forests

MDPI

Article

Estimating Regional Forest Carbon Density Using Remote Sensing and Geographically Weighted Random Forest Models: A Case Study of Mid- to High-Latitude Forests in China

Yuan Zhou [1], Geran Wei [1], Yang Wang [2], Bin Wang [1,3,*], Ying Quan [1], Zechuan Wu [1], Jianyang Liu [1], Shaojie Bian [1], Mingze Li [1], Wenyi Fan [1,3] and Yuxuan Dai [4]

[1] Key Laboratory of Sustainable Forest Ecosystem Management—Ministry of Education, School of Forestry, Northeast Forestry University, Harbin 150040, China; quanying@nefu.edu.cn (Y.Q.)
[2] Heilongjiang Forestry Vocational-Technical College, Mudanjiang 157011, China
[3] Engineering Consulting & Design Institute, Northeast Forestry University, Harbin 150040, China
[4] Shandong Nuclear Power Equipment Manufacturing Co., Ltd., Yantai 265100, China
* Correspondence: wangbin@nefu.edu.cn

Abstract: In the realm of global climate change and environmental protection, the precise estimation of forest ecosystem carbon density is essential for devising effective carbon management and emission reduction strategies. This study employed forest inventory, soil carbon, and remote sensing data combined with three models—Random Forest (RF), Geographically Weighted Regression (GWR), and the innovative Geographically Weighted Random Forest (GWRF) model—integrated with remote sensing technology to develop a framework for assessing the regional spatial distribution of the forest vegetation carbon density (FVC) and forest soil carbon density (FSC). The findings revealed that the GWRF model outperformed the other models in estimating both the FVC and FSC. The data indicated that the FVC in Heilongjiang Province ranged from 4.91 t/ha to 72.39 t/ha, with an average of 40.88 t/ha. In contrast, the average FSC was 182.29 t/ha, with a range of 96.01 t/ha to 255.09 t/ha. Additionally, the forest ecosystem carbon density (FEC) varied from 124.36 t/ha to 302.18 t/ha, averaging 223.17 t/ha. Spatially, the FVC, FSC, and FEC exhibited a consistent growth trend from north to south. The results of this study demonstrate that machine learning models that consider spatial relationships can improve predictive accuracy, providing valuable insights for the future spatial modeling of forest carbon storage.

Keywords: forest vegetation carbon density; forest soil carbon density; remote sensing; spatial distribution; GWRF model

check for updates

Academic Editor: Vladimír Šebeň

Received: 14 December 2024
Revised: 30 December 2024
Accepted: 7 January 2025
Published: 9 January 2025

Citation: Zhou, Y.; Wei, G.; Wang, Y.; Wang, B.; Quan, Y.; Wu, Z.; Liu, J.; Bian, S.; Li, M.; Fan, W.; et al. Estimating Regional Forest Carbon Density Using Remote Sensing and Geographically Weighted Random Forest Models: A Case Study of Mid- to High-Latitude Forests in China. *Forests* **2025**, *16*, 96. https://doi.org/10.3390/f16010096

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Forests sequester approximately 45% of organic carbon and account for two-thirds of the annual carbon uptake in terrestrial ecosystems, playing a pivotal role in carbon dioxide (CO_2) absorption and mitigating global warming [1]. Atmospheric carbon captured by forest ecosystems is ultimately stored in vegetation and soil carbon pools, which are interdependent and collectively constitute the primary reservoir of forest carbon. The accurate estimation of forest ecosystem carbon (FEC) storage or carbon density is essential for understanding carbon dynamics, combating climate change, and guiding conservation strategies [2]. Despite this, few studies have simultaneously estimated the forest vegetation carbon storage density (FVC) and the forest soil carbon storage density (FSC), leading to unclear spatial patterns of FEC at the regional scale.

Traditional forest inventory methods, while yielding precise data for FEC estimation, are resource-intensive and spatially discontinuous [3]. Recently, remote sensing technology has emerged as a key tool for estimating FEC due to its capability to provide real-time, dynamic data and the large-scale monitoring of temporal and spatial changes in forest resources [4]. Nonetheless, the complex relationship between remote sensing variables and ground-measured data presents challenges. Developing models that effectively reduce estimation uncertainties and enhance inversion accuracy remains a critical issue in current research.

In recent years, high-resolution remote sensing data have been integrated with machine learning models, such as Random Forest (RF) [5], and geostatistical methods, such as Geographically Weighted Regression (GWR) [6], to estimate FEC at a regional scale [7]. RF models, however, fail to account for spatial relationships between variables, while GWR models struggle with nonlinear relationships and interactions among dependent and independent variables. Addressing these spatial and interaction complexities represents a significant advancement in improving the accuracy of forest carbon stock estimates, thus offering new possibilities for carbon stock estimation. The emerging Geographically Weighted Random Forest (GWRF) model combines the strengths of RF and GWR, capturing nonlinear data relationships and elucidating local variable effects, which enhances understanding of the fundamental processes driving relationships between variables within a model [8]. To date, GWRFs have been applied in various domains, including crop yield prediction [9], atmospheric PM 2.5 forecasting [10], and the spatial distribution of diabetes prevalence [11].

A technical framework utilizing remote sensing was developed to estimate the regional-scale spatial distribution of FEC. Initially, remote sensing data, combined with forest inventory data, was employed to estimate the spatial distribution of the FVC in Heilongjiang Province for 2015. Subsequently, the estimated FVC served as a variable to estimate the spatial distribution of the FSC within the same region and period. Finally, FEC was derived by summing the FVC and FSC. This approach enabled a comparative analysis of the three modeling techniques—RF, GWR, and GWRF—to identify the most effective model for estimating the FVC and FSC.

2. Study Area

The study area for this research included the Heilongjiang forests in China (121°11′~135°05′ E, 43°26′~53°33′ N), covering four prominent forest ecological zones: Daxing'an Mountain, Xiaoxing'an Mountain, Zhangguangcai Mountain, and Wanda Mountain, with a total area of 473,000 km^2 (Figure 1). The terrain is predominantly mountainous, with elevations ranging from 50 to 1500 m. The average annual temperatures range from −5 to 6 °C, while the total annual precipitation varies between 390 and 660 mm. The growing season extends from May to September. The primary forest ecosystems consist of temperate coniferous and deciduous broad-leaved forests, with notable tree species including *Larix gmelinii*, *Pinus koraiensis*, *Quercus mongolica*, and *Betula platyphylla*.

