agronomy

Article

The Effect of Bioclimatic Covariates on Ensemble Machine Learning Prediction of Total Soil Carbon in the Pannonian Biogeoregion

Dorijan Radočaj *, Mladen Jurišić and Vjekoslav Tadić

Faculty of Agrobiotechnical Sciences Osijek, Josip Juraj Strossmayer University of Osijek, Vladimira Preloga 1, 31000 Osijek, Croatia; mjurisic@fazos.hr (M.J.); vtadic@fazos.hr (V.T.)
* Correspondence: dradocaj@fazos.hr; Tel.: +385-31-554-965

Abstract: This study employed an ensemble machine learning approach to evaluate the effect of bioclimatic covariates on the prediction accuracy of soil total carbon (TC) in the Pannonian biogeoregion. The analysis involved two main segments: (1) evaluation of base environmental covariates, including surface reflectance, phenology, and derived covariates, compared to the addition of bioclimatic covariates; and (2) assessment of three individual machine learning methods, including random forest (RF), extreme gradient boosting (XGB), and support vector machine (SVM), as well as their ensemble for soil TC prediction. Among the evaluated machine learning methods, the ensemble approach resulted in the highest prediction accuracy overall, outperforming the individual models. The ensemble method with bioclimatic covariates achieved an R^2 of 0.580 and an RMSE of 10.392, demonstrating its effectiveness in capturing complex relationships among environmental covariates. The results of this study suggest that the ensemble model consistently outperforms individual machine learning methods (RF, XGB, and SVM), and adding bioclimatic covariates improves the predictive performance of all methods. The study highlights the importance of integrating bioclimatic covariates when modeling environmental covariates and demonstrates the benefits of ensemble machine learning for the geospatial prediction of soil TC.

Keywords: WorldClim; GEMAS; remote sensing; environmental covariates; hyperparameter tuning

check for updates

Citation: Radočaj, D.; Jurišić, M.; Tadić, V. The Effect of Bioclimatic Covariates on Ensemble Machine Learning Prediction of Total Soil Carbon in the Pannonian Biogeoregion. *Agronomy* **2023**, *13*, 2516. https://doi.org/10.3390/agronomy13102516

Academic Editor: Xingang Xu

Received: 6 September 2023
Revised: 27 September 2023
Accepted: 28 September 2023
Published: 29 September 2023

1. Introduction

Soil carbon, as a vital component of terrestrial ecosystems, has a crucial role in supporting biodiversity and sustaining agricultural productivity [1]. The distribution and dynamics of soil total carbon (TC) are influenced by a variety of biotic and abiotic factors, making accurate geospatial predictions of TC across large geographic regions a challenging but essential aim [2]. Successful soil TC prediction models are indispensable for informed decision-making in land management and climate change mitigation efforts [3]. Previous research indicated the relationship of bioclimatic variables with soil TC levels, as these factors quantify climate effects essential in shaping soil carbon dynamics and storage in terrestrial ecosystems [4]. Bioclimatic variables represent key climatic factors that directly influence biological processes, and their interactions with soil properties are fundamental in determining the distribution and accumulation of organic carbon in the soil [5]. Understanding the intricate relationships between bioclimatic variables and soil TC levels is essential for unraveling the complex mechanisms governing soil carbon sequestration and turnover, as well as for predicting the impacts of climate change on soil carbon stocks [6]. With ongoing climate change, shifts in bioclimatic variables are expected to have significant implications for soil carbon storage. Changes in temperature and precipitation patterns can alter the balance between carbon inputs and outputs from the soil, potentially leading to changes in soil carbon stocks [7].

Machine learning techniques have lately acquired popularity in the geospatial prediction of soil characteristics due to their capacity to handle complicated, non-linear interactions and evaluate large datasets [8,9]. Among these strategies, ensemble machine learning techniques based on effective machine learning algorithms, such as support vector machines (SVM), random forests (RF), and extreme gradient boosting (XGB), have shown to be effective tools for improving prediction accuracy [8,10]. In order to ensure that all features of the data are taken into account, ensemble machine learning approaches combine predictions from many base models, each of which was trained on a different subset of the data or with distinct variations in the feature space [11]. The ensemble model produces a more thorough and reliable representation of the underlying soil property patterns, improving predictive performance, by aggregating predictions from these distinct models [12]. Complex non-linear interactions between vegetation or soil characteristics and environmental factors including elevation, land use, climate, and topography characteristics are frequently present [4,13]. Due to their adaptability, ensemble models can accurately represent the regional variety of soil characteristics.

The Pannonian biogeoregion, located in Central Europe, is characterized by diverse landscapes ranging from fertile plains to hilly terrains and supports a significant extent of agricultural activities. Consequently, understanding the spatial distribution and dynamics of soil TC in this region is crucial for implementing effective land management practices and mitigating the impacts of climate change [14]. Nevertheless, while ensemble machine learning techniques have shown promise in predicting soil carbon properties, little research has focused on investigating the role of bioclimatic covariates in these models within the context of the Pannonian biogeoregion.

The objectives of this study were to provide (1) evaluation of base environmental covariates, including surface reflectance, phenology, and derived covariates, and the addition of bioclimatic covariates in the prediction of soil TC; and (2) assessment of three individual machine learning methods, including RF, XGB, and SVM, as well as their ensemble for soil TC prediction. By providing the answers to these objectives, this study aims to evaluate the effect of bioclimatic covariates on the accuracy and robustness of ensemble machine learning models for predicting total soil carbon in the Pannonian biogeoregion.

2. Materials and Methods

2.1. Study Area and Soil Data

The study area covers the Pannonian biogeoregion, a distinct geographic area located primarily in Central Europe, covering parts of several countries, including Hungary, Slovakia, Serbia, Romania, Czechia, and Croatia, according to the European Environment Agency [15] (Figure 1). The Pannonian biogeoregion is typically classified under the Köppen–Geiger climate classification as Cfb climate, signifying a temperate, humid climate with warm summers and cool winters [16]. The Pannonian Plain, which forms a significant part of the Pannonian Biogeoregion, contributes to the formation of this climate due to its lowland topography and proximity to the Carpathian Mountains. The plains allow for the accumulation of heat during the summer months, while the presence of the Carpathian Mountains to the north and northwest serves as a barrier to cold polar air masses during winter, resulting in milder temperatures compared to regions further east at similar latitudes.

A total of 145 soil samples from the Geochemical mapping of agricultural and grazing land soil (GEMAS) project with the soil TC data collected during 2008–2009 were used in the study [17]. The harmonized soil sampling data were downloaded from the World Soil Information Service (WoSIS) of the International Soil Reference and Information Centre (ISRIC) database using the Web feature service (WFS) [18]. The descriptive statistics of input soil TC values are displayed with boxplots in Figure 2, representing variation per country and USDA soil taxonomy great groups from OpenLandMap [19] in the study area. The initial dataset was randomly split to training in test data using the split-sample method, where the dataset was divided into a training set (70%) and a test set (30%). The training set was used to train the machine learning models, while the independent test set was used to evaluate the prediction accuracy of tested methods. This procedure and split ratio of 70:30 for training and test data were successfully used in previous digital soil mapping studies [20,21].

