land

MDPI

Article

Digital Mapping of Soil Organic Carbon Using Machine Learning Algorithms in the Upper Brahmaputra Valley of Northeastern India

Amit Kumar [1,2,*], Pravash Chandra Moharana [3], Roomesh Kumar Jena [4], Sandeep Kumar Malyan [5], Gulshan Kumar Sharma [6], Ram Kishor Fagodiya [7], Aftab Ahmad Shabnam [1], Dharmendra Kumar Jigyasu [1], Kasthala Mary Vijaya Kumari [1] and Subramanian Gandhi Doss [2]

1 Central Muga Eri Research and Training Institute, Lahdoigarh, Jorhat 785700, Assam, India; jigyasudk.csb@gov.in (D.K.J.); aftab.csb@gov.in (A.A.S.); cmertilad.csb@nic.in (K.M.V.K.)
2 Central Sericultural Research and Training Institute, Mysuru 570008, Karnataka, India; sgdoss.csb@nic.in
3 ICAR-National Bureau of Soil Survey and Land Use Planning, Nagpur 440033, Maharashtra, India; pravash.moharana@icar.gov.in
4 ICAR-Indian Institute of Water Management, Bhubaneswar 751023, Odisha, India; roomesh.jena@icar.gov.in
5 Department of Environmental Studies, Dyal Singh Evening College, University of Delhi, New Delhi 110003, India; sandeepkmalyan@gmail.com
6 ICAR-Indian Institute of Soil and Water Conservation, Research Centre, Kota 324002, Rajasthan, India; gulshan.sharma@icar.gov.in
7 ICAR-Central Soil Salinity Research Institute, Karnal 132001, Haryana, India; ram.iari4874@gmail.com or ram.fagodiya@icar.gov.in
* Correspondence: amitkumar.csb@gov.in or amit_bio80@yahoo.com

Abstract: Soil Organic Carbon (SOC) is a crucial indicator of ecosystem health and soil quality. Machine learning (ML) models that predict soil quality based on environmental parameters are becoming more prevalent. However, studies have yet to examine how well each ML technique performs when predicting and mapping SOC, particularly at high spatial resolutions. Model predictors include topographic variables generated from SRTM DEM; vegetation and soil indices derived from Landsat satellite images predict SOC for the Lakhimpur district of the upper Brahmaputra Valley of Assam, India. Four ML models, Random Forest (RF), Cubist, Extreme Gradient Boosting (XGBoost), and Support Vector Machine (SVM), were utilized to predict SOC for the top layer of soil (0–15 cm) at a 30 m resolution. The results showed that the descriptive statistics of the calibration and validation sets were close enough to the total set data and calibration dataset, representing the complete samples. The measured SOC content varied from 0.10 to 1.85%. The RF model's performance was optimal in the calibration and validation sets (R^2c = 0.966, $RMSE_c$ = 0.159%, R^2v = 0.418, $RMSE_v$ = 0.377%). The SVM model, on the other hand, had the next-lowest accuracy, explaining 47% of the variation (R^2c = 0.471, RMSEc = 0.293, R^2v = 0.081, RMSEv = 0.452), while the Cubist model fared the poorest in both the calibration and validation sets. The most-critical variable in the RF model for predicting SOC was elevation, followed by MAT and MRVBF. The essential variables for the Cubist model were slope, TRI, MAT, and Band4. AP and LS were the most-essential factors in the XGBoost and SVM models. The predicted OC ranged from 0.44 to 1.35%, 0.031 to 1.61%, 0.035 to 1.71%, and 0.47 to 1.36% in the RF, Cubist, XGBoost, and SVM models, respectively. Compared with different ML models, RF was optimal (high accuracy and low uncertainty) for predicting SOC in the investigated region. According to the present modeling results, SOC may be determined simply and accurately. In general, the high-resolution maps might be helpful for decision-makers, stakeholders, and applicants in sericultural management practices towards precision sericulture.

Keywords: environmental covariates; predictive mapping; random forest; sericulture soil; digital SOC map

Citation: Kumar, A.; Moharana, P.C.; Jena, R.K.; Malyan, S.K.; Sharma, G.K.; Fagodiya, R.K.; Shabnam, A.A.; Jigyasu, D.K.; Kumari, K.M.V.; Doss, S.G. Digital Mapping of Soil Organic Carbon Using Machine Learning Algorithms in the Upper Brahmaputra Valley of Northeastern India. *Land* **2023**, *12*, 1841. https://doi.org/10.3390/land12101841

Academic Editors: Nick B. Comerford, Abdul M. Mouazen, Antonio Comparetti and Santo Orlando

Received: 19 July 2023
Revised: 30 August 2023
Accepted: 18 September 2023
Published: 27 September 2023

1. Introduction

Soil Organic Carbon (SOC) imparts a significant role in greenhouse gas dynamics by acting as a sink and source based on the prevailing conditions and soil management. Soil is the most-extensive carbon storage system in the terrestrial ecosystem. Thus, the global scientific community requires an understanding of the dynamics of soil organic carbon. As a result, there is a growing demand for SOC information worldwide [1,2]. The best depiction of SOC information is high-resolution soil maps, which are difficult to prepare at the global level simultaneously, and SOC information needs to be continuously upgraded as and when the region-specific information is available. Recently, soil mapping has been achieved through ground-based surveys; the spatial distribution of the soil properties of a landscape is difficult to specify at an appropriate resolution through these traditional mapping methods. Thus, developing a more-accurate and -reliable methodology and process is highly required to forecast the soil properties of specific soil types or locations.

In line with this, Digital Soil Mapping (DSM) is a highly promising methodology combining two advanced technologies, i.e., machine learning and remote sensing, such as Hyper spectral, multispectral, and radar [3,4]. DSM is in high demand globally to map region-specific soil qualities and is influential in achieving various sustainability goals, including land-use management. DSM's vital machine learning (ML) techniques are artificial neural networks, decision trees, linear models, multivariate adaptive regression splines, regression trees, and support vector machines [5,6]. The statistical relation among the soil, environmental variables, satellite information, topographic characteristics such as digital elevation models, etc., is the merit of these methodologies, which have many advantages over the traditional methods of soil mapping. Due to these advantages, there has been a surge in DSM investigations, especially regarding the spatial variation in soil properties observed in recent decades [7,8]. Hengl et al. [9] used DSM at a resolution of 1 km in Africa to forecast soil parameters such as organic carbon, pH, sand, silt, and clay fractions. Hengl et al. [10] used DSM at a resolution of 250 m to estimate SOC, pH, texture, and bulk density on a global scale. The *GlobalSoilMap* consortium is one of the best platforms to utilize the historical, as well as the recent information of the different regions and convert it into synchronized global information for further utilization in policy formulation, infrastructural development, biodiversity conservation, disaster risk assessment, etc. [11,12].

Northeastern India is rich in flora and fauna, has specific climates, high endemism, and is one of the most-vulnerable ecoregions of the globe [13]. The Brahmaputra Valley region has a unique status among the different ecosystems of Northeastern India. For example, the world's most-expensive golden silk, muga, is endemic to this region and has high traditional and ethical values among tribal and non-tribal inhabitants [14]. Muga silkworm rearing is an entirely outdoor activity, i.e., rearing is carried out on the host plant. Som is the primary host plant for muga silk production. Assam contributes > 90% of the muga silk production [14]. Lakhimpur is the highest muga-silk-producing district among all the producing districts in the world. The Brahmaputra Valley, especially the upper part, loses Soil Organic Carbon (SOC) due to enhanced land-use land-cover changes, deforesttion, flooding, leftover fallow land, etc. SOC is highly required to understand the carbon pool dynamic under various agricultural and forest ecosystems. Machine learning can serve as an essential tool to investigate and estimate the SOC at the landscape level by using lower physical inputs than grid-based physical mapping. It also reduces the uncertainty level, which is higher in grid-based physical mapping. Keeping this view, the current investigation was planned using 160 endemic silk (muga and eri)-specific soil organic datasets of the Lakhimpur district of Assam, which bears the first rank in Muga silk production globally. The aims of the investigation were (i) to prepare a digital SOC map with different ML algorithms, (ii) to evaluate the model efficiency, and (iii) to identify the most-influencing environmental covariates in SOC distribution under silk-producing soils.

2. Materials and Methods

2.1. Study Area

Lakhimpur is located in the northeast corner of Assam State and between 26°48′ and 27°53′ northern latitude and 93°42′ and 94°20′ east longitude (Figure 1). Papump and Siang (Arunachal Pradesh), Dhemji and Majuli (Assam), and Bishwanath (Assam) are situated on north, east, south, and west, respectively. The Brahmaputra and Subansiri Rivers flow east and south of Lakhimpur. Lakhimpur's total geographical area is about 2277 km^2. The rural and urban areas are 2257 and 20 km^2, respectively. Lakhimpur's climate is subtropical and has hot, humid summers and cold winters. The mean annual and soil temperatures are 24.5 °C and 22 °C, respectively. The mean annual precipitation ranges from 2000–4000 mm. The dominant soil of Lakhimpur is alluvial. Lakhimpur contributes to all major types of silk and is the highest in muga silk production among all the districts in India.

