **2023**, *12*, 1841. [CrossRef]
14. Nozari, S.; Pahlavan-Rad, M.R.; Brungard, C.; Heung, B.; Borůvka, L. Digital soil mapping using machine learning-based methods to predict soil organic carbon in two different districts in the Czech Republic. *Soil Water Res.* **2024**, *19*, 32–49. [CrossRef]
15. Lima, A.A.J.; Lopes, J.C.; Lopes, R.P.; de Figueiredo, T.; Vidal-Vázquez, E.; Hernández, Z. Soil Organic Carbon Assessment Using Remote-Sensing Data and Machine Learning: A Systematic Literature Review. *Remote Sens.* **2025**, *17*, 882. [CrossRef]
16. Yan, K.; Wang, D.; Feng, Y.; Hou, S.; Zhang, Y.; Yang, H. Digital mapping of soil organic carbon in a plain area based on time-series features. *Ecol. Indic.* **2025**, *171*, 113215. [CrossRef]
17. Saurette, D.D.; Heck, R.J.; Gillespie, A.W.; Berg, A.A.; Biswas, A. Sample size optimization for digital soil mapping: An empirical example. *Land* **2024**, *13*, 365. [CrossRef]
18. Wang, J.; Zhang, Z.; Wang, Y.; Qin, C.-Z.; Chen, X.; Zhang, Y.; Hu, Z. How to solve small sample size problems in time-series soil organic carbon mapping: New insights from the Third Law of Geography. *Geoderma* **2025**, *460*, 117402. [CrossRef]
19. Li, T.; Cui, L.; Wu, Y.; McLaren, T.I.; Xia, A.; Pandey, R.; Liu, H.; Wang, W.; Xu, Z.; Song, X.; et al. Soil organic carbon estimation via remote sensing and machine learning techniques: Global topic modeling and research trend exploration. *Remote Sens.* **2024**, *16*, 3168. [CrossRef]
20. Ma, T.; Wei, T.; Qin, C.Z.; Zhu, A.-X.; Qi, F.; Liu, J.; Zhao, F.; Pan, H. In-situ recommendation of alternative soil samples during field sampling based on environmental similarity. *Earth Sci. Inform.* **2020**, *13*, 39–53. [CrossRef]
21. Cui, H.; Chang, X.; Gang, S. Optimizing Data Preprocessing and Hyperparameter Tuning for Soil Organic Carbon Content Prediction Using Large Language Models: A Case Study of the Black Soil and Windblown Sandy Soil Regions in Northeast China. *Appl. Sci.* **2026**, *16*, 3349. [CrossRef]
22. Zhang, W.; Wan, H.; Zhou, M.; Wu, W.; Liu, H. Soil total and organic carbon mapping and uncertainty analysis using machine learning techniques. *Ecol. Indic.* **2022**, *143*, 109420. [CrossRef]
23. Kakhani, N.; Alamdar, S.; Kebonye, N.M.; Amani, M.; Scholten, T. Uncertainty quantification of soil organic carbon estimation from remote sensing data with conformal prediction. *Remote Sens.* **2024**, *16*, 438. [CrossRef]
24. Hollmann, N.; Müller, S.; Eggensperger, K.; Hutter, F. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv* **2022**, arXiv:2207.01848.

25. Hollmann, N.; Müller, S.; Purucker, L.; Krishnakumar, A.; Körfer, M.; Bin Hoo, S.; Schirrmeister, R.T.; Hutter, F. Accurate predictions on small data with a tabular foundation model. *Nature* **2025**, *637*, 319–326. [CrossRef] [PubMed]
26. Brus, D.J. Sampling for digital soil mapping: A tutorial supported by R scripts. *Geoderma* **2019**, *338*, 464–480. [CrossRef]
27. Zepp, S.; Heiden, U.; Bachmann, M.; Möller, M.; Wiesmeier, M.; van Wesemael, B. Optimized bare soil compositing for soil organic carbon prediction of topsoil croplands in Bavaria using Landsat. *ISPRS J. Photogramm. Remote Sens.* **2023**, *202*, 287–302. [CrossRef]
28. Zhu, A.X.; Liu, J.; Du, F.; Zhang, S.J.; Qin, C.Z.; Burt, J.; Behrens, T.; Scholten, T. Predictive soil mapping with limited sample data. *Eur. J. Soil Sci.* **2015**, *66*, 535–547. [CrossRef]
29. Hutchinson, M.F. *ANUSPLIN Version 4.36 User Guide*; Centre for Resource and Environmental Studies: Canberra, Australia, 2006.