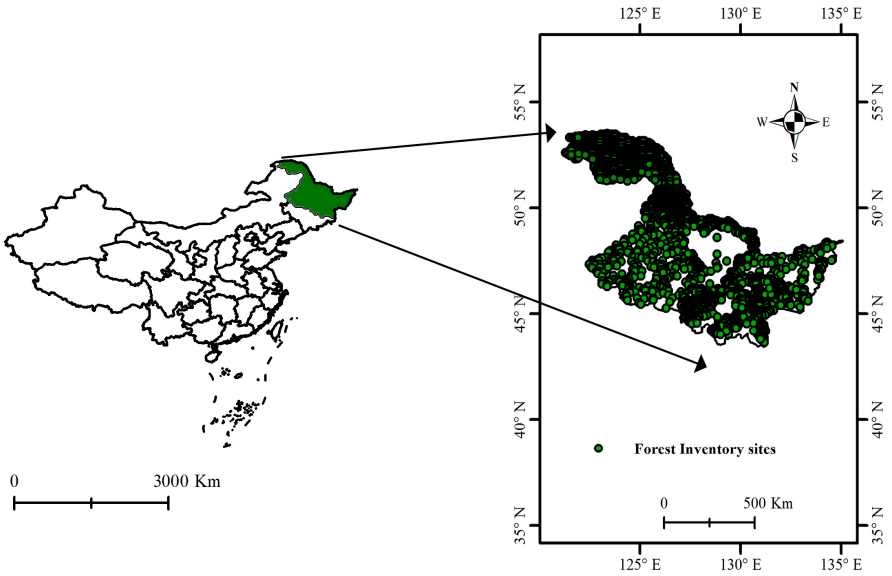

Figure 1. The green area represents the study area, and the green dots indicate the forest inventory sites.

3. Materials and Methods

3.1. Datasets

3.1.1. Forest Inventory Data

The dataset comprised 3074 sites from the 9th National Forest Resources Inventory (2015) (Figure 1). For each site, the data included details on the individual trees (e.g., DBH, height, and species), site location, site area (0.06 ha), forest type, age group, forest health, and management practices. Using individual tree data and a compatible biomass model for the key tree species in the Heilongjiang Province [12], biomass was estimated for each tree and categorized into four components: leaf, branch, stem, and root. Site biomass, the sum of all trees' biomass, was converted to carbon density (t/ha) using a conversion coefficient of 0.47 [13].

3.1.2. Soil Carbon Density Data

Additionally, samples from 569 soil carbon density sites were collected from 2012 to 2015 in Heilongjiang Province (Figure 2). Each site, measuring 30 m × 30 m, included three soil profiles with a depth of 1 m (Figure 3). Site locations were selected to represent a range of forest types, topographical features, and elevations (100 m to 1000 m), encompassing 12 primary forest types and various slope positions and directions. The soil carbon density (t/ha) for each site represents the average soil organic carbon content at 1 m depth across the three profiles.

3.1.3. Remote Sensing Data

Landsat 8 OLI data, with a 30 m resolution (visible and near-infrared bands) from 2015, were utilized to extract vegetation indices, spectral information, change components, and texture details, sourced from the Google Earth Engine (GEE) platform (https://earthengine.google.com/, accessed on 1 November 2024) (Google, Mountain View, CA, USA). Only images with less than 5% cloud coverage during the 2015 growing season were selected. The dataset underwent geometric, radiometric, and atmospheric calibration.

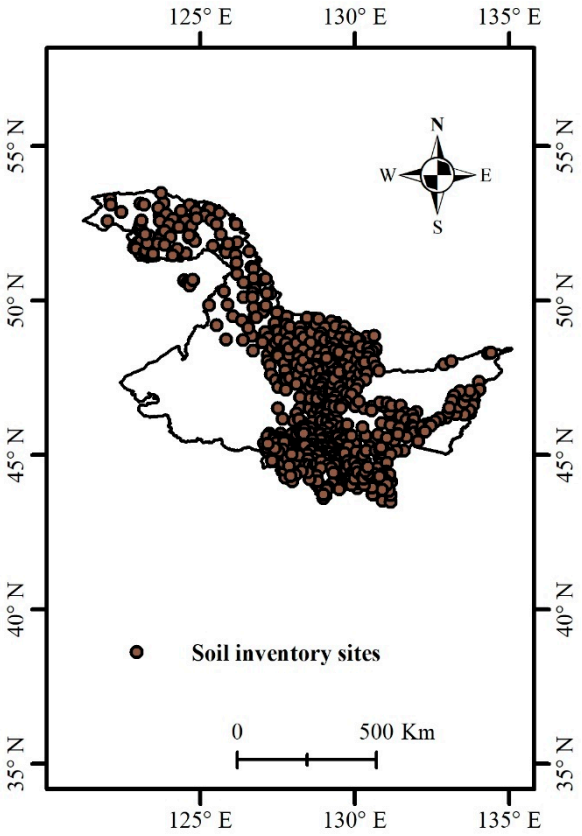

Figure 2. The brown dots indicate the soil inventory sites.

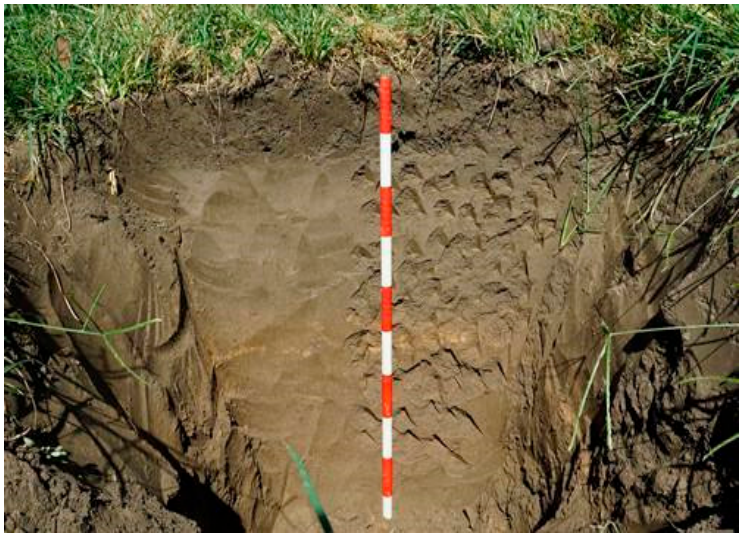

Figure 3. Soil profile field photograph.

3.1.4. Digital Elevation Model Data

Digital Elevation Model (DEM) data, with a 30 m resolution, were derived from the ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer) sensor onboard the Terra satellite. These were obtained from the Geospatial Data Cloud website (https://www.gscloud.cn/search, accessed on 1 November 2024) (Computer Network Information Center, Chinese Academy of Sciences, Beijing, China), which allowed for the derivation of altitude, slope, and aspect information.

3.1.5. Land Cover Classification Data

Land cover classification data were sourced from NASA's MCD12Q1 product, part of the International Geosphere-Biosphere Programme (IGBP) land cover classification scheme (https://ladsweb.modaps.eosdis.nasa.gov/search/, accessed on 1 November 2024). The MCD12Q1 Version 6 data product, based on supervised classifications of the MODIS Terra and Aqua reflectance data, provides annual global land cover classifications from 2001 to 2020 at a 500 m resolution. It distinguishes 17 land cover categories, including 11 natural vegetation types, 3 developed or built-up areas, and 3 non-vegetated categories. This study focused exclusively on the forested regions.

3.2. Methods or Methodology

The technical process of this study involved several key steps (Figure 4): First, features were extracted from the Landsat 8 OLI data, including vegetation indices and texture information. Following this, recursive feature elimination with RF was used to select the most important features for model prediction. The models for estimating forest vegetation and soil carbon storage were then constructed using RF, GWR, and GWRF. Finally, the models were evaluated through cross-validation and error analysis, leading to comprehensive results analysis.

