Prediction of Soil Organic Carbon Content Using Sentinel-1/2 and Machine Learning Algorithms in Swamp Wetlands in Northeast China

Honghua Zhang, Luhe Wan, and Yang Li

***Abstract*—Soil organic carbon (SOC) is a sensitive indicator of climate change, and small changes in the soil carbon pool will affect the carbon balance. Accurate and robust SOC quantitative prediction is of great significance to studying the carbon budget of swamp wetlands and its response to climate change. In this study, a new framework was proposed and assessed for predicting the SOC content based on Sentinel-2 (S2), Sentinel-1 (S1), and the digital elevation model (DEM) together with the extreme gradient boosting with random forest (XGBRF) model. The determination coefficient (R^2), root mean square error (RMSE), mean absolute error (MAE), and Lin's concordance correlation coefficient (LCCC) were applied to assess the performances of the models. The results revealed that the prediction performance of the XGBRF regression model was much better than that of extreme gradient boosting and random forest regression models. Compared with single sensor data, using multisensor data to predict the SOC content yielded more accurate results. The XGBRF model based on S1, S2, and DEM fusion yielded the highest prediction accuracy (R^2_testing = 0.6639, RMSE = 1.3236 g/kg, MAE = 1.2546 g/kg, LCCC = 0.7621). Regarding the importance of the variables, the S1 and S2 features were major contributors to the SOC content prediction (41% and 52%, respectively), followed by the topographic variables extracted from the DEM (7%). The proposed framework can be used for SOC prediction based on a small sample dataset, and it provides a method for long-term and rapid monitoring of the SOC contents in wetlands.**

***Index Terms*—Extreme gradient boosting with random forest (XGBRF), machine learning (ML), Sentinel-1/2, soil organic carbon (SOC), swamp wetlands.**

I. Introduction

NATURAL wetlands have a high biological production and low decomposition rates, which enable wetland soils to store large amounts of organic carbon. The IPCC (2000) cited the statistical results of the German Advisory Council on Global Change (WBGU) (1998), showing that the carbon storage per unit area of wetlands was three times that of tropical forests, making it the highest carbon storage per unit area among various terrestrial ecosystems. Furthermore, over 90% of the carbon storage of wetland ecosystems was found to be stored in the soil. Wetlands play an important role in the global carbon cycle and carbon balance, and the transformations of wetland carbon source and carbon sink functions are one of the important factors affecting global climate change [1], [2], [3]. However, wetland ecosystems are relatively fragile, and slight changes in climate can affect their carbon budgets. Soil organic carbon (SOC) is sensitive to climate change, and it can be used as an indicator to measure the dynamic change in the soil carbon storage in wetlands [4], [5], [6]. Therefore, accurate estimation of wetland SOC is helpful in predicting the feedback between wetland ecosystems and climate change and is significant in maintaining the carbon balance of such ecosystems.

Traditionally, the SOC content has been estimated through field investigation, sampling, and laboratory measurement. This traditional method is accurate but expensive and time-consuming, so it is difficult to use the traditional method for estimation of SOC storage on a large scale. Due to the characteristics of data availability and large-scale monitoring, remote sensing technology has become the key to the quantitative prediction of SOC [7]. For swamp wetlands characterized by large areas and inconvenient sampling, the traditional large-scale data acquisition technique is unrealistic. Thus, it is necessary to combine classical soil investigation techniques with advanced remote sensing technology to study the spatial distribution of SOC in swamp wetlands.

Remote sensing technology has been proven to be an effective means of obtaining soil properties. Optical images and multispectral data were initially used for detecting the properties of soils. The spectrum ranges from visible-near infrared (VIS-NIR) to shortwave infrared (SWIR) [8], [9], [10]. The sensors are carried on satellites, aircraft, and unmanned aerial vehicles (UAVs) [11], [12], [13], [14]. Previous studies have demonstrated that VIS-NIR-SWIR spectroscopy can successfully be applied to predict soil properties such as the texture, total carbon content, total nitrogen content, and PH. When combined with appropriate models, the prediction accuracy can be further improved [15], [16]. Compared with other optical remote sensing data, Sentinel-2 (S2) data have attracted much attention. The 13 bands

Manuscript received 1 January 2023; revised 11 March 2023 and 3 April 2023; accepted 25 May 2023. Date of publication 31 May 2023; date of current version 14 June 2023. This work was supported by the National Natural Science Foundation of China under Grant 42071079 and Grant 41671100. *(Corresponding author: Luhe Wan.)*

Honghua Zhang is with the Institute of Geographical Science, Harbin Normal University, Harbin 150025, China, and also with the Institute of Mining Engineering, Heilongjiang University of Science and Technology, Harbin 150022, China (e-mail: zhanghonghua@hrbnu.edu.cn).

Luhe Wan and Yang Li are with the Institute of Geographical Science, Harbin Normal University, Harbin 150025, China (e-mail: wanluhe@hrbnu.edu.cn; liyang@usth.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3281732

of S2 cover all bands from visible light to SWIR. Visible spectra (490, 560, and 665 nm) are important predictors of SOC remote sensing retrieval [7], [17]. SWIR region can detect important soil chemical properties, which are related to the quantitative prediction of SOC [18]. Therefore, S2 remote sensing images may be a very useful data source for predicting SOC content.

However, since optical sensors are susceptible to atmospheric radiation, cloud coverage, and rainy weather, it is still challenging to quantitatively predict the SOC content using only optical sensors. In contrast, synthetic aperture radar (SAR) sensors can observe the Earth all day long and are not affected by cloudy and rainy weather. The echo signal received by SAR can record the amplitude and phase information of radar waves reflected by ground objects. The SAR sensor can capture the relationship between the soil and vegetation, which provides an opportunity to predict the chemical properties of soils successfully and monitor soil changes continuously [19]. Several studies have confirmed the usefulness of SAR data for predicting soil properties such as SOC and bulk density [20], [21]. Sentinel-1 (S1) data have performed well in the prediction of wetland soil properties [20]. Previous studies have further found that compared with a single sensor, multisensor data fusion may improve the prediction accuracy of the SOC [22], [23].

SOC content prediction is achieved by establishing the relationships between environmental covariates and the SOC content. Remote sensing technology and machine learning (ML) models provide a good guarantee for SOC prediction. The advancements of multispectral sensors and SAR sensors provide more available environmental covariates. The development of ML models provides more model options for SOC prediction. Through a literature search, it was found that the geographically weighted regression [24], support vector regression [25], enhanced regression tree [26], and the random forest (RF) algorithm [27] have been more widely used in SOC prediction. Significantly, several studies have reported that tree-based models, such as the RF [28] and extreme gradient boosting (XGBoost) [29], [30], have better SOC prediction performances. The extreme gradient boosting with random forest (XGBRF) model is an advanced hybrid integration model that combines the advantages of the RF and XGBoost. It has been found that the XGBRF is an effective algorithm for dealing with classification problems, and its accuracy has been reported to be as high as 99.25% for specific datasets [31]. This algorithm may also achieve better results in SOC prediction research (especially for small sample datasets). However, there are few reports on the effectiveness and accuracy of this algorithm in solving regression problems, and there are no reports on the prediction of SOC using this algorithm.

The use of remote sensing and ML models to quantitatively invert the SOC has mainly been used for SOC mapping of agricultural land [30], and it has rarely been applied to swamp wetlands. Compared with farmland, swamp wetlands have a complex environment, dense vegetation cover, and relatively limited samples. Therefore, robust prediction models based on a small sample dataset are urgently needed for the quantitative prediction of SOC in swamp wetlands. In conclusion, in view of the difficult, time-consuming, and expensive nature of swamp wetland sampling, the rapid development of remote sensing techniques and ML algorithms may provide an opportunity to solve the problem of SOC prediction based on a small sample dataset. This study aimed to design a new framework that integrates S1, S2, digital elevation model (DEM), and the advanced XGBRF regression model to estimate the SOC content of swamp wetlands. The specific objectives were (1) to evaluate the feasibility of estimating the SOC in swamp wetlands using multispectral images, SAR data, and DEM (especially in the case of small sample datasets); (2) to compare the SOC prediction performance of the XGBRF model with those of other two models with better prediction performances (XGBoost and RF) under different data fusion scenarios; and (3) to estimate the relative importance of predictors from different data sources.

II. Materials and Methods

We designed a new framework that integrates multispectral data (S2), SAR data (S1), and DEM data, and used advanced ML models to predict the SOC contents of swamp wetlands. The research process comprised four steps: 1) obtaining images and SOC data; 2) preprocessing multisource data and extracting the predictor variables (a total of 46 predictor variables were extracted: 23 from S2, 19 from S1, and 4 from DEM); 3) training and evaluating the SOC prediction models (based on XGBoost, RF, and XGBRF) to identify the optimal model; and 4) obtaining spatial distribution maps of the SOC using the optimal model.

