land

MDPI

Article

Prediction of Soil Organic Carbon Content in Complex Vegetation Areas Based on CNN-LSTM Model

Zhaowei Dong [1,2], Liping Yao [2], Yilin Bao [2], Jiahua Zhang [1,2,3,*], Fengmei Yao [2], Linyan Bai [1,3] and Peixin Zheng [4]

[1] The Key Laboratory of Earth Observation of Hainan Province, Hainan Aerospace Information Research Institute, Sanya 572000, China; dongzhaowei23@mails.ucas.ac.cn (Z.D.); baily@aircas.ac.cn (L.B.)
[2] College of Earth and Planetary Sciences, University of Chinese Academy of Sciences, Beijing 100049, China; yaoliping22@mails.ucas.ac.cn (L.Y.); baoyilin21@mails.ucas.ac.cn (Y.B.); yaofm@ucas.ac.cn (F.Y.)
[3] Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
[4] Meteorological Information Center of Shanxi, Taiyuan 030002, China; z37533106@163.com
* Correspondence: zhangjh@radi.ac.cn; Tel.: +86-10-8217-8122

Abstract: Synthesizing bare soil pictures in regions with complex vegetation is challenging, which hinders the accuracy of predicting soil organic carbon (SOC) in specific areas. An SOC prediction model was developed in this study by integrating the convolutional neural network and long and short-term memory network (CNN-LSTM) algorithms, taking into consideration soil-forming factors such as climate, vegetation, and topography in Hainan. Compared with common algorithmic models (random forest, CNN, LSTM), the SOC prediction model based on the CNN-LSTM algorithm achieved high accuracy (R^2 = 0.69, RMSE = 6.06 g kg^{-1}, RPIQ = 1.96). The model predicted that the SOC content ranged from 5.49 to 36.68 g kg^{-1}, with Hainan in the central and southern parts of the region with high SOC values and the surrounding areas with low SOC values, and that the SOC was roughly distributed as follows: high in the mountainous areas and low in the flat areas. Among the four models, CNN-LSTM outperformed LSTM, CNN, and random forest models in terms of R^2 accuracy by 11.3%, 23.2%, and 53.3%, respectively. The CNN-LSTM model demonstrates its applicability in predicting SOC content and shows great potential in complex areas where obtaining sample data is challenging and where SOC is influenced by multiple interacting factors. Furthermore, it shows significant potential for advancing the broader field of digital soil mapping.

Keywords: soil organic carbon; deep learning; convolutional neural network; short-term and long-term memory

check for updates

Citation: Dong, Z.; Yao, L.; Bao, Y.; Zhang, J.; Yao, F.; Bai, L.; Zheng, P. Prediction of Soil Organic Carbon Content in Complex Vegetation Areas Based on CNN-LSTM Model. *Land* **2024**, *13*, 915. https://doi.org/10.3390/land13070915

Academic Editor: Chang Liang

Received: 12 May 2024
Revised: 16 June 2024
Accepted: 21 June 2024
Published: 23 June 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Soil organic carbon (SOC) is intricately connected to the worldwide carbon cycle and climate change [1]. SOC is one of the major controls of soil health and soil quality, as well as soil productivity, yield, and food security, so precise monitoring of SOC is crucial for evaluating soil health, improving agricultural methods, and studying variations in soil carbon reserves [2–4]. Accurately estimating SOC is crucial for comprehending the soil organic composition and promoting effective land use and sustainable development [5,6].

Digital soil mapping (DSM) is a technology that uses spatially available covariates and geographic information systems to describe the geographical and temporal distribution of soil properties [7]. It aims to map, assess, and monitor the condition and health of the soil [8–10]. Currently, soil-forming elements such as climate, soil, organism, and topography are widely considered in predicting soil properties using DSM [11–15]. Climatic factors play an essential role in soil formation and development, providing an environmental context for prediction [12]. Organism factors impact the accumulation and alteration of organic matter in the soil [13]. Topographic factors significantly influence hydrological

processes, affecting soil erosion and soil moisture distribution [14]. Soil properties indicate the material source of soil and aid in identifying soil heterogeneity [15].

Compared to areas where bare soil period images are easily accessible, Hainan is limited by vegetation shading, which makes it difficult to obtain bare soil images covering the entire study area, limiting the prediction accuracy of SOC and making DSM more challenging [16]. The results of previous studies confirmed that the DSM of complex areas using multispectral images and high-resolution digital elevation models (DEMs) as input data can achieve good results. However, there are still two main limitations in the current study [17,18]. First, most of the predictive inputs used in complex area studies are short-temporal or even single-temporal data, and the slow rate of change in soil properties makes it difficult for short-term data to accurately predict soil properties [19–21]. The lack of long time-series information may lead to an incomplete understanding of soil properties; second, in complex areas, existing models have limited learning ability for time-series inputs, which restricts the potential to quantify SOC content using multi-source remote sensing information.

Machine learning usually relies on sample and environmental covariate data to fit relationships [22]. However, it may not be beneficial for intricate areas to capture temporal and geographical features of input data. Deep learning models excel in characterizing massive, complicated, and abstract data by effectively extracting nonlinear relationships, making them highly adaptable and promising for applications in the field of DSM [22]. Convolutional neural networks (CNNs) and long short-term memory (LSTM) are two prominent network architectures in deep learning that specialize in handling spatial and temporal input, respectively [23,24]. Utilizing CNN allows for the precise extraction of intricate and abstract characteristics from soil spatial data to address the challenge of identifying complicated soil patterns and linkages [25,26]. LSTM excels in handling sequential time series data, and it can better utilize the rich surface information provided by long time series satellite images to analyze the long-term relationships existing in the time series data [27]. The convolutional neural network and long short-term memory (CNN-LSTM) model, which combines the two, can simultaneously process both spatial and temporal dimensional information, which helps to accurately capture the complex nonlinear relationship between soil properties and surface features and enhance the prediction accuracy of soil properties in complex areas. In recent years, many studies have applied deep learning and machine learning algorithms to DSM. Padarian et al. used CNN and environmental data to predict SOC depth throughout Chile, and the results showed that the prediction error of the CNN model was reduced by 30% compared to the traditional technique that used only covariate point information [25]. Liu et al. added volumetric weight and soil thickness based on readily accessible environmental variables such as vegetation elements and climate factors. Based on the soil attributes, such as vegetation elements and climate factors, the random forest (RF) algorithm simulated the SOC density of grassland in northern China, and the results showed that the R^2 reached 0.73 [28]. Some scholars used the long-term MODIS MCD12Q2 phenology variables to compare and study the effectiveness of the CNN model and the RF model in the prediction of the SOC in Anhui Province of China, and the results showed that the prediction accuracy of the CNN model was better than that of the RF model [29]. In addition, some studies have used LSTM to estimate soil properties based on hyperspectral data inputs, which can be regarded as a type of sequence data [26,30]. However, the CNN-LSTM model has not been used for SOC estimation in complex areas, and its performance in predicting SOC in such areas has yet to be assessed. Additionally, its ability to predict SOC content in complex areas compared to RF, CNN, and LSTM remains unexplored.

In summary, we developed an SOC content prediction method using long-time series satellite data combined with the CNN-LSTM model. Our objectives for this study are (1) to explore the potential of the CNN-LSTM model for DSM in complex regions; (2) to assess the comparative performance of this model against RF, CNN, and LSTM models in

such contexts; and (3) clarify the spatial correlation between soil-forming environmental variables and SOC content.

2. Materials and Methods

2.1. Study Area

The study area is situated on Hainan Island, Hainan Province, China (Figure 1), with latitudes between 18°09′ and 20°10′ N and longitudes from 108°37′ to 111°03′ E, covering a total area of 33,920 km^2 [31]. The region has a tropical monsoon climate with an average annual temperature range of 22 to 26 °C and an average annual rainfall of over 1600 mm. Precipitation decreases from east to west with notable year-to-year variations [32]. Hainan Island's topography is characterized by high elevation in the center and lower elevation in the surrounding regions, mostly mountains and plains, with elevations ranging from 0 to 1813 m. Rice and other crops are predominantly cultivated in the coastal lowlands, while natural forests dominate the mountainous areas in the south-central section [33]. Hainan is the only tropical province in China with over 60% forest coverage, which complicates the monitoring of SOC due to the extensive vegetation [34]. Therefore, an accurate estimation of SOC in Hainan is particularly necessary.

