forests

MDPI

Article

Tree Height–Diameter Model of Natural Coniferous and Broad-Leaved Mixed Forests Based on Random Forest Method and Nonlinear Mixed-Effects Method in Jilin Province, China

Qigang Xu [1], Fan Yang [2], Sheng Hu [3], Xiao He [4] and Yifeng Hong [1,*]

[1] East China Academy of Inventory and Planning, National Forestry and Grassland Administration, Hangzhou 310000, China; adslxqg@126.com
[2] Academy of Forestry Inventory and Planning, National Forestry and Grassland Administration, Beijing 100714, China; yang2170057@163.com
[3] Industrial Development and Planning Institute, National Forestry and Grassland Administration, Beijing 100010, China; hs356442192@163.com
[4] State Key Laboratory of Efficient Production of Forest Resources, Key Laboratory of Forest Management and Growth Modelling, National Forestry and Grassland Administration, Institute of Forest Resource Information Techniques, Chinese Academy of Forestry, Beijing 100091, China; hexiao@ifrit.ac.cn
* Correspondence: hongyifeng007@126.com

Abstract: Objective: The purpose of this article was to use the Random Forest method and nonlinear mixed-effects method to develop a model for determining tree height–diameter at breast height (DBH) for a natural coniferous and broad-leaved mixed forest in Jilin Province and to compare the advantages and disadvantages of the two methods to provide a basis for forest management practice. Method: Based on the Chinese national forest inventory data, the Random Forest method and nonlinear mixed-effects method were used to develop a tree height–DBH model for a natural coniferous and broad-leaved mixed forest in Jilin Province. Results: The Random Forest method performed well on both the fitting set and validation set, with an R^2 of 0.970, *MAE* of 0.605, and *RMSE* of 0.796 for the fitting set and R^2 of 0.801, *MAE* of 1.44 m, and *RMSE* of 1.881 m for the validation set. Compared with the nonlinear mixed-effects method, the Random Forest model improved R^2 by 33.83%, while the *MAE* and *RMSE* decreased by 67.74% and 66.44%, respectively, in the fitting set; the Random Forest model improved R^2 by 9.88%, while the *MAE* and *RMSE* decreased by 14.38% and 12.05%, respectively, in the validation set. Conclusions: The tree height–DBH model constructed based on the Random Forest method had higher prediction accuracy for a natural coniferous and broad-leaved mixed forest in Jilin Province and had stronger adaptability for higher-dimensional data, which can be used for tree height prediction in the study area.

Keywords: tree height–diameter model; Random Forest; nonlinear mixed-effects model; coniferous and broad-leaved forest

Citation: Xu, Q.; Yang, F.; Hu, S.; He, X.; Hong, Y. Tree Height–Diameter Model of Natural Coniferous and Broad-Leaved Mixed Forests Based on Random Forest Method and Nonlinear Mixed-Effects Method in Jilin Province, China. *Forests* **2024**, *15*, 1922. https://doi.org/10.3390/f15111922

Academic Editor: Russell G. Congalton

Received: 24 September 2024
Revised: 25 October 2024
Accepted: 28 October 2024
Published: 31 October 2024

1. Introduction

The tree height–diameter (H-D) model is one of the most useful tools in forest management. While measuring tree height in overcrowded and dense forests is time-consuming and difficult, DBH data can be accurately measured for all trees in a plot. Therefore, a tree H-D model is usually developed to predict the missing total height measurements for the remaining trees [1].

One of the most important factors of forest structure is the relationship between tree DBH and height [2]. As people's interest in the structure of natural mixed forests was growing, the demand for determining the relationship between tree height and diameter in mixed forests also rose. More comprehensive modeling methods were applied to the research on the H-D relationship. Walter et al. studied the relationship between tree height and diameter in unevenly aged pure beech forests using the mixed-effects model method

and introduced the concept of "pseudo-residuals" to test the model [3]. It was found that the mixed-effects statistical method could fit bivariate normal distribution data very well, and the fitted model was easy to explain. Especially for forestry data, the NLME (nonlinear mixed-effects) method often showed a strong predictive ability. However, it must also be pointed out that the tree-level nonlinear mixed-effects model needs data from several groups of plots to be capable of random effect calibration when making predictions [4].