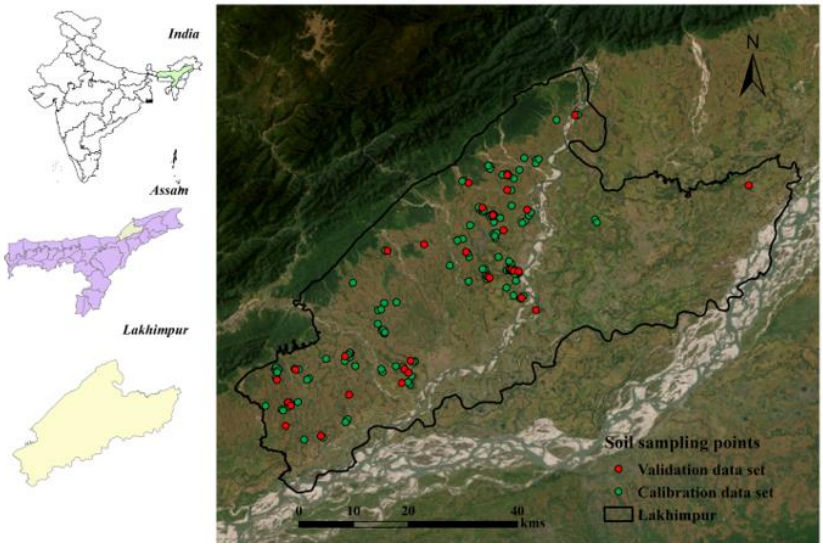

Figure 1. Spatial distribution of soil sampling points in the study area of the Upper Brahmaputra Valley in northeast India.

2.2. Soil Sampling and Analysis

The topsoil samples (0–15 cm) were collected through a v-notch following systematic randomized sampling from the study region. The initial cleaning was opted to be performed at each sampling point to ensure a debris-free soil sample. The composite samples were prepared in the field through the proportionate method (200 g from each sample) to represent the complete field condition. There were a total of 10 samples that contributed to each composite sample. The minimum quantity of the replicate was restricted to 500 g. One-hundred sixty composite samples were collected from the muga and eri sericulture farms. All the composite replicates were transported to the laboratory. Each replicate was air-dried in the shade and sieved with a 2 mm sieve. The wet oxidation method was used to estimate the SOC in the soil [15] as the OC recovery range was reported to be 60–86% with an average recovery of 76% [15].

2.3. Digital Soil Mapping Technique

The SCORPAN model deduces soil class/properties at a specific place indirectly from environmental factors such as age (a), climate (c), geographical position (n), organisms (o), parent material (p), and relief (r) [16]. The SCORPAN model provides the quantitative (empirical) relationship among the soil characteristics and variables that control spatial variability. The SCORPAN model follows Dokuchaev's [17] and Jenny's [18] theories.

$$S = f(s, c, o, r, p, a, n) \quad (1)$$

Soil can be used as the input data for the SCORPAN model in point data format, existing soil maps, or spectral characteristics (remotely sensed). Environmental covariates are raster data processed by a Geographic Information System (GIS) and are digital and geo-graphically explicit. The SCORPAN model makes quantifying the connections between spatially explicit digital environmental variables and the predicted soil properties easier. Uncertainty estimation under the SCORPAN model is more straightforward than other traditional methods.

2.4. Selection of Environmental Covariates

Forty-nine pedogenesis-related environmental covariates (ECs) were collected from multiple sources, i.e., indexes and data products determined from the satellites. These ECs were categorized as climate, organism, parent material, position, relief, remote sensing, and soil. The 30 m spatial resolution and 16 d revisit duration of the Landsat 8 Operational Land Imager have been extensively used in DSM [8]. The Landsat 8 Surface Reflectance Level 1 Tier 1 was used to calculate related indices using land surface reflectance data (https://www.usgs.gov, accessed on 25 May 2023). The spectral indices such as the Normalized Difference Vegetation Index (NDVI) and Landsat data (1–11 bands) were used to predict the SOC and used as variables in the model. A digital elevation model (30 m spatial resolution) was produced by SRTM and used to derive relief covariates. By using SagaGIS Version 6.3.0, the terrain indices such as the Aspect (Asp), Elevation (Elev), Slope (Slp), Relative Slope Position (RSP), Topographic Wetness Index (TWI), Multi-Resolution index of Valley Bottom Flatness (MRVBF), Valley Depth (VD), Convergence Index (CI), Channel Network Base Level (CNBL), Channel Network Distance (CND), LS-factor, Plan Curvature (PlaC), Profile Curvature (PrC), Total Catchment Area (TCA), Topographic Positioning Index (TPI), and Topographic Roughness Index (TRI) were generated. All 19 bioclimatic variables were retrieved from WorldClim products (http://www.worldclim.com/version2, accessed on 25 May 2023) for each sampling point to use in the model. All collected covariates were aggregated/disaggregated through average resampling/bilinear resampling into a 30 × 30 m grid. Quantitative spatial models were developed using these aggregated ECs. All the data used in the prediction modelling were in raster format. The sampling coordinates were converted into UTM WGS84 Zone 46 N before use.

Feature selection was carried out by a multicollinearity test. Based on the SCORPAN conceptual model of soil development, it was possible to generate 49 environmental covariates as predictors from three data sources, i.e., geomorphometric/Digital Elevation Model (DEM), remote sensing, and climatic variables, where the spatial variability of the soil organic carbon was explained. Since the number of covariates was initially 49, there could be a high correlation among them. Therefore, multicollinearity (a common limitation of modeling) might have existed. Consequently, the Variance Inflation Factor (VIF), implemented by using the SPSS software (IBM SPSS Statistics 20.0), was used to evaluate the distributions and relationships between all environmental covariates. In brief, the VIF assesses how much the variance of an estimated regression coefficient increases when the predictors are correlated. This approach tries to remove some irrelevant covariates step by step from the dataset. The covariates were selected based on the VIF values, ranging from 1 to less than 5. The final feature selection step was performed to considerably reduce the number of covariates without significantly decreasing model prediction accuracy. Based on the VIF value, 28 environmental covariates were selected from multiple sources (Table 1).

2.5. Machine Learning Techniques

Four ensemble models (Random Forest (RF), regression tree (Cubist), Extreme Gradient Boosting (XGBoost), and Support Vector Machine (SVM)) were used for the prediction mapping of the SOC in the present investigation. These ensemble models are constructed by integrating multiple learners/algorithms. The ensemble model was constructed by the ensemble learning/method, which integrates multiple algorithms. The respective trend

analysis was conducted in the R 4.1.2 software [19], and maps were reproduced in ArcGIS 10.3 software.

Table 1. Different environmental covariates used in the models.

Groups	Predictor	Abbreviation	Resolution	Description
Terrain indices	Elevation (m)	Elev	30 m	Vertical distance above sea level
	Slope	Slp	30 m	Inclination of the land surface from the horizontal
	Relative Slope Position	RSP		Relative slope position
	Topographic Wetness Index	TWI	30 m	Ratio of local catchment area to slope
	Multi-Resolution index of Valley Bottom Flatness	MRVBF	30 m	Measure of flatness and lowness
	Valley Depth	VD	30 m	Relative position of the valley
	Channel Network Base Level	CNBL	30 m	Calculates the distance to a channel network base level
	Channel Network Distance	CND	30 m	Calculates the distance to a channel network
Spectral indices	Normalized Difference Vegetation Index	NDVI	30 m	Amount of vegetation
	Landsat data (11 bands)	Band1–11	30 m	Landsat OLI spectral band
Climate	Annual Precipitation	AP	1 km	Bioclimatic variables (BIO1)
	Mean Annual Temperature	MAT	1 km	Bioclimatic variables (BIO12)

2.5.1. Random Forest

Random forest consists of multiple decisions (Classification and Regression Trees (CARTs)) based on the binary rule. Each input is considered a tree algorithm to define a relationship between dependent and independent variables. The essential parameters of random forest, i.e., the number of variables (Mtry) and the Number of trees (Ntree), can be selected and adjusted each time during the decision process to obtain the optimum result. The Ntree and Mtry were kept in the range of 1–30 and 100–3000, respectively, as defined parameters (Table 2). The final prediction of RF is the accumulation of the results of all the individual trees (algorithms). Out-Of-Bag (*OOB*) estimates the error by bootstrapping the original datasets. The Mean-Squared Error (MSE_{OOB}) is determined by combining the *OOB* predictions from all trees [20]:

$$MSE_{OOB} = \frac{1}{n}\sum_{i=1}^{n}\left(z_i - \hat{z}_i^{OOB}\right)^2 \quad (2)$$

where $\hat{Z}_i^{OOB}$ is the average of all *OOB* predictions. The percentage of explained variance (Var_{ex}) is determined after the MSE_{OOB} is normalized because it depends on the unit of the response variable:

$$Var_{ex} = 1 - \frac{MSE_{OOB}}{Var_z} \quad (3)$$

where Var_z denotes the answer variable's overall variance; thus, RF is handy for predicting the output of ecological variables under a composite environmental system. The "ranger" function in the "caret" package is called in R 4.1.2 for random forest prediction.