A. Study Area

The study area is located in the Khingan Range in the northern part of Heilongjiang Province, China. It is a concentrated distribution area of permafrost wetlands and is one of the most important wetland distribution areas in the subarctic region. In the study area, swamp wetlands and permafrost coexist and have a symbiotic relationship, so they are more sensitive to climate change. Surface water, permafrost meltwater, precipitation, and other water sources constitute a diversified water supply mechanism. The high latitude, high altitude, and permafrost constitute a cold control system. The above jointly constitutes a cold and wet geographical environment, which is conducive to the development of swamp wetlands. In the wide river valley, flat terrace, and platform areas, the soil water is supersaturated, forming a large area of swamp and peatland. The vegetation species in the swamp wetland are mainly Dahurian larch, Dusi bilberry, narrow-leaved eucalyptus, and sphagnum moss. Due to the cold climate, lush vegetation, and relatively small amount of evaporation in this region, it is difficult for microorganisms to decompose plant residues, which is conducive to the accumulation of SOC. In this study, two typical swamp wetlands distribution areas were selected as the study areas: Huzhong and Heihe (Fig. 1). Huzhong (52°02′–52°12′ N, 123°09′–123°26′ E) is located in the Greater Khingan Mountains, in the middle reaches of the Huma River, and there is predominant continuous permafrost in this region. Heihe (49°59′–50°10′ N, 126°34′–126°51′ E) is located in the eastern foothills of the Small Khingan Mountains

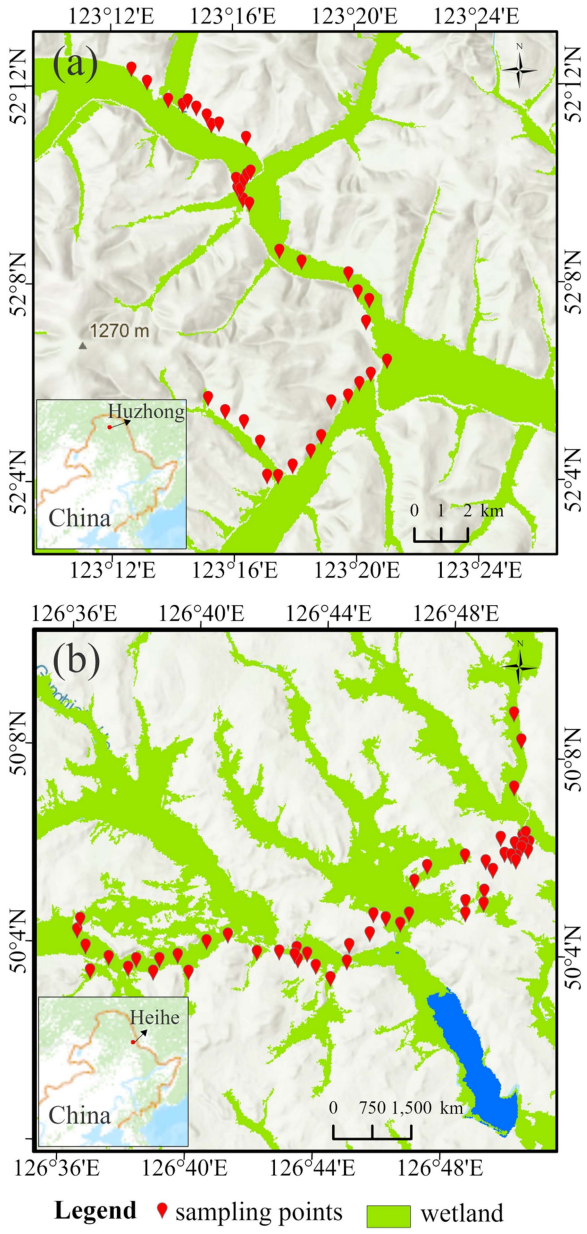


Fig. 1. Locations of the study areas and soil sampling points. (a) Huzhong. (b) Heihe.

and the upper reaches of the Gongbeila River, and is a typical forest wetland ecosystem with sparse island permafrost.

B. Soil Sample Collection

We conducted field investigations and sampling in the two study areas from August to September 2021. Considering the representativeness of the different types of surface cover and the sampling points, we collected topsoil (0–20 cm) samples from 89 locations in forest swamp, shrub swamp, and herb swamp areas. The geographic coordinates of each sampling point were accurately recorded using a handheld global positioning system (GPS) instrument. Fresh soil samples were brought back to the laboratory. First, the samples were naturally air dried, the nonsoil material (e.g., plant roots and stones) was removed, and the samples were ground and sieved (0.25 mm) to prepare the soil samples for analysis. The SOC content was determined using the potassium dichromate external heating method. The SOC contents of the samples varied from 6.4248 g/kg to 68.9247 g/kg. The mean value was 38.2399 g/kg, and the standard deviation was 17.1701 g/kg. To improve the fitting accuracy, the SOC contents were converted to the natural logarithm (LnSOC) in all of the prediction models, and finally, the prediction results were converted back to the actual SOC contents.

C. Data Acquisition and Processing

The predictor variables for the prediction of the SOC were extracted from S2, S1, and DEM data. These predictor variables from different sources were unified into the UTM/WGS84 projection coordinate system and converted into grid data (10 m resolution). In addition, due to the different dimensions and orders of magnitude of the data, all of the predictor variables were standardized before being input into the ML models. The data processing platforms used in this process were ArcGIS 10.6, ENVI5.3, and SNAP.

1) Processing of Sentinel-2 Imagery: S2 Multispectral Imaging (MSI) Level-2A images were used to retrieve the SOC contents in the two study areas. The Level-2A data were bottom of atmosphere corrected reflectance data that had been processed using radiation calibration and atmospheric correction. The acquisition dates of the two images used in the study were August 31 and September 18, 2021, which were close to the collection date of the soil samples. The S2 images were downloaded from the data sharing website of the European Space Agency. In total, 10 bands that have been widely used to evaluate soil properties were selected from the 13 S2 MSI bands: B2, B3, B4, B5, B6, B7, B8, B8a, B11, and B12. Vegetation and soil indexes may have strong correlations with different physical and chemical properties of soil. In this research, nine vegetation indexes and four soil index variables (Table I) were selected to predict the SOC contents of the swamp wetlands. A total of 23 prediction variables, including 10 multispectral bands, nine vegetation indexes, and four soil radiation indexes, were extracted from the S2 MSI for SOC prediction. All of the bands were resampled to a 10 m resolution. The data resampling and index calculations were completed using ENVI5.3.

2) Processing of Sentinel-1 Imagery: The S1 remote sensing images used in this study were interference wideswath (IW) mode with ground range detected (GRD) format, which were obtained on August 28 and 30, 2021. We selected dual polarized data, including VV and VH. After radiometric calibration, terrain correction, and other processing in SNAP8.0, the amplitude information of the SAR images was converted into the backscatter coefficient. Nineteen predictive variables were derived from the S1 images, including two dual polarization bands (VH and VV), three transformed bands (VH/VV, VH-VV, and (VH+VV)/2), and 14 textural features obtained from the VV and VH using the gray level co-occurrence matrix (GLCM) algorithm (VH_Mean, VH_Variance, VH_Homogeneity, VH_Contrast, VH_Dissimilarity,

TABLE I
VEGETATION AND SOIL INDEXES EXTRACTED FROM SENTINEL-2 IMAGES

Vegetation and soil indexes	Formulas
Normalized Difference Vegetation Index (NDVI) [32]	(B8-B4)/(B8+B4)
Ratio Vegetation Index (RVI) [33]	B8/B4
Green Normalized Difference Vegetation Index (GNDVI) [34]	(B8-B3)/(B8+B3)
Enhanced Vegetation Index-2 (EVI-2) [35]	2.5*(B8-B4)/(B8+2.4*B4+1)
Normalized Difference Index using S2 Bands 4 and 5 (NDI45) [36]	(B5-B4)/(B5+B4)
Soil Adjusted Vegetation Index (SAVI) [37]	(1+L)* (B8-B4)/(B8+B4+L) L=0.5 in most conditions
Inverted Red-Edge Chlorophyll Index (IRECI) [38]	(B7-B4)/(B5/B6)
Modified Chlorophyll Absorption in Reflectance Index (MCARI) [39]	[(B5-B4)-0.2*(B5-B3)]* (B5-B8)
Normalized Difference Moisture Index (NDMI) [40]	(B8-B11)/(B8+B11)
Soil Brightness Index (SBI) [41]	$\sqrt{((B4)^2+(B8)^2)}$
Normalized Difference Tillage Index (NDTI) [42]	(B11-B12)/(B11+B12)
Clay Mineral Ratio (CMR) [43]	B11/B12
Bare Soil Index (BSI) [44]	((B11+B4)-(B8+B2))/ ((B11+B4)+(B8+B2))*100+100

* Note: The bands of S2 are B2 (blue), B3 (green), B4 (red), B5 (red-edge 1), B6 (red-edge 2), B7 (red-edge 3), B8 (near-infrared), B8a (narrow-NIR), B11 (short-wavelength infrared (SWIR1)), and B12 (SWIR2).

VH_Entropy, VH_Correlation, VV_Mean, VV_Variance, VV_Homogeneity, VV_Contrast, VV_Dissimilarity, VV_Entropy, and VV_Correlation). The band transformation and texture feature extraction were completed using ENVI5.3.

3) Terrain Data Processing: The topographic variables were extracted from the Shuttle Radar Topography Mission (SRTM) DEM (30 m resolution). The DEM data were geocoded in the WGS84/EGM96 projection and downloaded from http://earthexplorer.usgs.gov/ in the Geo-TIFF format. The DEM was projected to the UTM/WGS84 coordinate system and resampled to a 10 m spatial resolution. Four topographic variables were calculated in ArcGIS 10.6, including the topographic wetness index (TWI), slope, elevation, and aspect.