Models generated based on more dimensions and a larger amount of data had better performance. However, the nonlinear mixed-effects model had high requirements for modeling data. It was necessary to consider whether the data conformed to a normal distribution and the collinearity of independent variables. When fitting the model, the selection of initial values was also a major challenge. Machine learning methods did not have such issues. They had lower requirements for data and could contain more variables. Machine learning methods were gradually becoming a hot topic in the field of forest modeling. Chen et al. used an integrated neural network to conduct tree-level H-D modeling for six main tree species in central Canada [5]. More than 30 potentially important stand structures, sites, and climate variables were added to the model. The results showed that the model developed by this method performed excellently. Zhang et al. used traditional methods, Bayesian methods without prior information, and Bayesian methods with prior information to compare and analyze the estimation effect of the tree height–diameter relationship [6]. The results showed that the fitting results of traditional methods, Bayesian methods without prior information, and Bayesian methods with prior information were similar. However, the credible interval of Bayesian methods was more concentrated than that of traditional methods, and the parameter interval with prior information was more than 59% smaller than that of traditional methods. Shen et al. used a multilayer neural network to construct a tree height–diameter at breast height model in Guangdong Province. The research results showed that the neural network model was more accurate than the mixed-effects model [7]. Ozcelik et al. used a multilayer neural network and a mixed-effects model to develop a tree height–diameter model for Crimean juniper trees in southwestern Turkey. The results showed that both the nonlinear mixed-effects regression and backpropagation neural network modeling methods could produce accurate results. Compared with traditional nonlinear regression, the root mean square error (*RMSE*) of both was reduced by more than 20%. It was also pointed out that the backpropagation neural network seemed to be the method with the best generalization ability. In addition, from a practical perspective, compared with the mixed-effects model, its advantage is that it does not need to calibrate prior information when making predictions [8].

Among the many machine learning algorithms, the Random Forest method (RF) is an integrated decision tree algorithm. It fits models in a data-driven manner, has relatively low requirements for variables, is not restricted by statistical assumptions, and can include more variables. Consequently, in the field of forest modeling, the Random Forest algorithm is also receiving increasing attention [9]. Ou et al. used four machine learning methods (Random Forest, Boosted Regression Trees, cubist, and Multivariate Adaptive Regression Splines) to develop the individual tree basal area increment (BAI) growth model for a mixed forest in Northeast China. They found that the Random Forest method was an effective and powerful modeling method for predicting the growth of individual trees' BAI [10]. Jevšenak et al. used the Random Forest method to develop a BAI growth model based on the data of the national forest inventory of Slovenia. The results showed that the Random Forest method can provide similar verification statistical results to those of the previous traditional methods used in research reports [11].

In the field of forest management in Jilin Province, the main method used is the traditional regression model. Methods with stronger predictive abilities are currently in demand. Based on the national forest resources inventory data of natural coniferous and broad-leaved mixed forests in Jilin Province, this article used a total of 363 plots and 1305 sets of observations. The Random Forest method and the nonlinear mixed-effects method were used to develop and compare the tree height–diameter model of natural

mixed forests in Jilin Province in order to identify the model with the best performance and analyze the advantages and disadvantages of the two methods. Our research results can provide help for forest management in Jilin Province.

2. Materials and Methods

2.1. Modeling Data

We used the following variables to represent stand density and stand competition status: $rSDI$ [12], BAL_j, $BAL_j\_interspecies$, $BAL_j\_intraspecies$, and $BA\_proportion$. The variable descriptions are shown in Table 1.

Table 1. Descriptions of stand density and stand competition variables.

Variable	Description	
$rSDI$	$N(\frac{D_0}{D_g})^{1.605}$	(1)
BAL_j	$\frac{1}{S}\sum_{i=1}^{n}(1_{D_i \le D_j} BA_i)$	(2)
$BAL_j\_interspecies$	$\frac{1}{S}\sum_{i=1}^{n}(1_{D_i \ge D_j \& Species_i \ne Species_j} BA_i)$	(3)
$BAL_j\_intraspecies$	$\frac{1}{S}\sum_{i=1}^{n}(1_{D_i \ge D_j \& Species_i = Species_j} BA_i)$	(4)
$BA\_proportion$	$\frac{\sum_{i=1}^{n} BA_{i,i \in species}}{\sum_{j=1}^{N} BA_{j,j \in stand}}$	(5)

Note: D represents the diameter at breast height (DBH) of an individual tree in the stand; D_0 represents the standard DBH, which is equal to the number of trees in fully stocked stands with an average diameter of 10 cm in China [13]; D_g is the average stand diameter (by basal area); BA is the basal area of an individual tree in the stand.