2.5.2. Cubist

Cubist is based on the linear regression model [6]. Furthermore, it recently gained importance in DSM investigations in various ecological and geological regions [21,22]. These models (linear regression) have a highly classified relationship with the end value [23]. This linear regression model is used as the leaf nodes of the other regression tree algorithms in the Cubist machine learning techniques. Cubist produces multivariate models, which are composed of various rules. Thus, the prediction model under the Cubist environment is a selection of the rules. In the regression tree (Cubist), the number of model trees

(committees) and the number of nearest neighbors (neighbors) were defined as 1–100 and 0–9, respectively (Table 2). This study used the "Cubist" package in the R 4.1.2 software.

Table 2. Hyperparameters of machine learning algorithms used in this study.

ML Algorithms	Random Forest (RF)		Cubist (Regression Tree)		Extreme Gradient Boosting (XGBoost)						Support Vector Machine (SVM)		
Hyperparameters	**Mtry**	**Ntree**	**Committees**	**Neighbors**	**Booster**	**Max_Depth**	**Min_Child _Weight**	**Colsample _Bytree**	**Subsample**	**eta**	**Kernel Type**	**C**	**σ**
Defined Parameters	1–30	100–3000	1–100	0–9	gbtree	3–10	0–5	0.5–1	0.5–1	0.01–0.5	RBF	0.01–100	0.01–100
Definition	the number of input variables	the number of trees	the number of model trees	the number of nearest neighbors	the type of model	the depth of the tree	the minimum sum of weights of all observations	the number of variables supplied to a tree	the number of samples supplied to a tree	learning rate	the kernel function	the penalty parameter	the bandwidth parameter

2.5.3. Extreme Gradient Boosting

The Extreme Gradient Boosting (XGBoost) algorithm is generally used to enhance the performance of the regression tree and K classification. XGBoost improves the calculation speed and reduces the chance of overestimation by simplifying the objective functions. Automatic simultaneous computation during the training step is another merit of XGBoost. Supplemental training strategies were used to extend "strong" learners from the "weak" learners by "boosting". Under XGBoost's training steps, simultaneous computations for the functions are performed automatically [24]. The XGBoost algorithm's parameters, i.e., the algorithm type, the depth of the tree, the sum of the weights of all observations, the sample number provided to a tree, the variables used in tree construction, and the learning rate used in this investigation, are given in Table 2.

2.5.4. Support Vector Machines

Support Vector Machines (SVMs) are considered as common predictors for any multivariate function up to a specified accuracy. SVMs are used to solve regression problems through linear and nonlinear models. The linear model detects the noise in the datasets, and the nonlinear model converts the input space into a larger dimension [25]. Under the SVM classification and regression process, a set of connected supervised learning algorithms was used. These algorithms are the universal predictors of multivariate functions with high accuracy. In the present investigation, the range of components, input data, kernel type function, parameter for penalty, bandwidth parameters, etc., were used (Table 2).

2.6. Model Evaluation

In this study, four evaluation metrics were calculated to analyze the model performance: the coefficient of determination (R2), Mean Error (ME), Lin's Concordance Correlation Coefficient (CCC), and Root-Mean-Squared Error (RMSE). The RMSE and R2 were used to measure the accuracy and stability, respectively. A smaller RMSE and higher R2 value depict the accuracy and stability of the model, respectively. The sample set in an

80:20 ratio was used for calibration and validation. The RMSEc and RMSEv were used to express the fitting accuracy of the models' calibration validation, respectively. These validation metrics are calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(p_i - o_i)^2}{\sum_{i=1}^{n}(\hat{p}_i - \hat{o}_i)^2} \quad (4)$$

$$ME = \frac{1}{n}\sum_{i=1}^{n}(p_i - o_i) \quad (5)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2} \quad (6)$$

where p_i and o_i are the predicted and observed values and $\hat{p}_i$ and $\hat{o}_i$ are the means of the predicted and observed values.

$$CCC = \frac{2\rho\sigma_0\sigma_p}{{\sigma_0}^2 + {\sigma_p}^2 + \left(\mu_0 - \mu_p\right)^2} \quad (7)$$

μ_0 and μ_p are the means of the observed and predicted values, and ${\sigma_0}^2$ and ${\sigma_p}^2$ are corresponding variances. ρ is the Pearson correlation coefficient between the observed and predicted values.

2.7. Uncertainty Assessment

The current study used non-parametric bootstrap methods to analyze the spatially detailed quantification of the best model's SOC prediction uncertainty. It ensured that the sample size and probability distribution of the old and new bootstrap datasets were the same. Fifty bootstrapped datasets were created for the calibration. Through the final evaluation indication and predicted outcome, respectively, it was presumed that the validation dataset's average accuracy was good. Maps of the 0.05 and 0.95 quantiles were created by the estimate. In other words, there were uncertainty levels for the SOC predictions for each pixel position in the research region. The 90% confidence interval's upper and lower bounds were mapped. Using the 5th and 95th percentiles of prediction, the mean of the SOC contents in each pixel and the 90% CI were determined. For the top-performing model, three maps of the SOC were created: the mean forecast, lower confidence interval (5%), and higher confidence interval (95%).

The Prediction Interval Coverage Probability (PICP) criteria were used to assess the uncertainties of the prediction [26,27]. The following confidence intervals were utilized in this study: 99, 97.5, 95, 90, 80, 60, 40, 20, 10, and 5%. One should expect the PICP value or proportion to be close to the corresponding confidence level to determine whether the indicated uncertainties have been effectively computed [28,29]. The uncertainty is at its lowest at the PICP, which is approximately 100(1 − a) %, like a 90% confidence interval. The accuracy of the PICP was established by an estimate of 90% for a 90% prediction interval. The standard deviations of the accuracy metrics (R2, RMSE, and ME) made from the 50 bootstrapped datasets were also employed to indicate, to a certain degree, the stability and uncertainty of the predictions.

3. Results

3.1. Descriptive Statistics

Eighty percent of the calibration set and twenty percent of the validation set, each comprising 128 and 32 sample points, were divided from the 160-sample dataset. The calibration set's coefficient of variation, skewness, and kurtosis values were proximal to the total set data and represented a complete sample, according to the descriptive statistics (Table 3), performed for the calibration and validation sets. The measured SOC content was found in the 0.10–1.85% range. The considerable variation between the minimum

and maximum SOC envisaged that the surface soil of the upper Brahmaputra Valley is susceptible to the environment, management practices, and soil disturbance as flooding is a yearly event in the studied region. The higher cumulative variation (47.36%) indicates that the SOC had high spatial variation and a semi-homogenous geographical distribution in the investigated region. This might be due to the topographical variation, crop management practices, and land-use and land-cover changes. A similar argument was devised by Jena et al. [30] and Moharana et al. [4,8] in their respective studies in the northeastern regions. Based on the investigation, contemporary statistical procedures must be prioritized to capture the geographical variability in the SOC as it is a stochastic variable highly influenced by the soil formation and the regional climate.

Table 3. Descriptive statistics of soil organic carbon in the study area of Upper Brahmaputra Valley in northeast India.

Dataset	n	Min	Max	Mean	SE	Median	SD	CV	Skewness	Kurtosis
Total	160	0.10	1.85	0.81	0.03	0.75	0.38	47.36	0.28	−0.66
Calibration	128	0.10	1.54	0.81	0.03	0.81	0.37	45.78	0.04	−0.88
Validation	32	0.29	1.85	0.78	0.07	0.63	0.42	54.37	1.05	0.27

SD, Standard Deviation; SE, Standard Error; CV, Coefficient of Variation.