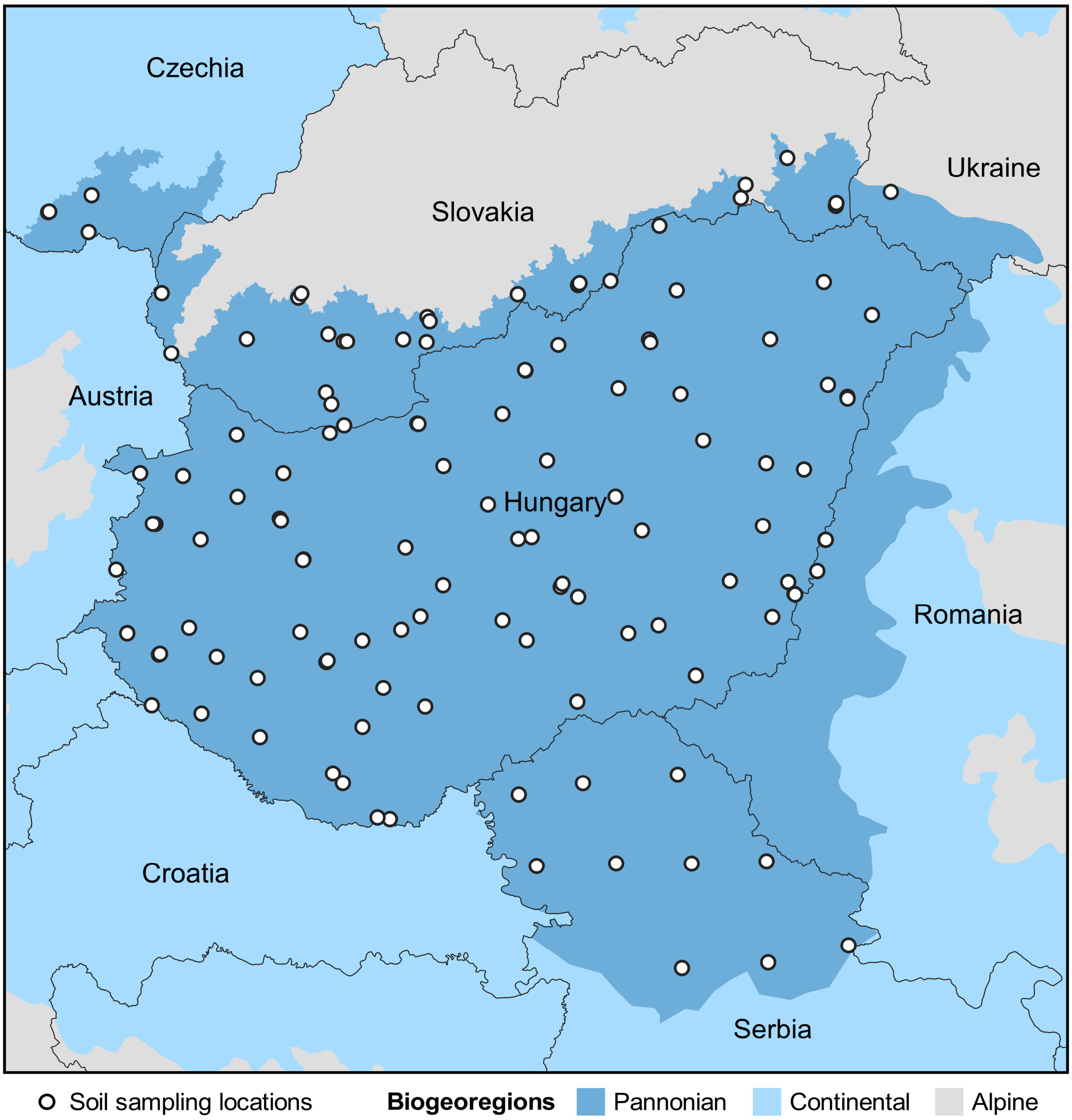

Figure 1. The study area containing the Pannonian biogeoregion in Central Europe.

2.2. Climate and Base Environmental Covariates

The integration of various environmental covariates in the predictive model allowed modeling of the complex interactions between soil TC and its surrounding environmental factors, including climate, vegetation, topography, and other derived data. This approach was recommended by several studies based on the machine learning prediction of soil properties [8,22,23]. To evaluate the effect of bioclimatic covariates on the prediction accuracy of soil TC, the machine learning prediction was performed in two instances based on the environmental covariate selection, including (1) base environmental covariates, consisting of surface reflectance, phenology, and derived covariates (Table 1); and (2) the combination of bioclimatic (Table 2) and base environmental covariates. The environmental covariates were assigned to individual soil sample points using the Google Earth Engine reducer. The yearly medians of the selected environmental covariates for the years 2008 and 2009 were used in the study to match the temporal frame of GEMAS field soil sampling.

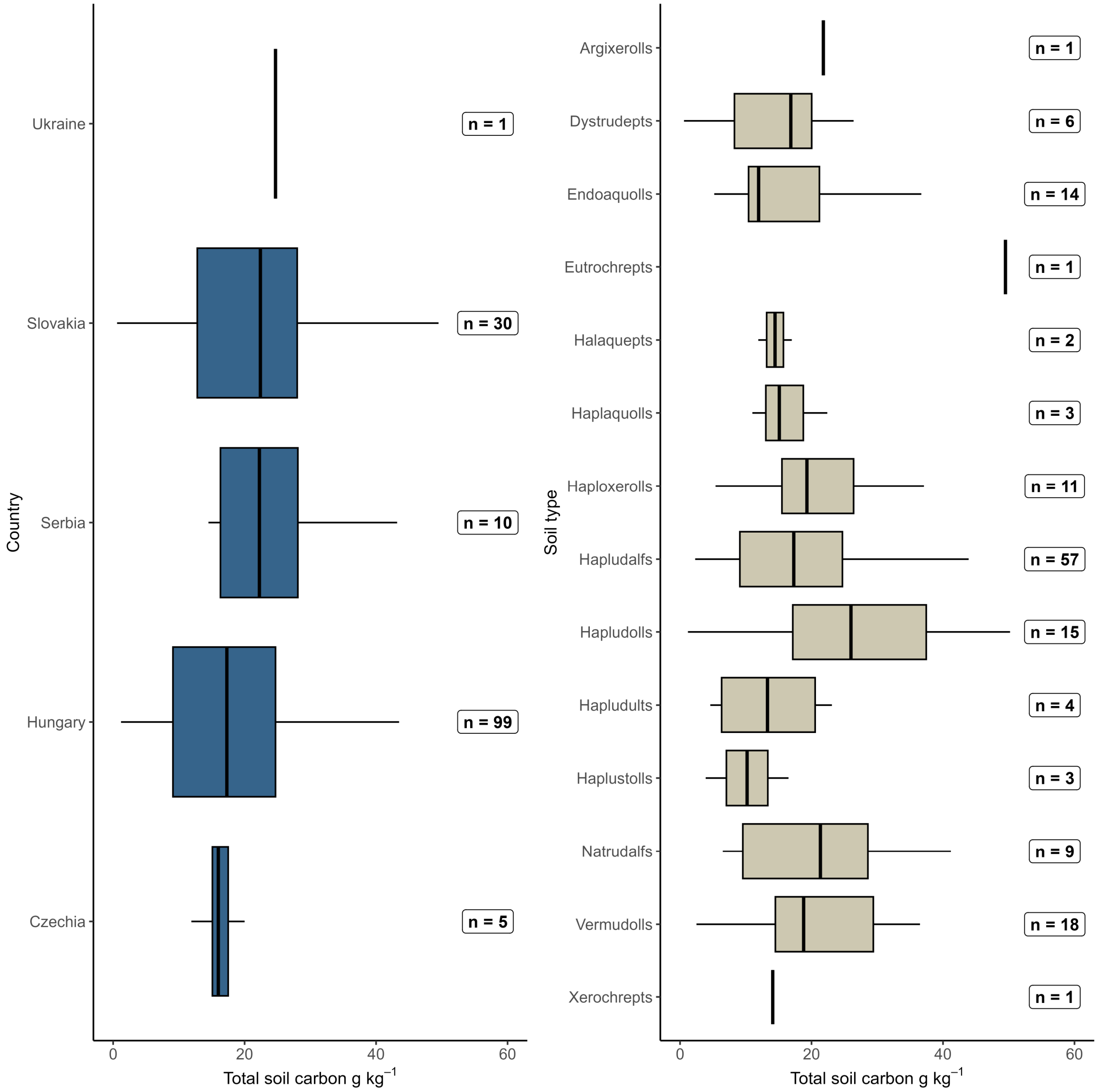

Figure 2. The boxplots representing soil TC variation per country and USDA soil taxonomy great groups from OpenLandMap [19] in the study area.

Surface reflectance data from multiple bands in the visible and near-infrared range are useful for characterizing vegetation cover and health, since healthy and dense vegetation is typically associated with higher carbon sequestration in soils due to enhanced photosynthesis and organic matter inputs [6]. The reflectance data from different bands can be used to derive vegetation indices, which are indicators of vegetation vigor and often correlate with higher soil TC content [24]. Vegetation phenology covariates represent different stages of vegetation growth and development throughout the year, providing insights into the timing of peak vegetation activity, periods of active photosynthesis, and vegetation senescence [25]. Variations in vegetation phenology impact the input of organic matter into the soil and affect the decomposition rates, thus influencing total soil carbon content [26]. Among the derived covariates, elevation data are crucial in geospatial modeling as they affect temperature, precipitation, and topographic features [27], which in turn impact vegetation distribution and soil TC dynamics [28]. Additionally, land surface temperature provided insights into soil heat flux and temperature fluctuations, which influence microbial activity and soil TC decomposition rates. Biophysical variables and gross primary productivity are additional metrics related to vegetation activity and productivity, which also influence soil TC [29].

Table 1. Base environmental covariates used for the prediction of soil TC.

Covariate Group	Individual Covariate	Data Source	Reference
Surface reflectance	Surface reflectance in 620–670 nm band (B01)	MOD09A1	[30]
	Surface reflectance in 841–876 nm band (B02)		
	Surface reflectance in 459–479 nm band (B03)		
	Surface reflectance in 545–565 nm band (B04)		
	Surface reflectance in 1230–1250 nm band (B05)		
	Surface reflectance in 1628–1652 nm band (B06)		
	Surface reflectance in 2105–2155 nm band (B07)		
Phenology covariates	Greenup	MCD12Q2	[31]
	MidGreenup		
	Peak		
	Maturity		
	MidGreendown		
	Senescence		
	Dormancy		
	Area under enhanced vegetation index 2 curve (EVI_Area)		
Derived covariates	Elevation	SRTM	[32]
	USDA soil taxonomy great groups (soil_type)	OpenLandMap	[19]
	Land surface temperature during day (LST_Day)	MOD11A1	[33]
	Land surface temperature during night (LST_Night)		
	Normalized difference vegetation index (NDVI)	MOD13A2	[34]
	Enhanced vegetation index (EVI)		
	Leaf area index (LAI)	MOD15A2H	[35]
	Fraction of absorbed photosynthetically active radiation (FAPAR)		
	Gross primary productivity (GPP)	MOD17A2H	[36]
	Net photosynthesis (NetPsy)		

Table 2. Bioclimatic environmental covariates used for the prediction of soil TC from the WorldClim dataset [37].

