remote sensing

MDPI

Article

Exploring the Impacts of Data Source, Model Types and Spatial Scales on the Soil Organic Carbon Prediction: A Case Study in the Red Soil Hilly Region of Southern China

Qiuyuan Tan [1], Jing Geng [1,2,*], Huajun Fang [3,4], Yuna Li [5] and Yifan Guo [3]

[1] School of Geospatial Engineering and Science, Sun Yat-sen University, Zhuhai 519082, China
[2] Key Laboratory of Natural Resources Monitoring in Tropical and Subtropical Area of South China, Ministry of Natural Resources, Zhuhai 519082, China
[3] Key Laboratory of Ecosystem Network Observation and Modeling, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
[4] The Zhongke-Ji'an Institute for Eco-Environmental Sciences, Ji'an 343000, China
[5] College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: gengj9@mail.sysu.edu.cn

Abstract: Rapid and accurate mapping of soil organic carbon (SOC) is of great significance to understanding the spatial patterns of soil fertility and conducting soil carbon cycle research. Previous studies have dedicated considerable efforts to the spatial prediction of SOC content, but few have systematically quantified the effects of environmental covariates selection, the spatial scales and the model types on SOC prediction accuracy. Here, we spatially predicted SOC content through digital soil mapping (DSM) based on 186 topsoil (0–20 cm) samples in a typical hilly red soil region of southern China. Specifically, we first determined an optimal covariate set from different combinations of multiple environmental variables, including multi-sensor remote sensing images (Sentinel-1 and Sentinel-2), climate variables and DEM derivatives. Furthermore, we evaluated the impacts of spatial resolution (10 m, 30 m, 90 m, 250 m and 1000 m) of covariates and the model types (three linear and three non-linear machine learning techniques) on the SOC prediction. The results of the performance analysis showed that a combination of Sentinel-1/2-derived variables, climate and topographic predictors generated the best predictive performance. Among all variables, remote sensing covariates, especially Sentinel-2-derived predictors, were identified as the most important explanatory variables controlling the variability of SOC content. Moreover, the prediction accuracy declined significantly with the increased spatial scales and achieved the highest using the XGBoost model at 10 m resolution. Notably, non-linear machine learners yielded superior predictive capability in contrast with linear models in predicting SOC. Overall, our findings revealed that the optimal combination of predictor variables, spatial resolution and modeling techniques could considerably improve the prediction accuracy of the SOC content. Particularly, freely accessible Sentinel series satellites showed great potential in high-resolution digital mapping of soil properties.

Keywords: soil organic carbon; digital soil mapping; Sentinel; covariates selection; model comparison; resolution

Citation: Tan, Q.; Geng, J.; Fang, H.; Li, Y.; Guo, Y. Exploring the Impacts of Data Source, Model Types and Spatial Scales on the Soil Organic Carbon Prediction: A Case Study in the Red Soil Hilly Region of Southern China. *Remote Sens.* **2022**, *14*, 5151. https://doi.org/10.3390/rs14205151

Academic Editor: Peng Fu

Received: 31 August 2022
Accepted: 13 October 2022
Published: 15 October 2022

1. Introduction

Soil organic carbon (SOC) is the largest carbon reservoir in the terrestrial ecosystem [1], which plays a vital role in monitoring the active C-exchange in the global carbon cycle [2,3]. Furthermore, SOC is a key indicator of soil fertility for vegetation and crop growth [4]. Consequently, the precise knowledge concerning the contents and spatial distribution of SOC contents in cropland soils is of great significance to, e.g., soil quality, ecological balance and food security in agricultural ecosystems [5,6].

Traditionally, soil information is obtained based on field soil sampling and laboratory chemical analysis, which is costly and time-consuming, especially at global, national or

regional scales [7,8]. With the rapid advancement in remote sensing technology during the recent decades, the use of satellite imagery in digital soil mapping (DSM) has provided the feasibility to spatially extend point soil observations to a larger scale prediction [6,9,10]. On the basis that soil molecules can absorb or reflect light at certain spectral bands, the relationships between SOC content and spectral reflectance ranging from the visible to shortwave-infrared have been widely delineated [11]. In contrast to optical sensors, synthetic aperture radar (SAR) offers unique capabilities that have the advantages of all-weather and all-day data acquisition and can provide enriched scattering information of ground objects [12]. However, the potential of radar remote sensing data in improving knowledge of regional SOC mapping has not been fully explored, especially in southern China, where optical images are frequently contaminated by cloud cover and suffer from unavailability due to rainy weather [13,14].

Apart from the application of multi-source remote sensing data in DSM, many efforts have been made in sifting and developing effective environmental covariates according to the targeted soil properties and landscapes [15,16]. Generally, in pedology, soil–landscape is characterized by soil-forming factors of parent material, climate, biology, topography and time [17,18] and has high spatial variability [19]. Scholars concentrated on how these soil-forming factors affected soil properties, especially SOC. For example, Wang et al. [20] proposed that climate factors are highly related to soil moisture and have a profound effect on the decomposition and accumulation of SOC by affecting plant growth and net primary productivity. Likewise, Martin et al. [21] considered that topographic factors determine the vertical distribution of water heat and influence the decomposition and transformation of SOC. Overall, it is critical to identify key environmental variables affecting SOC distribution, which are involved as model inputs for SOC prediction, but how to discern the optimal combination of environmental covariates to improve the modeling accuracy remains largely unclear.

Moreover, a large number of case studies demonstrated the importance of multi-scale landscape characteristics and remote sensing sensors for predicting soil variations [22–26]. For example, Behrens et al. [17] and Miller et al. [27] observed that mapping in soil–landscape modeling using multiple spatial scales of predictors improved prediction accuracies, with RMSE decreasing from 16.1% to 11.2% and adjusted R^2 increasing from negligible to 70%, respectively. Previous studies reasonably assumed that covariates such as terrain attributes and remote sensing sensors derived at finer spatial resolutions would perform better in SOC mapping, but several studies have shown the contrary [28–30]. For instance, Zhou et al. [28] combined Landsat-8, Sentinel-2 and Sentinel-3 with different spatial resolutions to map the national distribution of SOC content in Switzerland. The result revealed that the best predictions for SOC content were achieved by all available predictors at a resolution of 100 m rather than 20 m. Until now, there is still no consensus on which resolution of the variables has the best prediction accuracy. Whether finer-scale predictors will lead to superior predictive capability in soil properties mapping still needs to be addressed.

In addition, numerous techniques, including statistical, geostatistical, hybrid and machine learning, were widely applied in SOC prediction based on the DSM [18]. The research of Owusu et al. [31] showed that stepwise multiple linear regression (SMLR) could be used to depict the relationships between SOC contents and a set of covariates. Many similar studies also used various linear modeling approaches, such as partial least squares regression (PLSR) [32,33] and multiple linear regression (MLR) [34] in SOC prediction. Moreover, taking into account the superior capabilities of machine learning in modeling the non-linear relationship between soil and environmental factors, many scholars have applied machine learning algorithms, e.g., random forest [35], support vector machine [36], neural network [37], boosted regression tree [38], XGBoost [39] and Cubist [40] to model and map SOC contents. Overall, the quantitative comparison of the linear statistical methods and various emerging non-linear machine learning algorithms to spatially predict SOC distribution is scarce and needs to be further evaluated.

The red soil hilly region in southern China covers an area of 102 million ha and is primarily distributed in the tropical and subtropical areas of China [41]. This region is one of the most important agricultural production areas in China. Red soils are usually characterized by acidification, nutrient-deficient and low organic matter [42]. For decades, in addition to the fragile environmental conditions, irrational human exploitation has led to severe soil erosion, strong soil acidity, fertility degradation and soil contamination in the region [43]. Hence, more attention should be paid to soil properties prediction and mapping in this region, which can lay a scientific basis for reasonable agricultural management.

Therefore, this study aimed to investigate the spatial distribution of SOC content in a red soil hilly region of southern China through DSM by linear and non-linear modeling techniques. In the process, the effects of environmental covariates selection, spatial scale and model types on the SOC prediction accuracy were systematically assessed. The specific objectives of this study were to (1) identify the optimal combination of environmental covariates from a set of predictor variables, including multi-temporal Sentinel-1 SAR variables, Sentinel-2 multispectral variables, climate variables and terrain attributes; (2) investigate the variations in prediction accuracy with the changing spatial resolution (10 m, 30 m, 90 m, 250 m and 1000 m) of covariates and determine the optimal prediction scale; (3) compare the predictive performances among three linear statistical models (MLR, PLSR and SMLR) and three widely used machine-learners (RF, BRT and XGBoost); and finally, apply the best model using the optimal covariates set at the optimal spatial resolution to derive the spatial distribution of SOC content in the study area.

2. Materials and Methods

2.1. Study Area

Taihe County is located in south-central Jiangxi Province, southern China, and covers an area of 2667 km^2. The altitude ranges from 9 m to 1129 m, with an average of 174 m (Figure 1). As a typical subtropical red soil hilly area, the elevation is higher in the southeast and western regions, while the central region with low altitude belongs to the hinterland of the Jitai Basin. It is located in a typical subtropical monsoon climate, with an annual average temperature of 18.6 °C and annual precipitation of 1726 mm. Croplands are mainly distributed in the central basin and cover an area of approximately 740 km^2, accounting for 28% of the total area. Red soil, belonging to Ferralsols, is the dominant type of soil distributed widely in the region.

2.2. Soil Samples Collection and Processing

A total of 186 topsoil (0–20 cm) samples in croplands were collected by random sampling in December 2020. At each sampling point, five sub-samples were mixed into a representative composite sample. The location information was recorded by a portable GPS. After mixing and packaging, all the samples were sent to the laboratory for further processing. The plant residues and stones were first removed from soil samples and air-dried in the laboratory for half a month. Then, the dried samples were ground, homogenized and sieved through a 2 mm sieve. SOC content was analyzed by dry combustion in an elemental analyzer (Vario EL III, Elementar, Langenselbold, Germany).

2.3. Environmental Variables

Based on soil forming factors of soil–landscape (e.g., climate, biology, topography), we collected four types of available environmental variables, including Sentinel-1 SAR remote sensing variables, Sentinel-2 multispectral remote sensing variables, terrain attributes and climate variables (Table 1). These environmental covariates were unified to the WGS84 UTM Zone 50N projection.