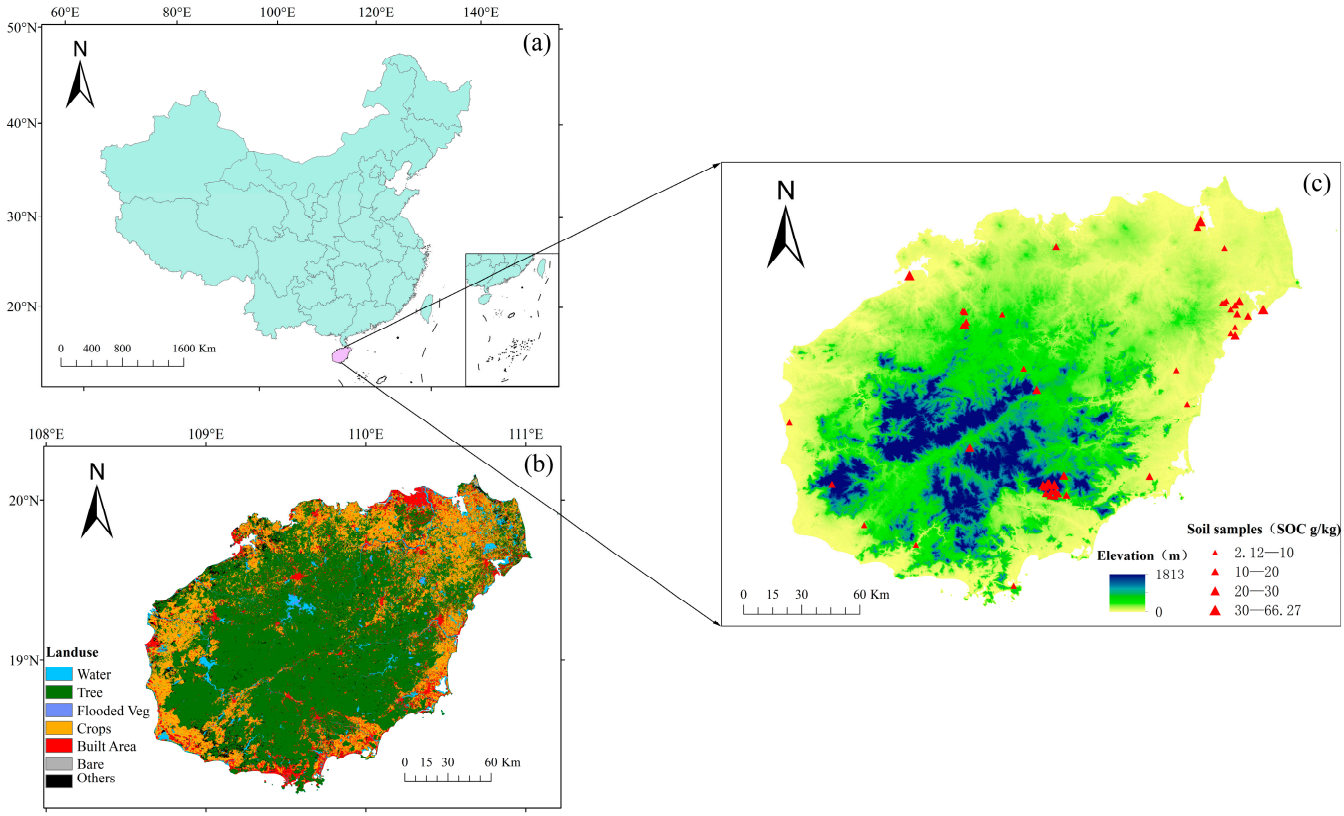


Figure 1. Location of the study area (**a**) and land use (**b**), along with elevation and distribution of sample points (**c**).

2.2. Data Collection and Pre-Processing

The CLORPT soil formation model explains the process of soil formation and its final characteristics and predicts soil properties through several key factors that influence soil formation. The study thoroughly incorporated climate, vegetation, soil, and topography based on the CLORPT model [11]. Given the slow rate of soil change and a long time series of environmental information aids in characterizing the spatial details of soil properties, environmental covariate data before and after five years of soil sampling (2005 to 2015) were selected as input data in this study (Table 1) [35]. This study produced monthly averages

over ten years using meteorological and vegetation data. After conducting covariance and independence analysis, variables exhibiting a Pearson correlation coefficient greater than 0.8 with any other variable were considered redundant and subsequently removed; only factors that showed a significant connection ($p < 0.05$) with SOC content were kept [36]. Seventeen environmental factors were selected as predictors of SOC.

Table 1. Input variables used in this study.

Soil Forming Factor	Input Variables	Abbreviation	Temporal Resolution	Spatial Resolution
Climate	Precipitation	PR	Monthly average	1 km
	Temperature	TA	Monthly average	1 km
Vegetation	Normalized difference vegetation index	NDVI	Monthly average	500 m
	Enhanced vegetation index	EVI	Monthly average	500 m
	Leaf area index	LAI	Monthly average	500 m
Soil	Sand content	Sand	Fixed	250 m
	Silt content	Silt	Fixed	250 m
	Clay content	Clay	Fixed	250 m
	Normalized differential water index	NDWI	Monthly average	500 m
Topographic	Elevation	Elevation	Fixed	30 m
	Slope	Slope	Fixed	30 m
	Aspect	Aspect	Fixed	30 m

The data collected in this study were resampled to a 500 m resolution using bilinear interpolation and then translated to the WGS 1984 UTM 50 N projection system as input.

2.2.1. Soil Sample Data

The study acquired the geographic locations and SOC contents of 73 soil samples from the study region through the Center for Resource and Environmental Science and Data (www.resdc.cn, accessed on 13 March 2023) of the Chinese Academy of Sciences [37]. These samples were taken from the top soil layer (0 to 20 cm), and data were collected around 2010.

2.2.2. Climate Data

The TerraClimate dataset combines detailed climate mean data from World-Clim (https://www.worldclim.org, accessed on 13 August 2023) with more precise monthly data from several sources to create monthly statistics of precipitation, maximum and minimum temperatures, wind speeds, and solar radiation [38]. The study acquired monthly average temperature and precipitation data from the TerraClimate dataset for the study area.

2.2.3. Vegetation Data

The data utilized comprised the normalized difference vegetation index (NDVI), the enhanced vegetation index (EVI), and the leaf area index (LAI) to describe the density and coverage of plants supported by soil [39,40]. All vegetation data were calculated from Terra MODIS (MOD09GA) imagery from the NASA Open Data Platform (https://earthdata.nasa.gov/, accessed on 20 August 2023) [41].

$$NDVI = \frac{(\rho_{nir} - \rho_{red})}{(\rho_{nir} + \rho_{red})} \quad (1)$$

$$EVI = \frac{(\rho_{nir} - \rho_{red})}{(\rho_{nir} + C_{1\times}\rho_{red} - C_2 \times \rho_{blue} + L)} \times G \quad (2)$$

$$LAI = \frac{NDVI}{K} \quad (3)$$

where ρ_{nir}, ρ_{red}, and ρ_{blue} represent the near-infrared, red, and blue bands, respectively, and the parameter coefficients C_1 = 6, C_2 = 7.5, L = 1, G = 2.5, and K are empirical coefficients.

2.2.4. Soil Attribute Data

Soil sand, silt, and clay data at a spatial resolution of 250 m for the research region were acquired via the SoilGrids platform (https://soilgrids.org, accessed on 03 September 2023). To maintain spatial resolution consistency in the data, the soil normalized differential water index (NDWI) was computed from Terra MODIS (MOD09GA) imagery obtained from the NASA Open Data Platform (https://earthdata.nasa.gov/, accessed on 20 August 2023) [41].

$$NDWI = \frac{(\rho_{green} - \rho_{nir})}{(\rho_{green} + \rho_{nir})} \quad (4)$$

where ρ_{green}, ρ_{nir} represent the green and near-infrared bands, respectively.

2.2.5. Topographic Data

Elevation model data were obtained from the shuttle radar topography mission-digital elevation model (SRTM-DEM) with a spatial resolution of 30 m. Slope and aspect were calculated in SAGA GIS.

3. Methodology

This study introduces a CNN-LSTM model designed to extract spatial feature information from topographic and soil data and capture time-varying features from climate, vegetation, and NDWI data. These are the CNN, LSTM, and CNN-LSTM models.

3.1. Convolutional Neural Network (CNN)

The CNN is a deep learning model designed for handling tasks involving grid-structured data. It extracts features from convolutional, pooling, and fully connected layers [42,43]. The structure is seen in Figure 2.

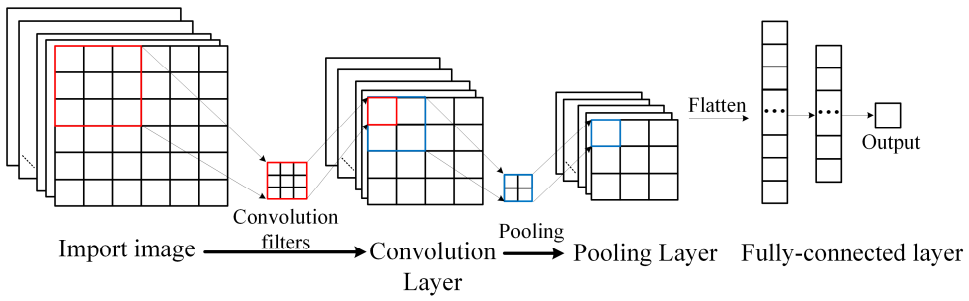


Figure 2. Schematic diagram of the CNN model structure.

The first layer of a CNN is a convolutional layer, which extracts local features of the data by sliding a specific convolutional kernel over the input data. Once the convolution operation is complete, an activation function (e.g., ReLU) is applied to introduce a nonlinear mapping, allowing the network to learn more complex features [44]. The CNN's subsequent layers usually employ a pattern of convolutional and pooling layers, which aid in diminishing the spatial distances of local features and decreasing computing expenses [44]. The fully connected layer includes an attention mechanism that focuses on processing important input variables by calculating pairwise covariances and assigning weights to highlight useful variables and suppress unimportant ones for SOC prediction [45].