Climate Parameter	Label	Description	Unit	Value Range in the Study Area	
				Min	Max
Air temperature	bio01	annual mean	°C	5.2	12.5
	bio02	mean diurnal range	°C	8.6	10.6
	bio03	isothermality	%	28	33
	bio04	seasonality	°C	70.4	81.0
	bio05	max of warmest month	°C	20.1	28.8
	bio06	max of coldest month	°C	–8.8	–1.3
	bio07	annual range	°C	28.6	33.1
	bio08	mean of wettest quarter	°C	9.3	20.6
	bio09	mean of driest quarter	°C	–4.2	18.1
	bio10	max of warmest quarter	°C	14.1	21.8
	bio11	max of coldest quarter	°C	–4.2	2.3

Table 2. *Cont.*

Climate Parameter	Label	Description	Unit	Value Range in the Study Area	
				Min	Max
Precipitation	bio12	annual total	mm	512	856
	bio13	total of warmest month	mm	61	123
	bio14	total of coldest month	mm	23	48
	bio15	seasonality	CV	19	43
	bio16	total of wettest quarter	mm	164	317
	bio17	total of driest quarter	mm	72	150
	bio18	total of warmest quarter	mm	164	317
	bio19	total of coldest quarter	mm	76	172

The 19 bioclimatic covariates from the WorldClim v2 database (1970–2000) represent various aspects of air temperature and precipitation, such as their annual means, quarterly extremes, and seasonal variations [37] (Figure 3). Air temperature influences the rate of soil organic matter decomposition, which affects the buildup and turnover of soil TC. Higher temperatures generally lead to increased microbial activity and faster decomposition of organic matter, potentially reducing soil carbon content, while cooler temperatures may slow down decomposition processes and result in higher soil carbon retention [38]. Precipitation is a critical driver of primary productivity and plant growth, affecting the input of organic matter into the soil through litterfall and root turnover, contributing to soil TC accumulation [39]. Areas with higher precipitation may support more vegetation and subsequently higher soil carbon content, while excessively wet or dry conditions can influence the decomposition rates of soil organic matter, consequentially impacting soil TC levels.

2.3. Ensemble Machine Learning Prediction and Accuracy Assessment

Geospatial prediction of soil TC was performed by three individual machine learning methods, including RF, XGB, and SVM as their ensemble, resulting in a total of four prediction approaches for each set of input environmental covariates. The tuning hyperparameters were fine-tuned using a grid search approach, with the selection of optimal hyperparameters performed according to the lowest root mean square error (RMSE). All training data columns were preprocessed with standardization and division with the standard deviation prior to machine learning prediction.

The RF combined multiple decision trees to create a robust and accurate prediction model, with each tree being trained on a bootstrap sample of the training data [40]. During the decision-making process, the model aggregated the predictions of individual trees to arrive at the final output. Three tuning hyperparameters were used for the RF predictions, including (1) the number of predictors randomly selected at each node of the decision tree during the construction of the RF (mtry); (2) the method used to select the best split at each node of the decision tree (splitrule); and (3) the minimum number of samples required to create a terminal node during the tree-building process (min.node.size). Similarly to RF, XGB sequentially built decision trees, while each one aimed to correct the errors of the previous trees based on a gradient boosting framework, where the trees are built in a manner that focuses on the instances that were previously mispredicted, resulting in an iterative improvement of the model's performance [41]. As was the case for RF, three tuning hyperparameters were used for XGB prediction: (1) the number of boosting iterations (nrounds); (2) maximum depth of each individual tree (max_depth); and (3) learning rate determined by the step size at each boosting round when updating the model weights (eta). Unlike the previous two tree-based methods, SVM aimed to determine an optimal hyperplane that best separates the input data points [42]. The two tuning parameters for SVM were the regularization parameter that controls the relationship between maximizing the margin and reducing the prediction error on the training data (C) and the width of the radial basis function kernel which was utilized with SVM (sigma).

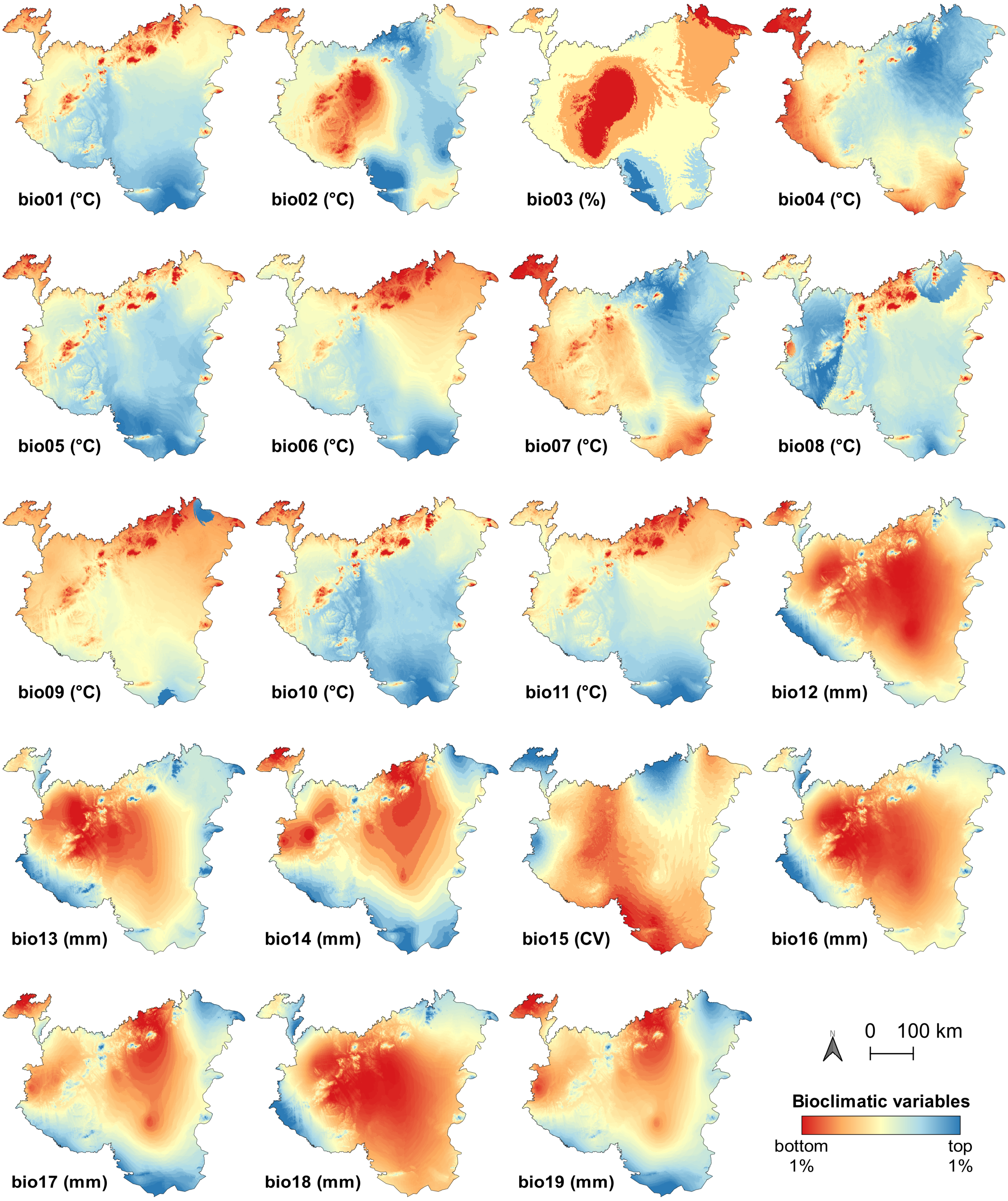

Figure 3. The display of 19 bioclimatic covariates in the study area from the WorldClim dataset.

To further enhance the geospatial prediction accuracy, an ensemble approach of RF, XGB, and SVM was adopted. The ensemble method combined predictions from multiple individual models to form a more robust and reliable prediction, aiming to mitigate the weaknesses of individual methods, potentially leading to improved predictive performance. In this study, this included bagging (RF), boosting (XGB), and support vector machine (SVM) approaches.

Two commonly used metrics for accuracy assessment, the coefficient of determination (R^2) and RMSE were used to evaluate the prediction accuracy of used machine learning prediction approaches. R^2 measures the proportion of variance in the observed data that is explained by the model, while RMSE quantifies the difference between predicted values and observed data points, providing an estimate of the model's prediction errors. A higher R^2 value suggested a better fit of the model to the observed data, implying higher predictive accuracy. Analogously, lower RMSE values indicated smaller prediction errors and better accuracy of the model.

3. Results and Discussion

3.1. Hyperparameter Tuning of Individual Machine Learning Methods

The machine learning prediction variants with all modeled environmental covariates (bioclimatic and base) and with only base covariates produced the same optimal tuning hyperparameters per individual method. Despite the similarity in optimal models for the prediction of soil TC regardless of the input environmental covariates, the variant with both bioclimatic and base covariates produced slightly lower RMSE for the optimal model (Figure 4), compared to the only base covariates (Figure 5). The hyperparameter tuning approach was recommended in previous studies for all three evaluated individual machine learning methods as it increased prediction accuracy for RF [43], XGB [44], and SVM [45].