The tree-level data for developing the tree height–diameter model came from the sample plot data of the national forest resources inventory (NFI) of natural coniferous and broad-leaved mixed forests in Jilin Province for 2014. In each plot, the tree heights of 2 to 5 individual trees representing the average tree height level of the plot were accurately measured. After removing outliers, there were 363 sample plots and 1305 sets of individual tree height–diameter observations. The data were divided into a fitting set (1044 observations) and a validation set (261 observations). The statistical data are shown in Table 2.

Table 2. Statistics of tree height (m) and DBH (cm).

Tree Variable	Fitting Data				Validation Data			
	Max	Min	Mean	Standard Deviation	Max	Min	Mean	Standard Deviation
D/cm	79.50	7.40	22.92	10.84	71.80	7.80	23.27	11.37
rSDI	344.14	25.03	154.80	49.00	289.45	25.03	158.15	46.10
BA_proportion	0.65	0.01	0.34	0.16	0.64	0.01	0.34	0.16
BAL/m^2	59.93	0.00	13.98	8.76	45.57	0.00	14.12	7.51
$BAL\_interspecies/m^2$	55.41	0.00	9.51	7.21	30.77	0.00	9.53	6.52
$BAL\_intraspecies/m^2$	25.67	0.00	4.48	4.05	25.45	0.00	4.58	4.05
Altitude/m	1860.00	175.00	811.66	290.97	1860.00	220.00	807.19	288.46
Soil thickness/cm	70.00	10.00	42.92	10.83	70.00	10.00	42.00	11.59
H/m	32.50	4.50	16.38	4.60	32.20	5.20	16.52	4.22

Climate data were sourced from ClimateAP (V3.00). ClimateAP is an application for dynamic local downscaling of historical and future climate data in the Asia–Pacific region [14]. Based on the latitude, longitude coordinates, and altitude information of the sample plots, the ClimateAP software could be used to extract seasonal and annual climate variables for each sample plot (the time interval extracted in this study was the average value from 1980 to 2010). The candidate climate factors are shown in Table 3.

Table 3. Descriptions of the candidate climatic variables.

Variable	Description
AHM	The humiture index
CMD	The Hargreaves moisture deficit index
DD_0	The number of days below 0 °C
DD_18	The number of days below 18 °C
DD18	The number of days above 18 °C
DD5	The number of days above 5 °C
EMT/°C	The extreme low temperature in the past 30 years
EXT/°C	The extreme high temperature in the past 30 years
EREF	The Hargreaves precipitation index
MAP/mm	The mean annual precipitation
MAT/°C	The mean annual temperature
MCMT/°C	The mean coldest month temperature
MWMT/°C	The mean warmest month temperature
NFFD	The number of frost days
PAS/mm	The snowfall from August of the previous year to July of the current year
TD/°C	The temperature difference between MWMT and MCMT

2.2. Tree Height–Diameter Model Based on Nonlinear Mixed-Effects Method

2.2.1. Selection of Climate Variables

Due to the large number of climate variables and their tendency to exhibit collinearity, a separate screening analysis of the climate variables was carried out. Principal component analysis (PCA) can be used as an exploratory method for assessing climate variability and is robust and reliable as an auxiliary technique when combined with other statistical techniques [15,16]. We first used the PCA method to analyze the data for all climate variables. Since the units of the climate variables were different, all variables were standardized before the PCA. The principal components that explained more than 80% of the variance were retained. For each principal component, variables with large loads were selected for further analysis. The variables that had a strong correlation with H and had the least multicollinearity among them were used as candidate options for modeling.

For competing relevant indicators (density index ($rSDI$), basal area proportion of the corresponding tree species of the sample tree in the sample plot ($BA\_proportion$), BAL, $BAL\_interspecies$, $BAL\_intraspecies$), and the two aspects of indicators representing the site conditions of the sample tree (*Altitude*, *Soil thickness*), the stepwise regression method was used for screening.