A CNN model is created in this study using two convolutional layers, two max-pooling layers, and two fully connected layers within the PyTorch framework. The study employs a convolutional layer with 2 × 2 convolutional kernels and a max-pooling layer with 2 × 2 filters. The study area's complexity may result in significant spatial heterogeneity

in the SOC distribution. Using a 2 × 2 convolutional kernel for predicting SOC content can decrease error rates in convolutional computation and effectively utilize correlations between neighboring images [46]. Additionally, maximal pooling is superior to other types of pooling operations to some degree [47].

3.2. Long Short-Term Memory (LSTM)

As shown in Figure 3, LSTM is a deep learning model designed for handling sequence data, especially well-suited for processing lengthy sequences and data with long-term correlations [48]. LSTM effectively addresses the issue of forgetting and prevents gradient vanishing and explosion when dealing with long sequences [49].

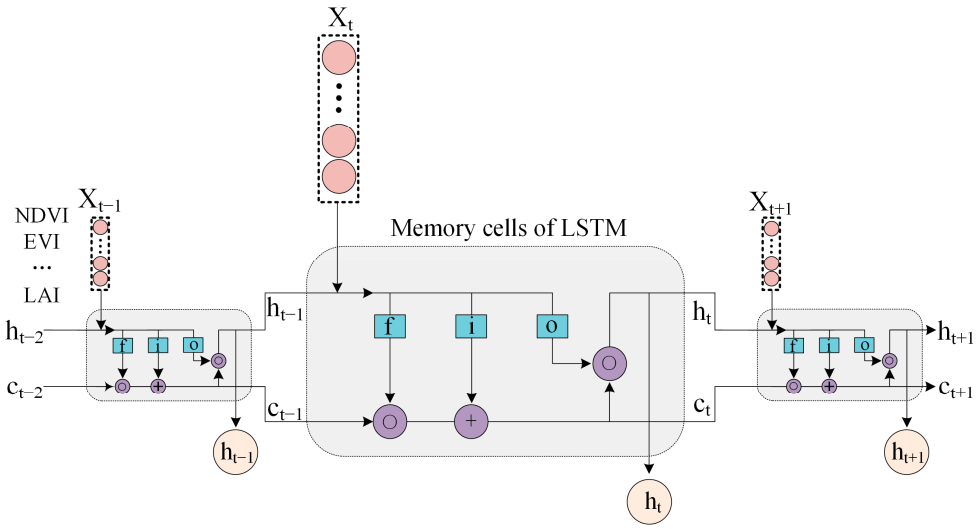


Figure 3. Schematic illustration depicting the architecture of the LSTM model. Where h_{t-2}, h_{t-1}, h_t, h_{t+1} represent the output from time t − 2 to t + 1; c_{t-2}, c_{t-1}, c_t, c_{t+1} represent the memorized information; X_{t-1}, X_t, X_{t+1} represent the input data; f stands for the forgetting gate; i stands for the input gate; and o stands for the output gate.

LSTM consists of memory cells (c), input gates (i), forgetting gates (f), and output gates (o), and each component adds unique weights, biases, and activation functions to the inputs to achieve certain functions [50]. The memory cell (c) functions as a unit responsible for transmitting and storing crucial information. The input gate (i) controls the input information at the present moment. The forgetting gate (f) determines which information to retain or discard by considering the output from the previous time step and the input from the current time step. The previous moment's output (h_{t-1}) and the current moment's input (X_t) are combined and processed through the forgetting gate (f), input gate (i), and output gate (o) to generate the memorized information of the current moment (c_t) and the output information (h_t). This output is then fed back as input for the next moment.

3.3. Convolutional Neural Network–Long Short-Term Memory Model (CNN-LSTM)

As shown in Figure 4, the CNN-LSTM is a deep learning model that integrates the characteristics of CNN and LSTM. The design concept aims to maximize the utilization of CNN's spatial processing and LSTM's temporal modeling to effectively collect input data features and enhance the model's capability to address intricate issues. The design concept aims to maximize the utilization of CNN's spatial processing and LSTM's temporal modeling to effectively collect input data features and enhance the model's capability to address intricate issues.

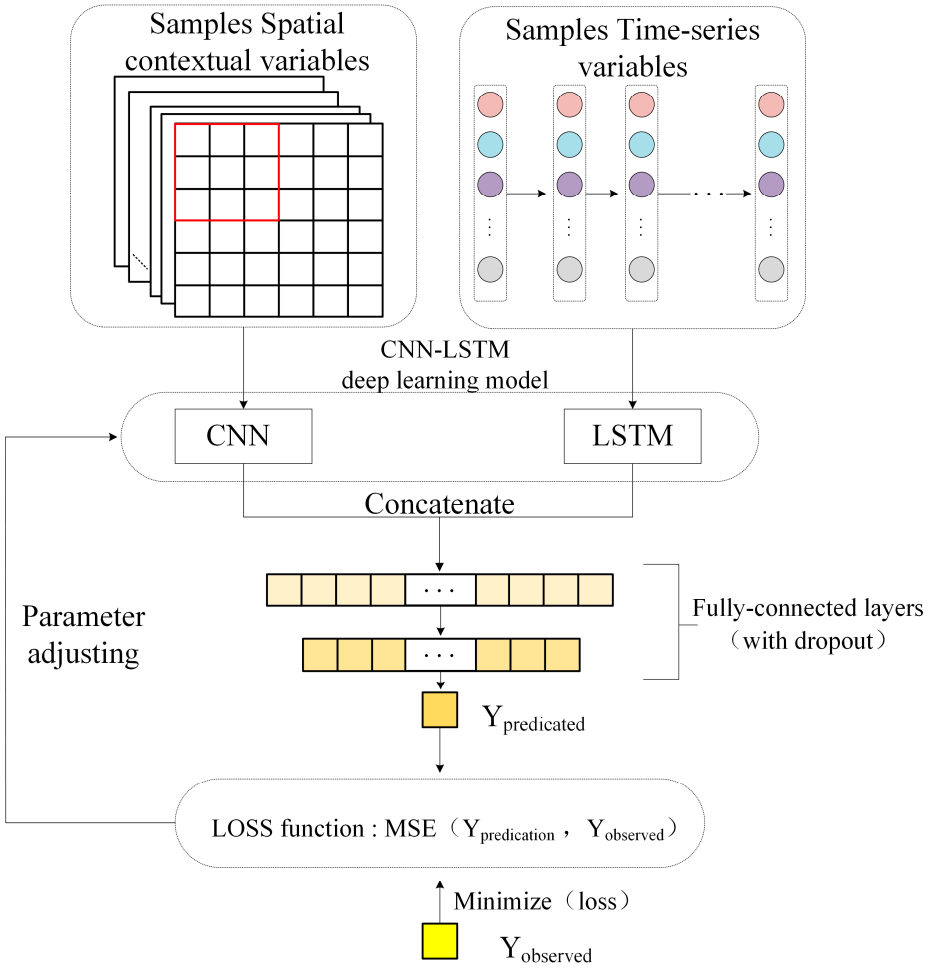


Figure 4. Iterative process flowchart of data in the CNN-LSTM model.

CNN processes topographic and soil data in the study area by convolution, pooling, and other operations to extract spatial features; LSTM analyzes climate vegetation data in the study area using its distinctive gating mechanism to obtain its temporal features. Subsequently, the results processed by CNN and LSTM are aggregated together and further computed through the fully connected layer to generate the final prediction results. Integrating the unique structural benefits of CNN and LSTM enhances the model's ability to capture the factors impacting SOC content and boosts prediction accuracy.

The accuracy of the CNN-LSTM model is limited by various factors. Adjustment of the learning parameter weights (W, U) and bias (b) can enhance prediction accuracy to a certain degree. In this study, the backpropagation algorithm is used to calculate the gradient of the loss function concerning the network parameters, and the gradient information is used to update the network parameters so that the loss function is minimized [23]. This study uses dropout in the fully connected layer and sets the dropout value to 0.25. By randomly dropping a small number of outputs during training, this technique not only effectively avoids the problem of overfitting but also prevents too much information loss that may lead to degradation of learning accuracy. This setup enhances the model's ability to generalize while ensuring effective learning from the data [51].

3.4. Correlation Analysis and Significance Test

This study used the correlation coefficient r to assess the level of linear correlation between the predictor and the predicted variables. The results of the analyses were conducted to determine statistical significance, with a threshold of $p < 0.05$ indicating significance.

$$r = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}} \quad (5)$$

where n represents the number of sample points, x_i is the value of the predicted quantity at point sample i, y_i is the measured SOC content of sample i, and $\overline{x}$, $\overline{y}$ are the average of n predicted and observed SOC values.

3.5. Evaluation of SOC Predictions

The sample data were split into a 2/3 training set and a 1/3 validation set. The root mean square error (RMSE), ratio of performance to interquartile distance (RPIQ), and coefficient of determination (R^2) were used as accuracy evaluation metrics. Among them, RMSE is used to evaluate the absolute error between predicted and true values, RPIQ is used to evaluate the robustness and generalization of the model, and R^2 is used to evaluate the fitting degree of the model. They are calculated as follows:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \quad (6)$$

$$RPIQ = \frac{IQ}{RMSE} \quad (7)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y_i})^2} \quad (8)$$

where n is the number of samples, y_i is the measured SOC content of sample i, and $\hat{y}_i$ is the predicted SOC content of sample i. IQ is the interquartile range, that is, the difference between the third quartile and the first quartile ($IQ = Q_3 - Q_1$). In general, well-performing models usually exhibit high R^2 and RPIQ values and low RMSE values.