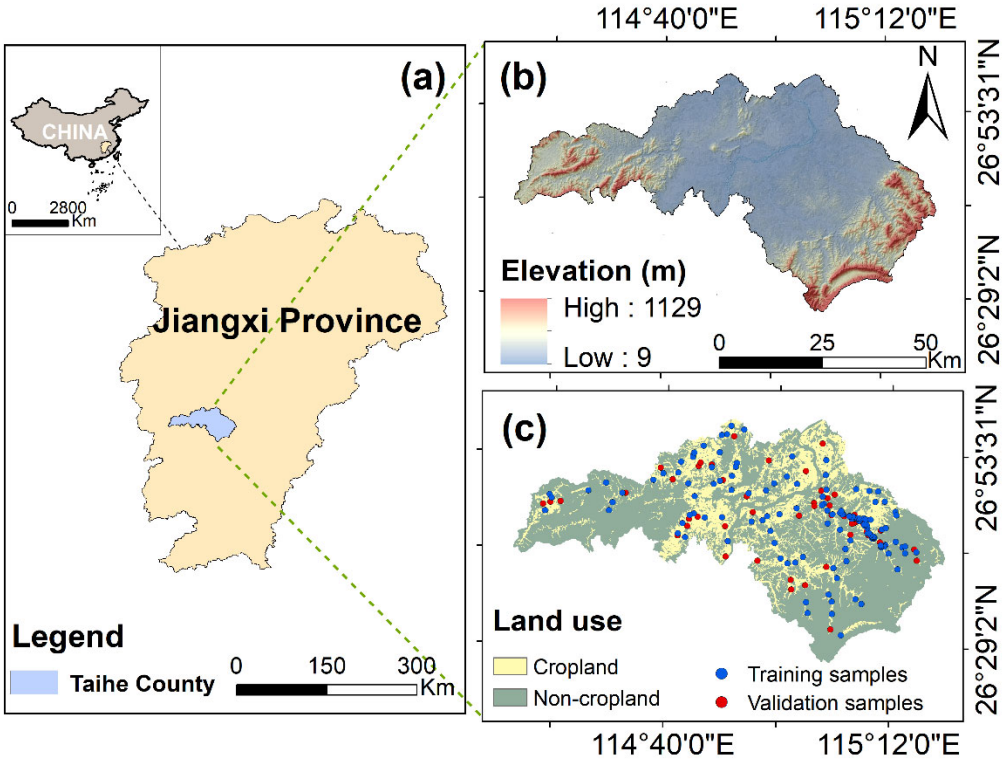

Figure 1. Study area: (**a**) location of the study area in China; (**b**) elevation distribution; (**c**) spatial distribution of cropland soil samples for model training and validation.

Table 1. Environmental variables used in this study.

Category	Variables	Description	Source	Resolution
SAR images	Multi-temporal backscattering coefficient predictors under VV/VH polarization	Processed from Sentinel-1 Level-1 Ground Range Detected (GRD) product (COPERNICUS/S1_GRD)	Google Earth Engine (GEE) platform (https://earthengine.google.com/, accessed on 14 March 2022)	10 m
Optical images	Surface reflectance predictors and vegetation indices (VIs)	Processed from Sentinel-2 Level-2A surface reflectance (SR) product (COPERNICUS/S2_SR)	GEE platform (https://earthengine.google.com/, accessed on 9 March 2022)	B2–B4, B8, VIs: 10 m B5–B7, B8A, B11, B12: 20 m
Terrain variables	Terrain attributes predictors	ASTER GDEM product and DEM derivatives processed with SAGA GIS	Geospatial Data Cloud website (GDC) (https://www.gscloud.cn/, accessed on 12 March 2022)	30 m
Climate variables	Bioclimatic predictors	WorldClim version 2.1 bioclimatic data	WorldClim (https://www.worldclim.org, accessed on 10 March 2022)	1 km

In this study, we first extracted the attribute values corresponding to each soil sample for all environmental variables under their original spatial resolutions, which were used as inputs for SOC modeling. In order to explore and compare the influence of the single or joint combination of environmental variables on the variation in SOC, the four types of environmental variables were regrouped into seven covariate sets (Table 2).

The optimal combination of environmental covariates was determined through the comparison of modeling performance. Subsequently, to further explore the influence of spatial resolutions on SOC prediction, all environmental variables filtered by the above process were converted into raster layers with spatial resolutions of 10 m, 30 m, 90 m, 250 m and 1000 m using the nearest neighbor resampling algorithm [44], during which the most

suitable spatial scale for modeling was determined based on accuracy metrics. The basic description and source information of all variables were shown as follows:

Table 2. Different combinations of environmental covariates used as model inputs for SOC prediction.

No.	Covariate Set	Predictor Variables
1	Set I	Sentinel-1 SAR images
2	Set II	Sentinel-2 multispectral images
3	Set III	Sentinel-1 and Sentinel-2 predictors
4	Set IV	Terrain attributes and climate variables
5	Set V	Sentinel-1 predictors, terrain and climate variables
6	Set VI	Sentinel-2 predictors, terrain and climate variables
7	Set VII	Sentinel-1/2-derived predictors, terrain and climate variables

2.3.1. SAR Remote Sensing Variables

Synthetic aperture radar (SAR) images were obtained from the C-band Sentinel-1 Level-1 Ground Range Detected (GRD) product (COPERNICUS/S1_GRD) on the Google Earth Engine (GEE) platform (https://earthengine.google.com/, accessed on 14 March 2022). The mode of this product is interferometric wide swath (IW), and the spatial resolution is 10 m. Thermal noise removal, radiometric calibration and terrain correction using SRTM 30 DEM were applied to preprocess the original images. The final terrain-corrected values of SAR data were converted to a decibels (dB) scale backscatter coefficient via log scaling ($10 \times \log10(x)$). The backscatter coefficient value under VV (vertical-vertical) and VH (vertical-horizontal) polarization of multi-temporal Sentinel-1 images were considered as the SAR variables in this study (Table 3).

Table 3. The basic information of the multi-temporal Sentinel-1 data used in this study.

Date	Imaging Mode	Polarization	Abbreviation
15 March 2020	IW	VV	S1_VV1
		VH	S1_VH1
8 April 2020	IW	VV	S1_VV2
		VH	S1_VH2
10 November 2020	IW	VV	S1_VV3
		VH	S1_VH3
22 November 2020	IW	VV	S1_VV4
		VH	S1_VH4
16 December 2020	IW	VV	S1_VV5
		VH	S1_VH5

2.3.2. Optical Remote Sensing Variables

Optical images were acquired from the Sentinel-2 Level-2A surface reflectance (SR) product (COPERNICUS/S2_SR) derived from the GEE cloud platform. The Sentinel-2 SR product was preprocessed with atmospheric correction and orthorectification and contained 12 spectral bands, including 10 m resolution visible (red, green, blue) and near-infrared bands, 20 m resolution red edge and short-wave infrared bands, and 60 m resolution aerosols and water vapor bands. The Sentinel-2 SR images from 3 March 2020 were collected with a cloud cover of less than 10%, and ten bands (B2, B3, B4, B5, B6, B7, B8, B8A, B11, B12) were selected as the optical predictor variables in this research. Additionally, previous studies showed that remote-sensing vegetation indices generated by optical bands were effective in predicting SOC [14,45]. Therefore, vegetation indices, including enhanced vegetation index (EVI), modified soil adjustment vegetation index (MSAVI) and normalized difference vegetation index (NDVI), were also calculated as optical predictors using Sentinel-2 bands. The calculation formula of EVI, MSAVI and NDVI are as follows:

$$\mathrm{EVI} = 2.5 \times \frac{\mathrm{NIR} - \mathrm{R}}{\mathrm{NIR} + 6\,\mathrm{R} - 7.5\,\mathrm{B} + 1} \quad (1)$$

$$\mathrm{MSAVI} = \frac{2\,\mathrm{NIR} + 1 - \sqrt{(2\,\mathrm{NIR} + 1)^2 - 8\,(\mathrm{NIR} - \mathrm{R})}}{2} \quad (2)$$

$$\mathrm{NDVI} = \frac{\mathrm{NIR} - \mathrm{R}}{\mathrm{NIR} + \mathrm{R}} \quad (3)$$

where NIR, R and B correspond to the near-infrared band (B8), red band (B4) and blue band (B2) of Sentinel-2 images, respectively.

2.3.3. Terrain Variables

Terrain variables were composed of elevation, slope, aspect, plain curvature (PLC), profile curvature (PRC), topographic wetness index (TWI), terrain ruggedness (TRI), vertical distance to channel network (VDCN), channel network base level (CNBL), valley depth (VD), relative slope position (RSP), multiresolution index of valley bottom flatness (MRVBF) and multiresolution index of ridge top flatness (MRRTF). Elevation was directly derived from the ASTER GDEM product at 30 m resolution, which was provided by the Geospatial Data Cloud website (GDC) (https://www.gscloud.cn/, accessed on 12 March 2022). The other terrain variables were calculated from DEM with SAGA GIS (http://saga-gis.org/, accessed on 12 March 2022).

2.3.4. Climate Variables

Climate variables were downloaded from WorldClim (https://www.worldclim.org, accessed on 10 March 2022). In this study, we used the WorldClim version 2.1 bioclimatic dataset, which was the average for the years 1970–2000 [44,46]. This dataset contained 19 bioclimatic variables with four spatial scales (10 min, 5 min, 2.5 min and 30 s). In order to better match other finer resolution environmental variables, we chose the 30 s grid (approximately 1 km) data covering the study area. The bioclimatic variables were derived from the monthly temperature and rainfall values to generate more biologically meaningful variables. Annual trends (e.g., mean annual temperature, annual precipitation), seasonality (e.g., annual range in temperature and precipitation) and extreme or limiting environmental factors (e.g., the temperature of the coldest and warmest month, and precipitation of the wet and dry quarters) could be presented by these climate covariates. Detailed information on the bioclimatic variables is shown in Table 4.

Table 4. The information on the bioclimatic variables.

Code	Name	Abbreviation
BIO1	Annual Mean Temperature	AMT
BIO2	Mean Diurnal Range (Mean of monthly (max temp–min temp))	MDR
BIO3	Isothermality (BIO2/BIO7) (×100)	ITM
BIO4	Temperature Seasonality (standard deviation × 100)	TS
BIO5	Max Temperature of Warmest Month	MTWM
BIO6	Min Temperature of Coldest Month	MTCM
BIO7	Temperature Annual Range (BIO5-BIO6)	TAR
BIO8	Mean Temperature of Wettest Quarter	MTWetQ
BIO9	Mean Temperature of Driest Quarter	MTDQ
BIO10	Mean Temperature of Warmest Quarter	MTWarQ
BIO11	Mean Temperature of Coldest Quarter	MTCQ
BIO12	Annual Precipitation	AP
BIO13	Precipitation of Wettest Month	PWM
BIO14	Precipitation of Driest Month	PDW
BIO15	Precipitation Seasonality (Coefficient of Variation)	PS
BIO16	Precipitation of Wettest Quarter	PWetQ
BIO17	Precipitation of Driest Quarter	PDQ
BIO18	Precipitation of Warmest Quarter	PWarQ
BIO19	Precipitation of Coldest Quarter	PCQ

2.4. Modeling Techniques

Three linear models, including multiple linear regression (MLR); partial least squares regression (PLSR); stepwise multiple linear regression (SMLR); and three non-linear machine learning models, including random forest (RF), boosted regression tree (BRT), extreme gradient boosting (XGBoost) were employed to model and map the distribution of SOC content in the study area.

2.4.1. Linear Models

As one of the most widely used linear models, multiple linear regression (MLR) can determine the functional relationship between predictive variables and target variables by least square fitting. The "stats" package of R statistical software version 4.1.3 was used to implement the MLR model in our research.