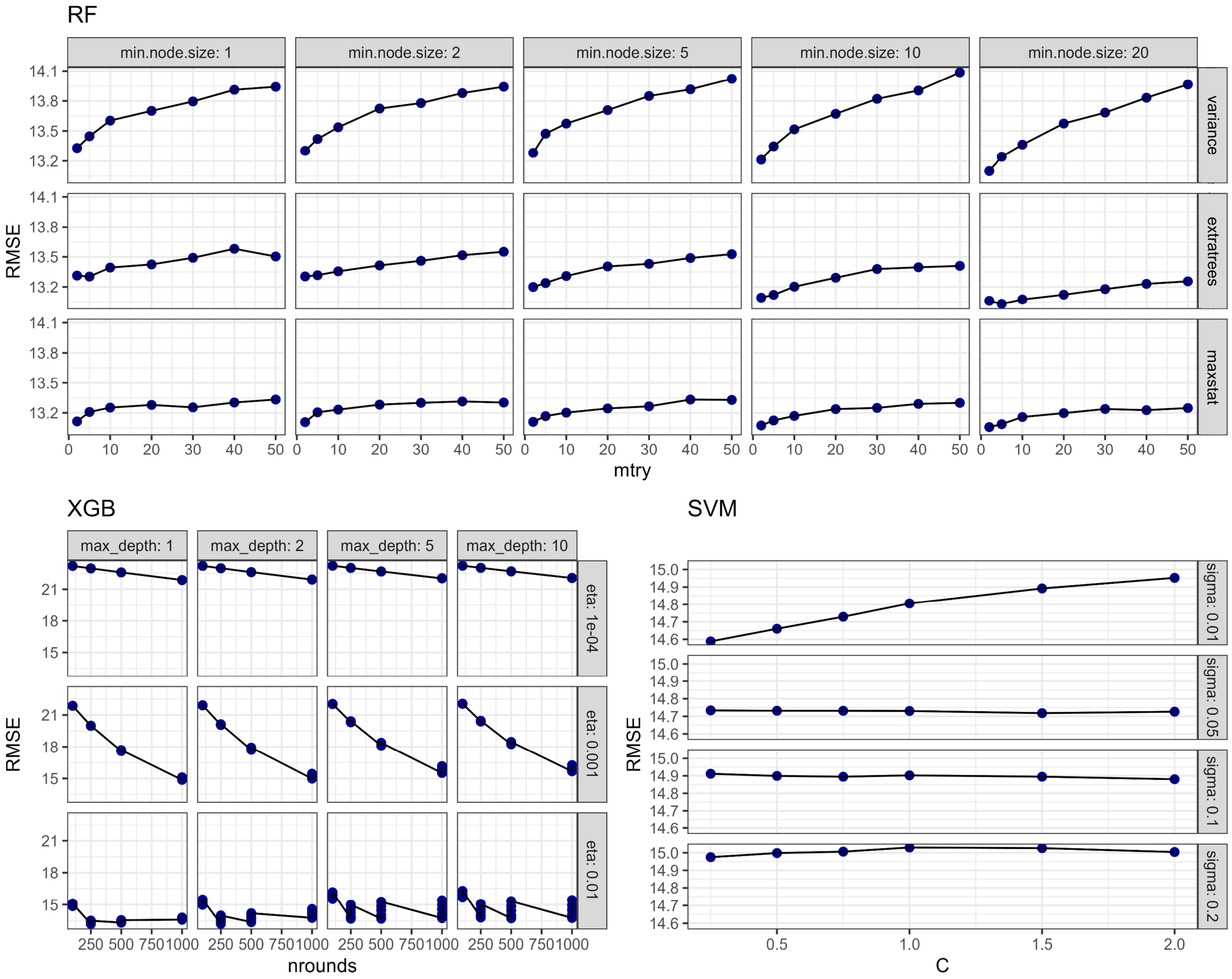

Figure 4. The optimal tuning hyperparameters for individual machine learning methods based on bioclimatic and base environmental covariates.

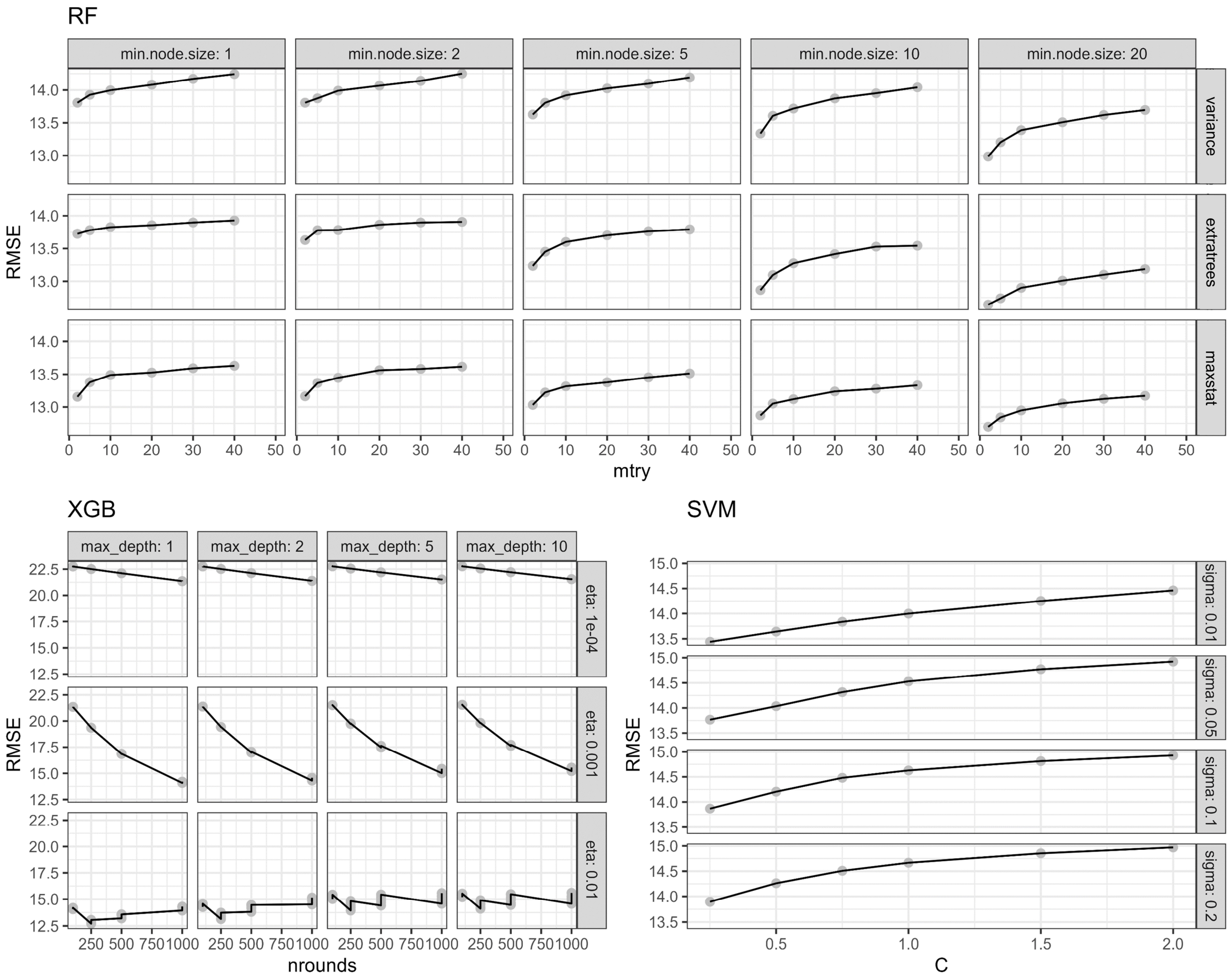

Figure 5. The optimal tuning hyperparameters for individual machine learning methods based on only base environmental covariates.

The optimized RF model with mtry = 5, splitrule = extratrees, and min.node.size = 20 produced a well-balanced ensemble with enhanced generalization capability. The mtry value of 5 promoted feature diversity and mitigated the risk of individual trees' overfitting, while the adoption of extratrees bolsted the model's robustness, enabling it to handle noisy and complex data effectively. Moreover, the specified 'min.node.size' facilitated the development of informative trees with deeper structures, enhancing the model's ability to capture intricate patterns in the data. The XGB with the optimal tuning hyperparameters of nrounds = 250, max_depth = 1, and eta = 0.01 produced a conservative splitting and robustness to imbalanced data. The max_depth hyperparameter setting restricted the decision trees in the XGB to shallow structures, preventing the model from becoming overly complex and reducing the risk of overfitting. The optimal prediction was performed in 250 boosting rounds, while the eta of 0.01 implied a relatively low step size during the boosting process for a stable convergence. The smaller optimal sigma value for SVM (sigma = 0.01) led to sharper and more localized decision boundaries around data points, enabling increased versatility in handling diverse data distributions and complex decision boundaries. Moreover, the optimal C value (C = 0.25) emphasized a simpler decision boundary with controlled margins, so the model exhibited improved generalization capabilities. This well-balanced regularization approach helped avoid overfitting, contributing to enhanced model robustness [46].