2.2.2. Base Model

The basic model was based on the research of Zang [17]. The Richard model was selected. The Richard model always performed well in simulating the relationship between tree height and diameter at breast height. The formula is as follows:

$$H = 1.3 + (a_0 \times (1 - exp(-b_0 \times D))^c + \varepsilon \quad (6)$$

where H represents the tree height; D represents the diameter at breast height; a_0, b_0 and c represent the parameter, ε represents the error term.

Due to the research object being a natural mixed forest, indicators related to the competition of sample trees, namely $rSDI$, $BA\_proportion$, BAL, $BAL\_interspecies$, $BAL\_intraspecies$, and two indicators representing the site conditions of the sample trees (*Altitude* and *Soil thickness*) were selected for stepwise regression screening and then entered into the model. Therefore, the model could be written as follows:

$$\mathrm{H} = f(\beta, D, Competition\ status, Site\ condition) + \varepsilon \quad (7)$$

where β is a vector of the fixed effect parameters, $Competition\ status$ is a variable group of the competition status of individual trees, and $Site\ condition$ is a variable group representing the site condition of individual trees. Other variables are as defined before.

2.2.3. Tree Species Data and Dummy Variable

The divided groups of *Pinus koraiensis* Sieb et Zucc., *Picea koraiensis* Nakai, *Larix gmelinii* (Rupr.) Kuzen., *Pinus sylvestris* var. *sylvestriformis* (Takenouchi) Cheng et C. D. Chu, and other coniferous tree species groups *Quercus mongolica* Fisch. ex Ledeb. group, *Betula platyphylla* Sukaczev group, *Fraxinus mandshurica* Rupr., *Juglans mandshurica* Maxim. and *Phellodendron amurense* Rupr. group, *Ulmus pumila* L. group, *Acer mono* Maxim. group, *Tilia tuan* Szyszyl. group, and *Populus* L. group and miscellaneous tree group, constituting a total of 14 tree species groups, were used to construct tree species dummy variables. After adding the tree species dummy variables, the model could be written as follows:

$$\mathrm{H} = f(\beta, D, Competition\ status, Site\ condition, S_m\ ) + \varepsilon \quad (8)$$

where S_m is the dummy variable of the tree species group. Among them, $S_1 = S_2 = \ldots = S_{13} = 0$ represents the *Pinus koraiensis* group, S_1 = 1 represents the *Picea koraiensis* group, and 0 represents other tree species groups, while S_2 = 1 represents the *Larix gmelinii* group. They were defined in turn according to the order introduced above. Other variables are as defined before.

2.2.4. Nonlinear Mixed-Effects Climate-Sensitive Model

In order to quantify the impact of the climate on *H-D* allometry, by reparameterizing the parameters in the basic *H-D* model and adding the selected climate variables to the model, the model could be written as follows:

$$\mathrm{H} = f(\beta, D, Competition\ status, Site\ condition, Climate\ , S_m) + \varepsilon \quad (9)$$

where *Climate* represents the vector of climate variables screened by PCA and correlation analysis. Other variables are as introduced previously.

Mixed-effects models have been proven to perform excellently in the field of forestry modeling. We chose to develop a mixed-effects model with the sample plot as a random effect. The new *H-D* mixed-effects model can be written as follows:

$$H_{ij} = f\left(\beta, D_{ij}, Competition\ status, Site\ condition, Climate, S_m + u_i\right) + \varepsilon_{ij} \quad (10)$$

$$u_i \sim N(0, \sigma^2_{plot})$$

where H_{ij} and D_{ij} are the tree height and DBH of the jth individual tree of the ith tree species, repectively. u_i is the random effect representing the sample plot. ε_{ij} is the random term. Other variables are as defined before.

When making model predictions, the corresponding parameter values of the random effect were calculated using the Empirical Best Linear Unbiased Prediction method (EBLUP). The formula is as follows:

$$\hat{u}_i = \hat{\Psi}\hat{Z}_i^T(\hat{Z}_i\hat{\Psi}\hat{Z}_i^T + \hat{R}_i)^{-1}e_i \quad (11)$$

where $\hat{u}_i$ is the estimated value of the random effect. $\hat{\Psi}$ is the q × q variance–covariance matrix representing the variation between groups. q is the number of random effects. $\hat{R}_i$ is the k × k variance–covariance matrix representing the variation within groups. k is the number of observations within a group. $\hat{Z}_i$ is the partial derivative matrix of the random effect. e_i is the residual vector between the measured value and the estimated value of the fixed effect model.