The frequency distribution of the calibration and validation dataset of the SOC content is depicted in Figure 2. The rug plot on the X-axis represents the locations and SOC contents of the particular sample-collection point. The bell-shaped curve of the SOC contents shows the normal distribution. Further, normalization has yet to be opted for as machine learning models do not require normalization of the data points [31]. The prediction models were fit using the original SOC concentration in the studied location. The probability densities and histograms are depicted by curves and bars, respectively. The uncertainties of the spatial autocorrelation were checked through the Global Moran's I. It was ensured that there was no spatial autocorrelation in the geographic patterns before fitting the machine learning techniques (prediction models).

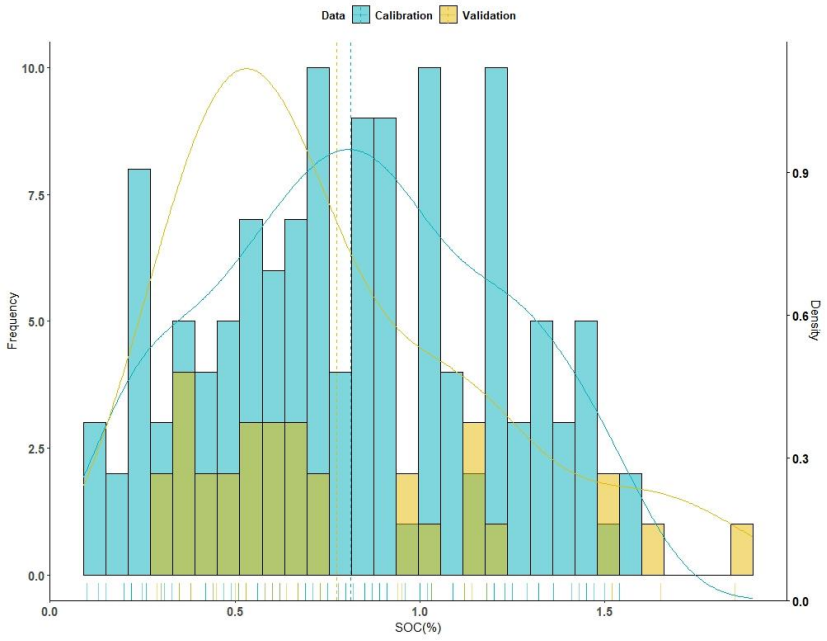

Figure 2. Soil Organic Carbon (SOC) content distribution in the calibration and validation dataset is represented by the histogram, density plot, and rug plot. Vertical dashed lines represent the mean values.

The descriptive statistics of the environmental covariates are presented in Table 4. Terrain attributes such as the Asp, CI, CND, LS-factor, MRRTF, MRVBF, RSP, Slp, TCA, and TPI were highly variable (CV > 35%). Landsat bands had a CV < 15%, indicating high data homogeneity. Our findings from the CV analysis suggest that the environmental covariates had a semi-homogeneous geographical distribution, which may be related to variations in the research area's soil-forming factors. As demonstrated by Falahatkar et al. [32] and Dharumarajan et al. [27], a soil attribute such as SOC is a stochastic variable supported by the continuation of the soil-formation processes and the change in the climate. As a result, it is essential to use contemporary statistical procedures to survey the geographical variability in the SOC.

Table 4. Descriptive statistics of environmental covariates used in the models.

Covariates	Min	Max	Mean	Median	SD	CV	Skewness	Kurtosis
Asp	0.00	6.28	3.18	3.14	1.86	58.32	−0.02	−1.09
CI	−47.18	51.21	5.39	3.84	14.96	277.65	0.33	1.46
CNBL	70.75	109.97	82.43	79.61	8.59	10.42	1.18	0.67
CND	0.00	22.51	7.03	5.98	5.37	76.44	0.75	0.02
Elev	59.67	116.40	88.53	86.13	11.17	12.62	0.34	−0.46
LS-Factor	0.02	9.90	2.67	2.56	1.62	60.83	1.31	3.54
MRRTF	0.00	4.42	0.65	0.33	0.83	128.10	2.06	4.23
MRVBF	0.00	5.80	0.73	0.30	1.10	150.04	2.51	6.26
NDVI	0.08	0.46	0.37	0.37	0.05	14.19	−1.15	4.44
RSP	0.00	0.43	0.11	0.10	0.09	83.05	1.17	1.26
Slp	0.00	0.26	0.10	0.10	0.05	50.57	0.31	0.11
TCA	959	4,892,340	78,638	2317	522,500	664	8	71
TPI	−11.19	10.49	0.85	1.26	2.98	352.22	−0.39	1.59
TRI	1.08	8.40	3.23	3.08	1.22	37.73	1.07	2.27
VD	18.95	100.15	59.45	59.20	17.64	29.68	0.09	−0.63
Band1	10,693.90	14,115.70	11,112.31	11,038.65	355.06	3.20	4.01	31.11
Band2	9613.74	13,974.50	10,121.27	10,015.80	448.33	4.43	4.22	33.29
Band3	8952.72	14,430.60	9615.18	9455.46	554.81	5.77	4.38	34.80
Band4	7539.60	15,172.70	8488.40	8311.01	796.62	9.38	4.02	30.28
Band5	14,978.70	21,984.60	18,392.15	18,427.75	1178.43	6.41	0.10	0.37
Band6	11,266.50	20,536.80	13,582.21	13,334.00	1183.53	8.71	1.79	7.28
Band7	7657.73	18,707.30	9506.82	9174.76	1252.43	13.17	2.90	17.33
Band8	8331.01	14,683.00	9091.32	8789.09	739.24	8.13	3.19	19.53
Band9	5024.32	5068.92	5048.36	5048.27	7.03	0.14	0.00	0.93
Band10	27,930.40	30,232.70	28,404.28	2343.90	268.27	0.94	2.56	13.60
Band11	24,840.80	26,183.80	25,195.66	25,178.80	178.82	0.71	1.66	6.24
AP	2554.71	3253.06	3019.87	3126.40	222.67	7.37	−0.59	−1.18
MAT	23.65	24.15	23.98	23.95	0.10	0.43	−0.24	−0.41

SD, Standard Deviation; CV, Coefficient of Variation.

The pairwise correlation coefficient (r) was computed to determine the relationship between the topographic and remote sensing covariates and the SOC (Figure 3). The correlation matrix showed that the SOC had a positive correlation with the elevation ($r = -0.108$, $p < 0.05$), NDVI ($r = 0.07$, $p < 0.01$), and CNBL ($r = 0.178$, $p < 0.05$). In contrast, a significant negative correlation was observed for the AP ($r = -0.091$, $p < 0.05$) and TPI ($r = -0.16$, $p < 0.05$). The remaining topographic attributes (Asp, MRVBF, and LS-factor) were not significantly correlated with the SOC in the upper Brahmaputra Valley regions. Among the Landsat 8 remote sensing data, Band1, 2, 6, and 7 were positively correlated with the SOC. In the upper Brahmaputra Valley, soil formed on the top of the slope and, then, moved outward or downward under the influence of gravity and deposited in the valley. Therefore, the elevation and slope always negatively correlated with the SOC. The present study's correlation coefficient was low, although a similar trend was observed. Anthropogenic activity in the valley may be the reason for the low SOC in that region.

This observation is in agreement with the work carried out by previous researchers [3] in similar landscapes.

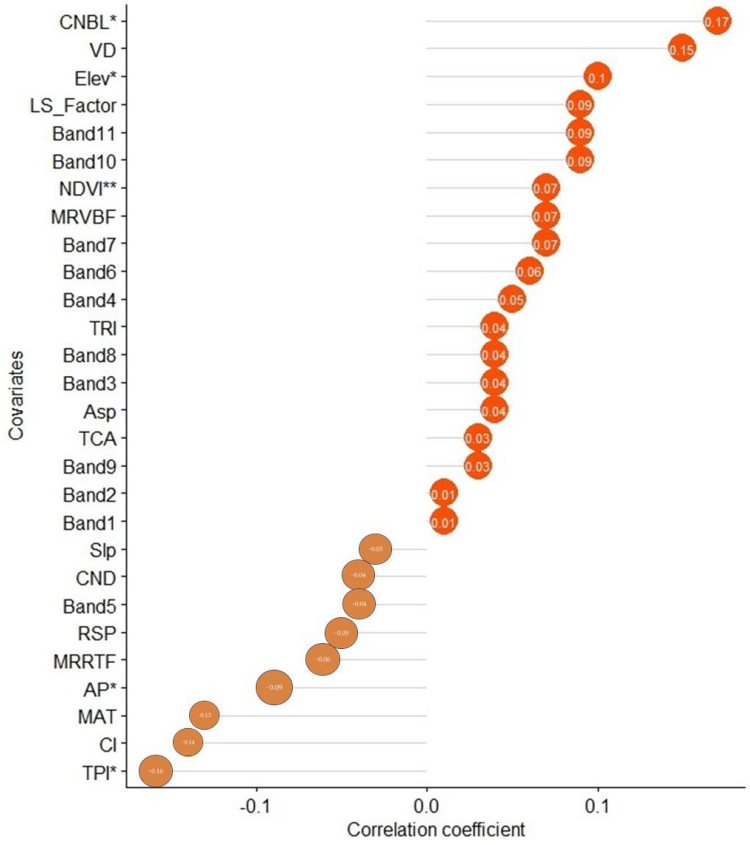

Figure 3. Correlation analysis between covariates and SOC in the study area of Upper Brahmaputra Valley in northeast India. Abbreviations for the variables are listed in Table 1. ** Correlation is significant at the 0.01 level. * Correlation is significant at the 0.05 level.