3.2. Prediction Accuracy of Ensemble and Individual Machine Learning Methods

The addition of bioclimatic covariate data enhanced the learning process of all evaluated methods, enabling them to capture more intricate relationships and improve their predictive performance, leading to higher R^2 and slightly lower RMSE (Table 3). The scatterplots of the predicted soil TC according to the test data are presented in Appendix A (Figures A1 and A2). The RF showed an increase of 0.052 in R^2 when bioclimatic covariates were included; the XGB exhibited a difference of 0.114, while SVM was the only method that produced higher accuracy with base covariates (Table 3). The ensemble model demonstrated an increase of 0.024. Similarly, the differences in RMSE further support the benefits of incorporating climate data. The RF model showed a reduction of 0.543 in RMSE, while the XGB model achieved a decrease of 1.011 when climate data was used. The SVM model exhibited a notable increase of 1.737, while the Ensemble model showed a reduction of 0.287. These negative differences suggest that the models' predictions were more accurate and closer to the true values when climate covariates were included. Overall, consistently higher R^2 and lower RMSE values with the inclusion of bioclimatic data strongly suggest that the inclusion of bioclimatic covariates improved the ability of evaluated machine learning methods to more accurately predict and explain the variability in total soil carbon. This confirms the observations from previous similar studies on a large scale [9,47].

Table 3. Prediction accuracy of the evaluated machine learning methods for the prediction of soil TC according to two environmental covariate selection variants.

Environmental Covariates	Method	R^2	RMSE
Bioclimatic and base covariates	RF	0.427	12.882
	XGB	0.408	12.280
	SVM	0.163	15.014
	Ensemble	0.580	10.392
Base covariates	RF	0.375	13.425
	XGB	0.294	13.291
	SVM	0.304	13.277
	Ensemble	0.548	10.679

When considering both prediction variants (with bioclimatic and base environmental covariates and with only base covariates), the ensemble approach resulted in the highest prediction accuracy of the evaluated methods. This strongly suggests that the ensemble method effectively captures complex relationships among environmental covariates, resulting in a relatively high degree of explained variance and lower prediction errors [48]. The ensemble method with bioclimatic covariates resulted in the best performance overall, achieving R^2 of 0.580 and RMSE of 10.392. This prediction accuracy is typical for the geospatial prediction of soil properties on a large scale according to the previous studies, which generally achieved prediction accuracy expressed by R^2 up to 0.5 [49–51]. Although RF, XGB, and SVM models also showed relative improvements with the addition of bioclimatic data, they significantly lagged the prediction accuracy of the ensemble approach. The decision tree-based methods, RF and XGB, produced relatively similar prediction accuracy in both environmental covariate selection variants, moderately outperforming SVM. The maps of the predicted soil TC according to evaluated machine learning methods and environmental covariate selection variants are presented in Figure 6.

3.3. Variable Importance of Climate and Base Environmental Covariates

Overall, the bioclimatic covariates produced moderate variable importance across evaluated machine learning methods (Figure 7). bio09 (mean temperature of driest quarter) and bio05 (mean temperature of warmest month) displayed substantial importance in the RF and SVM models, whereas bio16 (precipitation of wettest quarter) had higher importance in the XGB model. These discrepancies might be attributed to the distinct modeling approaches of each algorithm and their ability to capture unique relationships between these climate-related variables and total soil carbon levels [52].

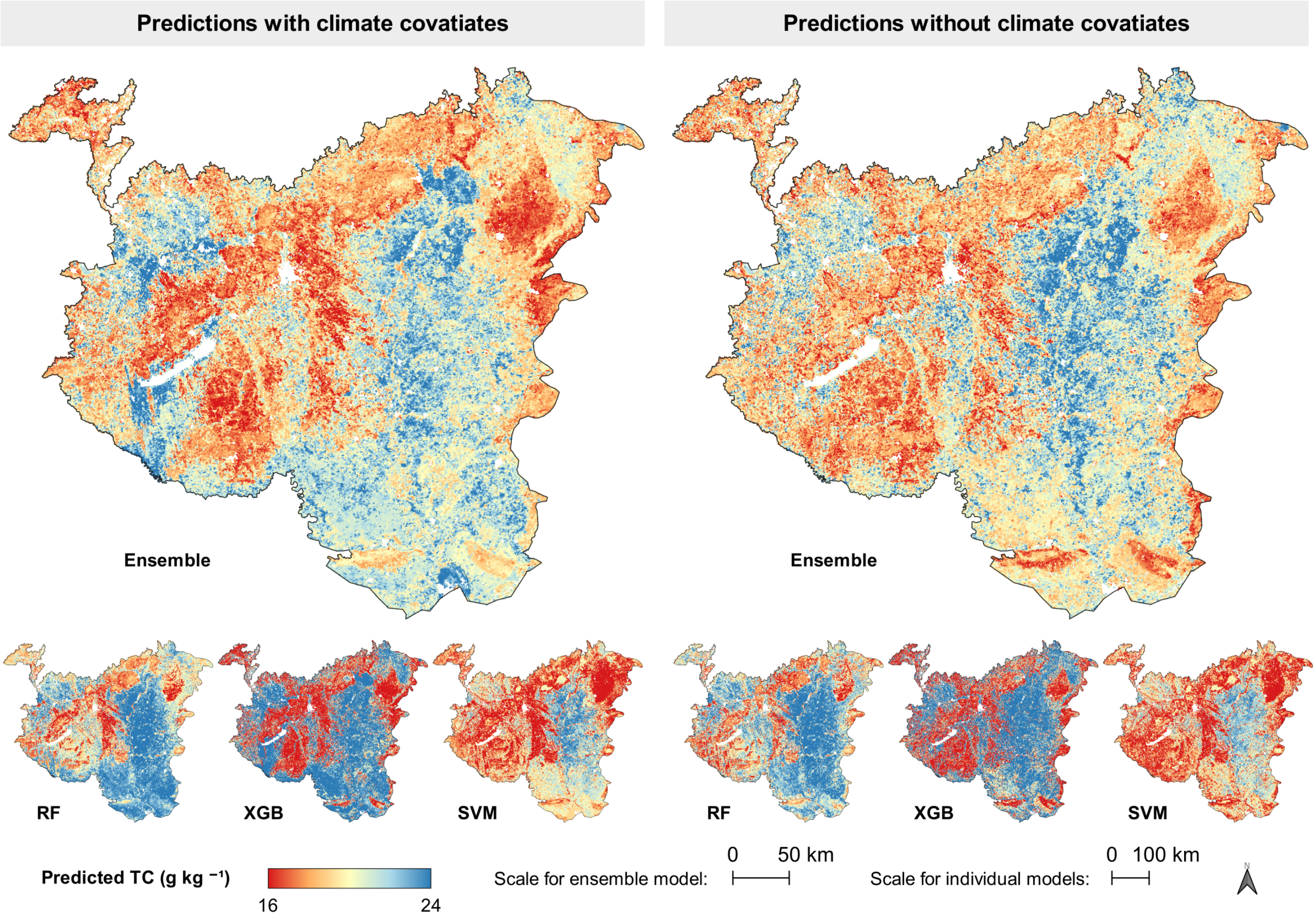
Predictions with climate covatiates
Predictions without climate covatiates
Ensemble
Ensemble
RF
XGB
SVM
RF
XGB
SVM
Predicted TC (g kg ⁻¹)
16
24
Scale for ensemble model:
0
50 km
Scale for individual models:
0
100 km

Figure 6. Maps of the predicted soil TC according to evaluated machine learning methods and environmental covariate selection variants.

The non-climate covariate groups (surface reflectance, phenology, and derived covariates) produced variations in the relative importance scores between the models. Elevation resulted as a derived covariate with the highest relative variable importance, ranking as one of the top covariates for SVM. This strongly suggests that variations in elevation have a significant impact on the distribution of soil TC, confirming the observations from multiple previous studies [53,54]. Additionally, vegetation-related covariates, such as NDVI and GPP, and phenology covariates, especially Senescence and Dormancy, also exhibited substantial importance scores across the individual models. These findings highlight the importance of vegetation dynamics and its interaction with total soil carbon levels [55]. The RF model attributed high importance to phenology covariates, including Senescence (during 2008), Dormancy (during 2008 and 2009), and MidGreendown (during 2008), indicating their strong influence on the RF model's predictions. The RF also strongly favored both biophysical variables (LAI and FAPAR), as well as NDVI. The XGB model assigned higher importance to MidGreendown during 2009 among the phenology covariates. However, it was much more exclusive towards the rest of the covariate groups, with the exception of surface reflectance in the green band (B04) during 2008. The SVM model highlighted MidGreendown and Midgreenup, with near-infrared (B05) during 2008 and 2009 as crucial predictors. However, it did not restrict its predictions to these covariates and produced several moderate variable importance values over all covariate groups, similarly to the RF.

These variations in importance scores reflect differences in modeling approaches and how the models capture relationships between covariates and total soil carbon [56].