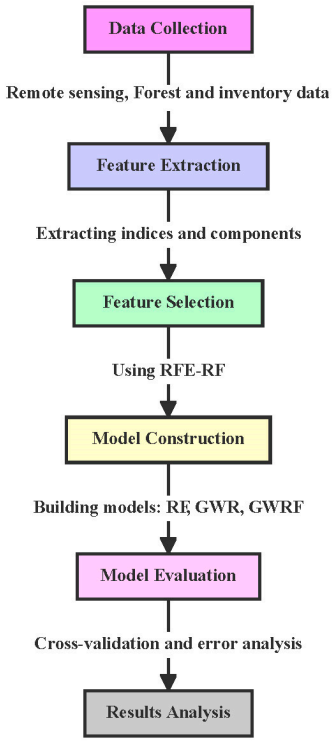

Figure 4. Technical process flowchart.

3.2.1. Feature Extraction and Feature Filtering

To develop the FVC estimation model, this study extracted features from the Landsat 8 OLI data, including original bands, their combinations, vegetation indices, change components, and texture information, as well as the slope, aspect, and elevation from the DEM. A total of 143 features were selected for inclusion in the model (Table S1).

Recursive Feature Elimination with Random Forest (RFE-RF) was utilized to refine the feature set. This technique determines the optimal number of variables by incrementally introducing them and assessing their impact on the model accuracy. RFE-RF provides variable importance scores while accounting for the multicollinearity among features [14]. The RFE-RF process involved the following steps: (1) Determining Variable Importance:

The RF method was applied to assess the importance of the original feature set. RF evaluated the variable importance by measuring the increase in prediction error when the values of a variable was randomly shuffled. A significant increase in error indicated the variable's importance; (2) Grouping Variables: Pearson correlation coefficients were calculated to identify correlations between variables. Variables with coefficients of 0.9 or higher were grouped, retaining only the most important variable from each cluster; (3) Re-evaluating Variable Importance: The RF method was used again to reassess the importance of the remaining variables; (4) Eliminating Least Important Variables: The least important variable was removed, and a 50-fold cross-validation process was repeated 10 times to assess the model prediction accuracy, with RMSE used as the evaluation criterion; (5) Iterative Removal: The process of removing the least important variables continued until only one variable remained.

Finally, RFE-RF produced model prediction accuracy metrics for the various numbers of variables, selecting the most crucial variable based on the lowest RMSE, thus achieving dimensionality reduction and identifying feature factors with significant impacts on the response variables. Additionally, "%IncMSE" (percent increase in mean squared error) was a dimensionless index employed to evaluate the variable importance in RF. It quantifies the increase in the model prediction error resulting from the random permutation of each variable: a higher value indicates greater importance of the variable. %IncMSE was used in this study to assess the variable importance. The "caret" and "randomForest" packages in R (4.1.3) were utilized for variable screening [15,16].

3.2.2. Construction of the FVC and FSC Estimation Model

Three representative models were selected to develop the forest vegetation and soil carbon storage estimation models: RF, GWR, and GWRF. The RF algorithm, a prominent non-parametric machine learning method, employed an ensemble of decision trees for classification and prediction. This integrated learning approach addressed the overfitting problem associated with single decision trees and automatically selected and scaled features, demonstrating good adaptability to complex data and robust tolerance to noise and anomalies [5]. In RF, each decision tree was built using different samples and features, ensuring variability and randomness in the results. During model training, parameter adjustment was essential for enhancing the model accuracy. The "randomForest" package in R was utilized to build the RF model [16].

The GWR model extended the ordinary least squares approach by incorporating spatial location into regression parameters [6]. GWR sequentially estimated the point parameters using locally weighted ordinary least squares, with the weights derived from the distance between the regression point and other observation points. This method allowed for the detection of spatial relationships among variables by analyzing how parameter estimates changed with spatial location. Although the GWR model's formula resembled that of global regression models, its parameters varied spatially [6,17] (Equations (1)–(3)). The "GWmodel" package in R was employed to construct the GWR model [18].

$$\gamma_i = \beta_{i0} + \sum_{k=1}^{n} \beta_{ik}(u_i, v_i)\chi_{ik} + \varepsilon_i \quad i = \{1, \ldots, n\} \tag{1}$$

In the model, γ_i represents the dependent variable at the ith point, (u_i, v_i) denotes the spatial coordinates, β_{i0} is the intercept, $\beta_{ik}(u_i, v_i)$ represents the coefficient of the kth independent variable, χ_{ik} is the value of the kth independent variable, and ε_i denotes the error term. The coefficients were determined using the following matrix formulation:

$$\hat{\beta}(u_i, v_i) = \left[X^T W(u_i, v_i) X\right]^{-1} X^T W(u_i, v_i)\gamma \tag{2}$$

In this formulation, $\hat{\beta}(u_i, v_i)$ represents the coefficient matrix, X and γ are the matrices for the independent and dependent variables, respectively, T denotes the matrix transpose operation, and $W(u_i, v_i)$ is a diagonal weighting matrix. $W(u_i, v_i)$ determines the influence of neighboring points on the regression point, with a Gaussian function used to compute its values.

$$w_{ij} = \exp\left[-\left(\frac{|d_{ij}|}{bw}\right)^2\right] \quad (3)$$

In this context, w_{ij} denotes the weight of the jth observation relative to the ith observation, d_{ij} represents the distance between the jth and ith points, and bw is the bandwidth, which defines the distance range for each local regression equation or the number of neighboring elements considered.

The GWRF model was initially introduced by Santos et al. [19] and later advanced by Georganos et al. [8], who integrated GWR with RF to fully realize the GWRF framework. This model builds upon the GWR concept while extending the global RF model, accommodating both spatial heterogeneity and nonlinear effects as well as interactions among variables. As a local model, GWRF considers nonlinear influences, variable interactions, and spatial autocorrelation, thereby mitigating the impact of spatial heterogeneity on the results. The GWRF model processes multivariate vector information to train RF by computing the local RF for each position, i.e., by applying distance-based weights to establish varying probabilities within the RF ensemble and modeling the spatial relationships among nearby observations. It incorporates the spatial position information of different features and synthesizes predictions from all decision trees through voting or averaging (Figure 5).

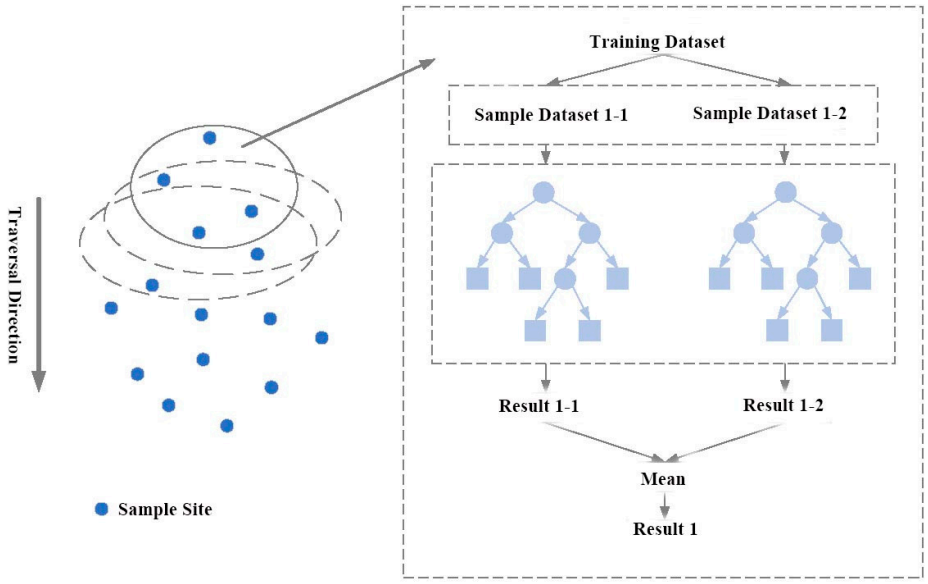

Figure 5. Illustration of the GWRF model principles.