Partial least squares regression (PLSR) combines the advantages of principal component regression, canonical correlation analysis and multiple linear regression methods. By projecting predictive variables and target variables into a new space, this method constructed a linear regression model, which effectively prevents the multicollinearity issue [33]. In this study, we used the "pls" package in R to construct the PLSR model.

Stepwise multiple linear regression (SMLR) is essentially a variable selection method that aims to remove the weakly correlated variables based on forward and backward Akaike information criteria [47]. We developed the SMLR model through the "stats" package in R.

2.4.2. Non-Linear Models

Random forest (RF) is an ensemble learning classifier based on multiple decision trees. The decision variables of a single tree are randomly selected to predict the target variables, and the final result is voted by each decision tree. The bootstrap sampling strategy was used to sample the training dataset, and the out-of-bag (OOB) samples omitted from the bootstrapped samples were used to estimate error and predictor importance [48]. Given the capability of estimating the importance of variables and the high stability, random forest is generally preferred in digital soil mapping [49]. In this paper, we used the "randomForest" package of R to construct the RF model. The two main tuning parameters, the number of input variables (mtry) and the number of trees (ntree), were used to optimize the RF model.

Boosted regression tree (BRT) is an additive regression model that combines the advantages of a regression tree and boosting algorithm. By adding a new tree to the previous trees to minimize the loss function, the boosting technique can keep the model from overfitting [50]. The BRT algorithm is an iterative process in which tree-based models are fitted iteratively using recursive binary splits to identify poorly modeled observations in existing trees until minimum model deviance is reached [51]. In our research, we used the "gbm" package of R to fit the BRT model. The number of trees (n.trees), the learning rate (shrinkage) and the max depth of each tree (interaction.depth) are the primary parameters for optimizing the BRT model.

Extreme gradient boosting (XGBoost) is an effective machine learning method, which is improved on the basis of the gradient-boosted decision tree (GBDT). Through the supplemental training strategies, XGboost extends a "strong" learner from a set of "weak" learners [39]. The key to this approach is using a regularization term to control the complexity of the model and avoid overfitting [52]. The "xgboost" package of R was used to construct the XGBoost model. The learning rate (eta), max depth of trees (max_depth), the number of samples supplied to a tree (colsample_bytree) and the max number of boosting iterations (nrounds) are the main hyper-parameters for the XGBoost model to be optimized.

2.5. Evaluation of Model Performance

In this study, we developed SOC prediction models based on the aforementioned three linear statistical methods and three non-linear machine learning algorithms using seven different combinations of environmental covariates at multiple spatial resolutions, including 10 m, 30 m, 90 m, 250 m and 1000 m. Covariate Set I contained only Sentinel-1 SAR

images, while Set II included only Sentinel-2 optical images. Set III was composed of both SAR and optical images. Set IV was composed of terrain attributes and climate variables. Set V included SAR images, terrain and climate auxiliary variables. Set VI included optical images, terrain and climate auxiliary variables. Meanwhile, Set VII contained all available environmental variables (Table 2). Through this process, the optimal combination of environmental variables, modeling technique and spatial scale was determined in order to derive the SOC map of the highest quality (Figure 2). The performances of models were evaluated by the coefficient of determination (R^2), the root means square error (RMSE) and mean absolute error (MAE). The experimental data of sample points were randomly divided into 70% as the training data and the remaining 30% as the validation data [53]. The spatial distribution of soil samples for model training and validation is depicted in Figure 1c. In this study, we used independent validation data to evaluate and compare the capability of predictive models.

$$R^2 = \frac{\sum_{i=1}^{n} \left(P_i - \overline{O}_i\right)^2}{\sum_{i=1}^{n} \left(O_i - \overline{O}_i\right)^2} \tag{4}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} \left(O_i - P_i\right)^2}{n}} \tag{5}$$

$$MAE = \frac{\sum_{i=1}^{n} |O_i - P_i|}{n} \tag{6}$$

where n represents the number of samples, i represents the single sample, O_i represents the observed value of SOC, P_i represents the predicted value of SOC, and $\overline{O}_i$ represents the average value of SOC observations.

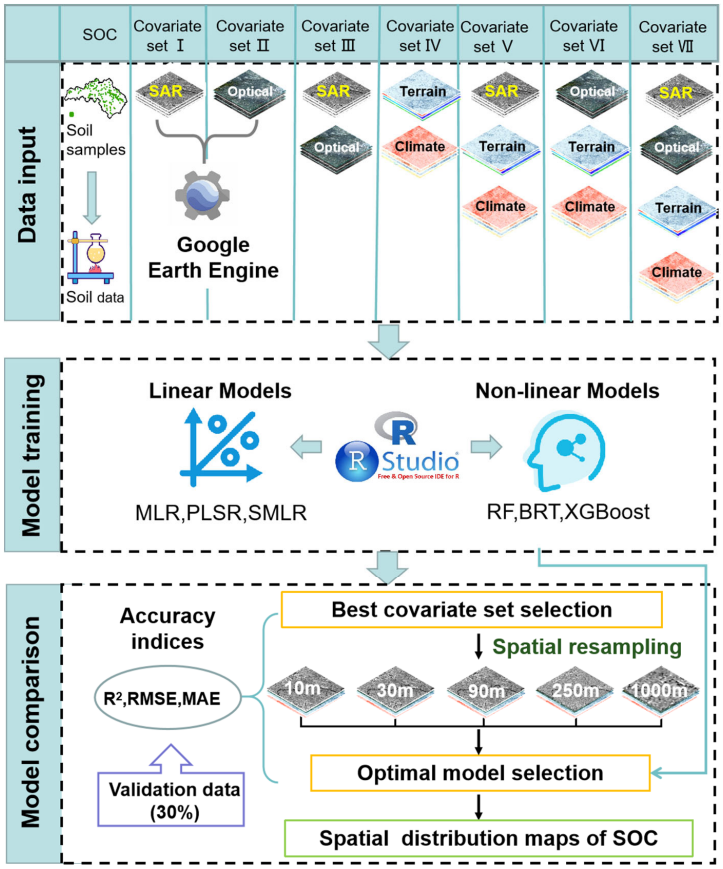

Figure 2. Workflow diagram for predicting SOC in this study.

2.6. Quantitative Spatial Analysis Technique

The standard deviational ellipse (SDE) method was introduced for the quantitative analysis of the spatial distribution pattern of predicted SOC contents. This technique was first proposed by Lefever in 1926 and was utilized to analyze the distribution characteristics of discrete data [54]. SDE was an excellent detector of the directionality and clustering of geographical features [55]. In SDE, the long axis of the ellipse represents the direction of SOC distribution, and the short axis represents the spatial clustering degree of SOC distribution (Figure 3). Generally, the difference between the long axis and short axis shows the directional strength of SDE. Meanwhile, the center of SDE indicated the center of gravity of SOC distribution. In this study, the standard deviational ellipse tool embedded in the ArcGIS 10.7 environment was used to implement SDE analysis.

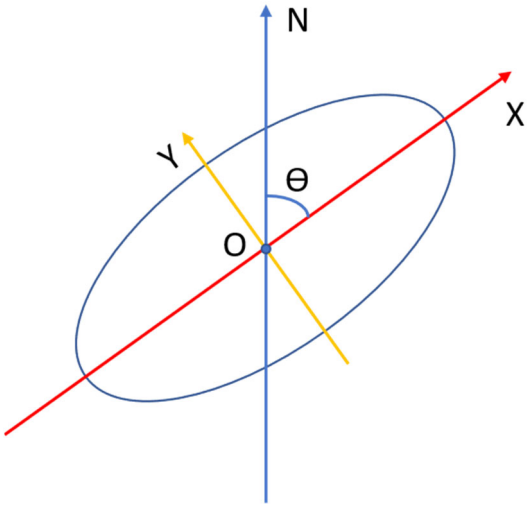

Figure 3. Standard deviational ellipse. Where X represents the direction of long axis, Y represents the direction of short axis, N represents North, O represents the center of SDE, and θ represents the rotation angle of SDE.

3. Results

3.1. Descriptive Statistics of SOC

The basic descriptive statistics of SOC are shown in Table 5. The content of the measured SOC in this study area ranged from 6.45 g/kg to 41.43 g/kg, with an average of 23.78 g/kg. The standard deviation of 5.42 g/kg and variation coefficient of 0.23 indicated a moderate variability of the total SOC observations. Similar ranges and average levels of SOC were observed in both the training and validation datasets. Moreover, the standard deviation and coefficient of variation in SOC suggested a similar distribution among the total, training and validation samples. According to the skewness value and Kolmogorov–Smirnov test, the distribution of SOC contents was confirmed to be a normal distribution.

Table 5. Basic descriptive statistics of the total, training and validation datasets for SOC.

	N	Min (g/kg)	Max (g/kg)	Mean (g/kg)	SD (g/kg)	CV	Skew	K-S
Total	186	6.45	41.43	23.78	5.42	0.23	0.07	0.20
Training	135	7.57	40.40	23.78	5.41	0.23	0.05	0.20
Validation	51	6.45	41.43	23.77	5.49	0.23	0.13	0.20

Notes: N is the number of various sample sets. Min, Max, SD, CV and Skew refer to the minimum, maximum, standard deviation, coefficient of variation and skewness, respectively. K-S represents the normal test of Kolmogorov–Smirnov at 95% significance level.

3.2. Comparison and Selection of Different Covariate Sets

Environmental covariates, including Sentinel-1 SAR images, Sentinel-2 optical images, terrain attributes and climate variables, were individually considered and assembled to evaluate the SOC prediction accuracy using six modeling techniques. Therefore, the optimal combination of environmental variables was determined for SOC prediction based on accuracy indicators. The result showed that different covariate sets as input predictors significantly affected the performance of SOC prediction (Figure 4). For instance, the covariate set II (only Sentinel-2 optical variables) generated better predictions than covariate set I (only Sentinel-1 SAR variables) under all six predictive models, indicating that multi-spectral reflectance information is relatively more important than SAR-derived backscattering coefficient predictors in SOC prediction models. Moreover, it was found that the joint application of Sentinel-1 and Sentinel-2 images (covariate set III) yielded generally higher prediction accuracy in contrast to employing a single satellite sensor, especially for the three machine learning models. For example, for the BRT model, the combination of Sentinel-1/2 increased R^2 from 0.11 in covariate set I and 0.15 in covariate set II to 0.22 in covariate set III. As we expected, the best predictive performance was obtained when Sentinel-1 images, Sentinel-2 images, terrain attributes and climate variables were all applied as model inputs. This suggests that covariate set VII was the optimal environmental covariate set for SOC prediction.