D. Machine Learning Models

Three ML techniques for predicting SOC are described in this section, namely XGBoost, RF, and XGBRF. The attribute values of the predictor variables were extracted from the raster data using the ENVI5.3 software (corresponding to the sampling points). The sample dataset was composed of the grid attribute values and SOC values of the sampling points, of which eighty percent were used to train the models and twenty percent were used to test the models. Python 3.9 and Scikit-learn software packages were used to establish the models and optimize the parameters of the three ML models.

XGBoost is a scalable end-to-end gradient boosting tree algorithm that can effectively deal with classification and regression problems [29]. The goal of the algorithm is to overcome the over-fitting problems and optimize the performance of the model [45]. XGBoost has two objective functions: loss function and regularization term [46]. The second derivative of the loss function is calculated using this algorithm, and the trend of gradient change is further considered to make the fitting faster and more accurate. The regularization term limits the number of leaf nodes through a penalty mechanism, thereby controlling the complexity of the model and preventing overfitting [47]. Parallel and distributed computing make the learning process faster.

The RF model, which is an ensemble learning algorithm containing multiple decision trees [48], is used to solve classification and regression problems [49], [50], [51]. Using bootstrap sampling technology, about 63.2% of the training dataset was randomly selected for the model training, and about 36.8% of the training dataset was used as the verification dataset to estimate the accuracy of the model (out-of-bag estimate). The RF algorithm overcomes the over-fitting problem of decision trees and has a good noise and outliers tolerance.

The XGBRF model is a hybrid ensemble model that integrates the XGBoost and RF algorithms. The XGBoost and RF algorithms are both advanced algorithms based on decision trees. XGBoost is an excellent boosting algorithm, and the RF is an excellent representative of a bagging algorithm. Serial boosting repeats the training by reweighting the incorrectly judged training samples to improve the accuracy of the basic estimators and reduce the deviation. Parallel bagging trains a variety of the basic estimators via sampling to reduce the variance. The XGBRF model actually integrates multiple RFs using the boosting algorithm to obtain classification or regression results [31]. The XGBRF makes use of the advantages of the XGBoost and RF to upgrade the accuracy of the model and to avoid over-fitting problems [52].

TABLE II
SCENARIOS WITH DIFFERENT VARIABLE COMBINATIONS

Scenario	Source of variables	Scenario	Source of variables
I	S2	IV	S1 and DEM
II	S1	V	S2 and S1
III	S2 and DEM	VI	S2, S1, and DEM

E. Model Performance Evaluation

We constructed six scenarios based on the different predictor variables extracted from the different data sources (Table II). Scenario I was constructed using S2-derived predictors; and Scenario II was constructed using S1-derived predictors. The other scenarios included two categories or three predictor variables from S2, S1, and DEM data. The purpose of the designed scenarios was to assess the impacts of the different variable combinations on the accuracy of the SOC prediction.

To estimate the SOC prediction performance of the models under different data fusion scenarios, we introduced four indexes for evaluating model accuracy: root mean square error (RMSE), coefficient of determination (R^2), mean absolute error (MAE), and Lin's concordance correlation coefficient (LCCC) [53]. With higher R^2 and LCCC values and lower RMSE and MAE values, the prediction performance of the model was better. The four indexes were calculated according to the following

equations:

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

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \bar{O})^2} \quad (2)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|P_i - O_i| \quad (3)$$

$$LCCC = \frac{2rS_PS_O}{(\bar{P} - \bar{O})^2 + S_P^2 + S_O^2} \quad (4)$$

where n is the number of soil samples; $P_i, O_i, \bar{P}$, and $\bar{O}$ represent the predicted value, observed value, mean of the predicted value, and mean of the observed value, respectively; S_P and S_O represent the standard deviations of the predicted and observed values, respectively; and r is Pearson's correlation coefficient.

To further evaluate the uncertainty of the models, we selected the coefficient of variation (CV) as the evaluation indicator. After 100 iterations of each model, 100 maps of SOC content were obtained. The standard deviations and means were calculated pixel by pixel, and then the CV map of the study area was obtained. The CV calculation formula is as follows:

$$CV = \frac{SD}{Mean} \times 100\% \quad (5)$$

where *SD* and *Mean* represent the standard deviation and mean of SOC content for each pixel after 100 iterations of the models, respectively.

To further test the reliability of the model predictions, we introduced the prediction interval coverage probability (PICP) to evaluate the probability that the observed values appeared in the prediction interval. The confidence interval (CI) and PICP were determined as follows:

$$CI_i = \bar{P}_i \pm \frac{S_{P_i}}{\sqrt{M}} \cdot u_{\frac{\alpha}{2}} \quad (6)$$

$$PICP = \left(\frac{1}{n} \cdot \sum_{i=1}^{n} C_i\right) \cdot 100\% \quad (7)$$

where $\bar{P}_i, S_{P_i}$ denote the means and standard deviations of the predicted values, respectively; M is the number of model iterations; n is the sample size; $u_{\frac{\alpha}{2}}$ is the standard normal distribution with respect to $\frac{\alpha}{2}$ upper quantile; α is 0.05 and the confidence probability is 95% in this study; and C_i represents a Boolean function, when the observed value was in the CI, C_i was 1, otherwise 0.

III. Results

A. Correlation of Predictor Variables and SOC

The Pearson′s correlation coefficients between the 46 predictors and the measured SOC content are presented in Table III. The results show that among the 23 predictors from the S2, B5 (red-edge 1) had the highest correlation with the SOC content, and B11 (SWIR1) and B12 (SWIR2) had strong a positive correlation with the SOC content. Compared with the other vegetation indices and soil indices, the MCARI and SBI had stronger correlations with the SOC content. Of the four terrain variables, the elevation had the strongest correlation, and the TWI exhibited a negative correlation. Among the 19 predictors from S1, VH, VH/VV, and VH-VV exhibited strong correlations, while VV and (VH+VV)/2 had a weak correlation with the SOC content. For the texture variables, the VH_Variance, VH_Contrast, VH_Dissimilarity, VH_Entropy, VV_Variance, VV_Contrast, VV_ Dissimilarity had strong positive correlations with SOC content, while VH_mean, VH_Homogeneity, VH_Correlation, and VV_Homogeneity negative correlations with the SOC content.

TABLE III
PEARSON'S CORRELATION COEFFICIENTS BETWEEN PREDICTOR VARIABLES AND OBSERVED SOC

predictor variables	correlation coefficients	predictor variables	correlation coefficients	predictor variables	correlation coefficients
B2	−0.176	RVI	−0.057	VH_Contrast	0.164
B3	0.167	SAVI	−0.135	VH_Dissimilarity	0.167
B4	0.237	BSI	0.143	VH_Entropy	0.208
B5	0.338	CMR	0.048	VHCorrelation	−0.129
B6	0.101	NDMI	−0.102	VH/VV	0.227
B7	0.068	NDTI	0.033	VH-VV	−0.257
B8	0.164	SBI	0.237	VV	0.094
B8a	0.100	Aspect	0.124	VV_Mean	0.011
B11	0.272	Elevation	0.229	VV_Variance	0.147
B12	0.220	Slope	0.044	VV_Homogeneity	−0.194
EVI-2	−0.125	TWI	−0.149	VV_Contrast	0.121
GNDVI	−0.007	(VH+VV)/2	−0.086	VV_Dissimilarity	0.173
IRECI	−0.062	VH	−0.220	VV_Entropy	0.096
MCARI	0.228	VH_Mean	−0.148	VV_Correlation	0.048
NDI45	−0.006	VH_Variance	0.132	–	–
NDVI	−0.135	VH_Homogeneity	−0.143	–	–

TABLE IV
SOC PREDICTION ACCURACIES OF THE THREE ML MODELS DIFFERENT SCENARIOS

Modeling technique	Scenarios	R^2_training (80%)	R^2_testing (20%)	RMSE (g/kg)	MAE (g/kg)	LCCC
XGBoost	I	0.9879	0.4738	1.4201	1.3122	0.6711
	II	0.9879	0.2698	1.5116	1.3829	0.6123
	III	0.9879	0.6558	1.3280	1.2301	0.7994
	IV	0.9879	−0.6111	1.8473	1.6058	0.4422
	V	0.9879	0.4933	1.4108	1.3096	0.7178
	VI	0.9879	0.3502	1.4767	1.3343	0.6638
RF	I	0.8513	0.4239	1.4434	1.3332	0.5612
	II	0.8617	0.4668	1.4235	1.3180	0.6177
	III	0.8490	0.4591	1.4271	1.3459	0.5856
	IV	0.8429	0.5068	1.4043	1.3068	0.6535
	V	0.8524	0.5158	1.3999	1.3177	0.6376
	VI	0.8555	0.5403	1.3879	1.3142	0.6600
XGBRF	I	0.8907	0.3882	1.4597	1.3459	0.5296
	II	0.9274	0.5648	1.3757	1.2535	0.6989
	III	0.8818	0.4767	1.4188	1.3275	0.6187
	IV	0.9153	0.4788	1.4178	1.2947	0.6860
	V	0.9169	0.5268	1.3946	1.3105	0.6389
	VI	0.9181	0.6639	1.3236	1.2546	0.7621