GWRF constructed a local RF model for each data point, focusing exclusively on nearby values. For each local RF, a spatial weight matrix was created to ensure that data points with higher weights were more likely to be selected during the decision tree's construction, accommodating the uneven distribution of sample points [20]. In the GWRF model, the area where each local RF operated was referred to as the neighborhood (or kernel), with the maximum distance between a data point and its neighborhood defined as the bandwidth.

In GWRF, parameters requiring adjustment included "mtry", "A", and bandwidth. For the local RF model, "mtry" denoted the number of candidate features randomly selected

at each split. The optimal "mtry" value was determined by assessing the RMSE on the out-of-bag data for various "mtry" values through ten-fold cross-validation repeated five times, with the value yielding the lowest RMSE selected as the optimal value. The fusion of the local (GWRF) and global (RF) models was controlled by the weight coefficient "A" [21]. A higher value of "A" increases the influence of the local model. The optimal "A" is determined by comparing the R^2 and RMSE values across a range of 0.1 to 1, with a step of 0.1. Additionally, selecting the appropriate bandwidth involved a trial-and-error process, where the optimal bandwidth was the one providing the highest R^2 value on the out-of-bag data, with the test range adjusted according to the sample distribution [20]. The "SpatialIML" package in R was utilized to implement GWRF [8].

3.2.3. Model Evaluation and Test

To assess the prediction accuracy of the various models, the following metrics were utilized: the coefficient of determination (R^2) (Equation (4)), root mean squared error (RMSE) (Equation (5)), relative root mean squared error (rRMSE%) (Equation (6)), and mean absolute error (MAE) (Equation (7)).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2} \tag{4}$$

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{n}} \tag{5}$$

$$\mathrm{rRMSE\%} = \frac{\mathrm{RMSE}}{\overline{y}} \times 100\% \tag{6}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\left(y_i - \hat{y}_i\right)\right| \tag{7}$$

Here, y_i represents the actual value of carbon storage, $\hat{y}_i$ denotes the estimated value, $\overline{y}_i$ is the mean value, and n indicates the number of samples.

4. Results

4.1. Feature Variable Screening Result for FVC Estimation

To balance accuracy and efficiency in the FVC estimation model, RMSE was utilized to identify the optimal number of variables, thus mitigating the risk of overfitting. As depicted in Figure 6a, the RMSE started to stabilize after the number of variables exceeded 11. Additionally, the %IncMSE values were used to assess and rank the importance of each variable in relation to forest vegetation carbon storage. Consequently, the top 11 variables with the highest %IncMSE values were selected for constructing the FVC estimation model, as illustrated in Figure 6b.

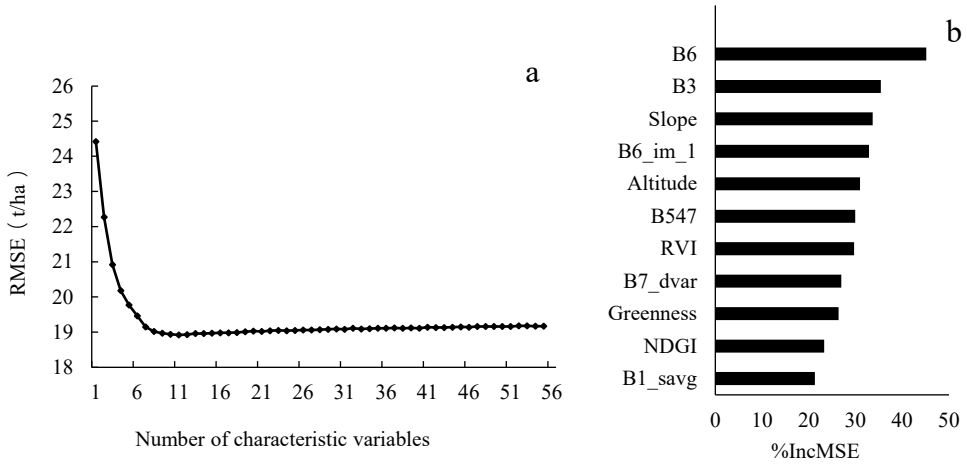

Figure 6. Screening results of FVC estimation model variables. (**a**) Variable selection outcomes; (**b**) Ranking of variable importance.

4.2. Evaluation of the FVC Estimate Results of Different Models

The spatial distribution of the FVC in the Heilongjiang Province for 2015 was estimated using three models: GWR, RF, and GWRF. For the RF model, key parameters were optimized through a systematic iterative approach to ensure maximum accuracy. The adjustment thresholds, step sizes, and optimal parameter values are detailed in Table S2.

In the GWR model, statistical values for the regression coefficients were provided, including the minimum, first quartile, mean, third quartile, and maximum values, as shown in Table S3.

For the GWRF model, the impact of varying the "mtry" and "A" values on model accuracy was assessed. The optimal "mtry" (4) and "A" (0.6) were selected based on minimizing R^2 and RMSE, as illustrated in Figures S1 and S2. Additionally, the "adaptive" kernel function was used to determine the optimal bandwidth, which was found to be 700 after several tests, as depicted in Figure S3.

Comparative analysis of the models' accuracy revealed that the GWRF model achieved the highest R^2 (0.41) and the lowest RMSE (18.42 t/ha), rRMSE% (47.99), and MAE (14.93 t/ha). Conversely, the GWR model exhibited the lowest R^2 (0.33) and the highest RMSE (19.52 t/ha), rRMSE% (50.87), and MAE (16.19 t/ha). Overall, the model accuracy ranked as follows: GWRF > RF > GWR. Detailed accuracy evaluation results are presented in Table 1 and Figure 7.

Table 1. Verification results for the three models for the FVC.

Models	Verification Accuracy			
	R^2	RMSE (t/ha)	rRMSE%	MAE (t/ha)
GWR	0.33	19.52	50.87	16.19
RF	0.38	18.73	48.80	15.33
GWRF	0.41	18.42	47.99	14.93

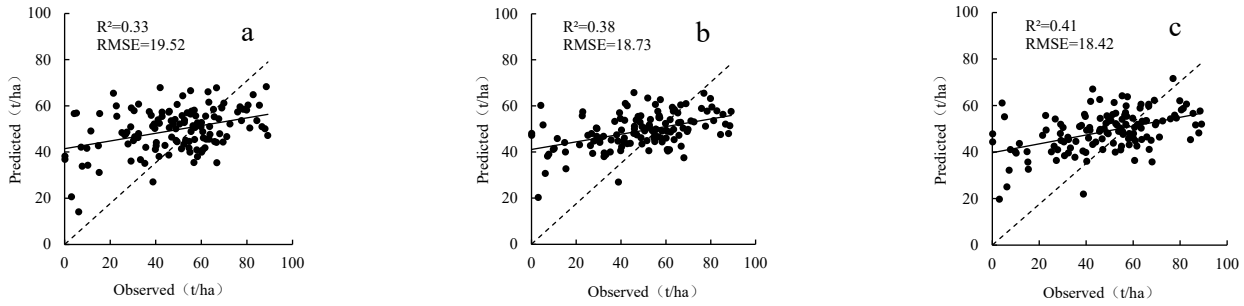

Figure 7. Verification results for the three models for the FVC. (**a**) GWR model results; (**b**) RF model results; (**c**) GWRF model results.