The data in this paper were the plot survey data of Jilin natural forest plots in 2014. Therefore, there was no time autocorrelation problem in the data. The formula of $\hat{R}_i$ is as follows:

$$R_i = \sigma^2 G_i^{0.5} \mathrm{I}_i G_i^{0.5} \quad (12)$$

where σ^2 is the residual vector of the model. G_i is the design matrix for explaining heteroscedasticity. In this paper, the power function form was adopted, which was $\mathrm{var}(\varepsilon_{ij}) = \hat{H}_{ij}^{2\gamma}$; I_i is the identity matrix.

The parameter estimation of the nonlinear mixed effect was based on the nlme module of R (Version 4.3.1) [18]. The algorithm was the default restricted maximum likelihood method.

2.3. Individual Tree H-D Model of Natural Coniferous and Broad-Leaved Mixed Forest Based on Machine Learning Methods

2.3.1. Random Forest

Random Forest is an ensemble learning algorithm that parallelizes individual decision trees. The Random Forest algorithm simultaneously adopts the ideas of resampling and combined prediction. The input variables of each sub-decision tree are randomly sampled from all the feature variables of the fitting set. There is no strong dependency between each individual decision tree. Each decision tree independently learns and makes predictions. Finally, through voting, the final classification result is reached according to the principle of the minority obeying the majority. Therefore, it is called Random Forest [9]. The Random Forest algorithm has the advantages of being insensitive to missing values and having an extremely strong model generalization ability. But at the same time, it also has the disadvantage of being prone to overfitting in high-dimensional data sets. However, this can be avoided by adjusting the optimal parameters.

2.3.2. Input Variable

Unlike NLME methods, the Random Forest algorithm had good adaptability to handling high-dimensional and collinear data. Moreover, the selected variable information was all effective information that had an impact on the tree height growth. Therefore, variable screening was not performed, and all variables were selected to develop a model: competition-related indicators (*rSDI*, *BA_proportion*, *BAL*, *BAL_interspecies*, *BAL_intraspecies*), indicators representing the site conditions of the sample trees (*Altitude* and *Soil thickness*), indicators representing climate conditions (*MAT*, *MWMT*, *MCMT*, *TD*, *MAP*, *AHM*, *DD_0*, *DD5*, *DD_18*, *DD18*, *NFFD*, *PAS*, *EMT*, *Eref*, *CMD*), and *DBH* as independent variables. The fitting and validation of the model were based on the Scikit-learn package in Python (Version 3.11.5) [19,20]. In this study, Random Forest algorithms performed one_hot encoding processing on discrete factor variables (tree species). That is, a factor variable of level N was expanded into N columns of attributes. Among these N column attributes of each sample observation value, a value of 1 indicated that the sample observation belonged to this category, and all other extended attributes were 0.

2.3.3. Parameter Tuning

In the process of parameter tuning of Random Forest, we used R_{cv}^2 of ten-fold cross-validation as an observation index to select the optimal parameters in the fitting set. The formula is as follows:

$$R_{cv}^2 = \frac{1}{k}\sum_{j=1}^{k}\left(1 - \frac{\sum_{i=1}^{n_j}(O_{ij} - P_{ij})^2}{\sum_{i=1}^{n_j}(O_{ij} - \overline{O}_j)}\right) \quad (13)$$

where k is the number of folds in cross-validation. In this paper, k was set to 10 folds. O_{ij} and P_{ij}, respectively, represent the ith observed value and model predicted value of the jth fold; $\overline{O}_j$ represents the average value of the observed values of the jth fold. n_j represents the sample number of the jth fold.

2.4. Model Evaluation

This paper used three indicators for model evaluation and inspection: the coefficient of determination (R^2), mean absolute error (*MAE*), and root mean square error (*RMSE*). The calculation formulas were as shown in Equations (14)–(16).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2} \tag{14}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|P_i - O_i| \tag{15}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(O_i - P_i)^2} \tag{16}$$

where O_i represents the observed value in the input data, P_i represents the predicted value of the model, and n represents the sample size of the input data. Both the *MAE* and *RMSE* were indicators for measuring the distance between the predicted values and observed values.