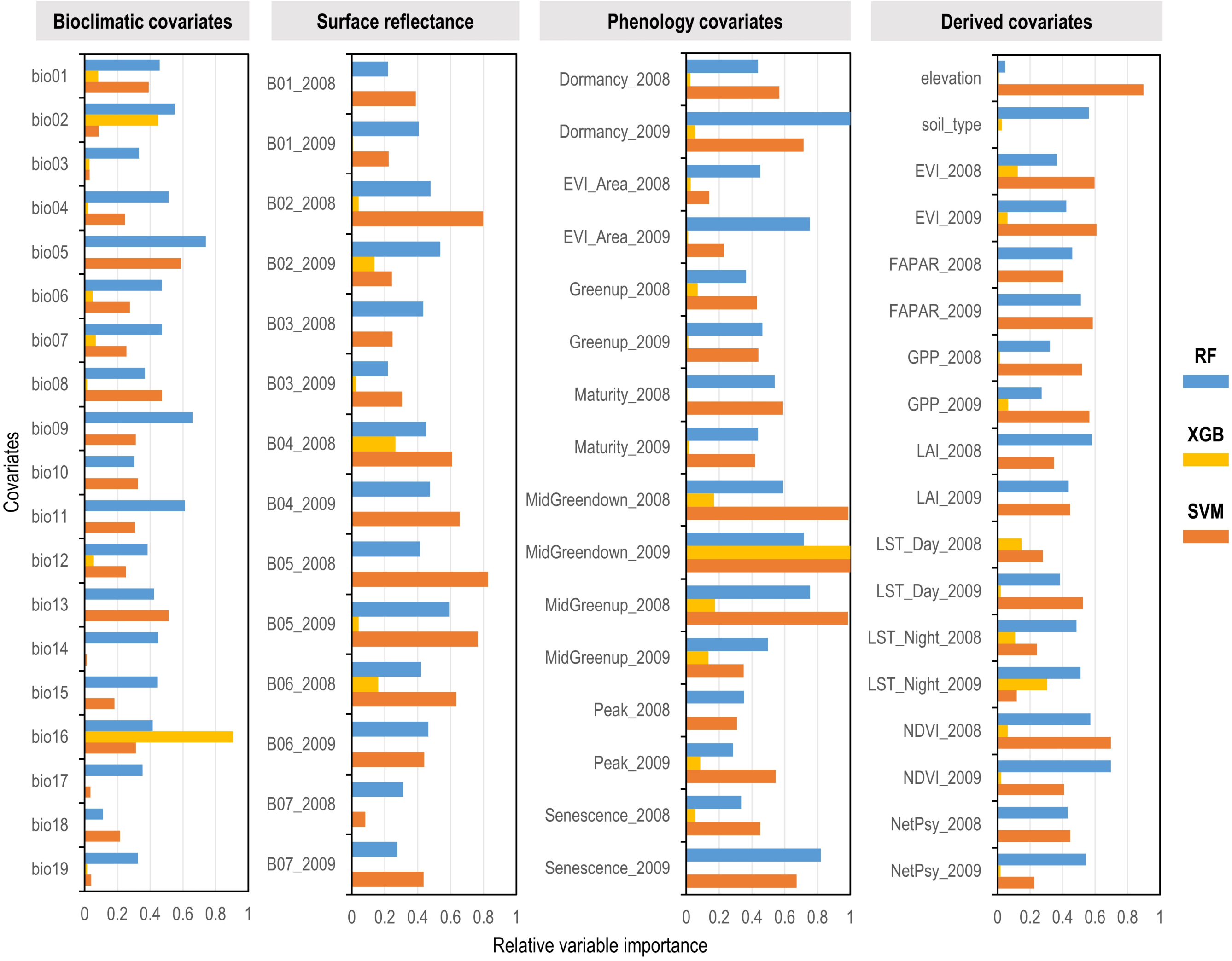

Figure 7. Relative variable importance of all input environmental covariates for evaluated individual machine learning methods.

4. Conclusions

The integration of diverse environmental covariates in the predictive model allowed for a comprehensive examination of the complex interactions between soil TC and its surrounding environmental factors, including climate, vegetation, topography, and other derived data. This evaluation demonstrated that the inclusion of bioclimatic covariates only slightly improved the prediction accuracy of all evaluated methods. Among the evaluated machine learning methods, the ensemble approach exhibited the highest prediction accuracy overall, outperforming the individual models. The ensemble method with bioclimatic covariates achieved an R^2 of 0.580 and an RMSE of 10.392, demonstrating its effectiveness in capturing complex relationships among environmental covariates. The variable importance analysis revealed that bioclimatic covariates displayed moderate importance across all evaluated machine learning methods. Specific climate-related variables, such as mean temperature of the driest and coldest quarters, were particularly influential in the RF and SVM models. On the other hand, the mean temperature of the wettest quarter exhibited higher importance in the XGB model. Non-climate covariates, including surface reflectance, phenology, and derived covariates, also showed variations in importance scores between the models. Elevation consistently ranked as a top covariate for all evaluated individual machine learning methods, emphasizing its significant impact on the distribution of soil TC.

In conclusion, this study provided answers to two main research aims: (1) the inclusion of bioclimatic covariates improved the ability of all evaluated machine learning methods to accurately predict and explain the variability in total soil carbon, leading to more robust and reliable predictions; and (2) the ensemble method, with its capability to capture complex relationships among environmental covariates, demonstrated superior predictive performance compared to individual machine learning methods. These findings contribute insights into the application of ensemble machine learning in geospatial prediction of soil properties and support informed decision-making in land management and environmental planning in the Pannonian biogeoregion and similar regions worldwide.

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

Funding: This research received no external funding.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://data.isric.org/geonetwork/srv/api/records/4a76ff8c-0f33-4327-a3d7-56b7c47005ba.

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

Appendix A

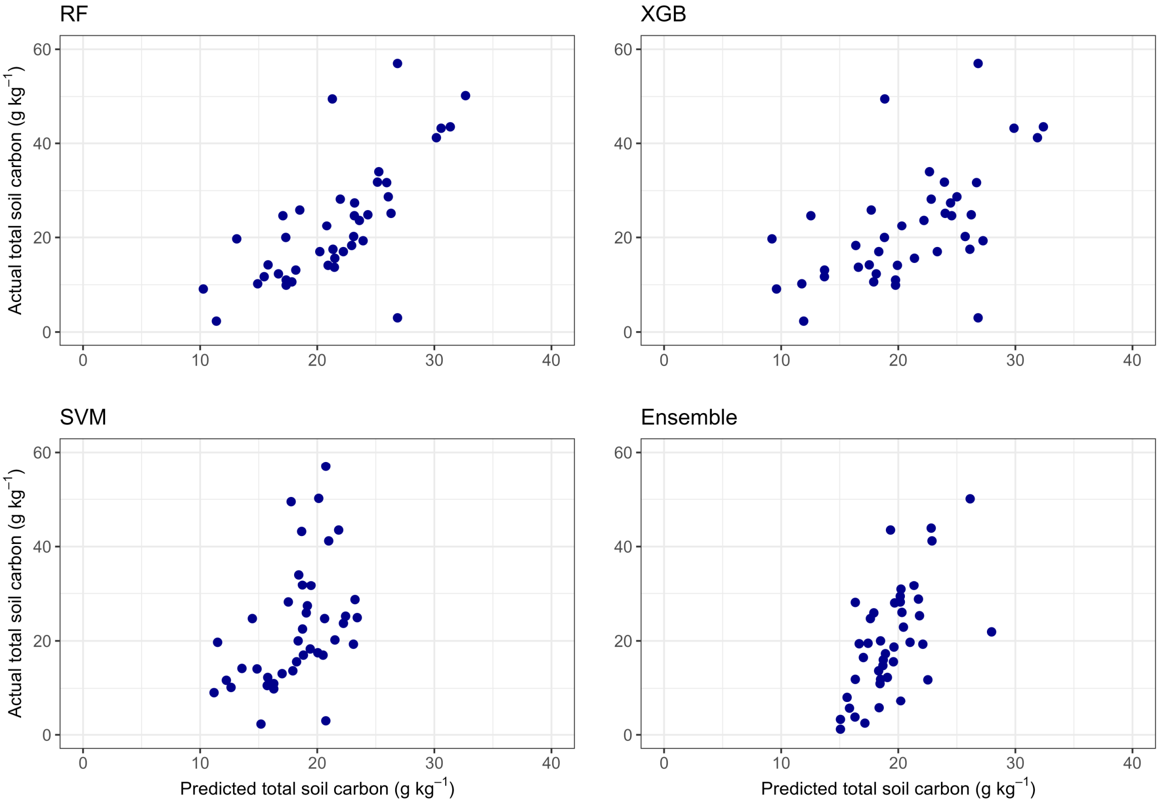

Figure A1. Scatterplots of the predicted and actual TC based on the bioclimatic and base environmental covariates.