4.3. Feature Variable Screening Results for FSC Estimation

For estimating the FSC, the FVC distribution, derived using the GWRF model, was incorporated as one of the variables. Additional variables, as detailed in Table S1, were also considered. Feature selection followed the RFE-RF approach. In this case, 13 key variables were identified (Figure 8a), and their importance was ranked, as shown in Figure 8b.

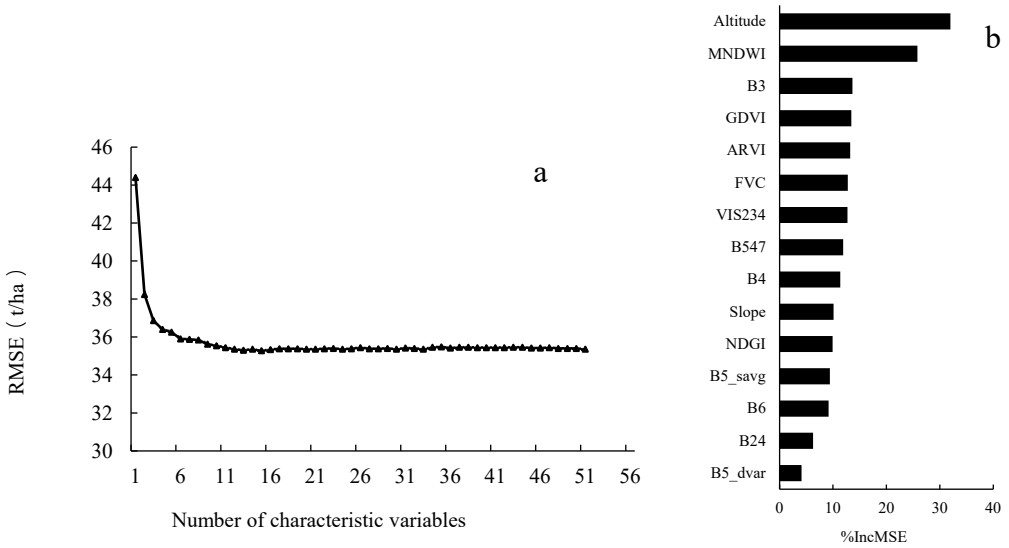

Figure 8. Screening results for the FSC estimation model characteristic variables. (**a**) Screening results of the characteristic variables; (**b**) Variable importance ranking.

4.4. Evaluation of the FSC Estimate Results of Different Models

The RF, GWR, and GWRF models were employed to estimate the FSC. For the RF model, the values of the five key parameters are detailed in Table S4. The regression coefficients for the GWR model are presented in Table S5. For the GWRF model, the parameters "mtry" and "A", as well as the bandwidth, are illustrated in Figures S4–S6, with "mtry" set to 12, "A" to 0.8, and bandwidth to 290.

Consistent with the FVC estimation results, the GWRF model demonstrated superior accuracy in estimating the FSC, while the GWR model showed the lowest accuracy (Table 2 and Figure 9). An additional comparison was conducted using the GWRF model without the FVC variable. Excluding the FVC led to a decrease of 0.05 in R^2, while RMSE, rRMSE%, and MAE increased by 2.82 t/ha, 0.79%, and 2.77 t/ha, respectively. These results are detailed in Figure 10.

Table 2. Verification results from the three models for the FSC.

Models	Verification Accuracy			
	R^2	RMSE (t/ha)	rRMSE%	MAE (t/ha)
GWR	0.42	62.88	17.63	47.82
RF	0.44	61.52	17.25	45.14
GWRF	0.53	56.68	15.90	40.36

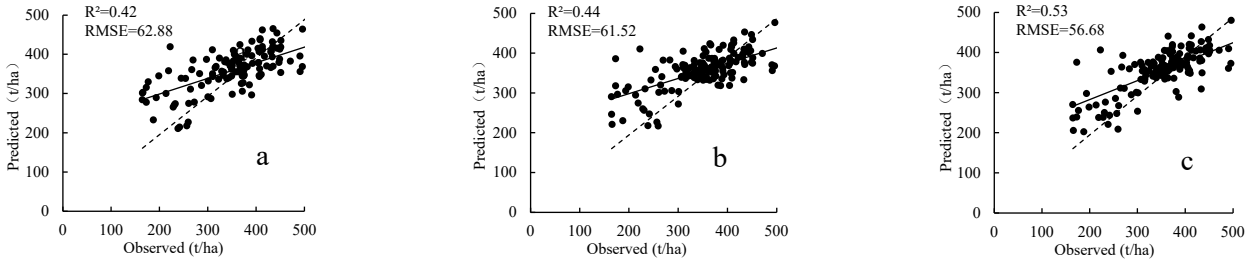

Figure 9. Verification results from the three models for the FSC. (**a**) GWR; (**b**) RF; (**c**) GWRF.

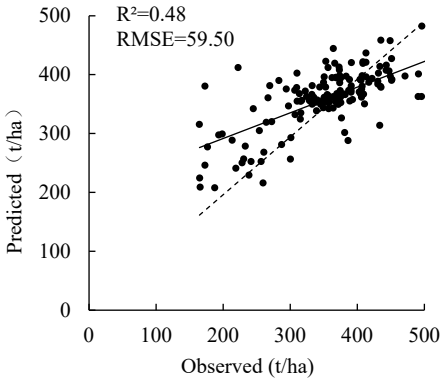

Figure 10. Verification results from the GWRF model after removing the variable FVC.

Additionally, an examination of the spatial autocorrelation of model residuals revealed that GWRF residuals exhibited a smaller Global Moran's I compared to those of GWR and RF, across distances ranging from 5 to 180 km for both the FVC and FSC estimations (Figure 11). This finding suggests that GWRF more effectively incorporates spatial autocorrelation information.

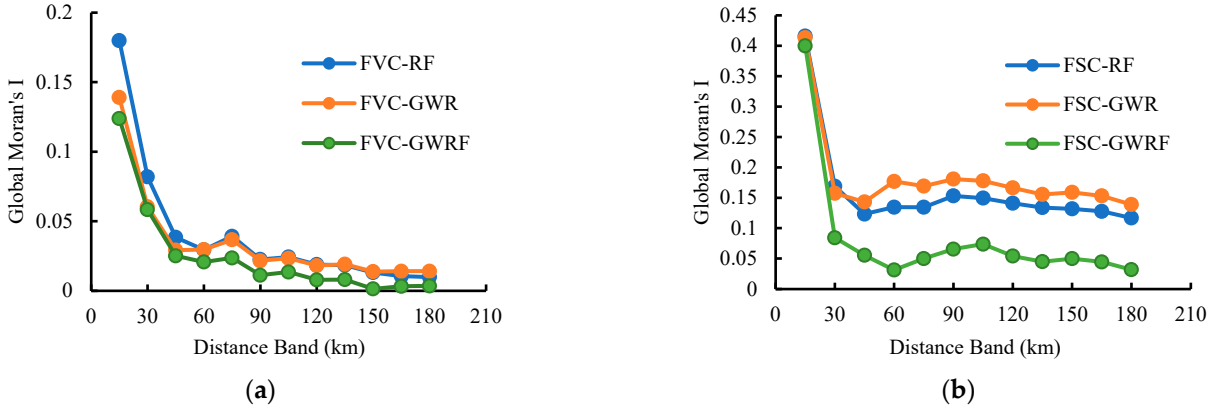

Figure 11. Changes in residual spatial autocorrelation coefficients for GWR, RF, and GWRF with increasing distance band. (**a**) FVC estimation; (**b**) FSC estimation.