3.6. Uncertainty Analysis of Mapping Results

To assess the uncertainty of the prediction model in mapping SOC content prediction results, we used the random forest algorithm to calculate the model uncertainty. Uncertainty reflects the robustness of the prediction model and is key to determining the confidence level of the predictions [52]. In this study, the 10-fold cross-validation method was used to determine the magnitude of uncertainty by predicting the SOC content 10 times and calculating the standard deviation. The smaller the uncertainty, the more stable the model proved to be.

4. Results

4.1. SOC Sample Characterization

From Table 2, the statistics of SOC content in the study area spanned a wide range (2.12 to 66.27 g kg^{-1}), with a total sample mean of 15.07 g kg^{-1} and a coefficient of variation of 74.64%. Most of the sampling sites were situated in the flat regions surrounding the research area, with a few sites in the central mountainous sections (Figure 1).

Table 2. Statistical description of soil samples.

Type	Minimum Value	Maximum Value	Mean Value	Median Value	Coefficient of Variation
Total sample	2.12 g kg^{-1}	66.27 g kg^{-1}	15.07 g kg^{-1}	10.65 g kg^{-1}	74.64%
Training Set	2.12 g kg^{-1}	66.27 g kg^{-1}	15.09 g kg^{-1}	10.75 g kg^{-1}	76.57%
Testing Set	3.24 g kg^{-1}	42.66 g kg^{-1}	15.04 g kg^{-1}	11.37 g kg^{-1}	67.14%

4.2. Correlation between Predictor Variables and SOC Content

All predictor variables and SOC content were analyzed for correlation, and only those that were statistically significant ($p < 0.05$) were retained; the results are shown in Figure 5. The pairs of variables that were not significantly correlated were mainly centered between LAI and clay, silt, slope, precipitation, and temperature, and most of the NDVI also showed non-significant correlations. Among the selected indicators, NDVI, temperature, and LAI data of different months showed a higher correlation with SOC content. This study found that data of the same type from different time periods exhibit highly significant positive correlations while showing weak correlations with other types of data. For example, NDVI values from different months displayed strong positive correlations (r all above 0.7) and weak correlations with other datasets (|r| all below 0.5). There may be consistent patterns in the variation of similar data kinds, while the patterns between distinct data types may not follow a linear trend [53].

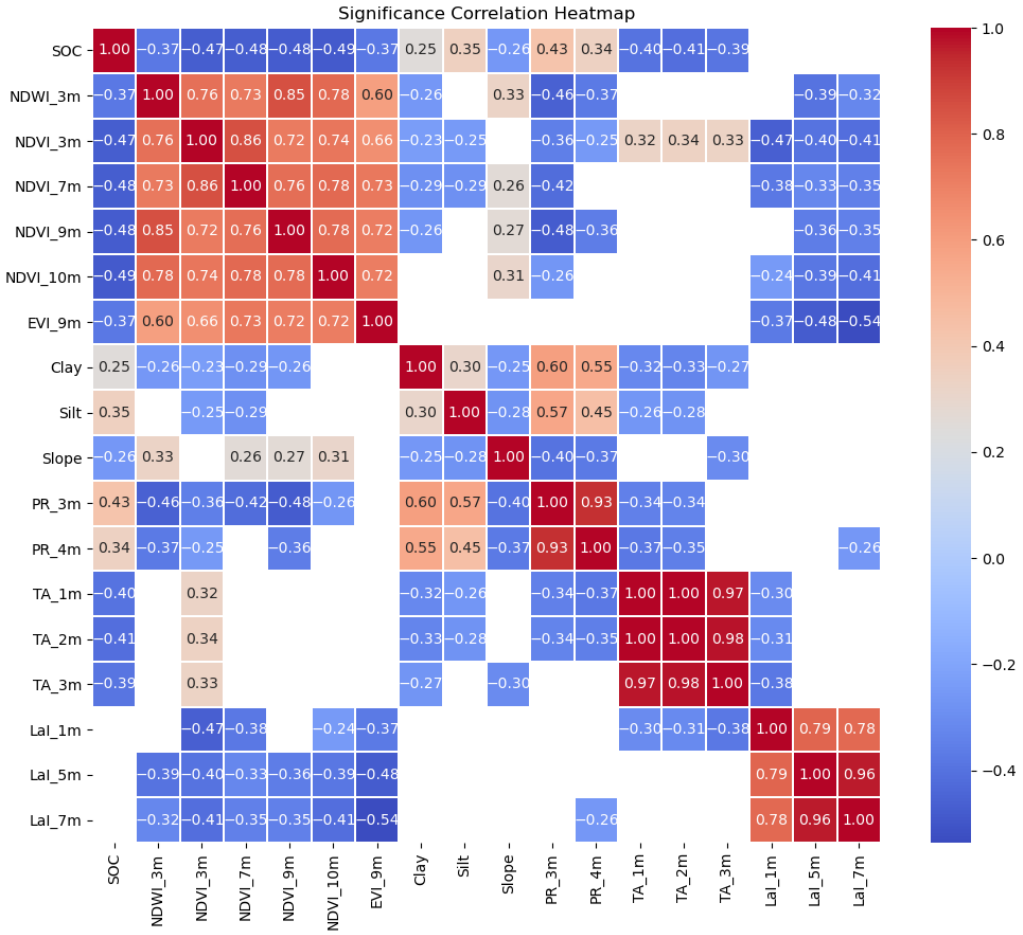


Figure 5. Examining the relationship between SOC in 2010 and the predicted variables in the study region. Dynamic variables are named by combining the variable name with the corresponding month. For example, EVI_9 m represents the EVI data for September following the fusion from 2005 to 2015. The definitions of all variable acronyms may be found in Table 1.

Among the inputs, NDVI showed the strongest association with SOC, with temperature following closely behind. Furthermore, there was a significant link between precipitation and SOC. The study revealed that most input variables were temporally distributed from March to September, coinciding with the seasonal rainfall and plant growing season. It has been shown that the importance of SOC controls varies with depth, with climate dominating the controls of shallow SOC [54], while Cen et al. found that carbon inputs from vegetation can also have a significant impact on surface SOC [55].

4.3. Comparison of Model Prediction Results

4.3.1. CNN-LSTM Model Prediction Results

As shown in Figure 6, the CNN-LSTM model prediction achieved high accuracy, which reached an R^2 of 0.69, RMSE of 6.06 g kg^{-1}, and RPIQ of 1.96. The fit to the actual values indicates that the model can effectively predict SOC content. The error is mainly in the overestimation of low SOC content.

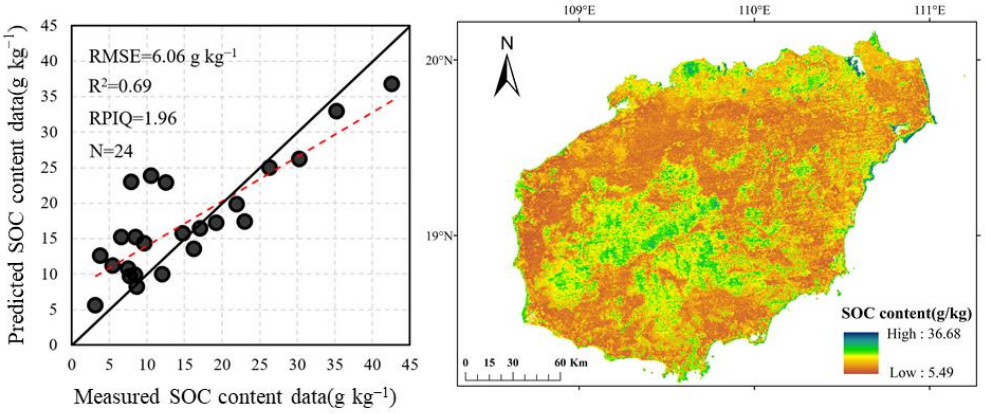


Figure 6. Spatial distribution map of prediction accuracy of the CNN-LSTM model for SOC.

The SOC prediction plot based on all covariates and the CNN-LSTM model is shown in Figure 6. The SOC content in the prediction map showed obvious spatial variability in distribution. The prediction results show that in the central part of the study area, the SOC is higher in the southern mountainous area, while the surrounding areas with gentle terrain have lower SOC content and a small number of high values are also distributed in the northern urban, mixed forest area. In the central and southern mountainous areas, the main land cover type is forest, with dense vegetation cover and complex topography; less human intervention makes the vegetation continuously input organic carbon to the soil, and the mountainous topography may slow down the water and soil loss, which makes the rate of accumulation of SOC in the soil higher. The surrounding gentle areas are dominated by arable land, and more agricultural activities, such as long-term plowing and fertilizer application, may accelerate the decomposition of the SOC content. Moreover, regions with gentle terrain are more prone to rainfall and wind erosion, further reducing SOC content.