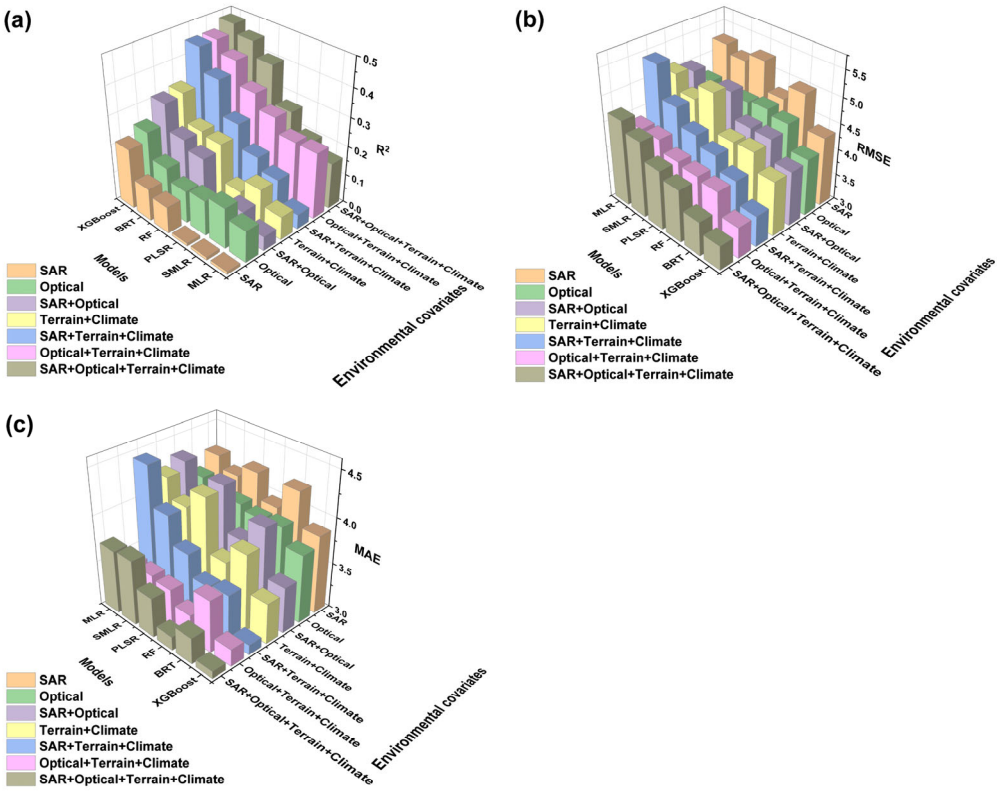

Figure 4. The R^2, RMSE and MAE among seven covariate sets in six models for SOC prediction. (**a**) Coefficient of determination (R^2), (**b**) root mean square error (RMSE) and (**c**) mean absolute error (MAE).

3.3. Assessment of Linear and Non-Linear Models Prediction at Multiple Resolutions

The performance results of covariate set VII (all available environmental variables) based on six models at five different resolutions are presented in Table 6. Meanwhile, the variation in prediction accuracy of linear (MLR, PLSR and SMLR) and non-linear (RF, BRT and XGBoost) models under five resolutions are shown in Figure 5. The comparative analysis of model performance illustrated that model types under multiple resolutions of

covariates significantly influenced the prediction accuracy of SOC contents. For instance, as shown in Table 6, the resolutions of covariates significantly controlled the model performance. Compared with coarser resolution, the finer resolution led to higher accuracy for each predictive model, and this can be described clearly by Figure 5, which depicted obvious change trends of the accuracy of modeling techniques at five spatial scales. When we focused on resolutions, all models had the best predictive accuracy at 10 m resolution, which suggested that the coarser-scale covariates had a worse explanatory ability for the variation in SOC. In addition, the significant effect of model types was also exhibited in Figure 5. More specifically, non-linear models performed better than linear models at all spatial scales. The prediction performance assessed by R^2, RMSE and MAE of validation set followed the order of XGBoost > BRT > RF > PLSR > SMLR > MLR. Furthermore, the best competitive model was XGBoost at the 10 m spatial resolution of all available predictive variables (XGBoost-10 m), and the highest accuracies performed by XGBoost-10 m were R^2 = 0.49, RMSE = 3.90 and MAE = 2.98.

Table 6. The performance of six modeling techniques at different spatial resolutions for SOC prediction using validation data. The most accurate result is shown in bold.

Model Type	Models	R^2	RMSE (g/kg)	MAE (g/kg)
Linear model	MLR			
	10 m	0.15	5.24	3.75
	30 m	0.09	5.93	4.36
	90 m	0.06	5.91	4.73
	250 m	0.04	6.42	4.85
	1000 m	0.02	6.16	4.70
	PLSR			
	10 m	0.27	4.67	3.44
	30 m	0.19	5.00	3.73
	90 m	0.15	5.02	3.75
	250 m	0.13	5.12	3.95
	1000 m	0.11	5.28	4.13
	SMLR			
	10 m	0.22	4.84	3.67
	30 m	0.15	5.22	3.97
	90 m	0.13	5.15	3.96
	250 m	0.09	5.64	4.25
	1000 m	0.06	5.55	4.38
Non-linear model	RF			
	10 m	0.39	4.56	3.28
	30 m	0.32	4.58	3.31
	90 m	0.29	4.86	3.50
	250 m	0.25	4.83	3.52
	1000 m	0.16	5.05	3.62
	BRT			
	10 m	0.42	4.28	3.21
	30 m	0.35	4.47	3.34
	90 m	0.32	4.52	3.47
	250 m	0.29	4.70	3.75
	1000 m	0.18	4.96	3.76
	XGBoost			
	10 m	**0.49**	**3.90**	**2.98**
	30 m	0.46	4.20	3.24
	90 m	0.34	4.43	3.37
	250 m	0.32	4.61	3.64
	1000 m	0.21	4.97	3.70

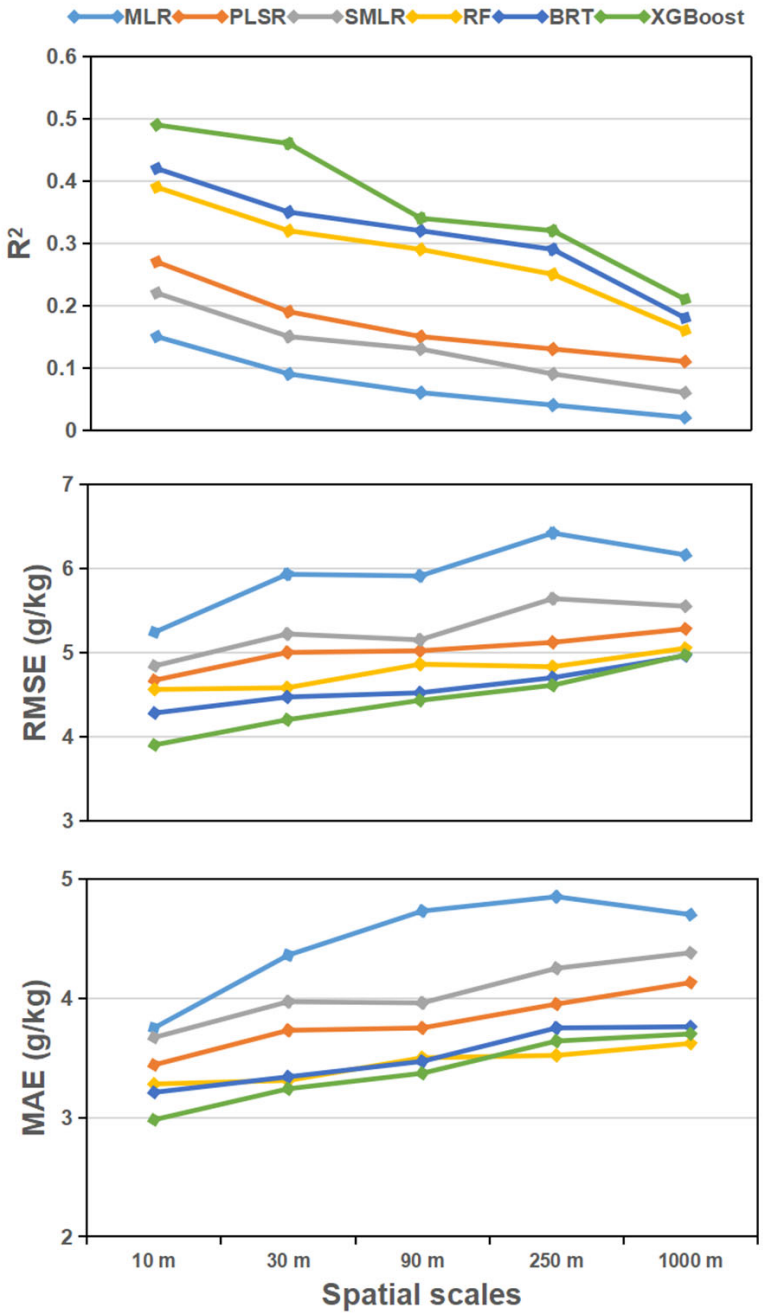

Figure 5. Accuracy comparison of models under five spatial scales.

3.4. Relative Importance of Environmental Covariates

The relative importance of adopted predictor variables was measured with the best model, namely the XGBoost at 10 m spatial scale (Figure 6). Among the top fifteen important variables, S2_B5 was the most important variable, explaining 9.50% of SOC variation. In addition, in the XGBoost model, eleven remote sensing variables (six optical + five SAR) occupied the top fifteen most important variables. Sentinel-2-derived predictors were the main explanatory variables for SOC prediction, with a relative importance of 37.86%, followed by climate (30.07%) and Sentinel-1 variables (29.40%). Notably, Sentinel-1 and Sentinel-2 variables together explained 67.26% of SOC variation, which suggested that remote sensing imagery had the most important effect on predicting SOC in the study area.

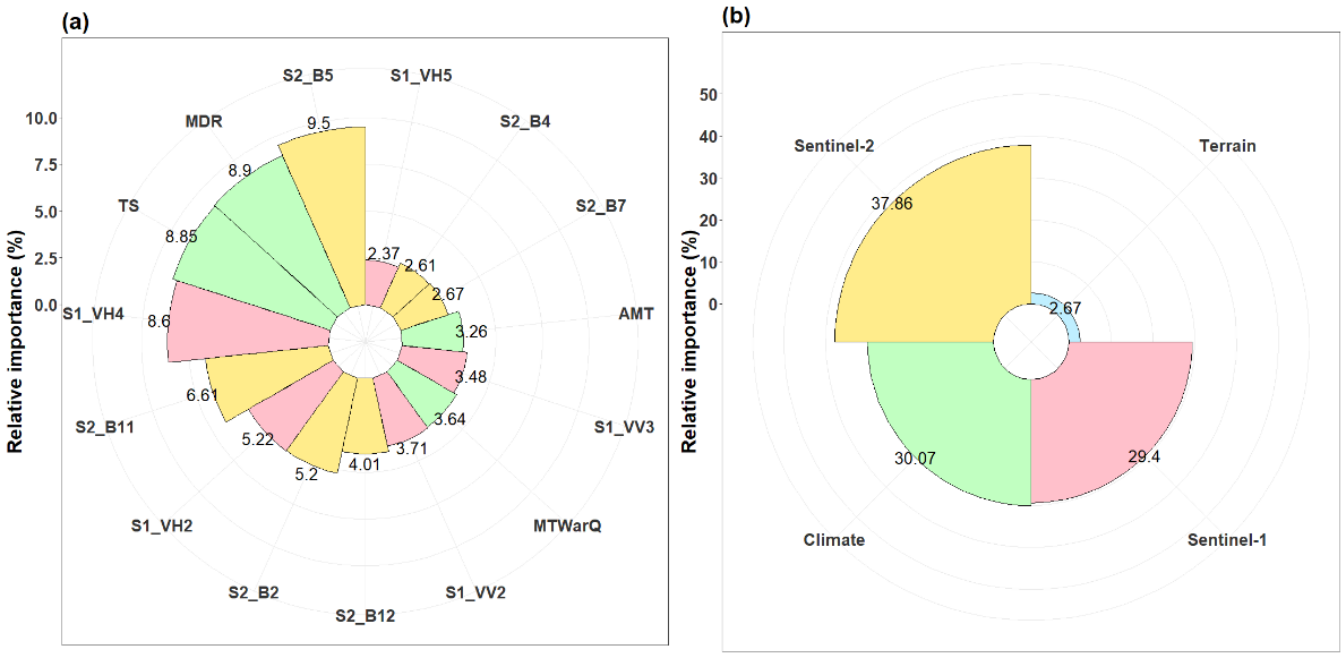

Figure 6. Relative variable importance estimation for the best model. (**a**) The relative importance of the top 15 most important variables, and (**b**) relative importance of different environmental variables groups.