B. Model Performance and Uncertainty

The accuracies of the XGBoost, RF, and XGBRF regression models in predicting the SOC content under six different scenarios are presented in Table IV, where R^2_training and R^2_testing denote the coefficients of determination of the training and testing datasets, respectively. The RMSE, MAE, and LCCC are all calculated from the testing datasets. The evaluation results

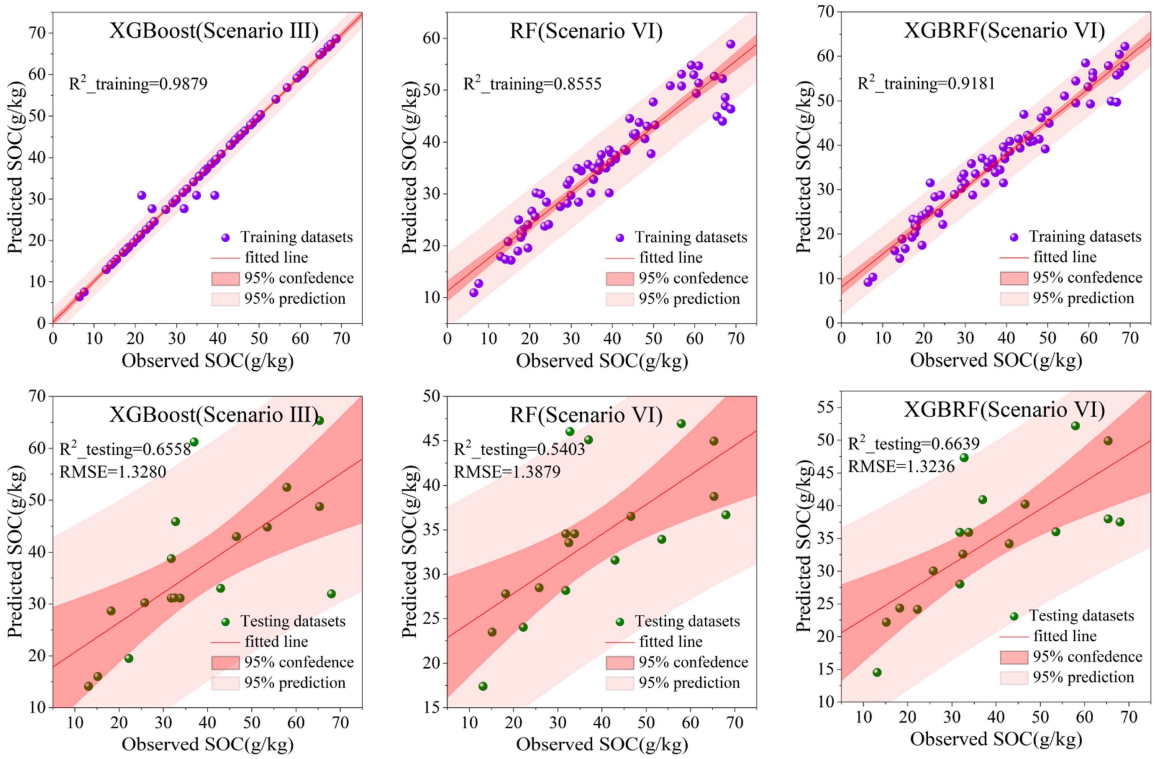


Fig. 2. Scatter plots of the observed SOC and the predicted SOC versus the training and testing datasets based on the XGBoost, RF, and XGBRF models.

reveal the influence of model selection and type and number of variables on the SOC prediction performance.

According to the prediction results of the XGBoost regression model, Scenario III was the best (R^2_testing = 0.6558; RMSE = 1.3280 g/kg; MAE = 1.2301 g/kg; LCCC = 0.7994), and its predictor variables were a combination of S2 and DEM data. In the other scenarios, the accuracy of the training set of the model was high, while the accuracy of the test set was low. The coefficient of determination of the test set of Scenario IV (R^2_training = 0.9879, R^2_testing = −0.6111) was even negative, which indicates that there was a serious over fitting phenomenon in the model under this scenario. This also revealed that the choice of feature variables was crucial to the predictive performance of the model. Overall, in this study, it was found that the prediction results of the XGBoost model exhibited great uncertainty.

The performance of the RF regression model was stable, the floating range of the R^2_testing values was 0.4239–0.5403, and Scenario VI had a better prediction performance. According to the prediction results of the RF model, the accuracy of the SOC prediction can be improved by the fusion of multisource data. The XGBRF regression model performed the most robustly, and the accuracy of Scenario VI was the highest (R^2_testing = 0.6639; RMSE = 1.3236 g/kg; MAE = 1.2546 g/kg; LCCC = 0.7621). There was no obvious over-fitting phenomenon in the predictions obtained using the RF and XGBRF models.

For a single type of predictor variables, for the XGBoost model, the prediction accuracies of the variables extracted from S2 were much higher than those for S1. For the RF and XGBRF models, the prediction accuracies of the variables extracted from S1 were higher than those for S2. The R^2_testing of the model with only DEM variables was less than 0.2, so it is not listed in Table III. When two data sources were used, compared with a single data source, the accuracy of the model fluctuated, improved, or decreased. When three types of prediction variable fusion models were used, the SOC prediction accuracy was effectively improved. For example, for the XGBRF model, the RMSE of Scenario VI decreased by 9.3% and the R^2_testing increased by 70.8% compared with Scenario I. Similar results were obtained for the RF model. These results further indicate the advantages of using multisensor data to predict the SOC content.

Fig. 2 presents scatter plots of the observed SOC and the predicted SOC for the training and testing dataset based on the XGBoost (Scenario III), RF (Scenario VI), and XGBRF (Scenario VI) models. The scatter plots of the training dataset indicate that the performance of the XGBoost (Scenario III) was obviously better than those of the RF (Scenario VI) and XGBRF (Scenario VI), but the performance of this model based on the testing dataset was not the best. By synthesizing the performances of the three models on the training and testing datasets, it was found that the XGBRF (Scenario VI) was the most robust of the three models.

To evaluate the uncertainty of the models, three models with higher prediction accuracy were selected, including XGBoost (Scenario III), RF (Scenario VI), and XGBRF (Scenario VI) models. In addition, we took Huzhong as an example to calculate the CV and evaluate the uncertainty of the models (Fig. 3). Fig. 3

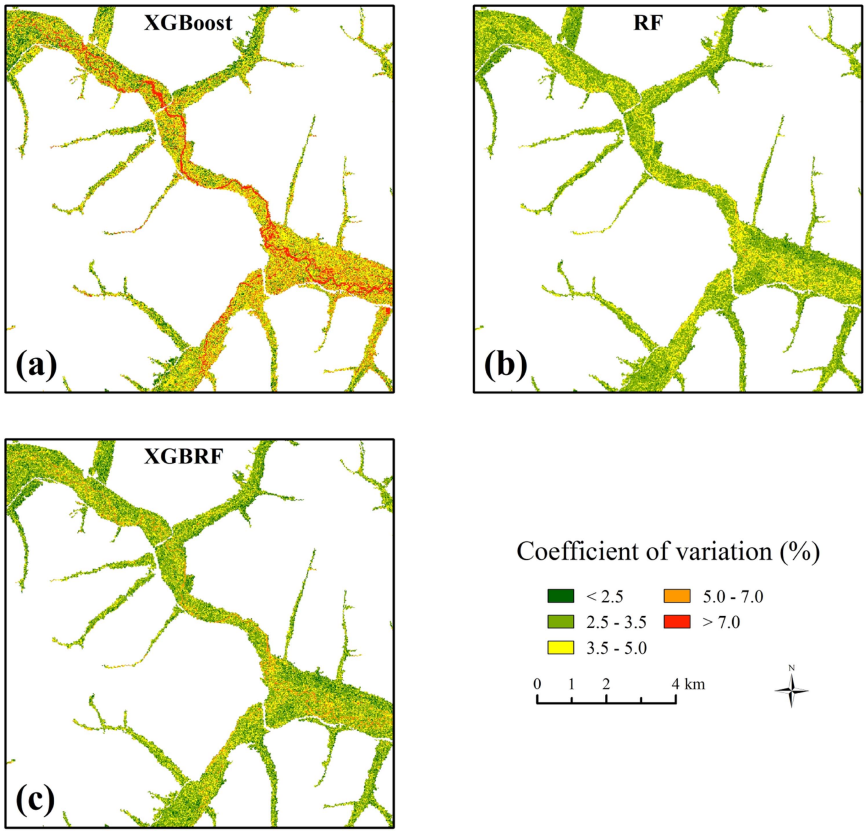


Fig. 3. CV maps for predicting SOC content based on the XGBoost, RF, and XGBRF models.

shows that the CV of XGBoost (Scenario III) was relatively higher (1.7–19.5%), while that of RF (Scenario VI) and XGBRF (Scenario VI) was relatively lower (2.0–7.1% and 1.8–8.9%, respectively). The mean CV values of XGBoost, RF, and XGBRF were 4.34%, 3.28%, and 3.30%, respectively. Therefore, compared with XGBoost, RF and XGBRF produced lower and uniform variations in the mapping region. In addition, we also calculated the PICP to evaluate the probability that the observed values were in the CI. The results showed that the PICP values of XGBoost, RF, and XGBRF were 84.6%, 71.8%, and 79.5%, respectively. None of them reached the expected value of 95%, but XGBoost and XGBRF were superior to RF. According to the values of CV and PICP, XGBRF was more robust than XGBoost and RF, and was more suitable for predicting SOC in this region.