4.3.2. Model Accuracy Comparison

The prediction accuracies of CNN, LSTM, and RF models are shown in Figure 7. Among these three models, the LSTM model (R^2 = 0.62, RMSE = 6.84 g kg^{-1}, RPIQ = 1.55) performs the best, followed by CNN (R^2 = 0.56, RMSE = 7.02 g kg^{-1}, RPIQ = 1.54), and the RF model (R^2 = 0.45, RMSE = 7.74 g kg^{-1}, RPIQ = 1.09) performs the worst in terms of both the absolute error and the goodness of fit. Overall, all three models share a common problem of overestimation at low values and underestimation at high values, with the CNN and RF models having particularly pronounced overestimation in the low-value region and the CNN and LSTM having more severe underestimation in the high-value region. The reason may be that the relatively simple structure of these models cannot effectively handle the complex nonlinear relationship between SOC and environmental covariates.

However, the deep learning model still achieves higher accuracy than RF even when some environmental covariates are missing, reflecting the advantage of the deep learning model in considering the changing features of spatial contextual information. The CNN-LSTM model achieved the optimal model accuracy, indicating that combining CNN and LSTM makes the model more powerful in extracting complex spatiotemporal information to deal with spatially and temporally complex features.

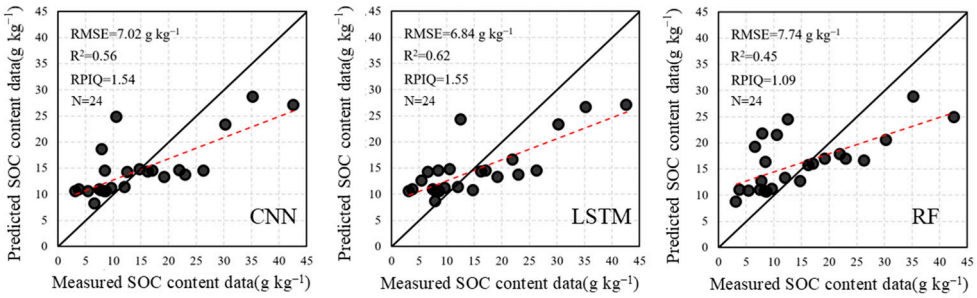


Figure 7. Accuracy comparison of SOC prediction models with CNN, LSTM, and RF methods.

4.3.3. SOC Distribution Characteristics

The spatial distribution of SOC content in Hainan based on CNN, LSTM, and RF models is shown in Figure 8. Combining Figures 7 and 8 with the CNN-LSTM model shows that the above model reveals clear shortcomings in the high SOC value part. Usually, the SOC is higher in the central and southern mountainous areas where forests are the main land cover type, while it is lower in the adjacent plains with mild terrain. However, there are some differences between these prediction maps: the CNN model can discriminate between high and low-value zones, but there is an obvious underestimation of high-value zones. The LSTM model has an overestimation of some low-value zones, and there is also a problem of not perceiving the boundary of the low and high-value zones in the south-central part of the study area, and the RF model performs poorly in general and cannot effectively distinguish the low-value zones from the high-value zones. The limitations of CNN, which focuses on capturing spatial features, and LSTM, which focuses on temporal features, prevent them from utilizing effective information more extensively, which may be the main reason for the inaccurate prediction. The RF model's approach of fitting prediction functions using existing data may need to adequately address the spatial heterogeneity in complex areas, resulting in unsatisfactory predictions for SOC.

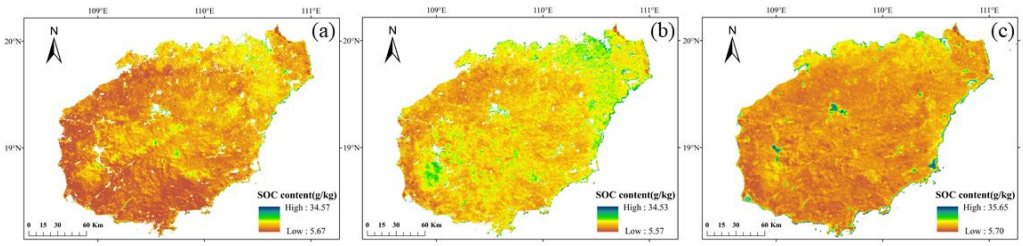


Figure 8. Spatial distribution of SOC content in Hainan, (**a**) CNN model, (**b**) LSTM model, (**c**) RF model.

The maps predicted using CNN and LSTM models appear to be more fragmented compared to CNN-LSTM and RF models. This may be due to the limited number of environmental variables that the model can handle, which results in the model preferring to fit the spatial features of some of the data. Compared to the other predicted maps, the maps generated using the CNN-LSTM model are smoother, suggesting that the model has a significant advantage when dealing with time-series data.

4.4. Uncertainty of Prediction

Figure 9 shows the uncertainty in the results of CNN-LSTM-based prediction of SOC content. We observe that the uncertainties are mainly concentrated in the lower range. Relatively high uncertainties were observed in the central and southwestern parts of the study area. In contrast, the uncertainty is generally lower in the surrounding regions of the study area, with higher uncertainty observed in only a very small portion of the region. It is noteworthy that when observed in conjunction with Figure 6, it can be observed that there are higher uncertainties in areas with high SOC content and lower uncertainties in areas with low SOC content.

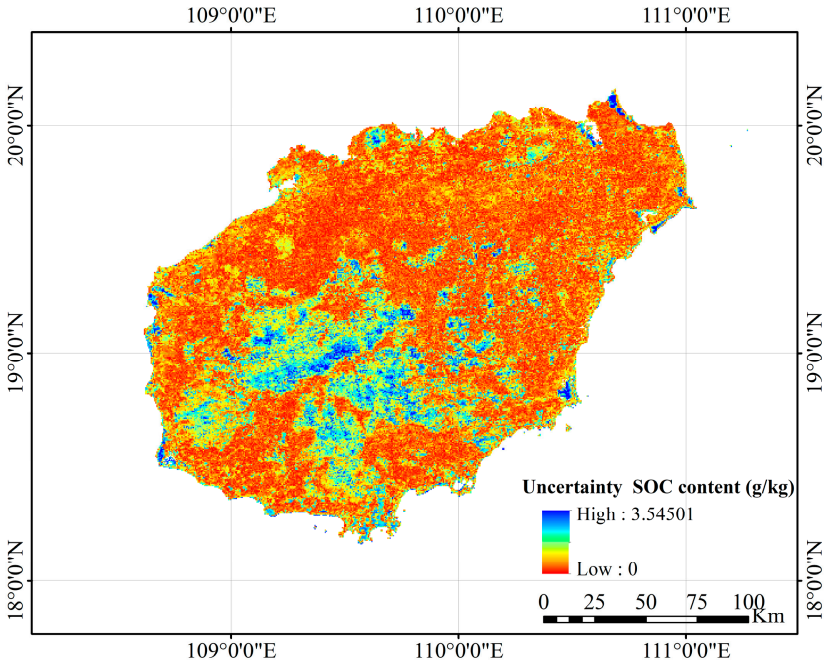


Figure 9. Uncertainty distribution of SOC content based on CNN-LSTM modeling.

4.5. Importance of Variables

The importance of different input variables in the CNN-LSTM model is shown in Figure 10. Our results show that the five input variables with the highest importance are Silt, Clay, TA_1m, TA_2m, and LaI_7m. In addition, we compared the importance of different types of input variables. Our results showed that vegetation had the highest importance among the input data we used (50.78%), followed by meteorology (30.70%). In our study area, Silt became the variable with the highest weight, possibly due to the region's considerable spatial variability of soil texture [56]. Soil texture can show considerable differences between regions and soil types. At the same time, the vegetation type as a whole has the highest weight. This is mainly attributed to the fact that the study area is located in the tropics and has a wide distribution of tropical rainforests, resulting in vegetation being a key factor influencing SOC content [57].

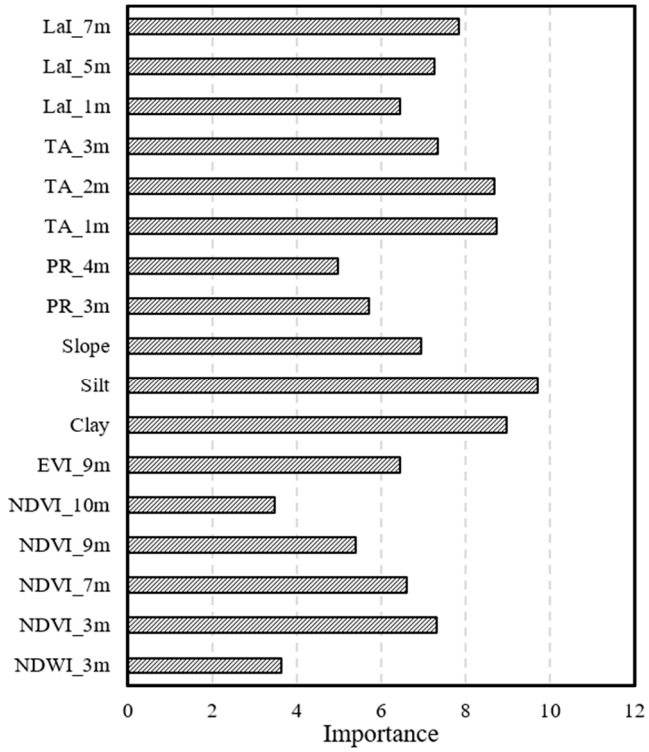


Figure 10. Importance of different input variables for CNN-LSTM models.