Moreover, MDR and TS were the primary bioclimatic variables affecting SOC distribution. Both MDR and TS were located in the top four most important variables. For bioclimatic variables, all the variables included in the top fifteen most important variables were related to temperature parameters, which indicated that temperature variables were more important than the precipitation variables in explaining the variation in SOC. In contrast, terrain factors accounted for only 2.67% of the relative importance of SOC prediction in the XGBoost model. This indicated a weak effect of topographic variables on SOC prediction, which may be ascribed to the fact that croplands were largely distributed in the flat basin in our study area (Figure 1c).

3.5. Spatial Prediction of SOC Contents

The spatial distribution of SOC contents in croplands of the study area predicted by the XGBoost model at five spatial resolutions is displayed in Figure 7. Broadly speaking, the five prediction maps shared similar patterns and showed a strong spatial variation in SOC contents in our study area. Specifically, there was a decreasing trend of SOC contents from southeast to northwest, with the lowest content in the north-central region near the Ganjiang River. In contrast, the highest predicted value of SOC content occurred in the southeast, where were mountainous regions. This was in accordance with the distribution trends of the observed SOC samples in this study. However, there was still a discrepancy in SOC contents prediction among five maps under various spatial scales. For example, the map of SOC contents modeled by XGBoost at 10 m resolution presented a lower minimum value (4.82 g/kg), whereas the range showed on the prediction map with 1000 m resolution was comparatively larger (8.52 g/kg to 40.11 g/kg). In addition, all prediction maps also had several slight differences in the average and standard deviation values of predicted SOC content. The best accuracy and finest map details were performed in the 10 m resolution prediction map of SOC.

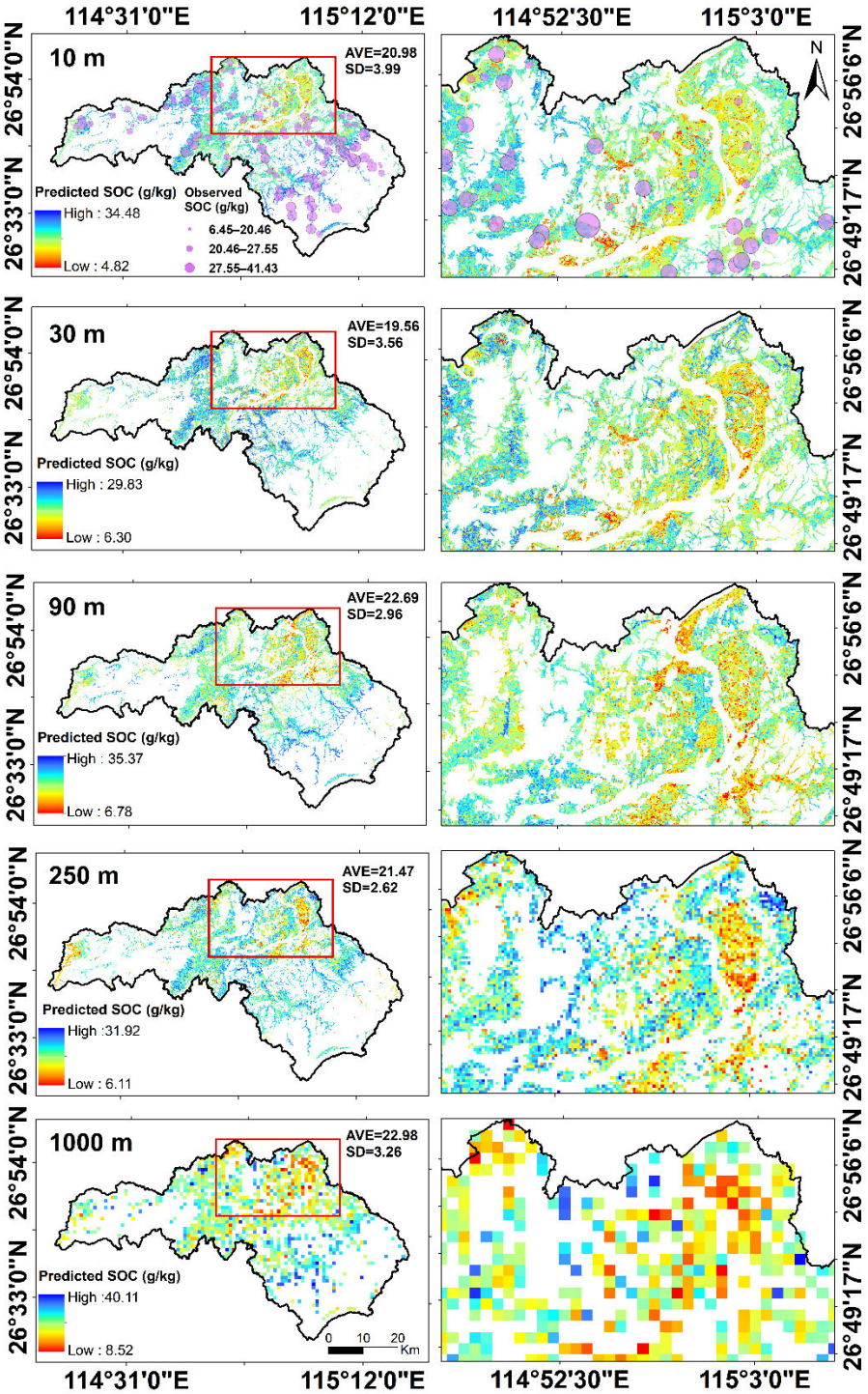

Figure 7. The distribution maps of cropland SOC contents predicted by the XGBoost technique at 10 m, 30 m, 90 m, 250 m and 1000 m resolution, respectively. The maps were masked by GlobeLand30 2020 land cover dataset to derive SOC contents of croplands in the study area. AVE and SD correspond to the average values and standard deviation values of predicted SOC, respectively.

In order to quantitatively compare the spatial distribution patterns of SOC contents predicted at the five resolutions, we divided SOC contents into three levels (low, middle and high) using the natural breaks (Jenks) method in ArcGIS 10.7 [56]. A standard deviational ellipse approach was applied to detect the direction and center of the distribution of SOC. As shown in Figure 8, the directions and centers were quite similar in general SOC content of the whole cropland region among all five spatial scales. The predicted SOC is maintained

in the southeast-to-northwest direction, which is consistent with the spatial prediction maps depicted in Figure 7. However, significant differences in SOC distribution patterns were presented at the specific lower, middle and higher levels of SOC content. For instance, in the region with low-level SOC content, the center of gravity at the 250 m spatial scale significantly deviated from others. Meanwhile, the directions of the spatial distribution of SOC maps at various content levels were different. The higher SOC content had a larger rotation angle value. On the whole, there are almost no differences in the direction and center of the distribution of general SOC at various spatial resolutions, while significant differences exist in the direction and center of the distribution of specific content levels of SOC.

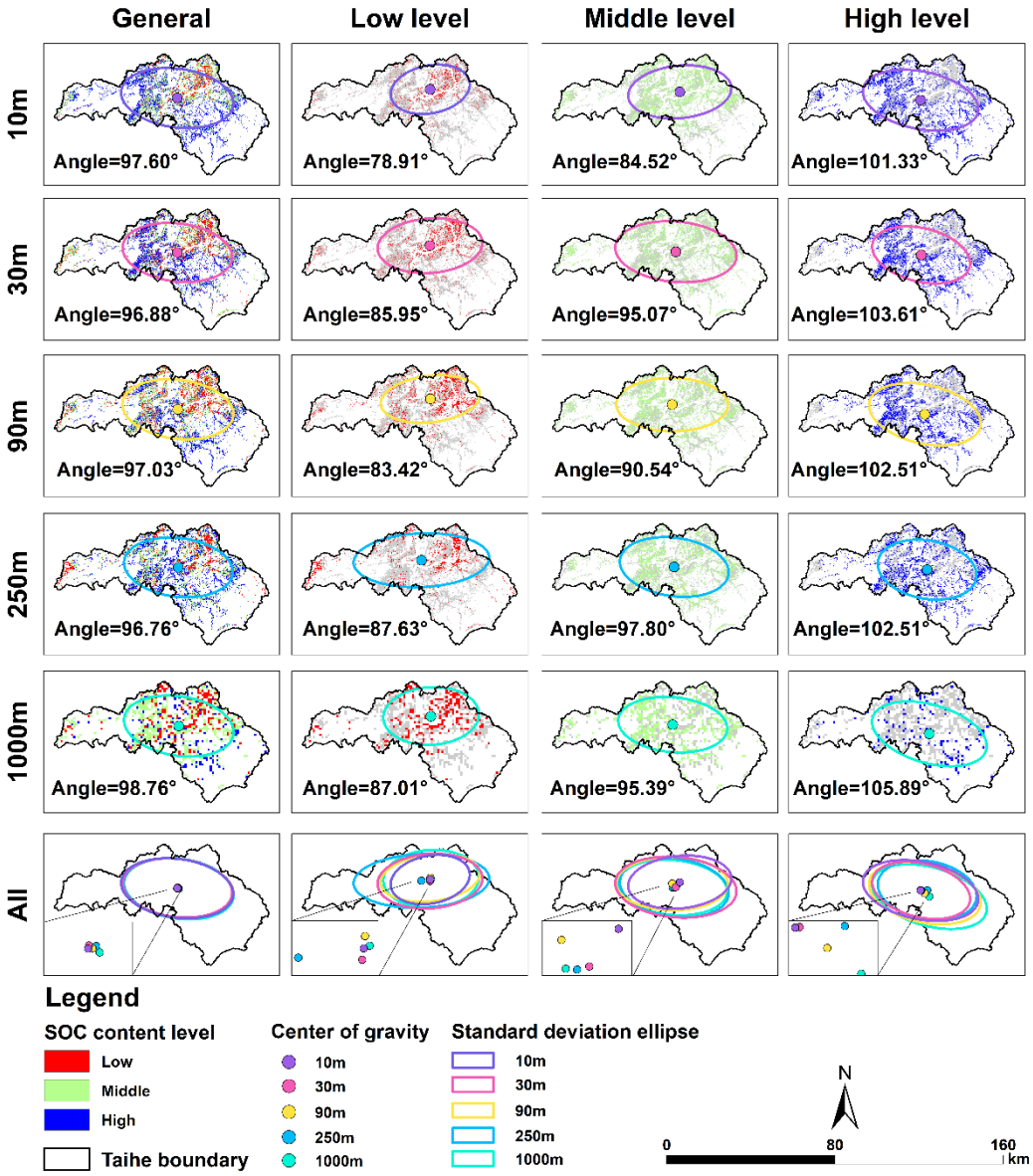

Figure 8. The standard deviational ellipses and centers of gravity of spatially predicted SOC at five different spatial resolutions for the whole, high-level, middle-level and low-level SOC content regions, respectively.