3.2. Evaluation of Prediction Models

The R^2, CCC, RMSE, and ME were computed, and the results are depicted in Table 5. The performance (goodness of fit and errors employed) of the RF, Cubist, SVM, and XGBoost models for the prediction of the SOC in the upper Brahmaputra Valley in the northeast state of India was evaluated. The XGBoost model had the highest accuracy and the lowest error on the calibration set (R^2c = 0.998, RMSEc = 0.022%) and had slightly lower accuracy on the test set (R^2v = 0.152, RMSEv = 0.424%). In contrast, the RF model performed well on both the calibration and validation sets (R^2c = 0.966, RMSEc = 0.159%, R^2v = 0.418, RMSEv = 0.377%). However, the SVM model had the next-lowest accuracy, accounting for just around 47% of the variance, and the Cubist model fared the poorest on both the calibration and validation sets (R^2_c = 0.471, $RMSE_c$ = 0.293, R^2_v = 0.081, $RMSE_v$ = 0.452). The RF model provided the lowest RMSE, i.e., 0.159 and 0.377, for the calibration and validation datasets. However, the lowest ME was observed for XGBoost, which was 0 on the calibration dataset and 0.054 on the validation dataset. The ME for the RF model was 0.001 and 0.136 for the calibration and validation datasets. The CCC values of 0.863 and 0.549 on the calibration and validation datasets were found for the RF model, suggesting good agreement between the predicted and observed values. Based on the highest R^2 for the validation value of the RF model, it showed the best ability to predict the SOC in the sericultural soil of this northeast region of India.

Table 5. Performance of various models for predicting soil carbon in the study area of Upper Brahmaputra Valley in northeast India.

Model	Calibration				Validation			
	R^2_c	CCC_c	$RMSE_c$	ME_c	R^2_v	CCC_v	$RMSE_v$	ME_v
RF	0.966	0.863	0.159	0.001	0.418	0.549	0.377	0.136
Cubist	0.396	0.571	0.291	0.039	0.230	0.314	0.485	0.062
SVM	0.471	0.453	0.293	0.015	0.081	0.175	0.452	0.049
XGBoost	0.998	0.990	0.022	0.000	0.152	0.190	0.424	0.054

R^2: coefficient of determination, ME: Mean Error, RMSE: Root-Mean-Squared Error, CCC: Lin's Concordance Correlation Coefficient.

The RF model's R^2 was not the highest. Still, its RMSE was the lowest when compared to the other models since its predictions were closer to the actual values than those of other models (the predicted–observed fit line was almost 1:1) (Figure 4). Additionally, the slope of the predicted–observed fit curve for the RF and Cubist models dramatically improved, demonstrating a considerable improvement in prediction accuracy, primarily due to the concentration of the scatter distribution. The scatter spots are nearer to the curve of the fitting.

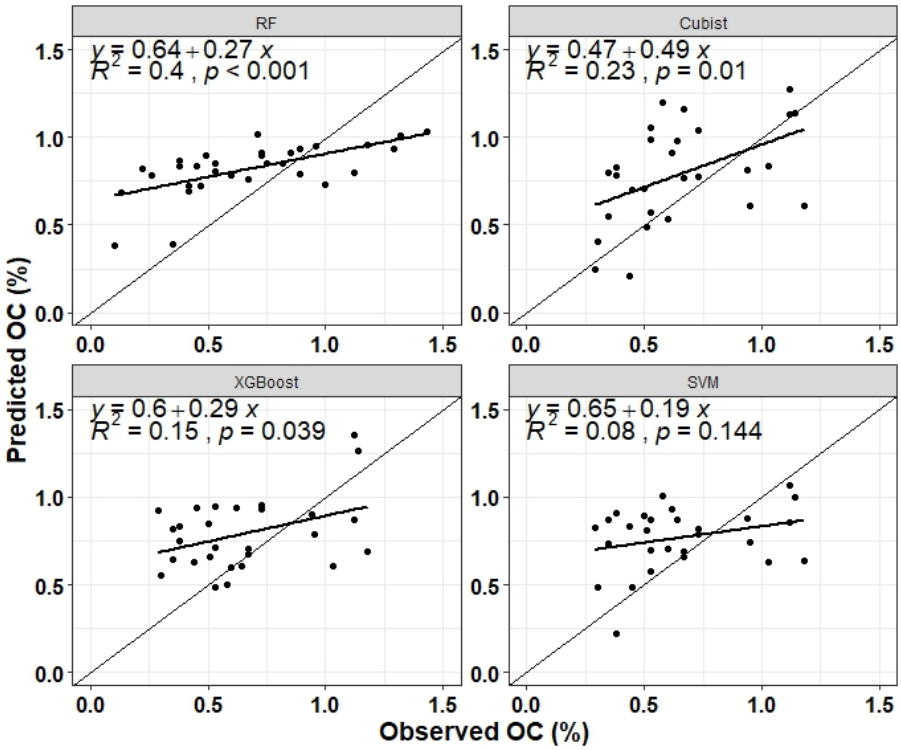

Figure 4. Predicted vs. observed parameters of soil organic carbon by various models in the study area of Upper Brahmaputra Valley in the northeast state of India.

3.3. Importance of Environmental Variables

Figure 5's comparative relative variable importance plots show that RF chose more variables than the Cubist, SVM, and XGBoost models. The order of the most-crucial factors in the RF model for predicting the SOC was Elev > MAT > Band 3 > Band1 > MRVBF. The most-crucial variables for utilizing a Cubist model to predict the variance of the SOC were Slp, TRI, MAT, and Band4. The AP and LS were the most-important factors in the XGBoost and SVM models. One of the primary benefits of RF models over other ML models is that the former assess the relative relevance of the covariates in the model, in contrast to Cubist, which maintains the model through stepwise selection only with highly

correlated predictive variables [33]. Even though there may be relationships between the predictive variables and soil, RF avoids removing those [34]. In a study conducted in the Zahak County of Iran, Pahlavan-Rad and Akbarimoghadam [35] discovered that the CNBL and elevation were the most-crucial covariates for predicting soil qualities. The elevation and slope were the most-significant factors for the SOC prediction for both the RF and Cubist models, followed by the other topography and vegetation factors. In the upper Brahmaputra Valley in the northeast state of India, flooding causes soil sediments to flow and gather in lower-elevation regions. These topography-driven erosional processes transport the SOC from higher elevations, where the SOC concentrations are often lower, to lower elevations, where the SOC concentrations are typically higher [30].

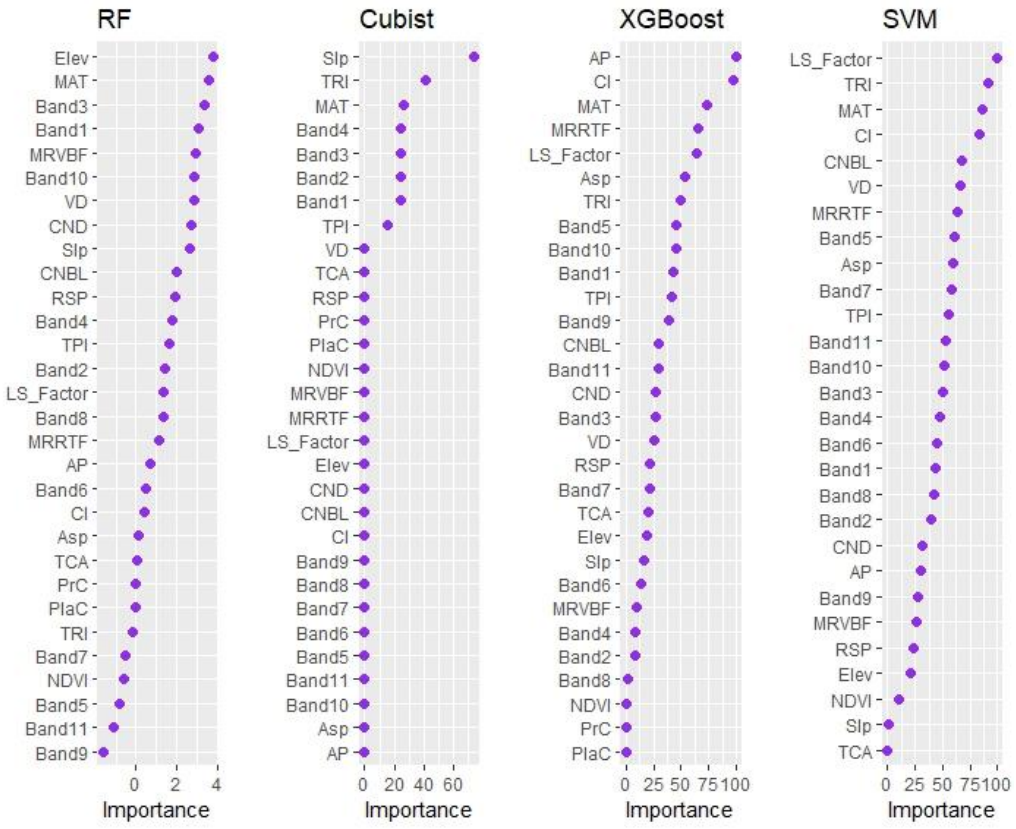

Figure 5. Importance variables in predicting soil organic carbon by various models in the study area of Upper Brahmaputra Valley in northeast India.