30. Huete, A.; Didan, K.; Miura, T.; Rodriguez, E.P.; Gao, X.; Ferreira, L.G. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sens. Environ.* **2002**, *83*, 195–213. [CrossRef]
31. Mathieu, R.; Pouget, M.; Cervelle, B.; Escadafal, R. Relationships between satellite-based radiometric indices simulated using laboratory reflectance data and typic soil color of an arid environment. *Remote Sens. Environ.* **1998**, *66*, 17–28. [CrossRef]
32. Chen, Q.; Vaudour, E.; Richer-de-Forges, A.C.; Arrouays, D. Spectral indices in remote sensing of soil: Definition, popularity, and issues. A critical overview. *Remote Sens. Environ.* **2025**, *329*, 114918. [CrossRef]
33. Pei, T.; Qin, C.Z.; Zhu, A.X.; Yang, L.; Luo, M.; Li, B.; Zhou, C. Mapping soil organic matter using the topographic wetness index: A comparative study based on different flow-direction algorithms and kriging methods. *Ecol. Indic.* **2010**, *10*, 610–619. [CrossRef]
34. Emadi, M.; Taghizadeh-Mehrjardi, R.; Cherati, A.; Danesh, M.; Mosavi, A.; Scholten, T. Predicting and mapping of soil organic carbon using machine learning algorithms in Northern Iran. *Remote Sens.* **2020**, *12*, 2234. [CrossRef]
35. Ben Ghorbal, A.; Grine, A.; Eid, M.M.; El-Kenawy, E.-S.M. Sustainable soil organic carbon prediction using machine learning and the ninja optimization algorithm. *Front. Environ. Sci.* **2025**, *13*, 1630762. [CrossRef]
36. Radočaj, D.; Jug, D.; Jug, I.; Jurišić, M. A comprehensive evaluation of machine learning algorithms for digital soil organic carbon mapping on a national scale. *Appl. Sci.* **2024**, *14*, 9990. [CrossRef]
37. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
38. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222. [CrossRef]
39. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: Unbiased boosting with categorical features. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 6639–6649.
40. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
41. Altman, N.S. An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* **1992**, *46*, 175–185. [CrossRef]
42. Wadoux, A.M.J.C.; Minasny, B.; McBratney, A.B. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. *Earth-Sci. Rev.* **2020**, *210*, 103359. [CrossRef]
43. Chang, C.W.; Laird, D.A.; Mausbach, M.J.; Hurburgh, C.R., Jr. Near-infrared reflectance spectroscopy–principal components regression analyses of soil properties. *Soil Sci. Soc. Am. J.* **2001**, *65*, 480–490. [CrossRef]
44. Wang, H.; Pei, J.; Yan, F.; Wang, W.; Yuan, S.; Wang, K.; Fu, W.; Gan, X.; Chen, J. Application of TabPFN model on the energy performance improvement of high-power multistage centrifugal pump. *Energy* **2025**, *336*, 138399. [CrossRef]
45. Xin, Z.; Zhang, J.; Lan, M.; Liu, Q. A novel small data prediction model of activity for multiple slag systems based on TabPFN and SHAP. *ISIJ Int.* **2025**, *65*, 2275–2280. [CrossRef]
46. Jiang, J.P.; Liu, S.Y.; Cai, H.R.; Zhou, Q.L.; Ye, H.J. Representation learning for tabular data: A comprehensive survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2026**, *48*, 6488–6508. [CrossRef] [PubMed]
47. Triantakonstantis, D.; Karakostas, A. Soil organic carbon monitoring and modelling via machine learning methods using soil and remote sensing data. *Agriculture* **2025**, *15*, 910. [CrossRef]
48. Zhu, A.; Liu, J.; Yang, X.; Zhao, F.; Brungard, C. Quantification of prediction uncertainty due to sample representation for digital soil mapping based on Third Law of Geography. *Pedosphere* **2026**. [CrossRef]
49. Wiesmeier, M.; Urbanski, L.; Hobley, E.; Lang, B.; von Lützow, M.; Marin-Spiotta, E.; van Wesemael, B.; Rabot, E.; Ließ, M.; Garcia-Franco, N.; et al. Soil organic carbon storage as a key function of soils—A review of drivers and indicators at various scales. *Geoderma* **2019**, *333*, 149–162. [CrossRef]
50. Liu, X.; Xu, H.; Cheng, Y.; Zhang, Y.; Li, Y.; Wang, F.; Shen, C.; Chen, B. Vegetable Productivity, Soil Physicochemical and Biochemical Properties, and Microbiome in Response to Organic Substitution in an Intensive Greenhouse Production System. *Agriculture* **2025**, *15*, 1493. [CrossRef]
51. Liu, X.; Wang, M.; Liu, Z.; Li, X.; Ji, X.; Wang, F. Spatial and temporal evolution of soil organic matter and its response to dynamic factors in the Southern part of Black Soil Region of Northeast China. *Soil Tillage Res.* **2025**, *248*, 106475. [CrossRef]
52. Dong, Y.; Wang, X.; Wang, S.; Li, B.; Liu, J.; Huang, J.; Li, X.; Zeng, Y.; Su, W. Enhancing soil organic carbon prediction by unraveling the role of crop residue coverage using interpretable machine learning. *Geoderma* **2025**, *455*, 117225. [CrossRef]
53. Cutting, B.J.; Atzberger, C.; Gholizadeh, A.; Robinson, D.A.; Mendoza-Ulloa, J.; Marti-Cardona, B. Remote quantification of soil organic carbon: Role of topography in the intra-field distribution. *Remote Sens.* **2024**, *16*, 1510. [CrossRef]

54. Lilburne, L.; Helfenstein, A.; Heuvelink, G.B.; Eger, A. Interpreting and evaluating digital soil mapping prediction uncertainty: A case study using texture from SoilGrids. *Geoderma* **2024**, *450*, 117052. [CrossRef]
55. Stumpf, F.; Schmidt, K.; Goebes, P.; Behrens, T.; Schönbrodt-Stitt, S.; Wadoux, A.; Xiang, W.; Scholten, T. Uncertainty-guided sampling to improve digital soil maps. *Catena* **2017**, *153*, 30–38. [CrossRef]