4. Discussion

4.1. Effects of Various Variable Combinations on SOC Prediction

The comparison and selection of environmental covariate sets in this study suggested that different variable combinations significantly affected the accuracies of SOC predic-

tion. The covariate set VII (SAR+ Optical+ Terrain+ Climate) had the highest accuracy in most predictive models. Therefore, the combination of SAR data, optical data, terrain and climate data were superior to either SAR or optical covariate set in SOC mapping. Similar improvements were also observed in previous studies, which demonstrated that the combination between different satellite sensors and various environmental variables could be successfully applied for effective SOC prediction [35,40]. Comprehensively considering more useful variables is of significance in improving SOC prediction [45]. Moreover, we found that set I (SAR bands only) was simulatively inferior to set II (optical bands only) in all models (Figure 4), which indicated that optical-based image data had stronger predictive power than SAR-based image data in this study. This was consistent with the study of Zhou et al. [35], who revealed that the optical bands were more important than SAR bands in the SOC prediction of the Heihe River Basin. However, this finding differs from previous studies, which for instance, proved that Sentinel-1 images were useful for SOC prediction [57]. Therefore, a competitive accuracy based on SAR data is expected to be further explored.

4.2. Comparison of Models Performance under Multi-Scales

The results showed that the overall prediction accuracy of non-linear models was generally better than that of linear models (Figure 5), indicating that SOC contents probably had non-linear relationships with covariates. This was consistent with previous studies [58,59], which confirmed that machine learning algorithms, i.e., RF, Cubist and BRT, outperformed linear models such as PLSR and SMLR. Therefore, it appears that non-linear models have sufficient potential to estimate the variability of SOC content [28,40,60]. However, several studies reported the opposing result that the PLSR model had shown better performance compared with other models [61,62]. The model's performance is usually affected by some other factors, such as the condition of the study area and the representativeness of sampling sites, and few models could be well fit for any research. Apart from the capability of different algorithms with various statistic functions, the input covariates under different conditions, i.e., the spatial resolution, also resulted in the difference in model performances for SOC prediction [63]. This was confirmed by our finding that the finer-resolution covariates had better model performances than the coarser covariates. Similar results were reported in existing studies. For example, Guo et al. [64] studied the selection of key terrain attributes with multi-scale for SOC prediction and concluded that the coarser grid sizes deteriorated the accuracy of terrain parameters and influenced the capability of soil properties prediction. Further, the approach that converted finer scales to coarser scales probably lost important information and gave rise to the worse performance of modeling techniques. However, some scholars also pointed out that the predictive models run by coarser spatial resolution variables can produce better prediction results than models based on higher spatial resolution variables [28,65]. However, their studies were mostly based on a national scale with high spatial variations, which can be strongly influenced by the scale effect [66]. By comparison, our study area has a smaller extent and lower spatial variation. Correspondingly, it is not difficult to understand that coarser resolutions are good at capturing the global characteristics of the landscape, while finer resolutions are ideal for soil properties prediction with relatively small spatial variation.

4.3. Analysis of the Relative Importance of Environmental Covariates

The relative variable importance demonstrated that the most influential factors were remote sensing data (Sentinel-2 and Sentinel-1 images) at 10 m spatial resolution. Sentinel-2 band 5 (red edge band) was especially confirmed as the most important predictor in the XGBoost model. As reported in much previous research, the red edge band (centered at 740 nm) of Sentinel-2/MSI is highly closed to the absorption feature spectrum of the N-H chemical bond, which characterizes some compounds of soil organic matter [67,68]. Similarly, Castaldi et al. [69] used a random forest algorithm to explore the relative importance of environmental covariates for SOC prediction in croplands area, and their results indi-

cated that Sentinel-2 red edge bands (band 4 and band 5) contributed the most important as well as SWIR bands (band 11 and band 12). In addition, our study also exhibited the capability and potential of Sentinel-1 data for an effective explanation of the variation in SOC. Many studies confirmed the possibility of SAR for predicting SOC through the relationship observed in soil–vegetation systems [70,71]. Therefore, both optical data and SAR data have great impacts on the prediction of SOC content. Moreover, bioclimatic variables introduced in this study also showed considerable influence on SOC prediction. Among them, temperature-dependent variables TS (temperature seasonality) and MDR (mean diurnal range) were the most important bioclimatic variables. Scholars conceived that temperature can have a positive effect on biota activities and further accelerate the decomposition of soil organic matter and the accumulation of SOC [44]. By contrast, terrain variables were found to have a weak effect on the variation in SOC in this study. However, numerous studies regarded that terrain factors greatly affected the distribution of SOC content and were the necessary variables for SOC prediction [64,72,73]. However, as the croplands of this research are mostly located in flat areas (Figure 1), the soil samples are limited to represent the overall terrain characteristics. Therefore, the effort of terrain variables for accurately mapping SOC is insufficient. This is also supported by Zhang et al. [11] and Song et al. [74], who noted that terrain attribute covariates played less important roles in SOC prediction than other environmental covariates in a flat terrain area.

4.4. Research Limitations and Future Work

Our study provided a quantification analysis framework of the effects of environmental covariates selection, spatial scales and model types on soil organic carbon prediction in croplands. Although the SOC prediction maps successfully explained the variation in SOC in our research, several limitations still need to be improved. First, as mentioned in many recent studies, soil parent materials and agronomic management factors have important influences on SOC modeling [75,76]. Thus, some environmental covariates which can represent the soil parent materials and agriculture management of SOC in cropland should be considered.

Second, in this study, in order to explore the impact of spatial scales of environmental covariates on SOC predictive performance, we transformed the spatial resolutions of environmental variables simply through the nearest neighbor resampling method. Although this method is commonly used in previous studies, the upscaling and downscaling techniques for resampling environmental covariates produced additional uncertainties. Thus, discrepancies in spatial resolution among multi-source data may generate high-uncertainty prediction results. For example, the original resolution of climate variables in this research is 1 km; however, both of the remote sensing data are much finer (10 m or 20 m), and uncertainties may occur during the resampling process especially downscaling from 1 km to 10 m. In fact, coarse-resolution climate variables cannot provide detailed information at finer scales [75]. In consideration of this issue, increasing studies have attempted to apply more advanced scale transformation algorithms. For instance, Wu et al. [77] effectively downscaled the land surface temperature (LST) using an improved multi-factor geographically weighted regression (MFGWR) algorithm. In future studies, more effective scale transformation methods should be developed and compared to assess the uncertainty of SOC prediction.

Moreover, despite the proven applicability of the current research framework for SOC prediction, the potential of the research framework proposed in this study for other soil properties prediction and mapping still needs to be evaluated in future work. The combination of multiple soil properties (e.g., soil total nitrogen, phosphorus, potassium and pH) mapped at high spatial resolution and with superior prediction accuracy is of great significance for soil health assessment and cropland resource management.

5. Conclusions

In this study, the spatial distribution of soil organic carbon was modeled and mapped through DSM in croplands of a red soil hilly region in southern China. In particular, the impacts of diverse environmental covariates selection, spatial scale and model types (linear and non-linear techniques) were comprehensively investigated. The main conclusions can be summarized as follows:

(1) For the covariate sets selection, the optimal set was the covariate set VII, which was the combination of Sentinel-1 SAR bands, Sentinel-2 optical bands, terrain attributes and climate variables. This revealed that the use of different satellite sensors and various environmental variables could be effectively applied for SOC prediction;

(2) Among all predictive models, the optimal model was the XGBoost model at 10 m resolution (R^2 = 0.49, RMSE = 3.90, MAE = 2.98). The prediction accuracy order was followed by XGBoost > BRT > RF > PLSR > SMLR > MLR, and the overall performance of non-linear machine learning models was better than the linear models;

(3) Environmental variables, especially remote sensing data, made significant contributions to SOC prediction in the XGBoost model. Our study highlighted the potential of Sentinel series satellite images, especially Sentinel-2 imagery, in SOC prediction. In addition, the spatial resolution of environmental covariates significantly affected the prediction of SOC. The finer resolution of auxiliary variables contributed to better model performance, and the best resolution was 10 m for all models;

(4) The spatial patterns of the distribution maps of SOC generated by XGBoost at various spatial scales (10 m, 30 m, 90 m, 250 m, 1000 m) were quite similar, which presented a decreasing trend of SOC contents from southeast to northwest, consistent with the distribution of observed SOC ones. However, the specific content level of SOC (low, middle, high) had significant differences in the direction and center of the spatial distribution of SOC as indicated by the standard deviational ellipse (SDE) method.

Overall, this study revealed that the prediction of SOC contents in the framework of digital soil mapping could be affected not only by the source and resolution of covariates but also by the model types, including linear and non-linear models. Therefore, future studies, especially in agroecosystems with similar environmental conditions, should consider the use of covariates with different sources and resolutions and choose a suitable model to increase SOC prediction accuracy.

Author Contributions: Conceptualization, J.G. and Q.T.; methodology, Q.T. and J.G.; software, Q.T.; validation, H.F., Y.L. and Y.G.; formal analysis, Q.T., J.G., H.F., Y.L. and Y.G.; investigation, H.F., Y.L. and Y.G.; resources, J.G. and H.F.; data curation, Q.T. and J.G.; writing—original draft preparation, Q.T. and J.G.; writing—review and editing, Q.T., J.G., H.F., Y.L. and Y.G.; visualization, Q.T.; supervision, J.G.; project administration, J.G.; funding acquisition, J.G. and H.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (32101301, 41977041), Guangdong Basic and Applied Basic Research Foundation (2020A1515110172), Jiangxi Provincial Science and Technology Special Project of Jinggangshan Agricultural High-tech Industrial Demonstration Zone, Foundation of President of the Zhongke-Ji'an Institute for Eco-Environmental Sciences (ZJIEES-2021-01, ZJIEES-2022-02), Science and Technology Project of Jinggangshan Agricultural High-tech Industrial Demonstration Zone (No. 202151).