5. Discussion

5.1. Advantages of CNN-LSTM Model

In comparison to conventional methods, the approach in this study provides the following advantages. First, the model in this study addresses spatial heterogeneity in complex areas and enhances the ability to understand the characteristics of the data by simultaneously processing information in both spatial and temporal dimensions, taking advantage of the joint use of spatiotemporal information in the research [27]. Compared to a previous SOC estimation study, including Hainan (R^2 = 0.53, RMSE = 6.74 g kg^{-1}), better fitting accuracy was achieved even without distinguishing land use types [58]. Mahmoudzadeh et al. [59] utilized numerous soil samples, auxiliary variables, and five machine learning techniques (K-approximation, RF, extreme gradient boosting, and support vector machine, etc.) to predict SOC for the western region of Iran (29,043 km^2) achieving R^2 values between 0.5 and 0.6. Therefore, it is difficult to estimate the spatial distribution of SOC accurately by relying only on machine learning models with relatively simple structures. The model in this study has a significant advantage in terms of accuracy compared to previous SOC prediction models.

In addition, the model has a significant advantage over CNN, LSTM, and RF methods for SOC prediction in complex areas. The model combines the advantages of CNN and LSTM. The LSTM model can handle more environmental covariates compared to CNN [58] and has a more complex structure and stronger feature learning and representation capabilities compared to the RF model [60]. By using CNN in concert, the LSTM model significantly improves the accuracy of SOC prediction compared with using either model alone.

5.2. Environmental Variables and SOC

By observing Figures 5 and 10 and disaggregating the statistics of the predicted results (Tables A1–A4), a strong correlation can be found between vegetation compared to SOC. The use of vegetation indices is effective in predicting SOC in comparison. Thus, vegetation indices show great potential in improving the accuracy of digital soil mapping. Critical pheno-

logical periods were used for SOC prediction in some studies [19,61], suggesting that plant growth may strongly influence soil carbon inputs. Therefore, long time series of vegetation may be beneficial in improving the accuracy of predicting SOC; the biggest challenge in SOC prediction using long-time vegetation (ten years as used in this study) is the effective extraction of temporal features contained in long time series. The LSTM model can address this challenge as it is specifically designed for time series data.

Studies have shown that changes in soil use type can also significantly alter SOC [62]. Other studies have shown that in a humid climate near the Mediterranean Sea, total soil carbon in the near-surface 10 cm of soil increased by 23% after 20 years of afforestation [63]. Wang et al. showed that in the hilly, semi-arid climate of the Loess Plateau of China, only at the soil surface of the 5-cm depth, different land the organic carbon content of the overburden is only significant [64]. This suggests that climatic conditions can have an important effect on SOC accumulation. In addition, topographic factors can have a profound effect on SOC, with soil erosion in rugged terrain altering the spatial distribution of SOC fixation effects [65] and vegetation regrowth proving to be a significant carbon sink, while erosion may be a source of carbon in erosion-sensitive areas.

5.3. Limitations and Potential Improvements

Although the proposed model has many advantages, as mentioned above, some disadvantages must be considered. There are some shortcomings in the sample data used in this paper. First, the sample dataset is small, resulting in limited feature and geospatial coverage, and a sufficiently large and well-designed number of sample points can more effectively capture geospatial and feature space. For example, Brungard et al. constructed a regionally integrated model for the Upper Colorado River Basin (432,000 km^2) in the western United States using 12,194 soil depth category observations with similar accuracy (56.1% to 75.0%) to the global model accuracy (62.8%) but with less uncertainty [66]. Holmes et al. used nearly 40,000 geographically aligned sites as sample data to focus on improved spatial distribution prediction of the soil coarse fragment layer using digital soil mapping methods in the southwestern region of Australia (1 million km^2), with an overall accuracy of 0.82 to 0.92 [67]. Loiseau et al. improved the spatial distribution prediction of the soil coarse fragment layer by examining the Mayenne region of France (~5000 km^2) by gradually reducing the number of training sites (from 7500 to 400, simulating different observation densities) to test the performance of the ordinary kriging and quartile random forest algorithms. In general, the performance metrics improve with increasing observation density [68]. The results indicate that the main limiting factor in making DSM predictions is the amount of data collected in the field.

Second, it is shown that clustering the sample sets can lead to close samples in the validation and calibration sets, leading to a higher test accuracy of the model than the actual accuracy. The use of methods such as leave-alone and cross-validation can mitigate the overestimation of model accuracy due to the sample sets, to some extent, but they are often ineffective when the data are significant. Therefore, future research should emphasize the acquisition of sample data to obtain higher prediction accuracy. In this paper, the test set range and coefficient of variation are smaller than the validation set, which means this prediction does not predict the data outside the valid range. However, it will also cause the extreme value to not be verified. In addition, some studies have shown that SoilGrids2.0 may produce some bias in the local area [69,70], but the addition of this data can improve the generalization ability of the model to a certain extent [71]; of course, we will also try to use regional data in the future [36].

Finally, seventeen covariates were used for SOC prediction in this paper, which is not in line with the principle of DSM's economy. However, we removed all the data with low correlation and those that did not pass the significance test, and the remaining covariates were retained in order to ensure prediction accuracy.

6. Conclusions

A CNN-LSTM model was created in this research to forecast SOC in Hainan using environmental factors. Utilizing CNN and LSTM concurrently can enhance the accuracy of SOC prediction by extracting spatiotemporal features more effectively than previous models, making it a promising tool for SOC prediction in complex regions. These findings highlight the potential of CNN-LSTM models to be applied in various geographic and environmental contexts. By accurately predicting SOC, they can make a significant contribution to soil management practices and inform decisions that promote sustainable land use and agricultural productivity. In addition, they provide valuable insights into the carbon cycle, which is critical for understanding and mitigating the effects of climate change. Future research should address the limitations identified in this study. Increasing the number and density of in situ observations is essential to improve prediction accuracy. Overall, its strong data processing and precise prediction capabilities make it a valuable tool in soil carbon cycle research and management, providing crucial data for understanding soil ecosystem structure and function for soil management and conservation.

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

Funding: This work was supported by the Finance Science and Technology Project of Hainan Province (No. ZDYF2021SHFZ063), the National Natural Science Foundation of China (No. 42071425), and the Special Educating Project of the Talent for Carbon Peak and Carbon Neutrality of the University of Chinese of Academy of Science (No. 20231106).

Data Availability Statement: The data supporting this study are available from the following resources in the public domain. Soil sample data were downloaded from the Center for Resource and Environmental Science and Data at www.resdc.cn, accessed on 13 March 2023. The Climate data are freely available from the TerraClimate dataset at http://thredds.northwestknowledge.net:8080/thredds/ncss/grid, accessed on 13 August 2023. The Vegetation and NDWI data were obtained from the NASA Open Data Platform at https://earthdata.nasa.gov/, accessed on 20 August 2023. Soil attribute data were freely downloaded from the SoilGrids platform at https://soilgrids.org, accessed on 03 September 2023. Topographic data were derived from the Earth Data platform at https://www.earthdata.nasa.gov/, accessed on 18 September 2023. In addition, all data used in this study are publicly available, which greatly enhances the portability of this study's methodology.

Acknowledgments: The authors are thankful for all the data providers. We also thank Adeel Ahmad for editing the English language.

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

Appendix A

Table A1. Elevation statistics of CNN-LSTM model prediction results.

Type	Minimum Value	Maximum Value	Mean Value	Percentage of Area
0~110 m	5.49 g kg^{-1}	36.68 g kg^{-1}	12.92 g kg^{-1}	52.52%
110~277 m	6.11 g kg^{-1}	33.73 g kg^{-1}	12.51 g kg^{-1}	24.91%
277~515 m	6.45 g kg^{-1}	30.18 g kg^{-1}	16.44 g kg^{-1}	12.51%
515~830 m	9.98 g kg^{-1}	32.73 g kg^{-1}	19.39 g kg^{-1}	7.26%
830~1813 m	12.48 g kg^{-1}	32.38 g kg^{-1}	22.20 g kg^{-1}	2.79%

Table A2. Slope statistics of CNN-LSTM model prediction results.

Type	Minimum Value	Maximum Value	Mean Value	Percentage of Area
0~19.55°	5.48 g kg^{-1}	34.82 g kg^{-1}	13.68 g kg^{-1}	9.64%
19.55~37.01°	5.71 g kg^{-1}	36.68 g kg^{-1}	12.87 g kg^{-1}	17.63%
37.01~52.72°	5.71 g kg^{-1}	34.70 g kg^{-1}	12.69 g kg^{-1}	22.15%
52.72~60.09°	5.74 g kg^{-1}	34.67 g kg^{-1}	13.12 g kg^{-1}	23.68%
60.09~89.04°	6.11 g kg^{-1}	33.73 g kg^{-1}	16.53 g kg^{-1}	26.88%

Table A3. Aspect statistics of CNN-LSTM model prediction results.