C. Spatial Distribution Maps of SOC

The XGBoost (Scenario III), RF (Scenario VI), and XGBRF (Scenario VI) models were found to have better performances than the others, so we predicted and mapped the SOC content of the swamp wetlands using these three models (Fig. 4). Twenty-seven predictors extracted from the S2 and DEM data were used in the XGBoost model; and all 46 variables were used in the RF and XGBRF models. The spatial distribution characteristics of the SOC content obtained using the RF and XGBRF models were very similar, while the spatial distribution of the SOC obtained using the XGBoost model was very different from the results obtained using the previous two models.

The statistical chart of the SOC content of the swamp wetlands in the two study areas is presented in Fig. 5. In Huzhong, the mean and standard deviation of the SOC content for the three models were 28.9687 g/kg and 10.2299 g/kg for the XGBoost, 33.4016 g/kg and 3.9651 g/kg for the RF, and 34.4360 g/kg and 5.5084 g/kg for the XGBRF, respectively. In Heihe, the mean and standard deviation of the SOC content for the three models were 32.7794 g/kg and 12.0219 g/kg for the XGBoost, 30.6720 g/kg and 4.3513 g/kg for the RF, and 30.4601 g/kg and 4.9982 g/kg for the XGBRF, respectively. For both the SOC spatial distribution maps and the statistical charts, the SOC content of the Huzhong swamp wetland was slightly higher than that of the Heihe swamp wetland. This may be related to their specific climatic and hydrological conditions. The cold and humid conditions in the Huzhong area are more conducive to the storage of SOC.

D. Importance of Predictor Variables

The relative importance of the variables was ranked for the XGBRF and RF models (Fig. 6) (importance expressed in percentage). There was a slight difference in the results of the relative importance ranking of the predictors for the two models, which revealed that there were differences in the dominant predictive variables in the different models. According to the

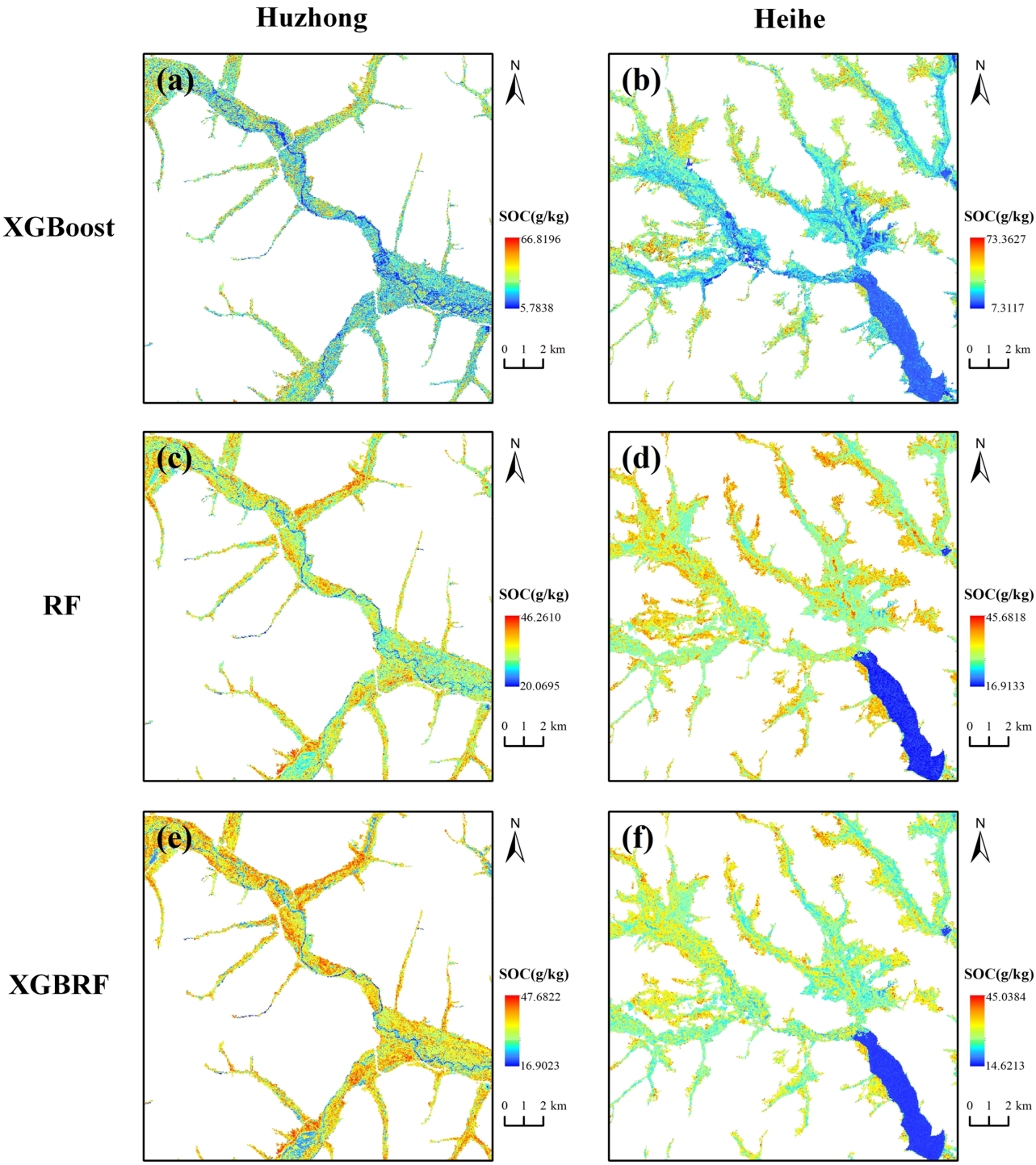


Fig. 4. Spatial distributions of SOC content in the swamp wetlands in the study areas based on the XGBoost, RF, and XGBRF models.

ranking results for the XGBRF model, the S1 and S2 features were the main explanatory variables of the SOC prediction, accounting for 41% and 52% of the total relative importance, respectively, followed by the topographic variables extracted from the DEM (7%). Among the 46 characteristic variables, the top five in terms of importance were B12 (5.43%), GNDVI (3.86%), NDTI (3.75%), VH_Dissimilarity (3.61%), and B7 (3.19%).

The ranking results of the RF model indicate that the S1 variables (47%) and S2 variables (42%) were also the most dominant, followed by the terrain variables (11%), and the variables with significant contributions were B5 (7.02%), VH_Homogeneity (6.07%), TWI (5.98%), GNDVI (5.09%), and VV_Homogeneity (4.49%). Significantly, in the XGBRF and RF models, the relative degrees of importance of the short-wave infrared bands and their derived variables were high,

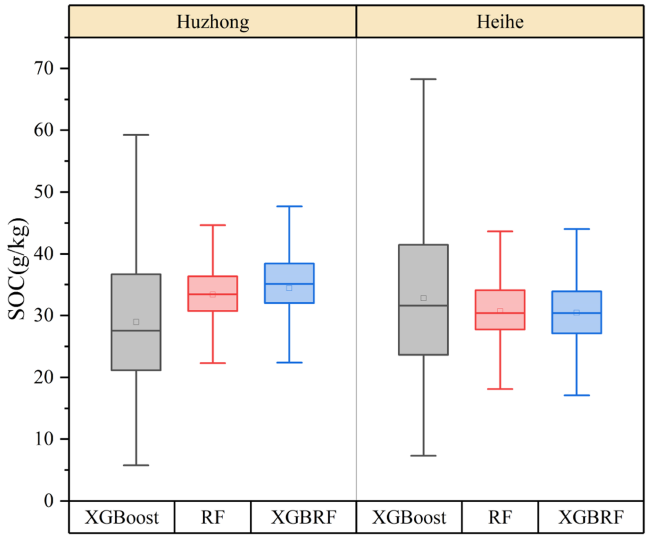
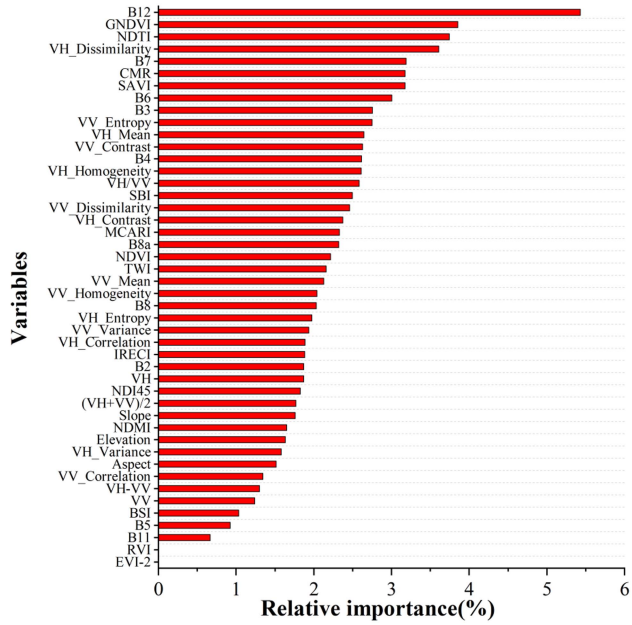
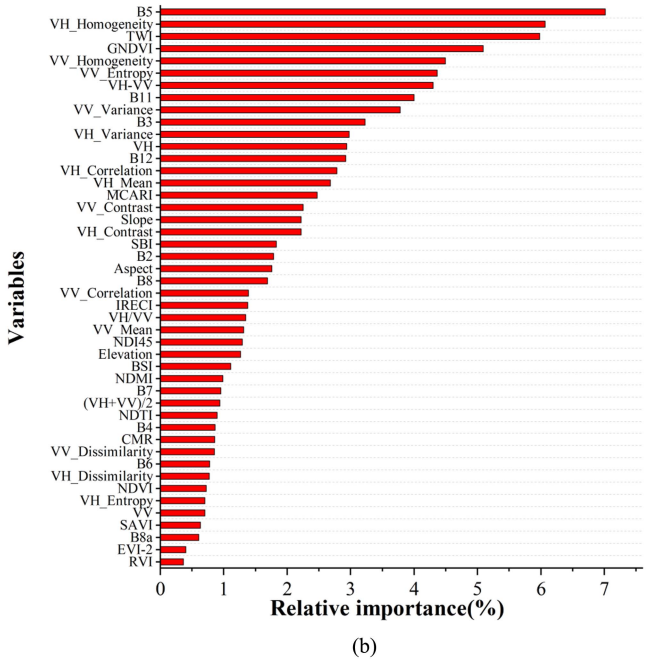