Acknowledgments: The authors would like to thank anonymous reviewers for their constructive suggestions and comments to improve the quality of this paper. Special gratitude should also be given to the European Space Agency for providing free access to the Sentinel-1/2 imagery.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Schmidt, M.W.I.; Torn, M.S.; Abiven, S.; Dittmar, T.; Guggenberger, G.; Janssens, I.A.; Kleber, M.; Kögel-Knabner, I.; Lehmann, J.; Manning, D.A.C.; et al. Persistence of soil organic matter as an ecosystem property. *Nature* **2011**, *478*, 49–56. [CrossRef] [PubMed]
2. Chappell, A.; Webb, N.P.; Butler, H.J.; Strong, C.L.; McTainsh, G.H.; Leys, J.F.; Rossel, R.A.V. Soil organic carbon dust emission: An omitted global source of atmospheric CO_2. *Glob. Change Biol.* **2013**, *19*, 3238–3244. [CrossRef] [PubMed]
3. Chen, Y.; Wei, T.; Ren, K.; Sha, G.; Guo, X.; Fu, Y.; Yu, H. The coupling interaction of soil organic carbon stock and water storage after vegetation restoration on the Loess Plateau, China. *J. Environ. Manag.* **2022**, *306*, 114481. [CrossRef] [PubMed]
4. Tessema, B.; Sommer, R.; Piikki, K.; Söderström, M.; Namirembe, S.; Notenbaert, A.; Tamene, L.; Nyawira, S.S.; Paul, B. Potential for soil organic carbon sequestration in grasslands in East African countries: A review. *Grassl. Sci.* **2020**, *66*, 135–144. [CrossRef]
5. Bationo, A.; Kihara, J.; Vanlauwe, B.; Waswa, B.; Kimetu, J. Soil organic carbon dynamics, functions and management in West African agro-ecosystems. *Agric. Syst.* **2007**, *94*, 13–25. [CrossRef]
6. Chen, D.; Chang, N.; Xiao, J.; Zhou, Q.; Wu, W. Mapping dynamics of soil organic matter in croplands with MODIS data and machine learning algorithms. *Sci. Total Environ.* **2019**, *669*, 844–855. [CrossRef] [PubMed]
7. Forkuor, G.; Hounkpatin, O.K.L.; Welp, G.; Thiel, M. High Resolution Mapping of Soil Properties Using Remote Sensing Variables in South-Western Burkina Faso: A Comparison of Machine Learning and Multiple Linear Regression Models. *PLoS ONE* **2017**, *12*, e0170478. [CrossRef] [PubMed]
8. Xu, Y.; Wang, X.; Bai, J.; Wang, D.; Wang, W.; Guan, Y. Estimating the spatial distribution of soil total nitrogen and available potassium in coastal wetland soils in the Yellow River Delta by incorporating multi-source data. *Ecol. Indic.* **2020**, *111*, 106002. [CrossRef]
9. Bhattarai, N.; Quackenbush, L.J.; Dougherty, M.; Marzen, L.J. A simple Landsat–MODIS fusion approach for monitoring seasonal evapotranspiration at 30 m spatial resolution. *Int. J. Remote Sens.* **2015**, *36*, 115–143. [CrossRef]
10. Zhou, T.; Zhao, M.; Sun, C.; Pan, J. Exploring the Impact of Seasonality on Urban Land-Cover Mapping Using Multi-Season Sentinel-1A and GF-1 WFV Images in a Subtropical Monsoon-Climate Region. *ISPRS Int. J. Geo Inf.* **2017**, *7*, 3. [CrossRef]
11. Zhang, Y.; Guo, L.; Chen, Y.; Shi, T.; Luo, M.; Ju, Q.; Zhang, H.; Wang, S. Prediction of Soil Organic Carbon based on Landsat 8 Monthly NDVI Data for the Jianghan Plain in Hubei Province, China. *Remote Sens.* **2019**, *11*, 1683. [CrossRef]
12. Lausch, A.; Bannehr, L.; Beckmann, M.; Boehm, C.; Feilhauer, H.; Hacker, J.; Heurich, M.; Jung, A.; Klenke, R.; Neumann, C.; et al. Linking Earth Observation and taxonomic, structural and functional biodiversity: Local to ecosystem perspectives. *Ecol. Indic.* **2016**, *70*, 317–339. [CrossRef]
13. Wang, X.; Zhang, Y.; Atkinson, P.M.; Yao, H. Predicting soil organic carbon content in Spain by combining Landsat TM and ALOS PALSAR images. *Int. J. Appl. Earth Obs. Geoinf. ITC J.* **2020**, *92*, 102182. [CrossRef]
14. Wang, H.; Zhang, X.; Wu, W.; Liu, H. Prediction of Soil Organic Carbon under Different Land Use Types Using Sentinel-1/-2 Data in a Small Watershed. *Remote Sens.* **2021**, *13*, 1229. [CrossRef]
15. Jeong, G.; Oeverdieck, H.; Park, S.J.; Huwe, B.; Ließ, M. Spatial soil nutrients prediction using three supervised learning methods for assessment of land potentials in complex terrain. *Catena* **2017**, *154*, 73–84. [CrossRef]
16. Song, X.; Liu, F.; Ju, B.; Zhi, J.; Li, D.; Zhao, Y.; Zhang, G. Mapping soil organic carbon stocks of northeastern China using expert knowledge and GIS-based methods. *Chin. Geogr. Sci.* **2017**, *27*, 516–528. [CrossRef]
17. Behrens, T.; Schmidt, K.; Ramirez-Lopez, L.; Gallant, J.; Zhu, A.-X.; Scholten, T. Hyper-scale digital soil mapping and soil formation analysis. *Geoderma* **2014**, *213*, 578–588. [CrossRef]
18. Zhang, G.-L.; Liu, F.; Song, X.-D. Recent progress and future prospect of digital soil mapping: A review. *J. Integr. Agric.* **2017**, *16*, 2871–2885. [CrossRef]
19. Grimm, R.; Behrens, T.; Märker, M.; Elsenbeer, H. Soil organic carbon concentrations and stocks on Barro Colorado Island—Digital soil mapping using Random Forests analysis. *Geoderma* **2008**, *146*, 102–113. [CrossRef]
20. Wang, B.; Waters, C.; Orgill, S.; Gray, J.; Cowie, A.; Clark, A.; Liu, D.L. High resolution mapping of soil organic carbon stocks using remote sensing variables in the semi-arid rangelands of eastern Australia. *Sci. Total Environ.* **2018**, *630*, 367–378. [CrossRef]
21. Martin, M.; Orton, T.; Lacarce, E.; Meersmans, J.; Saby, N.; Paroissien, J.; Jolivet, C.; Boulonne, L.; Arrouays, D. Evaluation of modelling approaches for predicting the spatial distribution of soil organic carbon stocks at the national scale. *Geoderma* **2014**, *223*, 97–107. [CrossRef]
22. Sun, X.-L.; Wang, H.-L.; Zhao, Y.-G.; Zhang, C.; Zhang, G.-L. Digital soil mapping based on wavelet decomposed components of environmental covariates. *Geoderma* **2017**, *303*, 118–132. [CrossRef]
23. Zhu, H.; Hu, W.; Ding, H.; Lv, C.; Bi, R. Scale- and location-specific multivariate controls of topsoil organic carbon density depend on landform heterogeneity. *Catena* **2021**, *207*, 105695. [CrossRef]
24. Zhou, Y.; Chen, S.; Zhu, A.-X.; Hu, B.; Shi, Z.; Li, Y. Revealing the scale- and location-specific controlling factors of soil organic carbon in Tibet. *Geoderma* **2021**, *382*, 114713. [CrossRef]
25. Tian, H.; Zhang, J.; Zhu, L.; Qin, J.; Liu, M.; Shi, J.; Li, G. Revealing the scale-and location-specific relationship between soil organic carbon and environmental factors in China's north-south transition zone. *Geoderma* **2022**, *409*, 115600. [CrossRef]
26. Zhao, R.; Biswas, A.; Zhou, Y.; Zhou, Y.; Shi, Z.; Li, H. Identifying localized and scale-specific multivariate controls of soil organic matter variations using multiple wavelet coherence. *Sci. Total Environ.* **2018**, *643*, 548–558. [CrossRef]
27. Miller, B.A.; Koszinski, S.; Wehrhan, M.; Sommer, M. Impact of multi-scale predictor selection for modeling soil properties. *Geoderma* **2015**, *239–240*, 97–106. [CrossRef]