Type	Minimum Value	Maximum Value	Mean Value	Percentage of Area
North	5.92 g kg^{-1}	34.62 g kg^{-1}	14.25 g kg^{-1}	14.32%
Northeast	5.49 g kg^{-1}	34.43 g kg^{-1}	14.02 g kg^{-1}	12.83%
East	5.82 g kg^{-1}	34.43 g kg^{-1}	13.86 g kg^{-1}	11.56%
Southeast	5.74 g kg^{-1}	34.67 g kg^{-1}	14.00 g kg^{-1}	11.98%
South	5.71 g kg^{-1}	34.43 g kg^{-1}	13.78 g kg^{-1}	13.02%
Southwest	5.73 g kg^{-1}	34.71 g kg^{-1}	13.82 g kg^{-1}	12.64%
West	5.71 g kg^{-1}	36.68 g kg^{-1}	13.83 g kg^{-1}	11.50%
North	5.49 g kg^{-1}	34.82 g kg^{-1}	14.01 g kg^{-1}	12.15%

Table A4. Land-use of CNN-LSTM model prediction results.

Type	Minimum Value	Maximum Value	Mean Value	Percentage of Area
Evergreen Broadleaf Forests	6.95 g kg^{-1}	33.68 g kg^{-1}	19.39 g kg^{-1}	16.92%
Woody Savannas	6.11 g kg^{-1}	30.22 g kg^{-1}	12.98 g kg^{-1}	27.53%
Savannas	5.49 g kg^{-1}	36.68 g kg^{-1}	12.92 g kg^{-1}	35.21%
Grasslands	5.82 g kg^{-1}	33.16 g kg^{-1}	12.19 g kg^{-1}	5.86%
Croplands	5.72 g kg^{-1}	30.80 g kg^{-1}	11.38 g kg^{-1}	5.23%
Urban and Built-up Lands	5.71 g kg^{-1}	30.46 g kg^{-1}	14.60 g kg^{-1}	1.30%
Cropland/Natural Vegetation Mosaics	5.95 g kg^{-1}	32.04 g kg^{-1}	12.24 g kg^{-1}	6.70%

References

1. Prăvălie, R.; Nita, I.A.; Patriche, C.; Niculiță, M.; Birsan, M.V.; Roșca, B.; Bandoc, G. Global changes in soil organic carbon and implications for land degradation neutrality and climate stability. *Environ. Res.* **2021**, *201*, 111580. [CrossRef] [PubMed]
2. Fujisaki, K.; Chevallier, T.; Chapuis-Lardy, L.; Albrecht, A.; Razafimbelo, T.; Masse, D.; Ndour, Y.B.; Chotte, J.L. Soil carbon stock changes in tropical croplands are mainly driven by carboninputs: A synthesis. *Agric. Ecosyst. Environ.* **2018**, *259*, 147–158. [CrossRef]
3. Georgiou, K.; Jackson, R.B.; Vindušková, O.; Abramoff, R.Z.; Ahlström, A.; Feng, W.; Harden, J.W.; Pellegrini, A.F.; Polley, H.W.; Soong, J.L.; et al. Global stocks and capacity of mineralassociated soil organic carbon. *Nat. Commun.* **2022**, *13*, 3797. [CrossRef]
4. Witzgall, K.; Vidal, A.; Schubert, D.I.; Höschen, C.; Schweizer, S.A.; Buegger, F.; Pouteau, V.; Chenu, C.; Mueller, C.W. Particulate organic matter as a functional soil component for persistent soil organic carbon. *Nat. Commun.* **2021**, *12*, 4115. [CrossRef]
5. Enck, B.F.; Campos, M.C.C.; Pereira, M.G.; de Souza, F.G.; Santos, O.A.Q.; Diniz, Y.V.d.F.G.; Martins, T.S.; Cunha, J.M.; Lima, A.F.L.d.; Souza, T.A.F.d. Forest–Fruticulture Conversion Alters Soil Traits and Soil Organic Matter Compartments. *Plants* **2022**, *11*, 2917. [CrossRef]
6. Ramos, F.T.; Dores, E.F.d.C.; Weber, O.L.d.S.; Beber, D.C.; Campelo Jr, J.H.; Maia, J.C.d.S. Soil organic matter doubles the cation exchange capacity of tropical soil under no-till farming in Brazil. *J. Sci. Food Agric.* **2018**, *98*, 3595–3602. [CrossRef]
7. McBratney, A.B.; Santos, M.M.; Minasny, B. On digital soil mapping. *Geoderma* **2003**, *117*, 3–52. [CrossRef]
8. Chen, S.; Arrouays, D.; Mulder, V.L.; Poggio, L.; Minasny, B.; Roudier, P.; Libohova, Z.; Lagacherie, P.; Shi, Z.; Hannam, J.; et al. Digital mapping of GlobalSoilMap soil properties at a broad scale: A review. *Geoderma* **2022**, *409*, 115567. [CrossRef]
9. Arrouays, D.; Poggio, L.; Guerrero, O.A.S.; Mulder, V.L. Digital soil mapping and GlobalSoilMap. Main advances and ways forward. *Geoderma Reg.* **2020**, *21*, e00265. [CrossRef]
10. Zeraatpisheh, M.; Jafari, A.; Bodaghabadi, M.B.; Ayoubi, S.; Taghizadeh-Mehrjardi, R.; Toomanian, N.; Kerry, R.; Xu, M. Conventional and digital soil mapping in Iran: Past, present, and future. *Catena* **2020**, *188*, 104424. [CrossRef]

11. Jenny, H. *Factors of Soil Formation: A System of Quantitative Pedology*; Courier Corporation: North Chelmsford, MA, USA, 1994.
12. Gibson, A.; Hancock, G.; Bretreger, D.; Cox, T.; Hughes, J.; Kunkel, V. Assessing digital elevation model resolution for soil organic carbon prediction. *Geoderma* **2021**, *398*, 115106. [CrossRef]
13. Jin, F.; Yang, W.; Fu, J.; Li, Z. Effects of vegetation and climate on the changes of soil erosion in the Loess Plateau of China. *Sci. Total Environ.* **2021**, *773*, 145514. [CrossRef]
14. 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]
15. Baldock, J.A. Composition and cycling of organic carbon in soil. In *Nutrient Cycling in Terrestrial Ecosystems*; Springer: Berlin/Heidelberg, Germany, 2007; pp. 1–35.
16. Searle, R.; McBratney, A.; Grundy, M.; Kidd, D.; Malone, B.; Arrouays, D.; Stockman, U.; Zund, P.; Wilson, P.; Wilford, J.; et al. Digital soil mapping and assessment for Australia and beyond: A propitious future. *Geoderma Reg.* **2021**, *24*, e00359. [CrossRef]
17. Phiri, D.; Simwanda, M.; Salekin, S.; Nyirenda, V.; Murayama, Y.; Ranagalage, M. Sentinel-2 data for land cover/use mapping: A review. *Remote Sens.* **2020**, *12*, 2291. [CrossRef]
18. Uuemaa, E.; Ahi, S.; Montibeller, B.; Muru, M.; Kmoch, A. Vertical accuracy of freely available global digital elevation models (ASTER, AW3D30, MERIT, TanDEM-X, SRTM, and NASADEM). *Remote Sens.* **2020**, *12*, 3482. [CrossRef]
19. Van der Putten, W.H.; Bardgett, R.D.; Bever, J.D.; Bezemer, T.M.; Casper, B.B.; Fukami, T.; Kardol, P.; Klironomos, J.N.; Kulmatiski, A.; Schweitzer, J.A.; et al. Plant–soil feedbacks: The past, the present and future challenges. *J. Ecol.* **2013**, *101*, 265–276. [CrossRef]
20. Melillo, J.M.; Frey, S.D.; DeAngelis, K.M.; Werner, W.J.; Bernard, M.J.; Bowles, F.P.; Pold, G.; Knorr, M.A.; Grandy, A.S. Long-term pattern and magnitude of soil carbon feedback to the climate system in a warming world. *Science* **2017**, *358*, 101–105. [CrossRef] [PubMed]
21. Fathololoumi, S.; Vaezi, A.R.; Alavipanah, S.K.; Ghorbani, A.; Saurette, D.; Biswas, A. Improved digital soil mapping with multitemporal remotely sensed satellite data fusion: A case study in Iran. *Sci. Total Environ.* **2020**, *721*, 137703. [CrossRef]
22. Behrens, T.; Schmidt, K.; MacMillan, R.A.; Viscarra Rossel, R.A. Multi-scale digital soil mapping with deep learning. *Sci. Rep.* **2018**, *8*, 15244. [CrossRef]
23. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
24. DiPietro, R.; Hager, G.D. Deep learning: RNNs and LSTM. In *Handbook of Medical Image Computing and Computer Assisted Intervention*; Elsevier: Amsterdam, The Netherlands, 2020; pp. 503–519.
25. Padarian, J.; Minasny, B.; McBratney, A.B. Using deep learning for digital soil mapping. *Soil* **2019**, *5*, 79–89. [CrossRef]
26. Singh, S.; Kasana, S.S. Estimation of soil properties from the EU spectral library using long short-term memory networks. *Geoderma Reg.* **2019**, *18*, e00233. [CrossRef]
27. Zhang, L.; Cai, Y.; Huang, H.; Li, A.; Yang, L.; Zhou, C. A CNN-LSTM model for soil organic carbon content prediction with long time series of MODIS-based phenological variables. *Remote Sens.* **2022**, *14*, 4441. [CrossRef]
28. 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]
29. Yang, L.; Cai, Y.; Zhang, L.; Guo, M.; Li, A.; Zhou, C. A deep learning method to predict soil organic carbon content at a regional scale using satellite-based phenology variables. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *102*, 102428. [CrossRef]
30. Singh, S.; Kasana, S.S. Quantitative estimation of soil properties using hybrid features and RNN variants. *Chemosphere* **2022**, *287*, 131889. [CrossRef] [PubMed]
31. Luo, K. *Collection of Hainan Tropical Agricultural Zoning*; Science Press: Beijing, China, 1985. (In Chinese)
32. Sun, R.; Wu, Z.; Chen, B.; Lan, G.; Qi, D. Change features of precipitation events in Hainan Island from 1959 to 2013. *Chin. J. Trop. Crops* **2017**, *38*, 2176–2182.
33. Liu, Q.; Yang, D.; Cao, L.; Anderson, B. Assessment and prediction of carbon storage based on land use/land cover dynamics in the tropics: A case study of hainan island, China. *Land* **2022**, *11*, 244. [CrossRef]
34. Han, G.; Chen, J.; He, C.; Li, S.; Wu, H.; Liao, A.; Peng, S. A web-based system for supporting global land cover data production. *ISPRS J. Photogramm. Remote Sens.* **2015**, *103*, 66–80. [CrossRef]
35. O'brien, R.M. A caution regarding rules of thumb for variance inflation factors. *Qual. Quant.* **2007**, *41*, 673–690. [CrossRef]
36. Liu, F.; Wu, H.; Zhao, Y.; Li, D.; Yang, J.L.; Song, X.; Shi, Z.; Zhu, A.X.; Zhang, G.L. Mapping high resolution national soil information grids of China. *Sci. Bull.* **2022**, *67*, 328–340. [CrossRef] [PubMed]
37. Xu, L.; Yu, G.; He, N.; Wang, Q.; Gao, Y.; Wen, D.; Li, S.; Niu, S.; Ge, J. Carbon storage in China's terrestrial ecosystems: A synthesis. *Sci. Rep.* **2018**, *8*, 2806. [CrossRef] [PubMed]
38. Abatzoglou, J.T.; Dobrowski, S.Z.; Parks, S.A.; Hegewisch, K.C. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. *Sci. Data* **2018**, *5*, 170191. [CrossRef] [PubMed]
39. 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]
40. Huete, A.; Didan, K.; Miura, T.; Rodriguez, E.P.; Gao, X.; Ferreira, L.G. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sens. Environ.* **2002**, *83*, 195–213. [CrossRef]
41. Gao, B.C. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sens. Environ.* **1996**, *58*, 257–266. [CrossRef]

42. Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 6999–7019. [CrossRef]
43. O'shea, K.; Nash, R. An introduction to convolutional neural networks. *arXiv* **2015**, arXiv:1511.08458.
44. Meng, X.; Bao, Y.; Luo, C.; Zhang, X.; Liu, H. SOC content of global Mollisols at a 30 m spatial resolution from 1984 to 2021 generated by the novel ML-CNN prediction model. *Remote Sens. Environ.* **2024**, *300*, 113911. [CrossRef]
45. Zheng, K.; Yang, D.; Liu, J.; Cui, J. Recognition of teachers' facial expression intensity based on convolutional neural network and attention mechanism. *IEEE Access* **2020**, *8*, 226437–226444. [CrossRef]
46. Pan, L.; Ji, B.; Wang, H.; Wang, L.; Liu, M.; Chongcheawchamnan, M.; Peng, S. MFDNN: Multi-channel feature deep neural network algorithm to identify COVID19 chest X-ray images. *Health Inf. Sci. Syst.* **2022**, *10*, 4. [CrossRef]
47. Scherer, D.; Müller, A.; Behnke, S. Evaluation of pooling operations in convolutional architectures for object recognition. In *International Conference on Artificial Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 92–101.
48. Van Houdt, G.; Mosquera, C.; Nápoles, G. A review on the long short-term memory model. *Artif. Intell. Rev.* **2020**, *53*, 5929–5955. [CrossRef]
49. Sherstinsky, A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Phys. D Nonlinear Phenom.* **2020**, *404*, 132306. [CrossRef]
50. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
51. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
52. Goidts, E.; Van Wesemael, B.; Crucifix, M. Magnitude and sources of uncertainties in soil organic carbon (SOC) stock assessments at various scales. *Eur. J. Soil Sci.* **2009**, *60*, 723–739. [CrossRef]
53. Li, X.; Shang, B.; Wang, D.; Wang, Z.; Wen, X.; Kang, Y. Mapping soil organic carbon and total nitrogen in croplands of the Corn Belt of Northeast China based on geographically weighted regression kriging model. *Comput. Geosci.* **2020**, *135*, 104392. [CrossRef]
54. Jobbágy, E.G.; Jackson, R.B. The vertical distribution of soil organic carbon and its relation to climate and vegetation. *Ecol. Appl.* **2000**, *10*, 423–436. [CrossRef]
55. Cen, Y.; Lou, Y.; Gao, Z.; Liu, W.; Zhang, X.; Sun, G.; Li, Y. Vegetation carbon input moderates the effects of climate change on topsoil organic carbon in China. *Catena* **2023**, *228*, 107188. [CrossRef]
56. Fukumasu, J.; Jarvis, N.; Koestel, J.; Kätterer, T.; Larsbo, M. Relations between soil organic carbon content and the pore size distribution for an arable topsoil with large variations in soil properties. *Eur. J. Soil Sci.* **2022**, *73*, e13212. [CrossRef]
57. Bao, Y.; Yao, F.; Meng, X.; Fan, J.; Zhang, J.; Liu, H.; Mouazen, A.M. Dynamic modeling of topsoil organic carbon and its scenarios forecast in global Mollisols regions. *J. Clean. Prod.* **2023**, *421*, 138544. [CrossRef]
58. Wang, L.; Wang, X.; Kooch, Y.; Song, K.; Zheng, S.; Wu, D. Remote estimation of soil organic carbon under different land use types in agroecosystems of Eastern China. *Catena* **2023**, *231*, 107369. [CrossRef]
59. 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]
60. Shinde, P.P.; Shah, S. A review of machine learning and deep learning applications. In Proceedings of the 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 16–18 August 2018; pp. 1–6.
61. Yang, L.; Song, M.; Zhu, A.X.; Qin, C.; Zhou, C.; Qi, F.; Li, X.; Chen, Z.; Gao, B. Predicting soil organic carbon content in croplands using crop rotation and Fourier transform decomposed variables. *Geoderma* **2019**, *340*, 289–302. [CrossRef]
62. Guo, L.B.; Gifford, R.M. Soil carbon stocks and land use change: A meta analysis. *Glob. Chang. Biol.* **2002**, *8*, 345–360. [CrossRef]
63. Del Galdo, I.; Six, J.; Peressotti, A.; Francesca Cotrufo, M. Assessing the impact of land-use change on soil C sequestration in agricultural soils by means of organic matter fractionation and stable C isotopes. *Glob. Chang. Biol.* **2003**, *9*, 1204–1213. [CrossRef]
64. Wang, Y.; Fu, B.; Lü, Y.; Song, C.; Luan, Y. Local-scale spatial variability of soil organic carbon and its stock in the hilly area of the Loess Plateau, China. *Quat. Res.* **2010**, *73*, 70–76. [CrossRef]
65. Wang, Y.; Fu, B.; Lü, Y.; Chen, L. Effects of vegetation restoration on soil organic carbon sequestration at multiple scales in semi-arid Loess Plateau, China. *Catena* **2011**, *85*, 58–66. [CrossRef]
66. Brungard, C.; Nauman, T.; Duniway, M.; Veblen, K.; Nehring, K.; White, D.; Salley, S.; Anchang, J. Regional ensemble modeling reduces uncertainty for digital soil mapping. *Geoderma* **2021**, *397*, 114998. [CrossRef]
67. Holmes, K.W.; Griffin, E.A.; van Gool, D. Digital soil mapping of coarse fragments in southwest Australia: Targeting simple features yields detailed maps. *Geoderma* **2021**, *404*, 115282. [CrossRef]
68. Loiseau, T.; Arrouays, D.; Richer-de Forges, A.C.; Lagacherie, P.; Ducommun, C.; Minasny, B. Density of soil observations in digital soil mapping: A study in the Mayenne region, France. *Geoderma Reg.* **2021**, *24*, e00358. [CrossRef]
69. Richer-De-Forges, A.C.; Arrouays, D.; Poggio, L.; Songchao, C.; Lacoste, M.; Minasny, B.; Libohova, Z.; Roudier, P.; Mulder, V.L.; Nédélec, H.; et al. Hand-feel soil texture observations to evaluate the accuracy of digital soil maps for local prediction of soil particle size distribution: A case study in Central France. *Pedosphere* **2023**, *33*, 731–743. [CrossRef]

70. Tifafi, M.; Guenet, B.; Hatté, C. Large differences in global and regional total soil carbon stock estimates based on SoilGrids, HWSD, and NCSCD: Intercomparison and evaluation based on f ield data from USA, England, Wales, and France. *Glob. Biogeochem. Cycles* **2018**, *32*, 42–56. [CrossRef]
71. Hengl, T.; Mendes de Jesus, J.; Heuvelink, G.B.; Ruiperez Gonzalez, M.; Kilibarda, M.; Blagotić, A.; Shangguan, W.; Wright, M.N.; Geng, X.; Bauer-Marschallinger, B.; et al. SoilGrids250m: Global gridded soil information based on machine learning. *PLoS ONE* **2017**, *12*, e0169748. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.