Fig. 5. Statistical chart of the SOC in the swamp wetlands in the study areas.



(a)



(b)

Fig. 6. Relative degrees of importance of the predictor variables. (a) XGBRF. (b) RF.

accounting for 13% and 11%, respectively. The S1 characteristic variables played a significant role (relative importance was greater than 40%). The relative degrees of importance of the topographical wetness index (TWI) were 2.16% and 5.98%. All of these facts indicate the effectiveness of the shortwave infrared bands, SAR data, and topographic factors in predicting the SOC.

IV. Discussion

A. Performances of SOC Prediction Models

The above-mentioned analysis results indicate that the SOC prediction precision was closely related to the selection of the ML algorithms, the types of data sources, and the fusion scenarios of the prediction variables. Overall, the prediction precision and stability of the XGBRF were better than those of the RF and XGBoost models. The results demonstrate that the XGBRF model, which combines the advantages of the RF and XGBoost, can effectively improve the SOC prediction accuracy. Our results are similar to the advantages of the XGBRF in solving classification problems [31]. It was also found that the RF model performed stably and effectively avoided over-fitting, which is consistent with previous research results [54], [55], [56]. In this study, the fitting accuracy of the XGBoost regression model for the training set was very high (R^2_training $= 0.9879$), while the performance in the test set was unstable. The R^2_testing values of the six different scenarios varied from -0.6111 to 0.6558, which is inconsistent with the excellent performance of the XGBoost reported in some studies [30]. The quantity and characteristics of the samples and the choice of the prediction variables may be the direct reasons for this inconsistency. Previous studies have also found that no model performs well in every case [57], so it is essential to use the measured SOC data for the study area to correct prediction models.

In this study, the fusion of S1, S2, and DEM data was of great significance to effectively predicting the SOC content. The performance of the model with fused optical imagery, SAR imagery, and DEM was better than those of the models that used optical imagery data alone. For example, the R^2_testing value of the XGBRF model increased from 0.3882 to 0.6639, and the R^2_testing value of the RF model increased from 0.4239 to 0.5403. Several recent studies have also reported the advantages of multisource data fusion in SOC prediction [22], [26], [58]. It was also found that the prediction accuracy of the model using S1 data alone was also high (R^2_training $= 0.9274$, and R^2-testing $= 0.5648$). This indicates the great potential of the use of SAR

data in SOC prediction, which has also been reported in previous studies [20], [59]. In the future, we can continue to tap the application potential of SAR data, especially in areas that are greatly affected by clouds and rainy weather.

In our study, the results showed that the XGBRF model has robust forecasting ability, especially in densely vegetated areas, good application results have been achieved, which provides a new idea for the application of this model in similar areas. In addition, a number of indicators were used to comprehensively evaluate the accuracy and reliability of the models, and then the best model and data fusion scheme were selected, which also provides a useful reference for the evaluation of SOC prediction model.

B. Variable Importance

In this section, the ranking of the relative degrees of importance of the predictive variables of the XGBRF model are discussed. Among the 23 prediction variables derived from the S2, SWIR1 (B11), and SWIR2 (B12) and their derivative variables, the NDTI and CMR played important roles in the SOC prediction. The sum of the importance of the four variables was as high as 13%. This reflects the fact that the SWIR spectrum is favorable for the detection of SOC, which is in accordance with the results of previous studies [60]. This is also consistent with the correlation results presented in Table III; that is, SWIR1 (B11) and SWIR2 (B12) have strong correlations with the SOC content, with correlation coefficients of 0.272 and 0.220, respectively. The Soil Adjusted Vegetation Index (SAVI) was an important predictor for the SOC retrieval, with an importance of 3.17%, reflecting its high sensitivity to the soil background, which is consistent with the results of previous research [30]. Given that vegetation growth is closely related to soil characteristics, vegetation indexes can capture the changes in soil properties and can be used as effective variables for SOC prediction. Among the multiple vegetation indexes, the Green Normalized Difference Vegetation Index (GNDVI) was the most sensitive variable (with a relative importance of 3.86%). The GNDVI was the calculation result for the Green and NIR bands. Compared with the NDVI (2.22%), it was more sensitive in the SOC prediction, which is in accordance with the findings of previous research [58].

In this study, among the four topographic variables extracted from the DEM data, the Topographic Wetness Index (TWI) played an important role in the SOC prediction, with a relative importance of 2.16%. TWI comprehensively considers the influence of terrain and soil characteristics on soil water distribution and can identify soil water gradient. Soil moisture is an important factor affecting SOC accumulation [54]. Our results also confirm that TWI is an effective variable for the quantitative prediction of SOC.

S1 images were used to predict soil properties by capturing the characteristics of short-term changes in vegetation. The features extracted from the S1 were proven to make an important contribution to improving the prediction accuracy of SOC. In particular, the GLCM texture features of the VV polarization and VH polarization were identified as ideal variables for predicting the SOC, with a contribution rate of 32%. Similarly, the results presented in Table III show that most of the GLCM texture variables have strong correlations with the SOC content, such as the VH_Contrast (0.164) and VH_Entropy (0.208). The findings of this study reveal that the predictor variables extracted from the optical, SAR data, and DEM data were effective in estimating the SOC.

It was found that the relative importance of shortwave infrared bands, SAVI, GNDVI, TWI, and GLCM texture features were higher, which also reflected that SOC content was affected by vegetation, topography, and soil properties. This finding also provided scientific support for the selection of prediction variables.

C. Uncertainty in Current Research

Although XGBRF with multisensor data fusion has been proven to be a good SOC prediction model, there are some uncertainties. First, the quality of remote sensing data determines the prediction accuracy of SOC [61]. However, due to the influence of clouds and the revisit cycle, the collection time of the soil samples did not completely coincide with the remote sensing imaging times of S2 and S1. As a consequence, we should further investigate whether the imaging times of the optical images and the SAR images have an impact on the SOC estimation. Second, the multisource data were derived from different platforms, and there may be some errors in the data conversion process, which could affect the subsequent modeling errors. Third, the results of the model accuracy evaluation have limitations. Owing to the limited sample size, no model tests under different sample size scenarios were performed. It is necessary to collect additional samples in other areas to complete the migration verification of the model.

V. Conclusion

In the study, we proposed and assessed an SOC prediction method based on optical images (S2), SAR data (S1), DEM data, and an advanced ML model (XGBRF). This method was applied to the prediction of the SOC content in swamp wetlands in northeastern China. Overall, the precision and robustness of the XGBRF model were superior to those of the RF and XGBoost models. The predictor variables derived from multisensor data were found to have better prediction performances than those derived from single sensors. The prediction accuracy of the XGBRF, with the fusion of S1, S2, and DEM data, was the highest (R^2_testing = 0.6639, RMSE = 1.3236 g/kg, MAE = 1.2546 g/kg, LCCC = 0.7621). In terms of the degrees of importance of the variables, the S1 and S2 features were the main explanatory variables of the SOC prediction (41% and 52%, respectively), followed by the topographic variables extracted from the DEM data (7%). Importantly, quantitative prediction of the SOC content in swamp wetlands can be achieved using the new framework developed in this study (in the case of small soil sample datasets), but its robustness still needs to be verified in a wider geographical area.

REFERENCES

[1] E. Gorham, "Northern peatlands: Role in the carbon cycle and probable responses to climatic warming," *Ecological Appl.*, vol. 1, no. 2, pp. 182–195, May 1991, doi: 10.2307/1941811.

[2] E. Maltby and P. Immirzi, "Carbon dynamics in peatlands and other wetland soils regional and global perspectives," *Chemosphere*, vol. 27, no. 6, pp. 999–1023, Sep. 1993, doi: 10.1016/0045-6535(93)90065-D.