4.5. Spatial Distribution of the FVC, FSC, and FEC

The GWRF model was employed to quantitatively estimate the spatial distribution of the FVC and FSC in the study area for 2015. The results indicated that the FVC in the Heilongjiang Province ranged from 4.91 t/ha to 72.39 t/ha, with an average of 40.88 t/ha, displaying a north-to-south increasing trend (Figure 12). Conversely, the average FSC was approximately 4.5 times greater than the FVC, reaching 182.29 t/ha, with values spanning from 96.01 t/ha to 255.09 t/ha. This variable also exhibited a north-to-south increasing trend, but with a more pronounced gradient (Figure 13). Summing the FVC and FSC on a pixel-by-pixel basis provided the FEC for the study area. The spatial variation of FEC mirrored that of both the FVC and FSC, ranging from a minimum of 124.36 t/ha to a maximum of 302.18 t/ha, with an average value of 223.17 t/ha (Figure 14).

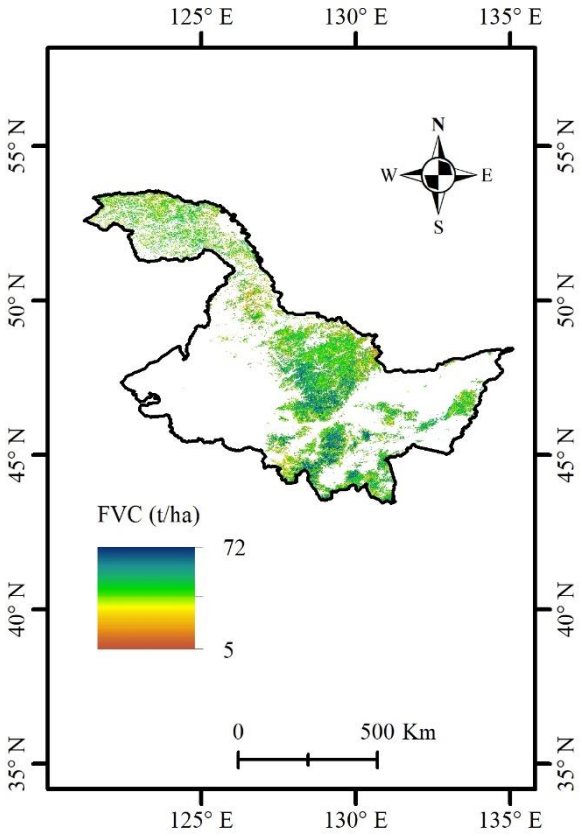

Figure 12. Spatial distribution of the FVC in the Heilongjiang Province for 2015.

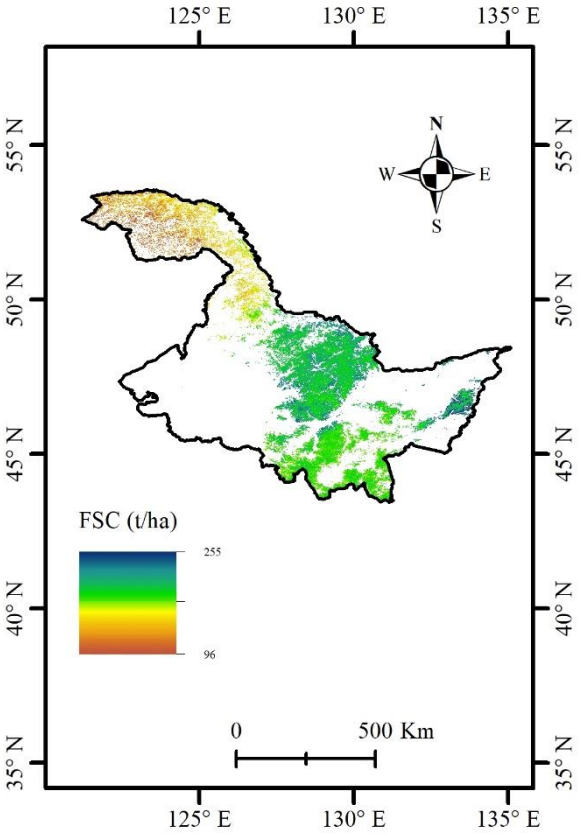

Figure 13. Spatial distribution of the FSC in the Heilongjiang Province for 2015.

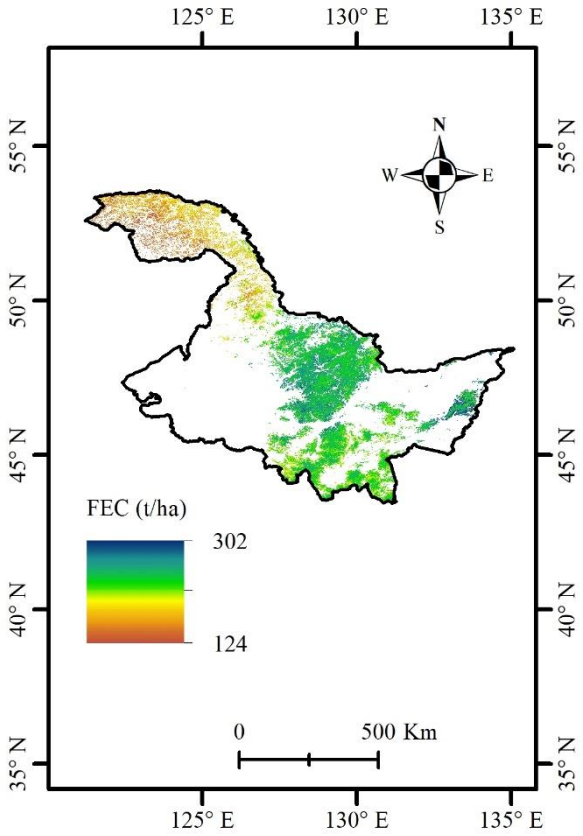

Figure 14. The spatial distribution of FEC in the Heilongjiang Province for 2015.

5. Discussion

In this study, remote sensing models for the FVC and FSC were constructed in three steps to estimate the spatial distribution of the FVC, FSC, and FEC in the Heilongjiang Province for 2015. The results demonstrate that the GWRF model outperformed both the RF and GWR models in accuracy for estimating the FVC and FSC. This superiority arose from the GWRF model's integration of the GWR and RF advantages, enabling it to capture spatial variation and exhibit strong nonlinear fitting capabilities. Additionally, the model's consideration of spatial positional influence, by assigning weights based on sample locations, further enhanced its accuracy and robustness [8].