3.4. Spatial Prediction of SOC

Significant discrepancies between the predictions are shown in Figure 6, which depicts the spatial predictions of the SOC made by the various models. Although the four models' cross-validation accuracy metrics were comparable, RF's predictions more closely matched the spatial pattern of the OC we anticipated in the research area. The SOC in the research area appeared to be underestimated by the SVM prediction. All of the maps for the prediction models displayed abrupt and gradual shifts throughout the research region. In the RF, Cubist, XGBoost, and SVM models, the predicted OC ranged from 0.44 to 1.35%, 0.031 to 1.61%, 0.035 to 1.71%, and 0.47 to 1.36%, respectively. It is impossible to say which model is the most-accurate without independently confirming these predictions; however, we selected the RF model as the "best" because the cross-validation accuracy metrics were similar, and the spatial predictions visually matched our perception of the terrain. The spatial patterns of the SOC in all models were reasonable, with large values in the study area's western region, which is dominated by forest and covered in dense vegetation, and minor values in its eastern regions, which have soils subject to significant erosion and crop cultivation. Low-elevation cultivable fields appear more uncertain in the upper

Brahmaputra Valley than in high-elevation cultivable lands. This can be explained by the variation in the management techniques used in the more-intensively farmed regions [30]. The covariates utilized in these predictive models may have needed to adequately account for the management aspects contributing to the spatial heterogeneity in the SOC content levels in these downstream-farmed regions. The high-elevation areas are less actively farmed, which may have less impact due to management and because the terrain features have a stronger hold over them. SOC concentrations are the outcome of the balance between carbon imports and outputs in soils; several variables, including local characteristics such as vegetation and topography and environmental conditions, affect this equilibrium [36]. According to Pahlavan-Rad et al. [37], topography was the most-significant covariate influencing the distribution of the SOC in our study.

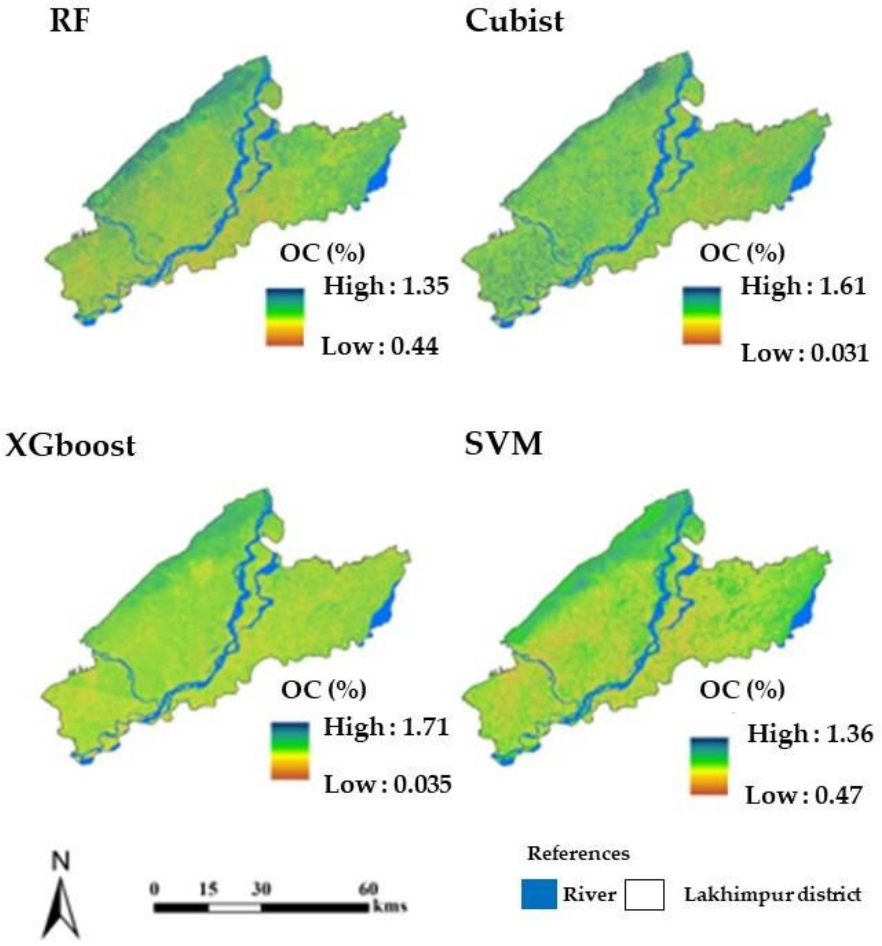

Figure 6. Distribution of soil organic carbon predicted by various model in the study area of Upper Brahmaputra Valley in northeast India.

3.5. Uncertainty Prediction

To quantify the spatial distribution of prediction uncertainty, bootstrapping was used. Figure 7 displays the level of uncertainty in the best model's (RF) predictions. The uncertainty analysis partially confirmed the trend of the ML algorithms' ability to predict the SOC. With a 90% confidence interval, uncertainty was found in the lower and upper predicted bounds. The effectiveness of the predicted uncertainties during testing was evaluated using the PICP technique. The uncertainty analysis exhibited the same trends as the ML algorithms' capacity to predict the SOC.

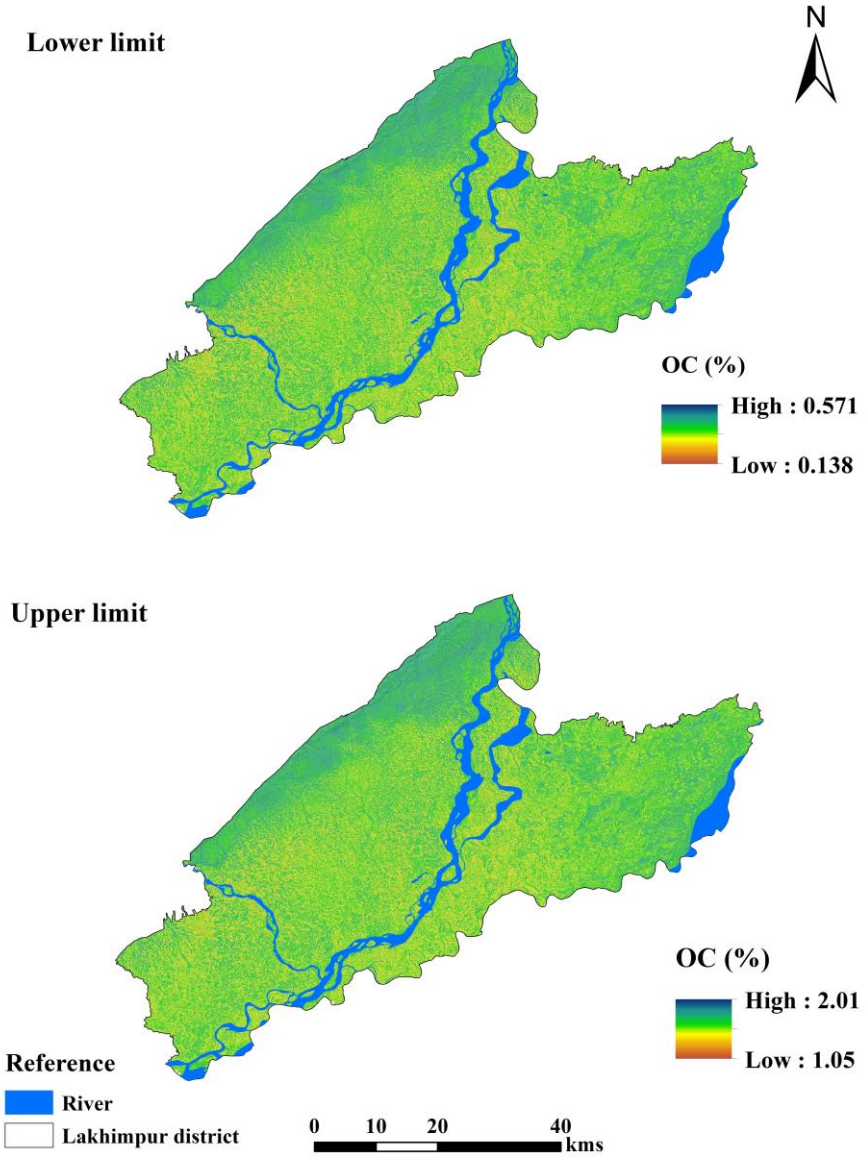

Figure 7. Distribution of lower and upper limits of 90% prediction interval of soil organic carbon predicted by the best model (RF) in the study area of Upper Brahmaputra Valley in northeast India.