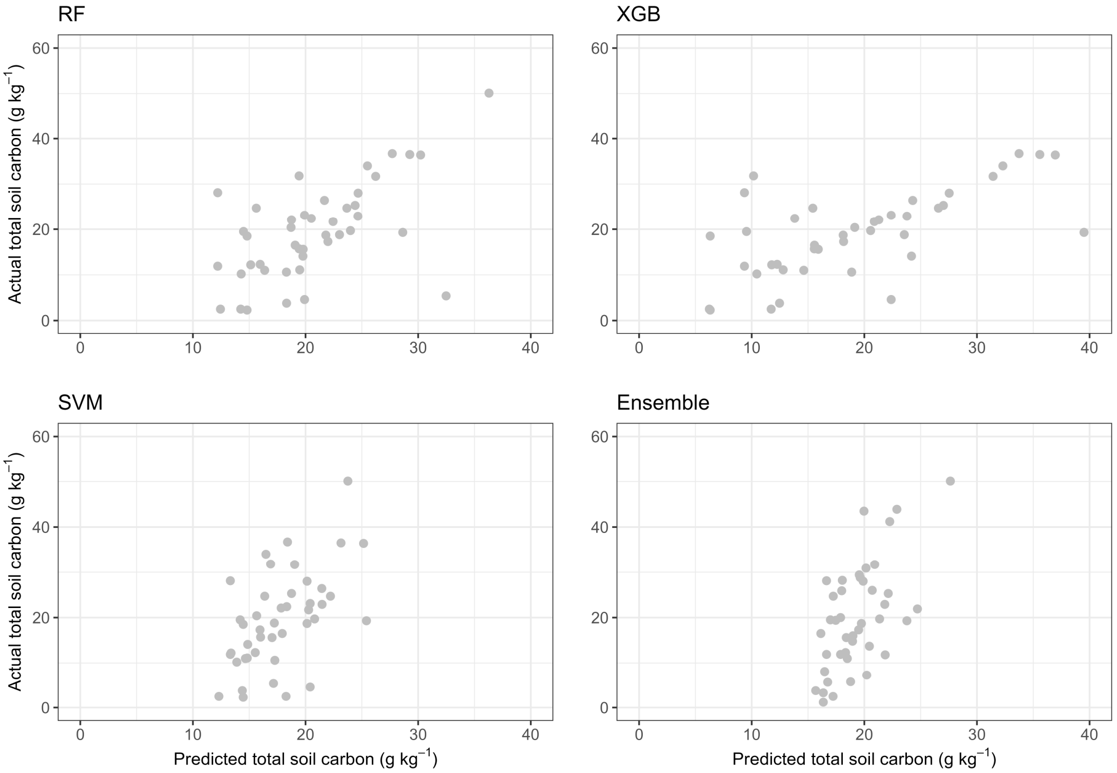

Figure A2. Scatterplots of the predicted and actual TC based only on base environmental covariates.

References

1. Bhattacharya, S.S.; Kim, K.-H.; Das, S.; Uchimiya, M.; Jeon, B.H.; Kwon, E.; Szulejko, J.E. A Review on the Role of Organic Inputs in Maintaining the Soil Carbon Pool of the Terrestrial Ecosystem. *J. Environ. Manag.* **2016**, *167*, 214–227. [CrossRef]
2. Keskin, H.; Grunwald, S.; Harris, W.G. Digital Mapping of Soil Carbon Fractions with Machine Learning. *Geoderma* **2019**, *339*, 40–58. [CrossRef]
3. Taylor, A.; Kalnins, A.; Koot, M.; Jackson, R.; Toloza, A.; Ahmed, H.S.; Goddard, R.; Blake, W.H. Portable Gamma Spectrometry for Rapid Assessment of Soil Texture, Organic Carbon and Total Nitrogen in Agricultural Soils. *J. Soils Sediments* **2023**, *23*, 2556–2563. [CrossRef]
4. Zeraatpisheh, M.; Garosi, Y.; Reza Owliaie, H.; Ayoubi, S.; Taghizadeh-Mehrjardi, R.; Scholten, T.; Xu, M. Improving the Spatial Prediction of Soil Organic Carbon Using Environmental Covariates Selection: A Comparison of a Group of Environmental Covariates. *Catena* **2022**, *208*, 105723. [CrossRef]
5. Tayebi, M.; Fim Rosas, J.T.; Mendes, W.D.S.; Poppiel, R.R.; Ostovari, Y.; Ruiz, L.F.C.; dos Santos, N.V.; Cerri, C.E.P.; Silva, S.H.G.; Curi, N.; et al. Drivers of Organic Carbon Stocks in Different LULC History and along Soil Depth for a 30 Years Image Time Series. *Remote Sens.* **2021**, *13*, 2223. [CrossRef]
6. Elbasiouny, H.; El-Ramady, H.; Elbehiry, F.; Rajput, V.D.; Minkina, T.; Mandzhieva, S. Plant Nutrition under Climate Change and Soil Carbon Sequestration. *Sustainability* **2022**, *14*, 914. [CrossRef]
7. Hombegowda, H.C.; van Straaten, O.; Köhler, M.; Hölscher, D. On the Rebound: Soil Organic Carbon Stocks Can Bounce Back to near Forest Levels When Agroforests Replace Agriculture in Southern India. *Soil* **2016**, *2*, 13–23. [CrossRef]
8. Hengl, T.; de Jesus, J.M.; Heuvelink, G.B.M.; Gonzalez, M.R.; 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]
9. Hengl, T.; Miller, M.A.E.; Križan, J.; Shepherd, K.D.; Sila, A.; Kilibarda, M.; Antonijević, O.; Glušica, L.; Dobermann, A.; Haefele, S.M.; et al. African Soil Properties and Nutrients Mapped at 30 m Spatial Resolution Using Two-Scale Ensemble Machine Learning. *Sci. Rep.* **2021**, *11*, 6130. [CrossRef]
10. Radočaj, D.; Jurišić, M.; Antonić, O.; Šiljeg, A.; Cukrov, N.; Rapčan, I.; Plaščak, I.; Gašparović, M. A Multiscale Cost–Benefit Analysis of Digital Soil Mapping Methods for Sustainable Land Management. *Sustainability* **2022**, *14*, 12170. [CrossRef]
11. Sagi, O.; Rokach, L. Ensemble Learning: A Survey. *WIREs Data Min. Knowl. Discov.* **2018**, *8*, e1249. [CrossRef]
12. Sylvain, J.-D.; Anctil, F.; Thiffault, É. Using Bias Correction and Ensemble Modelling for Predictive Mapping and Related Uncertainty: A Case Study in Digital Soil Mapping. *Geoderma* **2021**, *403*, 115153. [CrossRef]
13. Radočaj, D.; Vinković, T.; Jurišić, M.; Gašparović, M. The Relationship of Environmental Factors and the Cropland Suitability Levels for Soybean Cultivation Determined by Machine Learning. *Poljoprivreda* **2022**, *28*, 53–59. [CrossRef]