The workflow associated with this study is illustrated in Figure 1.

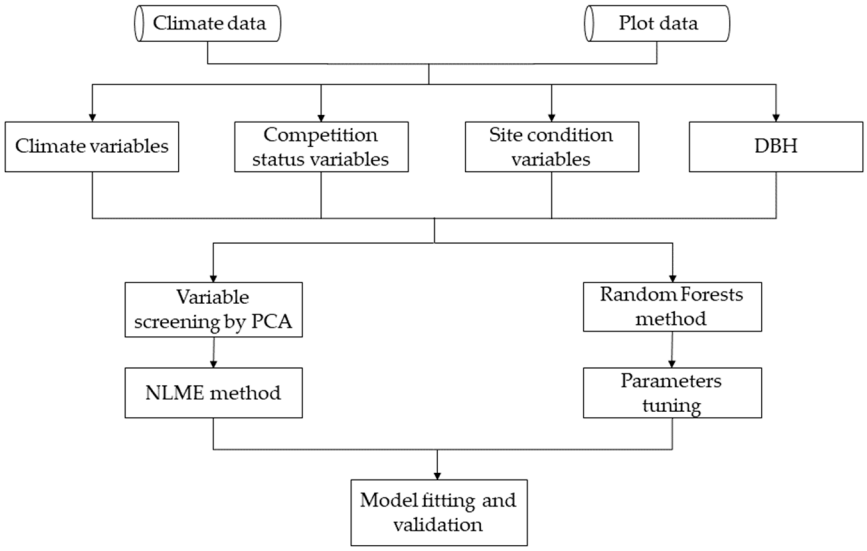

Figure 1. The workflow employed in this study to quantify the tree heights using NLME method and Random Forest method.

3. Results

3.1. Results of Nonlinear Mixed-Effects Model

3.1.1. Selection of Climate Variables

After standardizing the climate variables, principal component analysis was used. The results of the principal component analysis showed that the first two principal components explained 87.83% of the variation in all climate data.

The variable selection process of the principal component analysis was carried out according to the following principles: (1) Variables with an absolute value of load greater than 0.32 were selected. (2) If there was no variable with an absolute value of load greater than 0.32, the top three variables were selected. If there were variables with tied absolute values of load, all the tied variables were selected.

The results of each component load were shown in Table 4. For principal component 1, *DD5*, *MAT*, *MWMT*, and *DD_18* were selected for the next step. For principal component 2, *MCMT*, *TD*, and *EMT* were selected for the next step. The seven selected variables were subjected to Pearson correlation analysis with the dependent variable *H*. The analysis

results are shown in Table 5. The results showed that the seven climate variables were all linearly significant with tree height, *H*, and there was collinearity among the seven climate variables. We used the variance inflation factor, VIF, to gradually eliminate variables with high collinearity. The final results are shown in Table 6. Therefore, finally, three climate variables, *MCMT*, *TD*, and *EMT*, were selected to represent the climate and added to the model.

Table 4. PCA results for the climate variables.

	Comp.1	**Comp.2**	**Comp.3**
MAT	0.279	0.107	0.15
MWMT	0.276		0.203
MCMT	0.214	0.447	−0.134
TD	0.203	−0.399	0.352
MAP	−0.228	0.212	0.473
AHM	0.258	−0.117	−0.347
DD_0	−0.256	−0.287	−0.102
DD5	0.281		0.175
DD_18	−0.276	−0.136	−0.13
DD18	0.267		0.224
NFFD	0.252	0.266	0.163
PAS	−0.256	0.132	0.299
EMT	0.196	0.455	−0.245
EXT	0.272	−0.174	
Eref	0.23	−0.261	0.137
CMD	0.232	−0.266	−0.373
Cumulative variation	76.15%	87.83%	95.22%

Table 5. Pearson correlation coefficient matrices between H and climatic variables.