The significance of the FVC predictions for FSC estimations was notably observed (Figures 7c and 8), aligning with previous studies [22–24] that highlight the close relationship between soil and vegetation carbon storage. Both are interdependent in the carbon cycle [25]: vegetation sequesters CO_2 through photosynthesis and contributes to soil carbon via litterfall and root exudates, which in turn supports soil carbon stocks. Soil organic carbon provides essential nutrients and water, promoting plant growth and further influencing carbon sequestration. Thus, the interaction between vegetation and soil carbon storage forms a closely interconnected carbon cycle system within the ecosystem [26,27].

The results of our study indicate that in 2015, the FVC in the Heilongjiang Province averaged 40.88 t/ha, aligning closely with Chang et al.'s estimate of 40.0 t/ha for the same period [28]. Additionally, our estimated soil carbon density of 182.29 t/ha is consistent with previous research by Jiao and Hu [29], which reported soil carbon densities ranging from 150 to 200 t/ha for various forest types in the region. Furthermore, our estimate of a total forest carbon stock of 4681.44 Tg for 2015 corroborates the findings from Li's research [30], who observed comparable carbon stock estimates using extensive forest inventory data. These congruences validate the accuracy and reliability of our estimates.

The spatial distribution of the FVC, FSC, and FEC in the Heilongjiang Province in 2015 reveals a gradient from lower values in the north to higher values in the south. This trend can be attributed to three primary factors: (1) Climatic Conditions: The southern part of the Heilongjiang Province experiences relatively warm and humid conditions, which are conducive to forest growth. These conditions foster faster tree growth, higher forest coverage, and greater carbon storage. In contrast, the colder and drier northern region hampers forest development, resulting in lower carbon storage [30]; (2) Vegetation Types: The southern region predominantly features broad-leaved and mixed forests, which generally have higher carbon storage potential. In contrast, the northern region is characterized by coniferous forests, which tend to have lower carbon storage [31]; (3) Soil Conditions: Fertile soils in the southern region support robust plant growth, leading to increased forest carbon storage. Conversely, the poorer soil conditions in the north limit plant growth and carbon storage accumulation [32,33].

Nevertheless, several limitations and uncertainties affect our study. Firstly, as a local model, GWRF does not account for differences between the local models. Each local RF constructed by GWRF employs identical parameters and features, neglecting the potential benefits of optimizing the parameters and selecting the most relevant variables based on the local characteristics. To address this, methods such as RF variable selection [34], grid search [35], and recursive feature elimination [36] could be explored to enable local parameter tuning and feature selection. Secondly, this study relied solely on Landsat data to extract model features. Optical data may experience saturation effects on vegetation biomass or carbon storage, potentially leading to the overestimation of low vegetation carbon storage and the underestimation of high values. Incorporating additional data sources, such as LiDAR, could enhance the accuracy of the FVC and FSC estimations. Additionally, using the remote sensing-derived FVC as a variable for estimating the FSC

may introduce multiple layers of error transmission, which could increase uncertainty in the FSC predictions. Although this study presents a novel technical framework for estimating forest ecosystem carbon storage using remote sensing and the GWRF model, there is significant potential for the further enhancement of simulation accuracy by integrating additional remote sensing data sources such as LiDAR and SAR. By combining these diverse datasets, we can obtain information related to the forest's three-dimensional structure that is associated with forest carbon stocks. Future research should focus on developing integrated frameworks that incorporate these advanced remote sensing technologies to achieve more precise and reliable ecological assessments.

6. Conclusions

This study introduces a novel approach to estimating regional-scale forest carbon storage by integrating the strengths of the RF and GWR models through the GWRF model combined with remote sensing technology. The results demonstrate that the GWRF model achieved superior estimation accuracy for both the FVC and FSC, outperforming traditional RF and GWR models. Notably, the GWRF model, when incorporating the FVC variable, delivered more precise FSC estimates, highlighting the significant role of the FVC in FSC estimation. The application of the GWRF model enabled the successful quantification of the spatial distribution of the FVC and FSC in the Heilongjiang Province for 2015. The findings revealed a spatial increase in both the FVC and FSC from north to south, with the FSC showing a more pronounced rise. These results not only confirm the efficacy of the GWRF model in estimating forest carbon storage but also offer a scientific foundation for the improved management of regional-scale forest carbon resources.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/f16010096/s1, Table S1. Feature variables and calculation formula. Table S2. RF model parameter selection results for FVC. Table S3. The regression coefficient statistics of GWR model for FVC. Table S4. RF model parameter selection results for FSC. Table S5. The regression coefficient statistics of GWR model for FVC. Figure S1. The result of "mtry" parameter selection in GWRF model for FVC. Figure S2. The result of "A" parameter selection in GWRF model for FVC. Figure S3. The result of bandwidth selection in GWRF model for FVC. Figure S4. The result of "mtry" parameter selection in GWRF model for FSC. Figure S5. The result of "A" parameter selection in GWRF model for FSC. Figure S6. The result of bandwidth selection in GWRF model for FSC.

Author Contributions: Y.Z.: Methodology, Writing—Original Draft; G.W.: Software, Formal analysis; B.W.: Conceptualization, Writing—Review & Editing; Y.Q.: Formal analysis; Z.W.: Data Curation; J.L.: Software; S.B.: Visualization; M.L.: Writing—Review & Editing; W.F.: Conceptualization; Y.D.: Data curation; Y.W.: Data curation. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China, 2023YFD2201704; the Fundamental Research Funds for the Central Universities, 2572022DT03; Carbon neutrality special scientific Foundation project, HFW220100054.

Data Availability Statement: Dataset available on request from the authors.

Acknowledgments: We appreciate Xiaoyang Cui for providing the soil inventory data.