[3] S. C. Lee, C. J. Fan, Z. Y. Wu, and J. Y. Juang, "Investigating effect of environmental controls on dynamics of CO_2 budget in a subtropical estuarial marsh wetland ecosystem," *Environ. Res. Lett.*, vol. 10, no. 2, pp. 25005–25016, Feb. 2015, doi: 10.1088/1748-9326/10/2/025005.

[4] C. Schillaciet al., "Spatio-temporal topsoil organic carbon mapping of a semi-arid mediterranean region: The role of land use, soil texture, topographic indices and the influence of remote sensing data to modelling," *Sci. Total Environ.*, vol. 601–602, pp. 821–832, Dec. 2017, doi: 10.1016/j.scitotenv.2017.05.239.

[5] A. Lauschet al., "Linking remote sensing and geodiversity and their traits relevant to biodiversity—Part I: Soil characteristics," *Remote Sens.*, vol. 11, no. 20, Oct. 2019, Art. no. 2356, doi: 10.3390/rs11202356.

[6] S. M. O'Rourke, D. A. Angers, N. M. Holden, and A. B. McBratney, "Soil organic carbon across scales," *Glob. Change Biol.*, vol. 21, no. 10, pp. 3561–3574, Apr. 2015, doi: 10.1111/gcb.12959.

[7] F. Castaldi, S. Chabrillat, C. Chartin, V. Genot, A. R. Jones, and B. van Wesemael, "Estimation of soil organic carbon in arable soil in Belgium and Luxembourg with the LUCAS topsoil database," *Eur. J. Soil Sci.*, vol. 69, no. 4, pp. 592–603, Apr. 2018, doi: 10.1111/ejss.12553.

[8] J. M. Soriano-Disla, L. J. Janik, R. A. Viscarra Rossel, L. M. Macdonald, and M. J. McLaughlin, "The performance of visible, near-, and mid-infrared reflectance spectroscopy for prediction of soil physical, chemical, and biological properties," *Appl. Spectrosc. Rev.*, vol. 49, no. 2, pp. 139–186, 2014, doi: 10.1080/05704928.2013.811081.

[9] T. Angelopoulouet al., "Reflectance spectroscopy (Vis-NIR) for assessing soil heavy metals concentrations determined by two different analytical protocols, based on ISO 11466 and ISO 14869-1," *Water, Air, Soil Pollut.*, vol. 228, no. 11, Nov. 2017, Art. no. 436, doi: 10.1007/s11270-017-3609-9.

[10] N. Tziolas, N. Tsakiridis, E. Ben-Dor, J. Theocharis, and G. Zalidis, "A memory-based learning approach utilizing combined spectral sources and geographical proximity for improved VIS-NIR-SWIR soil properties estimation," *Geoderma*, vol. 340, no. 1, pp. 11–24, Apr. 2019, doi: 10.1016/j.geoderma.2018.12.044.

[11] É. F. M. Pinheiro, M. B. Ceddia, C. M. Clingensmith, S. Grunwald, and G. M. Vasques, "Prediction of soil physical and chemical properties by visible and near-infrared diffuse reflectance spectroscopy in the central Amazon," *Remote Sens.*, vol. 9, no. 4, Mar. 2017, Art. no. 293, doi: 10.3390/rs9040293.

[12] L. Liu, M. Ji, and M. Buchroithner, "Combining partial least squares and the gradient-boosting method for soil property retrieval using visible near-infrared shortwave infrared spectra," *Remote Sens.*, vol. 9, no. 12, Dec. 2017, Art. no. 1299, doi: 10.3390/rs9121299.

[13] A. Gholizadeh, D. Žižala, M. Saberioon, and L. Borůvka, "Soil organic carbon and texture retrieving and mapping using proximal, airborne and Sentinel-2 spectral imaging," *Remote Sens. Environ.*, vol. 218, pp. 89–103, Dec. 2018, doi: 10.1016/j.rse.2018.09.015.

[14] T. Angelopoulou, N. Tziolas, A. Balafoutis, G. Zalidis, and D. Bochtis, "Remote sensing techniques for soil organic carbon estimation: A review," *Remote Sens.*, vol. 11, no. 6, Mar. 2019, Art. no. 676, doi: 10.3390/rs11060676.

[15] E. Aldana-Jague, G. Heckrath, A. Macdonald, B. van Wesemael, and K. Van Oost, "UAS-based soil carbon mapping using VIS-NIR (480-1000 nm) multi-spectral imaging: Potential and limitations," *Geoderma*, vol. 275, pp. 55–66, Aug. 2016, doi: 10.1016/j.geoderma.2016.04.012.

[16] N. L. Tsakiridis, N. V. Tziolas, J. B. Theocharis, and G. C. Zalidis, "A GA-based stacking algorithm for predicting soil organic matter from vis–NIR spectral data," *Eur. J. Soil Sci.*, vol. 70, no. 3, pp. 578–590, May 2019, doi: 10.1111/ejss.12760.

[17] M. Nocitaet al., "Soil spectroscopy: An alternative to wet chemistry for soil monitoring," *Adv. Agronomy*, vol. 132, pp. 139–159, 2015, doi: 10.1016/BS.AGRON.2015.02.002.

[18] F. Castaldiet al., "Evaluating the capability of the Sentinel 2 data for soil organic carbon prediction in croplands," *ISPRS J. Photogrammetry Remote Sens.*, vol. 147, pp. 267–282, Dec. 2019, doi: 10.1016/j.isprsjprs.2018.11.026.

[19] M. Ottinger and C. Kuenzer, "Spaceborne L-band synthetic aperture radar data for geoscientific analyses in coastal land applications: A review," *Remote Sens.*, vol. 12, no. 14, Jul. 2020, Art. no. 2228, doi: 10.3390/rs12142228.

[20] R. M. Yang and W. W. Guo, "Using time-series Sentinel-1 data for soil prediction on invaded coastal wetlands," *Environ. Monit. Assessment*, vol. 191, Jun. 2019, Art. no. 462, doi: 10.1007/s10661-019-7580-3.

[21] J. Wanget al., "Estimating leaf area index and aboveground biomass of grazing pastures using Sentinel-1, Sentinel-2 and Landsat images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 154, pp. 189–201, 2019, doi: 10.1016/j.isprsjprs.2019.06.007.

[22] T. Zhou, Y. Geng, J. Chen, J. Pan, D. Haase, and A. Lausch, "High-resolution digital mapping of soil organic carbon and soil total nitrogen using DEM derivatives, Sentinel-1 and Sentinel-2 data based on machine learning algorithms," *Sci. Total Environ.*, vol. 729, no. 8, Apr. 2020, Art. no. 138244, doi: 10.1016/j.scitotenv.2020.138244.

[23] T. Zhouet al., "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.*, vol. 755, no. Part 2, Feb. 2020, Art. no. 142661, doi: 10.1016/j.scitotenv.2020.142661.

[24] E. M. Costa, W. S. Tassinari, H. S. K. Pinheiro, S. J. Beutler, and L. H. C. dos Anjos, "Mapping soil organic carbon and organic matter fractions by geographically weighted regression," *J. Environ. Qual.*, vol. 47, no. 4, pp. 718–725, Jul. 2018, doi: 10.2134/jeq2017.04.0178.

[25] M. Xu, X. Chu, Y. Fu, C. J. Wang, and S. H. Wu, "Improving the accuracy of soil organic carbon content prediction based on visible and near-infrared spectroscopy and machine learning," *Environ. Earth Sci.*, vol. 80, no. 8, Apr. 2021, Art. no. 326, doi: 10.1007/s12665-021-09582-x.

[26] T. Zhou, Y. Geng, J. Chen, M. Liu, and A. Lausch, "Mapping soil organic carbon content using multi-source remote sensing variables in the Heihe River Basin in China," *Ecological Indicator*, vol. 114, Jul. 2020, Art. no. 106288, doi: 10.1016/j.ecolind.2020.106288.

[27] H. Keskin, S. Grunwald, and W. G. Harris, "Digital mapping of soil carbon fractions with machine learning," *Geoderma*, vol. 339, pp. 40–58, Jan. 2019, doi: 10.1016/j.geoderma.2018.12.037.

[28] B. Wanget al., "High resolution mapping of soil organic carbon stocks using remote sensing variables in the semi-arid rangelands of eastern Australia," *Sci. Total Environ.*, vol. 630, pp. 367–378, Jul. 2018, doi: 10.1016/j.scitotenv.2018.02.204.

[29] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovary Data Mining*, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[30] T. T. Nguyenet al., "A novel intelligence approach based active and ensemble learning for agricultural soil organic carbon prediction using multispectral and SAR data fusion," *Sci. Total Environ.*, vol. 804, Jan. 2022, Art. no. 150187, doi: 10.1016/j.scitotenv.2021.150187.

[31] K. R. Bhatele and S. S. Bhadauria, "Glioma segmentation and classification system based on proposed texture features extraction method and hybrid ensemble learning," *Traitement du Signal*, vol. 37, no. 6, pp. 989–1001, Dec. 2020, doi: 10.18280/ts.370611.

[32] J. Rouse, R. H. Haas, J. A. Schell, and D. Deering, "Monitoring vegetation systems in the great plains with ERTS," *NASA Special Publication*, vol. 351, 1974, Art. no. 309.