The 90% confidence interval with lower and higher prediction bounds showed the degree of uncertainty. The PICP methodology is a method for assessing how well the anticipated uncertainties perform throughout testing. The percentage of observations within the associated prediction interval is the PICP. Simply assessing the coverage of the prediction intervals around different degrees of observed confidence allows for this. When the coverage probability and confidence level are closely monitored along the 1:1 line, as shown in Figure 8's graphs, it is obvious what to expect. According to Jena et al. [3], the probabilities above the 1:1 line show a slight overprediction of the uncertainty range. This may be because of the uncertainty approach. First, the sample density is where the forecast uncertainty comes from. Our study indicated more uncertainty in the region with a high elevation, few sampling locations, and generally dispersed fields. This resulted from the need for further knowledge regarding the connections between the soil and environmental covariates.

While this is happening, capturing the spatial variability of the environmental variables with more accuracy is possible, significantly reducing the deviation of the sample sites, particularly in the region of dramatic change for the environment and soil characteristics [38]. Second, the quality and amount of the environmental variables contribute

to the uncertainty. To predict the SOC, we gathered as many environmental variables as we could. The sample density should be increased; useful variables should be captured; spatial modelling techniques should be optimized to reduce uncertainty. The reliability and accuracy of the data products may be increased by minimizing sampling and modelling uncertainty using high-quality sample data and environmental variables, particularly in areas where soils and landscape changes occur quickly.

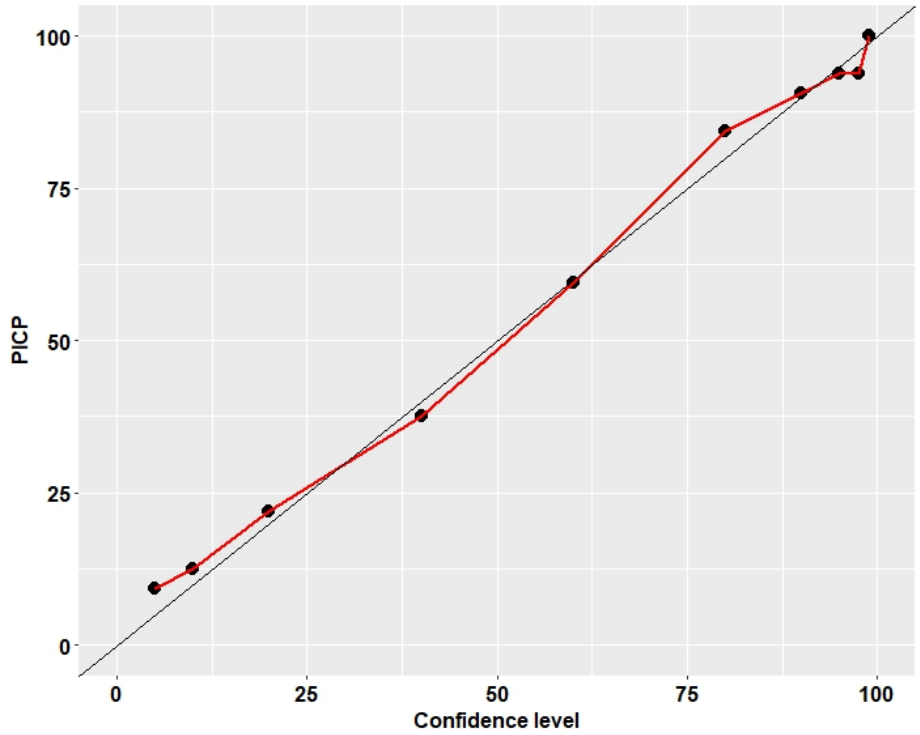

Figure 8. Based on the validation of the RF model for soil organic carbon between the Prediction Interval Coverage Probability (PICP) and confidence level (CI).

4. Conclusions

In this work, four tree-based machine-learning algorithms were examined. It calculated the SOC's spatial distribution patterns in northeastern India's upper Brahmaputra Valley. This study used optimum models, environmental covariates, and information to determine the spatial distribution of the SOC. A succinct summary of the conclusions is as follows:

1. RF had the best accuracy and the lowest uncertainty for predicting the regional SOC compared to XGBoost, SVM, and Cubist.
2. Compared to XGBoost, SVM, and Cubist, the RF showed higher R^2 and RMSE values for predicting SOC based on the validation data.
3. The order of the most-crucial factors in the RF model for predicting the SOC was Elev > MAT > Band 3 > Band1 > MRVBF. The most-crucial variables for utilizing a Cubist model to predict the variance of the SOC were Slp, TRI, MAT, and Band4. The AP and LS were the most-essential factors in the XGBoost and SVM models.
4. The predicted SOC ranged from 0.44 to 1.35%, 0.031 to 1.61%, 0.035 to 1.71%, and 0.47 to 1.36% with the RF, Cubist, XGBoost, and SVM models.

Based on the combination of DSM methodologies with currently available, high-resolution soil-formation environmental data, the results updated the accuracy of the regional variation of the SOC prediction at both the national scale and a detailed level. The SOC maps highlight probable reasons for the predicted uncertainty, which may be used to assess changes in soil quality following extensive cropping and direct further high-resolution DSM research. In general, fine-resolution soil maps are valuable to many

soil and environmental professionals and land managers in northeast India. As a result, we advocate employing the comparative technique used in this study area to map the SOC in other parts of India, given that the agroecological zones differ significantly throughout the northeast state of India.

Author Contributions: Software, P.C.M. and R.K.J.; validation, R.K.J. and P.C.M.; formal analysis, D.K.J. and A.A.S.; investigation, A.K. and D.K.J.; resources, P.C.M., K.M.V.K. and S.G.D.; data curation, D.K.J.; writing—A.K., R.K.F. and S.K.M. original draft preparation, A.K., P.C.M., G.K.S. and R.K.J.; reviewing and editing—G.K.S., R.K.F., S.K.M. and A.A.S.; visualization, S.G.D. and K.M.V.K.; supervision, S.G.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Most of the data are available in the tables and figures given in the manuscript.

Acknowledgments: The authors are highly thankful to the respective Directors for supporting the present investigation.