Conflicts of Interest: Yuxuan Dai is employed by Shandong Nuclear Power Equipment Manufacturing Co., Ltd., his employer's company was not involved in this study, and there is no relevance between this research and their company. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Friedlingstein, P.; O'Sullivan, M.; Jones, M.W.; Andrew, R.M.; Bakker, D.C.; Hauck, J.; Landschützer, P.; Le Quéré, C.; Luijkx, I.T.; Peters, G.P.; et al. Global Carbon Budget 2023. *Earth Syst. Sci. Data* **2023**, *15*, 5301–5369. [CrossRef]
2. Fang, J.; Yu, G.; Liu, L.; Hu, S.; Chapin, F.S., III. Climate change, human impacts, and carbon sequestration in China. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 4015–4020. [CrossRef] [PubMed]
3. Avitabile, V.; Herold, M.; Heuvelink, G.B.; Lewis, S.L.; Phillips, O.L.; Asner, G.P.; Armston, J.; Ashton, P.S.; Banin, L.; Bayol, N.; et al. An integrated pan-tropical biomass map using multiple reference datasets. *Glob. Change Biol.* **2016**, *22*, 1406–1420. [CrossRef]
4. Esteban, J.; McRoberts, R.E.; Fernández-Landa, A.; Tomé, J.L.; Næsset, E. Estimating forest volume and biomass and their changes using random forests and remotely sensed data. *Remote Sens.* **2019**, *11*, 1944. [CrossRef]
5. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
6. Fotheringham, A.S.; Charlton, M.E.; Brunsdon, C. Spatial variations in school performance: A local analysis using geographically weighted regression. *Geogr. Environ. Model.* **2001**, *5*, 43–66. [CrossRef]
7. Sun, W.; Liu, X. Review on carbon storage estimation of forest ecosystem and applications in China. *For. Ecosyst.* **2020**, *7*, 4. [CrossRef]
8. Georganos, S.; Grippa, T.; Niang Gadiaga, A.; Linard, C.; Lennert, M.; Vanhuysse, S.; Mboga, N.; Wolff, E.; Kalogirou, S. Geographical random forests: A spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto Int.* **2021**, *36*, 121–136. [CrossRef]
9. Khan, S.N.; Li, D.; Maimaitijiang, M. A geographically weighted random forest approach to predict corn yield in the US corn belt. *Remote Sens.* **2022**, *14*, 2843. [CrossRef]
10. Su, Z.; Lin, L.; Xu, Z.; Chen, Y.; Yang, L.; Hu, H.; Lin, Z.; Wei, S.; Luo, S. Modeling the effects of drivers on $PM_{2.5}$ in the Yangtze River Delta with geographically weighted Random Forest. *Remote Sens.* **2023**, *15*, 3826. [CrossRef]
11. Quiñones, S.; Goyal, A.; Ahmed, Z.U. Geographically weighted machine learning model for untangling spatial heterogeneity of type 2 diabetes mellitus (T2D) prevalence in the USA. *Sci. Rep.* **2021**, *11*, 6955.
12. Dong, L.; Zhang, L.; Li, F. A compatible system of biomass equations for three conifer species in Northeast, China. *For. Ecol. Manag.* **2014**, *329*, 306–317. [CrossRef]
13. Wang, B.; Li, M.; Fan, W.; Yu, Y.; Chen, J.M. Relationship between net primary productivity and forest stand age under different site conditions and its implications for regional carbon cycle study. *Forests* **2018**, *9*, 5. [CrossRef]
14. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389–422. [CrossRef]
15. Kuhn, M.J. Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, *28*, 1–26. [CrossRef]
16. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22.
17. Hu, T.; Sun, Y.; Jia, W.; Li, D.; Zou, M.; Zhang, M. Study on the estimation of forest volume based on multi-source data. *Sensors* **2021**, *21*, 7796. [CrossRef] [PubMed]
18. Lu, B.; Harris, P.; Charlton, M.; Brunsdon, C. The GWmodel R package: Further topics for exploring spatial heterogeneity using geographically weighted models. *Geo-Spat. Inf. Sci.* **2014**, *17*, 85–101. [CrossRef]
19. Santos, F.; Graw, V.; Bonilla, S. A geographically weighted random forest approach for evaluate forest change drivers in the Northern Ecuadorian Amazon. *PLoS ONE* **2019**, *14*, e0226224. [CrossRef]
20. Georganos, S.; Kalogirou, S. A forest of forests: A spatially weighted and computationally efficient formulation of geographical random forests. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 471. [CrossRef]
21. Aguirre-Gutiérrez, J.; Rifai, S.; Shenkin, A.; Oliveras, I.; Bentley, L.P.; Svátek, M.; Girardin, C.A.; Both, S.; Riutta, T.; Berenguer, E.; et al. Pantropical modelling of canopy functional traits using Sentinel-2 remote sensing data. *Remote Sens. Environ.* **2021**, *252*, 112122. [CrossRef]
22. Lal, R.J. Soil carbon sequestration impacts on global climate change and food security. *Science* **2004**, *304*, 1623–1627. [CrossRef]
23. Wang, Y.; Fu, B.; Lü, Y.; Chen, L. Effects of vegetation restoration on soil organic carbon sequestration at multiple scales in semi-arid Loess Plateau, China. *Catena* **2011**, *85*, 58–66. [CrossRef]
24. Wang, B.; Xu, G.; Ma, T.; Chen, L.; Cheng, Y.; Li, P.; Li, Z.; Zhang, Y. Effects of vegetation restoration on soil aggregates, organic carbon, and nitrogen in the Loess Plateau of China. *Catena* **2023**, *231*, 107340. [CrossRef]
25. Farrelly, D.J.; Everard, C.D.; Fagan, C.C.; McDonnell, K.P. Carbon sequestration and the role of biological carbon mitigation: A review. *Renew. Sustain. Energy* **2013**, *21*, 712–727. [CrossRef]
26. Cai, W.; He, N.; Li, M.; Xu, L.; Wang, L.; Zhu, J.; Zeng, N.; Yan, P.; Si, G.; Zhang, X.; et al. Carbon sequestration of Chinese forests from 2010 to 2060: Spatiotemporal dynamics and its regulatory strategies. *Sci. Bull.* **2022**, *67*, 836–843. [CrossRef] [PubMed]
27. Mo, L.; Zohner, C.M.; Reich, P.B.; Liang, J.; De Miguel, S.; Nabuurs, G.-J.; Renner, S.S.; van den Hoogen, J.; Araza, A.; Herold, M.; et al. Integrated global assessment of the natural forest carbon potential. *Nature* **2023**, *624*, 92–101. [CrossRef] [PubMed]

28. Chang, Z.; Fan, L.; Wigneron, J.-P.; Wang, Y.-P.; Ciais, P.; Chave, J.; Fensholt, R.; Chen, J.M.; Yuan, W.; Ju, W.; et al. Estimating Aboveground Carbon Dynamic of China Using Optical and Microwave Remote-Sensing Datasets from 2013 to 2019. *J. Remote Sens.* **2023**, *3*, 0005. [CrossRef]
29. Jiao, Y.; Hu, H. Carbon storage and its dynamics of forest vegetations in Heilongjiang Province. *J. Appl. Ecol.* **2005**, *16*, 2248–2252.
30. Li, X.; Huang, C.; Jin, H.; Han, Y.; Kang, S.; Liu, J.; Cai, H.; Hu, T.; Yang, G.; Yu, H.; et al. Spatio-temporal patterns of carbon storage derived using the InVEST model in Heilongjiang Province, Northeast China. *Front. Earth Sci.* **2022**, *10*, 846456. [CrossRef]
31. Xi, Z.; Chen, G.; Xing, Y.; Xu, H.; Tian, Z.; Ma, Y.; Cui, J.; Li, D. Spatial and temporal variation of vegetation NPP and analysis of influencing factors in Heilongjiang Province, China. *Ecol. Indic.* **2023**, *154*, 110798. [CrossRef]
32. Chen, K.; Wang, J.; He, Y.; Zhang, L. Estimations of forest carbon storage and carbon sequestration potential of key state-owned forest region in Daxing'anling, Heilongjiang province. *Ecol. Environ.* **2022**, *31*, 1725.
33. Wang, X.-C.; Wang, S.-D.; Yu, D.-P.; Zhou, L.; Dai, L.-M. Carbon storage and density of forest ecosystems in Heilongjiang Province, China. *Taiwan J. For. Sci.* **2012**, *27*, 229–238.
34. Genuer, R.; Poggi, J.-M.; Tuleau-Malot, C. Variable selection using random forests. *Pattern Recognit. Lett.* **2010**, *31*, 2225–2236. [CrossRef]
35. Probst, P.; Wright, M.N.; Boulesteix, A.L. Hyperparameters and tuning strategies for random forest. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2019**, *9*, e1301. [CrossRef]
36. Darst, B.F.; Malecki, K.C.; Engelman, C.D. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genet.* **2018**, *19*, 65. [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.