14. Nadeu, E.; Gobin, A.; Fiener, P.; van Wesemael, B.; van Oost, K. Modelling the Impact of Agricultural Management on Soil Carbon Stocks at the Regional Scale: The Role of Lateral Fluxes. *Glob. Chang. Biol.* **2015**, *21*, 3181–3192. [CrossRef]
15. European Environment Agency. Biogeographical Regions. Available online: https://www.eea.europa.eu/en/datahub/datahubitem-view/11db8d14-f167-4cd5-9205-95638dfd9618 (accessed on 30 July 2023).
16. Beck, H.E.; Zimmermann, N.E.; McVicar, T.R.; Vergopolan, N.; Berg, A.; Wood, E.F. Present and Future Koppen-Geiger Climate Classification Maps at 1-Km Resolution. *Sci. Data* **2018**, *5*, 180214. [CrossRef]
17. Négrel, P.; Ladenberger, A.; Reimann, C.; Birke, M.; Demetriades, A.; Sadeghi, M.; Albanese, S.; Andersson, M.; Baritz, R.; Batista, M.J.; et al. GEMAS: Geochemical Distribution of Mg in Agricultural Soil of Europe. *J. Geochem. Explor.* **2021**, *221*, 106706. [CrossRef]
18. Batjes, N.H.; Ribeiro, E.; van Oostrum, A. Standardised Soil Profile Data to Support Global Mapping and Modelling (WoSIS Snapshot 2019). *Earth Syst. Sci. Data* **2020**, *12*, 299–320. [CrossRef]
19. Hengl, T.; Nauman, T. *Predicted USDA Soil Great Groups at 250 m (Probabilities)*; Zenodo: Geneva, Switzerland, 2018.
20. Taghizadeh-Mehrjardi, R.; Nabiollahi, K.; Minasny, B.; Triantafilis, J. Comparing Data Mining Classifiers to Predict Spatial Distribution of USDA-Family Soil Groups in Baneh Region, Iran. *Geoderma* **2015**, *253–254*, 67–77. [CrossRef]
21. Saha, S.; Roy, J.; Pradhan, B.; Hembram, T.K. Hybrid Ensemble Machine Learning Approaches for Landslide Susceptibility Mapping Using Different Sampling Ratios at East Sikkim Himalayan, India. *Adv. Space Res.* **2021**, *68*, 2819–2840. [CrossRef]
22. 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]
23. 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]
24. Zhao, M.-S.; Rossiter, D.G.; Li, D.-C.; Zhao, Y.-G.; Liu, F.; Zhang, G.-L. Mapping Soil Organic Matter in Low-Relief Areas Based on Land Surface Diurnal Temperature Difference and a Vegetation Index. *Ecol. Indic.* **2014**, *39*, 120–133. [CrossRef]
25. Körner, C.; Möhl, P.; Hiltbrunner, E. Four Ways to Define the Growing Season. *Ecol. Lett.* **2023**, *26*, 1277–1292. [CrossRef] [PubMed]
26. Moore, C.E.; Brown, T.; Keenan, T.F.; Duursma, R.A.; van Dijk, A.I.J.M.; Beringer, J.; Culvenor, D.; Evans, B.; Huete, A.; Hutley, L.B.; et al. Reviews and Syntheses: Australian Vegetation Phenology: New Insights from Satellite Remote Sensing and Digital Repeat Photography. *Biogeosciences* **2016**, *13*, 5085–5102. [CrossRef]
27. Radočaj, D.; Jurišić, M.; Gašparović, M. A Wildfire Growth Prediction and Evaluation Approach Using Landsat and MODIS Data. *J. Environ. Manag.* **2022**, *304*, 114351. [CrossRef] [PubMed]
28. Gonçalves, D.R.P.; Mishra, U.; Wills, S.; Gautam, S. Regional Environmental Controllers Influence Continental Scale Soil Carbon Stocks and Future Carbon Dynamics. *Sci. Rep.* **2021**, *11*, 6474. [CrossRef]
29. Woltz, V.L.; Stagg, C.L.; Byrd, K.B.; Windham-Myers, L.; Rovai, A.S.; Zhu, Z. Above- and Belowground Biomass Carbon Stock and Net Primary Productivity Maps for Tidal Herbaceous Marshes of the United States. *Remote Sens.* **2023**, *15*, 1697. [CrossRef]
30. Vermote, E. *MODIS/Terra Surface Reflectance 8-Day L3 Global 500m SIN Grid V061*; NASA EOSDIS Land Processes DAAC: Missoula, MT, USA, 2021.
31. Friedl, M.; Gray, J.; Sulla-Menashe, D. *MODIS/Terra+Aqua Land Cover Dynamics Yearly L3 Global 500 m SIN Grid V061*; NASA EOSDIS Land Processes DAAC: Missoula, MT, USA, 2022.
32. NASA JPL. *NASA Shuttle Radar Topography Mission Global 1 Arc Second*; NASA JPL: Pasadena, CA, USA, 2013.
33. Wan, Z.; Hook, S.; Hulley, G. *MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 1 km SIN Grid V061*; NASA EOSDIS Land Processes DAAC: Missoula, MT, USA, 2021.
34. Didan, K. *MODIS/Terra Vegetation Indices 16-Day L3 Global 1 km SIN Grid V061*; NASA EOSDIS Land Processes DAAC: Missoula, MT, USA, 2021.
35. Myneni, R.; Knyazikhin, Y.; Park, T. *MODIS/Terra Leaf Area Index/FPAR 8-Day L4 Global 500 m SIN Grid V061*; NASA EOSDIS Land Processes DAAC: Missoula, MT, USA, 2021.
36. Running, S.; Zhao, M. *MODIS/Terra Gross Primary Productivity Gap-Filled 8-Day L4 Global 500 m SIN Grid V061*; NASA EOSDIS Land Processes DAAC: Missoula, MT, USA, 2021.
37. Fick, S.E.; Hijmans, R.J. WorldClim 2: New 1-km Spatial Resolution Climate Surfaces for Global Land Areas. *Int. J. Climatol.* **2017**, *37*, 4302–4315. [CrossRef]
38. Conant, R.T.; Ryan, M.G.; Ågren, G.I.; Birge, H.E.; Davidson, E.A.; Eliasson, P.E.; Evans, S.E.; Frey, S.D.; Giardina, C.P.; Hopkins, F.M.; et al. Temperature and Soil Organic Matter Decomposition Rates–Synthesis of Current Knowledge and a Way Forward. *Glob. Chang. Biol.* **2011**, *17*, 3392–3404. [CrossRef]
39. Sayer, E.J.; Heard, M.S.; Grant, H.K.; Marthews, T.R.; Tanner, E.V.J. Soil Carbon Release Enhanced by Increased Tropical Forest Litterfall. *Nat. Clim. Chang.* **2011**, *1*, 304–307. [CrossRef]
40. Belgiu, M.; Dragut, L. Random Forest in Remote Sensing: A Review of Applications and Future Directions. *Isprs J. Photogramm. Remote Sens.* **2016**, *114*, 24–31. [CrossRef]
41. Jia, Y.; Jin, S.; Savi, P.; Gao, Y.; Tang, J.; Chen, Y.; Li, W. GNSS-R Soil Moisture Retrieval Based on a XGboost Machine Learning Aided Method: Performance and Validation. *Remote Sens.* **2019**, *11*, 1655. [CrossRef]

42. Awad, M.; Khanna, R. Support Vector Machines for Classification. In *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*; Awad, M., Khanna, R., Eds.; Apress: Berkeley, CA, USA, 2015; pp. 39–66, ISBN 978-1-4302-5990-9.
43. Pham, B.T.; Qi, C.; Ho, L.S.; Nguyen-Thoi, T.; Al-Ansari, N.; Nguyen, M.D.; Nguyen, H.D.; Ly, H.-B.; Le, H.V.; Prakash, I. A Novel Hybrid Soft Computing Model Using Random Forest and Particle Swarm Optimization for Estimation of Undrained Shear Strength of Soil. *Sustainability* **2020**, *12*, 2218. [CrossRef]
44. Qiu, Y.; Zhou, J.; Khandelwal, M.; Yang, H.; Yang, P.; Li, C. Performance Evaluation of Hybrid WOA-XGBoost, GWO-XGBoost and BO-XGBoost Models to Predict Blast-Induced Ground Vibration. *Eng. Comput.* **2022**, *38*, 4145–4162. [CrossRef]
45. Hamrani, A.; Akbarzadeh, A.; Madramootoo, C.A. Machine Learning for Predicting Greenhouse Gas Emissions from Agricultural Soils. *Sci. Total Environ.* **2020**, *741*, 140338. [CrossRef] [PubMed]
46. Fernández, D.; Adermann, E.; Pizzolato, M.; Pechenkin, R.; Rodríguez, C.G.; Taravat, A. Comparative Analysis of Machine Learning Algorithms for Soil Erosion Modelling Based on Remotely Sensed Data. *Remote Sens.* **2023**, *15*, 482. [CrossRef]
47. Tan, Q.; Geng, J.; Fang, H.; Li, Y.; Guo, Y. Exploring the Impacts of Data Source, Model Types and Spatial Scales on the Soil Organic Carbon Prediction: A Case Study in the Red Soil Hilly Region of Southern China. *Remote Sens.* **2022**, *14*, 5151. [CrossRef]
48. Taghizadeh-Mehrjardi, R.; Schmidt, K.; Amirian-Chakan, A.; Rentschler, T.; Zeraatpisheh, M.; Sarmadian, F.; Valavi, R.; Davatgar, N.; Behrens, T.; Scholten, T. Improving the Spatial Prediction of Soil Organic Carbon Content in Two Contrasting Climatic Regions by Stacking Machine Learning Models and Rescanning Covariate Space. *Remote Sens.* **2020**, *12*, 1095. [CrossRef]
49. Radočaj, D.; Jurišić, M.; Rapčan, I.; Domazetović, F.; Milošević, R.; Plaščak, I. An Independent Validation of SoilGrids Accuracy for Soil Texture Components in Croatia. *Land* **2023**, *12*, 1034. [CrossRef]
50. Somarathna, P.D.S.N.; Malone, B.P.; Minasny, B. Mapping Soil Organic Carbon Content over New South Wales, Australia Using Local Regression Kriging. *Geoderma Reg.* **2016**, *7*, 38–48. [CrossRef]
51. Van Eynde, E.; Fendrich, A.N.; Ballabio, C.; Panagos, P. Spatial Assessment of Topsoil Zinc Concentrations in Europe. *Sci. Total Environ.* **2023**, *892*, 164512. [CrossRef] [PubMed]
52. Sun, Q.; Li, B.; Zhou, C.; Li, F.; Zhang, Z.; Ding, L.; Zhang, T.; Xu, L. A Systematic Review of Research Studies on the Estimation of Net Primary Productivity in the Three-River Headwater Region, China. *J. Geogr. Sci.* **2017**, *27*, 161–182. [CrossRef]
53. Shen, C.; Xiong, J.; Zhang, H.; Feng, Y.; Lin, X.; Li, X.; Liang, W.; Chu, H. Soil pH Drives the Spatial Distribution of Bacterial Communities along Elevation on Changbai Mountain. *Soil Biol. Biochem.* **2013**, *57*, 204–211. [CrossRef]
54. Tian, H.; Chen, G.; Zhang, C.; Melillo, J.M.; Hall, C.A.S. Pattern and Variation of C:N:P Ratios in China's Soils: A Synthesis of Observational Data. *Biogeochemistry* **2010**, *98*, 139–151. [CrossRef]
55. Clark, D.B.; Mercado, L.M.; Sitch, S.; Jones, C.D.; Gedney, N.; Best, M.J.; Pryor, M.; Rooney, G.G.; Essery, R.L.H.; Blyth, E.; et al. The Joint UK Land Environment Simulator (JULES), Model Description—Part 2: Carbon Fluxes and Vegetation Dynamics. *Geosci. Model Dev.* **2011**, *4*, 701–722. [CrossRef]
56. Koven, C.D.; Hugelius, G.; Lawrence, D.M.; Wieder, W.R. Higher Climatological Temperature Sensitivity of Soil Carbon in Cold than Warm Climates. *Nat. Clim. Chang.* **2017**, *7*, 817–822. [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.