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

References

1. Schillaci, C.; Acutis, M.; Vesely, F.; Saia, S. A simple pipeline for the assessment of legacy soil datasets: An example and test with soil organic carbon from a highly variable area. *Catena* **2019**, *1*, 110–122. [CrossRef]
2. Yigini, Y.; Panagos, P. Assessment of soil organic carbon stocks under future climate and land cover changes in Europe. *Sci. Total Environ.* **2016**, *557–558*, 838–850. [CrossRef] [PubMed]
3. Jena, R.K.; Moharana, P.C.; Dharumarajan, S.; Sharma, G.K.; Ray, P.; Deb Roy, P.; Ghosh, D.; Das, B.; Alsuhaibani, A.M.; Gaber, A.; et al. Spatial Prediction of Soil Particle-Size Fractions Using Digital Soil Mapping in the North Eastern Region of India. *Land* **2023**, *12*, 1295. [CrossRef]
4. Moharana, P.C.; Meena, R.L.; Nogiya, M.; Jena, R.K.; Sharma, G.K.; Sahoo, S.; Jha, P.K.; Aditi, K.; Vara Prasad, P.V. Impacts of Land Use on Pools and Indices of Soil Organic Carbon and Nitrogen in the Ghaggar Flood Plains of Arid India. *Land* **2022**, *11*, 1180. [CrossRef]
5. Taghizadeh-Mehrjardi, R.; Nabiollahi, K.; Kerry, R. Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. *Geoderma* **2016**, *266*, 98–110. [CrossRef]
6. Taghizadeh-Mehrjardi, R.; Minasny, B.; Sarmadian, F.; Malone, B.P. Digital mapping of soil salinity in Ardakan region, central Iran. *Geoderma* **2014**, *213*, 15–28. [CrossRef]
7. Liu, F.; Zhang, G.-L.; Song, X.; Li, D.; Zhao, Y.; Yang, J.; Wu, H.; Yang, F. High-resolution and three-dimensional mapping of soil texture of China. *Geoderma* **2020**, *361*, 114061. [CrossRef]
8. Moharana, P.C.; Dharumarajan, S.; Kumar, N.; Jena, R.K.; Pradhan, U.K.; Meena, R.M.; Sahoo, S.; Kumar, S.; Meena, R.L.; Tailor, B.; et al. Modelling and Prediction of Soil Organic Carbon using Digital Soil Mapping in the Thar Desert Region of India. *J. Indian Soc. Soil. Sci.* **2022**, *70*, 86–96. [CrossRef]
9. Hengl, T.; de Jesus, J.M.; MacMillan, R.A.; Batjes, N.H.; Heuvelink, G.B.M.; Ribeiro, E.; Samuel-Rosa, A.; Kempen, B.; Leenaars, J.G.B.; Walsh, M.G.; et al. SoilGrids1km—Global soil information based on automated mapping. *PLoS ONE* **2014**, *9*, e105992. [CrossRef]
10. Hengl, T.; Heuvelink, G.B.; Kempen, B.; Leenaars, J.G.B.; Walsh, M.G.; Shepherd, K.D.; Sila, A.; MacMillan, R.A.; de Jesus, J.M.; Tamene, L.; et al. Mapping soil properties of Africa at 250 m resolution: Random forests significantly improve current predictions. *PLoS ONE* **2015**, *10*, e0125814. [CrossRef]
11. Arrouays, D.; Grundy, M.G.; Hartemink, A.E.; Hempel, J.W.; Heuvelink, G.B.M.; Hong, S.Y.; Lagacherie, P.; Lelyk, G.; McBratney, A.B.; McKenzie, N.J.; et al. GlobalSoilMap: Toward a fine-resolution global grid of soil properties. *Adv. Agron.* **2014**, *125*, 93–134.
12. Dharumarajan, S.; Hegde, R.; Janani, N.; Singh, S.K. The need for digital soil mapping in India. *Geoderma Reg.* **2019**, *16*, e00204. [CrossRef]
13. Mishra, G.; Giri, K.; Jangir, A.; Francaviglia, R. Projected trends of soil organic carbon stocks in Meghalaya state of Northeast Himalayas, India. Implications for a policy perspective. *Sci. Total Environ.* **2020**, *698*, 134266. [CrossRef] [PubMed]
14. Jigyasu, D.K.; Kumar, A.; Shabnam, A.A.; Sharma, G.K.; Jena, R.K.; Das, B.; Naik, V.S.; Ahmed, S.A.; Kumari, K.M.V. Spatial Distribution of the Fertility Parameters in Sericulture Soil: A Case Study of Dimapur District, Nagaland. *Land* **2023**, *12*, 956. [CrossRef]
15. Walkley, A.; Black, I.A. An examination of the Degtjareff method for determining soil organic matter, and a proposed modification of the chromic acid titration method. *Soil. Sci.* **1934**, *37*, 29–38. [CrossRef]
16. McBratney, A.B.; Santos, M.M.; Minasny, B. On digital soil mapping. *Geoderma* **2003**, *117*, 3–52. [CrossRef]

17. Dokuchaev, V.V. *Russian Chernozems (Russkii Chernozems)*; Kaner, N., Ed.; US Department of Commerce: Springfield, VA, USA, 1976.
18. Jenny, H. *Factors of Soil Formation*; McGraw Hill: New York, NY, USA, 1941.
19. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019; Available online: https://www.R-project.org/ (accessed on 5 February 2023).
20. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22.
21. Emadi, M.; Taghizadeh-Mehrjardi, R.; Cherati, A.; Danesh, M.; Mosavi, A.; Scholten, T. Predicting and mapping of soil organic carbon using machine learning algorithms in Northern Iran. *Remote Sens.* **2020**, *12*, 2234. [CrossRef]
22. 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]
23. Mikkonen, H.G.; van de Graaff, R.; Clarke, B.O.; Dasika, R.; Wallis, C.J.; Reichman, S.M. Geochemical indices and regression tree models for estimation of ambient background concentrations of copper, chromium, nickel and zinc in soil. *Chemosphere* **2018**, *210*, 193–203. [CrossRef]
24. Fan, J.; Wang, X.; Wu, L.; Zhou, H.; Zhang, F.; Yu, X.; Lu, X.; Xiang, Y. Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China. *Energy Convers. Manag.* **2018**, *164*, 102–111. [CrossRef]
25. Zhang, H.T.; Gao, M.X. The Application of Support Vector Machine (SVM) Regression Method in Tunnel Fires. *Procedia Eng.* **2018**, *211*, 1004–1011. [CrossRef]
26. Lagacherie, P.; Arrouays, D.; Bourennane, H.; Gomez, C.; Martin, M.; Saby, N.P.A. How far can the uncertainty on a Digital Soil Map be known?: A numerical experiment using pseudo values of clay content obtained from Vis-SWIR hyperspectral imagery. *Geoderma* **2019**, *337*, 1320–1328. [CrossRef]
27. Dharumarajan, S.; Kalaiselvi, B.; Suputhra, A.; Lalitha, M.; Vasundhara, R.; Anil Kumar, K.S.; Nair, K.M.; Hegde, R.; Singh, S.K.; Lagacherie, P. Digital soil mapping of soil organic carbon stocks in Western Ghats, South India. *Geoderma Reg.* **2021**, *25*, e00387. [CrossRef]
28. Solomatine, D.P.; Shrestha, D.L. A novel method to estimate model uncertainty using machine learning techniques. *Water Resour. Res.* **2009**, *45*, W00B11. [CrossRef]
29. Malone, B.P.; McBratney, A.B.; Minasny, B. Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes. *Geoderma* **2011**, *160*, 614–626. [CrossRef]
30. Jena, R.K.; Bandyopadhyay, S.; Pradhan, U.K.; Moharana, P.C.; Kumar, N.; Sharma, G.K.; Roy, P.D.; Ghosh, D.; Ray, P.; Padua, S.; et al. Geospatial Modelling for Delineation of Crop Management Zones Using Local Terrain Attributes and Soil Properties. *Remote Sens.* **2022**, *14*, 2101. [CrossRef]
31. Lamichhane, S.; Adhikari, K.; Kumar, L. Use of multi-seasonal satellite images to predict SOC from cultivated lands in a Montane ecosystem. *Remote Sens.* **2021**, *13*, 4772. [CrossRef]
32. Falahatkar, S.; Hosseini, S.M.; Ayoubi, S.; Salmanmahiny, A. Predicting soil organic carbon density using auxiliary environmental variables in northern Iran. *Arch. Agron. Soil. Sci.* **2016**, *62*, 375–393. [CrossRef]
33. da Silva Chagas, S.; de Carvalho Junior, W.; Bhering, S.B.; Calderano Filho, B. Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. *Catena* **2016**, *139*, 232–240. [CrossRef]
34. Akpa, S.I.; Odeh, I.O.; Bishop, T.F.; Hartemink, A.E. Digital mapping of soil particle-size fractions for Nigeria. *Soil. Sci. Soc. Am. J.* **2014**, *78*, 1953–1966. [CrossRef]
35. Pahlavan-Rad, M.R.; Akbarimoghaddam, A. Spatial variability of soil texture fractions and pH in a flood plain (case study from eastern Iran). *Catena* **2018**, *160*, 275–281. [CrossRef]
36. Sahoo, U.K.; Singh, S.L.; Gogoi, A.; Kenye, A.; Sahoo, S.S. Active and passive soil organic carbon pools as affected by different land use types in Mizoram, Northeast India. *PLoS ONE* **2019**, *14*, e0219969. [CrossRef] [PubMed]
37. Pahlavan-Rad, M.R.; Dahmardeh, K.; Hadizadeh, M.; Keykha, G.; Mohammadnia, N.; Gangali, M.; Keikha, M.; Davatgar, N.; Brungard, C. Prediction of soil water infiltration using multiple linear regression and random forest in a dry flood plain, eastern Iran. *Catena* **2020**, *194*, 104715. [CrossRef]
38. Liang, Z.; Chen, S.; Yang, Y.; Zhao, R.; Shi, Z.; Rossel, R.A.V. National digital soil map of organic matter in topsoil and its associated uncertainty in 1980's China. *Geoderma* **2019**, *335*, 47–56. [CrossRef]