	DD5	**MAT**	**MWMT**	**DD_18**	**MCMT**	**TD**	**EMT**	**H**
DD5	1.000	-	-	-	-	-	-	-
MAT	0.982 ***	1.000	-	-	-	-	-	-
MWMT	0.989 ***	0.957 ***	1.000	-	-	-	-	-
DD_18	−0.966 ***	−0.996 ***	−0.935 ***	1.000	-	-	-	-
MCMT	0.690 ***	0.800 ***	0.624 ***	−0.830 ***	1.000	-	-	-
TD	0.780 ***	0.662 ***	0.839 ***	−0.613 ***	0.099 ***	1.000	-	-
EMT	0.625 ***	0.700 ***	0.552 ***	−0.712 ***	0.883 ***	0.089 ***	1.000	-
H	−0.273 ***	−0.254 ***	−0.257 ***	0.238 ***	−0.180 ***	−0.201 ***	−0.232 ***	1.000

Note: *** $p < 0.001$.

Table 6. The VIFs of the final climatic factors.

Climate Variable	***MCMT***	***TD***	***EMT***
VIF	4.531	1.010	4.522

After stepwise regression to screen the competition indicators and site indicators, the indicators finally entering the model were diameter (D), stand density index ($rSDI$), basal area proportion ($BA\_proportion$), and interspecific BAL ($BAL\_interspecies$). After the PCA and correlation analysis to screen the climate variables, the climate variables that were finally entered into the model were *MCMT*, *TD*, and *EMT*. Then, the screened variables and tree species dummy variables were put into different positions of parameters a, b, and c of the basic model, respectively. The optimal model with the best coefficient significance performance was as follows:

$$H_{ij} = 1.3 + \left(a_0 + a_1 rSDI + a_2 BAL + a_3 MCMT + a_4 EMT + \sum_{m=1}^{13} f_m S_m + u_i\right)\left[1 - e^{-b_0 D_{ij}}\right]^{c_0} + \varepsilon_{ij} \quad (17)$$

where $a_0 \sim a_6$ are the parameters to be estimated. The remaining variables are as previously described.

3.1.2. Fitting and Test Results of the Final NLME Model

The model accuracy test results showed that the nonlinear mixed-effects model performed well in both the fitting set and the validation set. The *MAE* and *RMSE* of the fitting set were 1.056 m and 1.446 m, respectively, and the *MAE* and *RMSE* of the validation set were 1.420 m and 1.926 m, respectively (Table 7).

Table 7. The parameter estimation results and model validation results of final H-D NLME model.

	Parameters	Parameter Definition	Equation (17)
Fixed effects parameters	a		17.872 (0.000)
	b		0.035 (0.000)
	c		0.926 (0.000)
	a_1	*rSDI*	0.022 (0.000)
	a_2	*BAL*	−4.493 (0.005)
	a_3	*MCMT*	0.686 (0.052)
	a_4	*EMT*	−0.461 (0.025)
	f_1	*Picea asperata Mast.*	0.211 (0.609)
	f_2	*Larix gmelinii*	2.689 (0.000)
	f_3	*Pinus koraiensis*	0.122 (0.781)
	f_4	*Pinus sylvestris*	−2.296 (0.013)
	f_5	*Other coniferous tree species*	−3.829 (0.000)
	f_6	*Quercusmongolica*	−1.707 (0.002)
	f_7	*Birch*	2.157 (0.000)
	f_8	*Fraxinus mandshurica&Juglans mandshurica&Phellodendron amurense*	0.980 (0.106)
	f_9	*Elm*	−0.436 (0.488)
	f_{10}	*Acer pictumThunb.&Acer triflorum&Acer mandshuricum*	−0.103 (0.852)
	f_{11}	*linden*	0.029 (0.944)
	f_{12}	*poplar*	0.554 (0.428)
	f_{13}	*Miscellaneous wood*	−1.557 (0.060)
Variance components	σ_{plot}		0.490
	γ		0.445
Model performance			
AIC			4837.658
Fitting set R^2			0.901
Fitting set *MAE* (m)			1.056
Fitting set *RMSE* (m)			1.446
Validation set R^2			0.791
Validation set *MAE* (m)			1.420
Validation set *RMSE* (m)			1.926

3.2. Results of Random Forest Model

The Random Forest model was developed based on the scikit-learn module in Python [21]. For the three main parameters, n_estimator (the number of trees in the forest), max_depth (the number of splits that each decision tree was allowed to make) and max_features (the size of the random subsets of features to consider when splitting a node), the grid search method was used for parameter tuning. The parameter tuning process was as follows: iterative fitting with 10-fold cross-validation with the fitting set data as the object, taking the highest R^2_{cv} as the optimal parameter, and calculating step by step to obtain the optimal parameter combination. The final optimal parameters were n_sitimators = 370; max_depth = 17; and max_features = 15. Then, a Random Forest model was developed based on the optimal parameters.