[33] C. J. Tucker, "Red and photographic infrared linear combinations for monitoring vegetation," *Remote Sens. Environ.*, vol. 8, no. 2, pp. 127–150, Jun. 1979, doi: 10.1016/0034-4257(79)90013-0.

[34] A. A. Gitelson, Y. J. Kaufman, and M. N. Merzlyak, "Use of a green channel in remote sensing of global vegetation from EOS-MODIS," *Remote Sens. Environ.*, vol. 58, no. 3, pp. 289–298, Dec. 1996, doi: 10.1016/S0034-4257(96)00072-7.

[35] Z. Jiang, A. R. Huete, K. Didan, and T. Miura, "Development of a two-band enhanced vegetation index without a blue band," *Remote Sens. Environ.*, vol. 112, no. 10, pp. 3833–3845, Oct. 2008, doi: 10.1016/j.rse.2008.06.006.

[36] J. Delegido, J. Verrelst, L. Alonso, and J. Moreno, "Evaluation of Sentinel-2 red-edge bands for empirical estimation of green LAI and chlorophyll content," *Sensors*, vol. 11, no. 7, pp. 7063–7081, Jul. 2011, doi: 10.3390/s110707063.

[37] A. R. Huete, "A soil-adjusted vegetation index (SAVI)," *Remote Sens. Environ.*, vol. 25, no. 3, pp. 295–309, Aug. 1988, doi: 10.1016/0034-4257(88)90106-X.

[38] W. J. Frampton, J. Dash, G. Watmough, and E. J. Milton, "Evaluating the capabilities of Sentinel-2 for quantitative estimation of biophysical variables in vegetation," *ISPRS J. Photogrammetry Remote Sens.*, vol. 82, pp. 83–92, Aug. 2013, doi: 10.1016/j.isprsjprs.2013.04.007.
[39] C. S. T. Daughtry, C. L. Walthall, M. S. Kim, E. B. de Colstoun, and J. E. McMurtrey, "Estimating corn leaf chlorophyll concentration from leaf and canopy reflectance," *Remote Sens. Environ.*, vol. 74, no. 2, pp. 229–239, Nov. 2000, doi: 10.1016/S0034-4257(00)00113-9.
[40] E. R. Hunt Jr. and B. N. Rock, "Detection of changes in leaf water content using near-and middle-infrared reflectances," *Remote Sens. Environ.*, vol. 30, no. 1, pp. 43–54, Oct. 1989, doi: 10.1016/0034-4257(89)90046-1.
[41] C. D. Elvidge and R. J. P. Lyon, "Influence of rock-soil spectral variation on the assessment of green biomass," *Remote Sens. Environ.*, vol. 17, no. 3, pp. 265–279, Jun. 1985, doi: 10.1016/0034-4257(85)90099-9.
[42] A. P. Van Deventer, A. D. Ward, P. H. Gowda, and J. G. Lyon, "Using thematic mapper data to identify contrasting soil plains and tillage practices," *Photogrammetric Eng. Remote Sens.*, vol. 63, no. 1, pp. 87–93, Jan. 1997, doi: 10.1117/12.277087.
[43] E. J. M. Carranza and M. Hale, "Mineral imaging with Landsat thematic mapper data for hydrothermal alteration mapping in heavily vegetated terrane," *Int. J. Remote Sens.*, vol. 23, no. 22, pp. 4827–4852, Nov. 2002, doi: 10.1080/01431160110115014.
[44] A. Rikimaru, P. S. Roy, and S. Miyatake, "Tropical forest cover density mapping," *Trop. Ecol.*, vol. 43, no. 1, pp. 39–47, Jan. 2002, doi: 10.1.1.465.8749.
[45] T. D. Phamet al., "Comparison of machine learning methods for estimating mangrove above-ground biomass using multiple source remote sensing data in the red river delta biosphere reserve, Vietnam," *Remote Sens.*, vol. 12, no. 8, Apr. 2020, Art. no. 1334, doi: 10.3390/rs12081334.
[46] A. Ibrahem Ahmed Osman, A. Najah Ahmed, M. F. Chow, Y. Feng Huang, and A. El-Shafie, "Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia," *Ain Shams Eng. J.*, vol. 12, no. 2, pp. 1545–1556, Jan. 2021, doi: 10.1016/j.asej.2020.11.011.
[47] T. D. Phamet al., "Estimating mangrove above-ground biomass using extreme gradient boosting decision trees algorithm with fused Sentinel-2 and ALOS-2 PALSAR-2 data in can gio biosphere reserve, Vietnam," *Remote Sens.*, vol. 12, no. 5, Feb. 2020, Art. no. 777, doi: 10.3390/rs12050777.
[48] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
[49] N. M. Baraka, J. Li, N. A. Mustapha, P. Uwamungu, and D. Al-Alimi, "Prediction on the fluoride contamination in groundwater at the Datong Basin, Northern China: Comparison of random forest, logistic regression and artificial neural network," *Appl. Geochemistry*, vol. 132, Jul. 2021, Art. no. 105054, doi: 10.1016/j.apgeochem.2021.105054.
[50] I. Khosravi and S. K. Alavipanah, "A random forest-based framework for crop mapping using temporal, spectral, textural and polarimetric observations," *Int. J. Remote Sens.*, vol. 40, no. 18, pp. 7221–7251, Apr. 2019, doi: 10.1080/01431161.2019.1601285.
[51] S. J. Forghani, M. R. Pahlavan-Rad, M. Esfandiari, and A. M. Torkashvand, "Spatial prediction of WRB soil classes in an arid floodplain using multinomial logistic regression and random forest models, south-east of Iran," *Arabian J. Geosci.*, vol. 13, no. 13, Jun. 2020, Art. no. 543, doi: 10.1007/s12517-020-05576-4.
[52] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, "A survey on ensemble learning," *Front. Comput. Sci.*, vol. 14, pp. 241–258, Aug. 2020, doi: 10.1007/s11704-019-8208-z.
[53] L. I.-K. Lin, "A concordance correlation coefficient to evaluate reproducibility," *Biometrics*, vol. 45, no. 1, pp. 255–268, Mar. 1989, doi: 10.2307/2532051.
[54] M. B. Siewert, "High-resolution digital mapping of soil organic carbon in permafrost terrain using machine learning: A case study in a sub-Arctic peatland environment," *Biogeosciences*, vol. 15, no. 6, pp. 1663–1682, Mar. 2018, doi: 10.5194/bg-15-1663-2018.
[55] Y. Khaledian and B. A. Miller, "Selecting appropriate machine learning methods for digital soil mapping," *Appl. Math. Model.*, vol. 81, no. 1–2, pp. 401–418, Dec. 2020, doi: 10.1016/j.apm.2019.12.016.
[56] J. Padarian, B. Minasny, and A. B. McBratney, "Machine learning and soil sciences: A review aided by machine learning tools," *SOIL*, vol. 6, no. 1, pp. 35–52, Feb. 2020, doi: 10.5194/soil-6-35-2020.
[57] S. Lamichhane, L. Kumar, and B. Wilson, "Digital soil mapping algorithms and covariates for soil organic carbon mapping and their implications: A review," *Geoderma*, vol. 352, pp. 395–413, Jun. 2019, doi: 10.1016/j.geoderma.2019.05.031.
[58] N. N. Leet al., "Learning from multimodal and multi-sensor earth observation dataset for improving estimates of mangrove soil organic carbon in Vietnam," *Int. J. Remote Sens.*, vol. 42, no. 18, pp. 6866–6890, Jun. 2021, doi: 10.1080/01431161.2021.1945158.
[59] R. M. Yang and W. W. Guo, "Modelling of soil organic carbon and bulk density in invaded coastal wetlands using Sentinel-1 imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 82, Oct. 2019, Art. no. 101906, doi: 10.1016/j.jag.2019.101906.
[60] T. D. Phamet al., "Improvement of mangrove soil carbon stocks estimation in north Vietnam using Sentinel-2 data and machine learning approach," *GISci. Remote Sens.*, vol. 58, no. 1, pp. 68–87, Nov. 2021, doi: 10.1080/15481603.2020.1857623.
[61] E. Vaudouret al., "The impact of acquisition date on the prediction performance of topsoil organic carbon from Sentinel-2 for croplands," *Remote Sens*, vol. 11, pp. 643–659, Sep. 2019, doi: 10.3390/rs11182143.

Honghua Zhang was born in Chaoyang, China, in 1979. She received the M.S. degree in mining engineering from the Heilongjiang University of Science and Technology, Harbin, China, in 2007. She is currently working toward the Ph.D. degree in geography with Harbin Normal University, Harbin, China.

Her research interests include environmental remote sensing monitoring, geographic information acquisition, and knowledge mining.

Luhe Wan received the bachelor's degree in geographical science and the master's degree in physical geography from Harbin Normal University, Harbin, China, in 1991 and 2000, respectively, and the Ph.D. degree in information management and information system from Harbin Institute of Technology, Harbin, China, in 2005.

He is a Professor and doctoral supervisor. Since July 2000, he has taught with Harbin Normal University and has been engaged in environmental remote sensing monitoring, spatial information mining, response of swamp wetland to environmental change, and so on.

Yang Li was born in Chifeng, China, in 1997. She received the master's degree in cartography and geographic information systems from Harbin Normal University, Harbin, China, in 2023.

Her research interests cover ecological environment monitoring and model simulation.