28. Zhou, T.; Geng, Y.; Ji, C.; Xu, X.; Wang, H.; Pan, J.; Bumberger, J.; Haase, D.; Lausch, A. Prediction of soil organic carbon and the C:N ratio on a national scale using machine learning and satellite data: A comparison between Sentinel-2, Sentinel-3 and Landsat-8 images. *Sci. Total Environ.* **2021**, *755*, 142661. [CrossRef] [PubMed]
29. Li, X.; Ding, J.; Liu, J.; Ge, X.; Zhang, J. Digital Mapping of Soil Organic Carbon Using Sentinel Series Data: A Case Study of the Ebinur Lake Watershed in Xinjiang. *Remote Sens.* **2021**, *13*, 769. [CrossRef]
30. Garosi, Y.; Ayoubi, S.; Nussbaum, M.; Sheklabadi, M. Effects of different sources and spatial resolutions of environmental covariates on predicting soil organic carbon using machine learning in a semi-arid region of Iran. *Geoderma Reg.* **2022**, *29*, e00513. [CrossRef]
31. Owusu, S.; Yigini, Y.; Olmedo, G.F.; Omuto, C.T. Spatial prediction of soil organic carbon stocks in Ghana using legacy data. *Geoderma* **2020**, *360*, 114008. [CrossRef]
32. Liu, S.; Yang, Y.; Shen, H.; Hu, H.; Zhao, X.; Li, H.; Liu, T.; Fang, J. No significant changes in topsoil carbon in the grasslands of northern China between the 1980s and 2000s. *Sci. Total Environ.* **2018**, *624*, 1478–1487. [CrossRef] [PubMed]
33. Guo, L.; Fu, P.; Shi, T.; Chen, Y.; Zhang, H.; Meng, R.; Wang, S. Mapping field-scale soil organic carbon with unmanned aircraft system-acquired time series multispectral images. *Soil Tillage Res.* **2020**, *196*, 104477. [CrossRef]
34. Olaya-Abril, A.; Parras-Alcántara, L.; Lozano-García, B.; Obregón-Romero, R. Soil organic carbon distribution in Mediterranean areas under a climate change scenario via multiple linear regression analysis. *Sci. Total Environ.* **2017**, *592*, 134–143. [CrossRef] [PubMed]
35. Zhou, T.; Geng, Y.; Chen, J.; Liu, M.; Haase, D.; Lausch, A. Mapping soil organic carbon content using multi-source remote sensing variables in the Heihe River Basin in China. *Ecol. Indic.* **2020**, *114*, 106288. [CrossRef]
36. John, K.; Isong, I.A.; Kebonye, N.M.; Ayito, E.O.; Agyeman, P.C.; Afu, S.M. Using Machine Learning Algorithms to Estimate Soil Organic Carbon Variability with Environmental Variables and Soil Nutrient Indicators in an Alluvial Soil. *Land* **2020**, *9*, 487. [CrossRef]
37. Were, K.; Bui, D.T.; Dick, Ø.B.; Singh, B.R. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. *Ecol. Indic.* **2015**, *52*, 394–403. [CrossRef]
38. Lamichhane, S.; Kumar, L.; Wilson, B. Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review. *Geoderma* **2019**, *352*, 395–413. [CrossRef]
39. 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]
40. Moura-Bueno, J.M.; Dalmolin, R.S.D.; Horst-Heinen, T.Z.; Grunwald, S.; Caten, A.T. Environmental covariates improve the spectral predictions of organic carbon in subtropical soils in southern Brazil. *Geoderma* **2021**, *393*, 114981. [CrossRef]
41. He, Z.; Zhang, M.; Wilson, M.J. Distribution and Classification of Red Soils in China. In *The Red Soils of China*; Springer: Dordrecht, The Netherlands, 2004; pp. 29–33. [CrossRef]
42. Han, Y.; Yi, D.; Ye, Y.; Guo, X.; Liu, S. Response of spatiotemporal variability in soil pH and associated influencing factors to land use change in a red soil hilly region in southern China. *Catena* **2022**, *212*, 106074. [CrossRef]
43. Ning, J.; Liu, J.; Kuang, W.; Xu, X.; Zhang, S.; Yan, C.; Li, R.; Wu, S.; Hu, Y.; Du, G.; et al. Spatiotemporal patterns and characteristics of land-use change in China during 2010–2015. *J. Geogr. Sci.* **2018**, *28*, 547–562. [CrossRef]
44. Zhang, X.; Xue, J.; Chen, S.; Wang, N.; Shi, Z.; Huang, Y.; Zhuo, Z. Digital Mapping of Soil Organic Carbon with Machine Learning in Dryland of Northeast and North Plain China. *Remote Sens.* **2022**, *14*, 2504. [CrossRef]
45. He, X.; Yang, L.; Li, A.; Zhang, L.; Shen, F.; Cai, Y.; Zhou, C. Soil organic carbon prediction using phenological parameters and remote sensing variables generated from Sentinel-2 images. *Catena* **2021**, *205*, 105442. [CrossRef]
46. Fick, S.E.; Hijmans, R.J. WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **2017**, *37*, 4302–4315. [CrossRef]
47. Wang, Z.; Du, Z.; Li, X.; Bao, Z.; Zhao, N.; Yue, T. Incorporation of high accuracy surface modeling into machine learning to improve soil organic matter mapping. *Ecol. Indic.* **2021**, *129*, 107975. [CrossRef]
48. Khanal, S.; Fulton, J.; Klopfenstein, A.; Douridas, N.; Shearer, S. Integration of high resolution remotely sensed data and machine learning techniques for spatial prediction of soil properties and corn yield. *Comput. Electron. Agric.* **2018**, *153*, 213–225. [CrossRef]
49. Yang, L.; He, X.; Shen, F.; Zhou, C.; Zhu, A.-X.; Gao, B.; Chen, Z.; Li, M. Improving prediction of soil organic carbon content in croplands using phenological parameters extracted from NDVI time series data. *Soil Tillage Res.* **2020**, *196*, 104465. [CrossRef]
50. Arabameri, A.; Yamani, M.; Pradhan, B.; Melesse, A.; Shirani, K.; Bui, D.T. Novel ensembles of COPRAS multi-criteria decision-making with logistic regression, boosted regression tree, and random forest for spatial prediction of gully erosion susceptibility. *Sci. Total Environ.* **2019**, *688*, 903–916. [CrossRef] [PubMed]
51. Yang, R.-M.; Zhang, G.-L.; Liu, F.; Lu, Y.-Y.; Yang, F.; Yang, F.; Yang, M.; Zhao, Y.-G.; Li, D.-C. Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. *Ecol. Indic.* **2016**, *60*, 870–878. [CrossRef]
52. Ahirwal, J.; Nath, A.; Brahma, B.; Deb, S.; Sahoo, U.K.; Nath, A.J. Patterns and driving factors of biomass carbon and soil organic carbon stock in the Indian Himalayan region. *Sci. Total Environ.* **2021**, *770*, 145292. [CrossRef] [PubMed]
53. Keskin, H.; Grunwald, S.; Harris, W.G. Digital mapping of soil carbon fractions with machine learning. *Geoderma* **2019**, *339*, 40–58. [CrossRef]

54. Lefever, D.W. Measuring Geographic Concentration by Means of the Standard Deviational Ellipse. *Am. J. Sociol.* **1926**, *32*, 88–94. [CrossRef]
55. Huang, J.; Song, L.; Yu, M.; Zhang, C.; Li, S.; Li, Z.; Geng, J.; Zhang, C. Quantitative spatial analysis of thermal infrared radiation temperature fields by the standard deviational ellipse method for the uniaxial loading of sandstone. *Infrared Phys. Technol.* **2022**, *123*, 104150. [CrossRef]
56. Jenks, G.F. *Optimal Data Classification for Choropleth Maps*; Department of Geographiy, University of Kansas Occasional Paper: Lawrence, KS, USA, 1977.
57. Poggio, L.; Gimona, A. Assimilation of optical and radar remote sensing data in 3D mapping of soil properties over large areas. *Sci. Total Environ.* **2017**, *579*, 1094–1110. [CrossRef]
58. Xu, S.; Wang, M.; Shi, X. Hyperspectral imaging for high-resolution mapping of soil carbon fractions in intact paddy soil profiles with multivariate techniques and variable selection. *Geoderma* **2020**, *370*, 114358. [CrossRef]
59. Chen, L.; Ren, C.; Li, L.; Wang, Y.; Zhang, B.; Wang, Z.; Li, L. A Comparative Assessment of Geostatistical, Machine Learning, and Hybrid Approaches for Mapping Topsoil Organic Carbon Content. *ISPRS Int. J. Geo Inf.* **2019**, *8*, 174. [CrossRef]
60. Morellos, A.; Pantazi, X.-E.; Moshou, D.; Alexandridis, T.; Whetton, R.; Tziotzios, G.; Wiebensohn, J.; Bill, R.; Mouazen, A.M. Machine learning based prediction of soil total nitrogen, organic carbon and moisture content by using VIS-NIR spectroscopy. *Biosyst. Eng.* **2016**, *152*, 104–116. [CrossRef]
61. Chen, Y.; Wang, J.; Liu, G.; Yang, Y.; Liu, Z.; Deng, H. Hyperspectral Estimation Model of Forest Soil Organic Matter in Northwest Yunnan Province, China. *Forests* **2019**, *10*, 217. [CrossRef]
62. Guo, P.; Li, T.; Gao, H.; Chen, X.; Cui, Y.; Huang, Y. Evaluating Calibration and Spectral Variable Selection Methods for Predicting Three Soil Nutrients Using Vis-NIR Spectroscopy. *Remote Sens.* **2021**, *13*, 4000. [CrossRef]
63. Adhikari, K.; Hartemink, A.E. Digital Mapping of Topsoil Carbon Content and Changes in the Driftless Area of Wisconsin, USA. *Soil Sci. Soc. Am. J.* **2015**, *79*, 155–164. [CrossRef]
64. Guo, Z.; Adhikari, K.; Chellasamy, M.; Greve, M.B.; Owens, P.R.; Greve, M.H. Selection of terrain attributes and its scale dependency on soil organic carbon prediction. *Geoderma* **2019**, *340*, 303–312. [CrossRef]
65. Xu, Y.; Smith, S.E.; Grunwald, S.; Abd-Elrahman, A.; Wani, S.P. Evaluating the effect of remote sensing image spatial resolution on soil exchangeable potassium prediction models in smallholder farm settings. *J. Environ. Manag.* **2017**, *200*, 423–433. [CrossRef] [PubMed]
66. Jelinski, D.E.; Wu, J. The modifiable areal unit problem and implications for landscape ecology. *Landsc. Ecol.* **1996**, *11*, 129–140. [CrossRef]
67. Katuwal, S.; Knadel, M.; Moldrup, P.; Norgaard, T.; Greve, M.H.; De Jonge, L.W. Visible–Near-Infrared Spectroscopy can predict Mass Transport of Dissolved Chemicals through Intact Soil. *Sci. Rep.* **2018**, *8*, 11188. [CrossRef] [PubMed]
68. Castaldi, F.; Chabrillat, 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.* **2019**, *11*, 2121. [CrossRef]
69. Castaldi, F.; Hueni, A.; Chabrillat, S.; Ward, K.; Buttafuoco, G.; Bomans, B.; Vreys, K.; Brell, M.; van Wesemael, B. Evaluating the capability of the Sentinel 2 data for soil organic carbon prediction in croplands. *ISPRS J. Photogramm. Remote Sens.* **2019**, *147*, 267–282. [CrossRef]
70. Yang, R.-M.; Guo, W.-W.; Zheng, J.-B. Soil prediction for coastal wetlands following Spartina alterniflora invasion using Sentinel-1 imagery and structural equation modeling. *Catena* **2019**, *173*, 465–470. [CrossRef]
71. Yang, R.-M.; Guo, W.-W. Using time-series Sentinel-1 data for soil prediction on invaded coastal wetlands. *Environ. Monit. Assess.* **2019**, *191*, 462. [CrossRef] [PubMed]
72. Kalambukattu, J.G.; Kumar, S.; Raj, R.A. Digital soil mapping in a Himalayan watershed using remote sensing and terrain parameters employing artificial neural network model. *Environ. Earth Sci.* **2018**, *77*, 203. [CrossRef]
73. Mahmoudzadeh, H.; Matinfar, H.R.; Taghizadeh-Mehrjardi, R.; Kerry, R. Spatial prediction of soil organic carbon using machine learning techniques in western Iran. *Geoderma Reg.* **2020**, *21*, e00260. [CrossRef]
74. Xiaodong, S.; Feng, L.; Zhang, G.; Decheng, L.; Yuguo, Z.; Jinling, Y. Mapping soil organic carbon using local terrain attributes: A comparison of different polynomial models. *Pedosphere* **2017**, *27*, 681–693.
75. Yang, R.-M.; Liu, L.-A.; Zhang, X.; He, R.-X.; Zhu, C.-M.; Zhang, Z.-Q.; Li, J.-G. The effectiveness of digital soil mapping with temporal variables in modeling soil organic carbon changes. *Geoderma* **2022**, *405*, 115407. [CrossRef]
76. Ning, L.; Cheng, C.; Lu, X.; Shen, S.; Zhang, L.; Mu, S.; Song, Y. Improving the Prediction of Soil Organic Matter in Arable Land Using Human Activity Factors. *Water* **2022**, *14*, 1668. [CrossRef]
77. Wu, J.; Zhong, B.; Tian, S.; Yang, A.; Wu, J. Downscaling of Urban Land Surface Temperature Based on Multi-Factor Geographically Weighted Regression. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 2897–2911. [CrossRef]