The tree height–diameter model of the Jilin natural forest was developed by using Random Forest. The test results of the model on the fitting set and the validation set were

as follows: The Random Forest method performed excellently on both the fitting set and the validation set. The R^2 of the fitting set was 0.970, the R^2 of the validation set was 0.801, the *MAE* was 1.44 m, and the *RMSE* was 1.881 m. Compared with the NLME model, in the training, the R^2 of Random Forest was increased by 33.83%, and the *MAE* and *RMSE* were decreased by 67.74% and 66.44%, respectively; in the performance of the Random Forest model on the validation set, R^2 was increased by 9.88%, and the *MAE* and *RMSE* were decreased by 14.38% and 12.05%, respectively (Table 8).

Table 8. Validation result of H-D model based on 3 machine learning methods and NLME method.

	Random Forest	NLME Method
Training set R^2	0.970	0.901
Training set *MAE* (m)	0.605	1.056
Fitting set *RMSE* (m)	0.796	1.446
Validation set R^2	0.801	0.791
Validation set *MAE* (m)	1.440	1.420
Validation set *RMSE* (m)	1.881	1.926

The residual plots of the finally fitted NLME model and the Random Forest model are shown in Figure 2. The results showed that the scatter points were randomly distributed on both sides of the 0-axis. No obvious heteroscedasticity trend was observed for the NLME model. There was an obvious trend in the lower graph, where the residuals in the Random Forest model were not uniformed distributed. The Random Forest model had a better predictive ability, and the scatter points were more closely distributed.

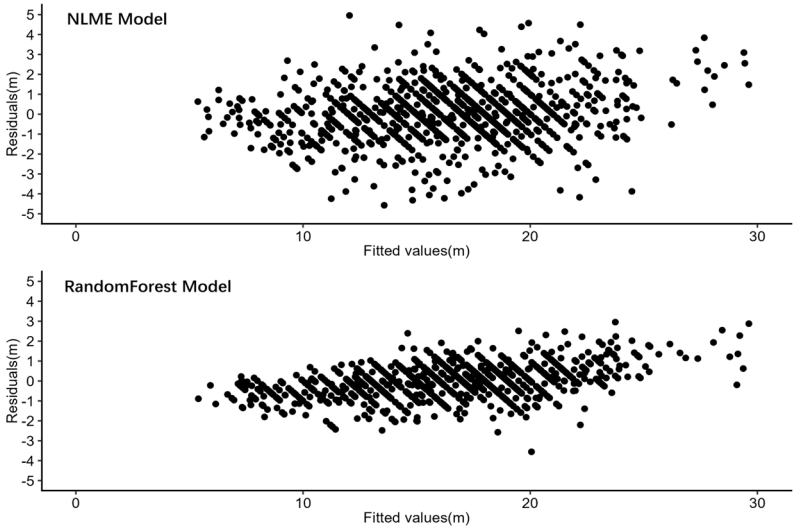

Figure 2. Residuals for tree height–diameter models (NLME model; Random Forest model).

4. Discussion

4.1. Nonlinear Mixed-Effects Model

The test results of using the nonlinear mixed-effects method to develop the tree height–diameter model for natural mixed forests in Jilin Province showed that the model had an excellent extrapolation ability. The nonlinear mixed-effects method had always performed well while developing the model. Sharma et al. used Czech national forest survey data to develop a nonlinear mixed-effects HDR (height–diameter ratio) model. The results showed that the R^2 of the HDR model for each tree species ranged from 0.8574 to 0.9605 [4]. Ciceu et al. used the nonlinear mixed-effects method to construct a tree height–diameter model for an unevenly aged mixed forest of Norway spruce in Romania. The research results showed that the addition of random effects increased the prediction accuracy of the tree height by 50 cm [22]. Meng et al. used the nonlinear mixed-effects

method to construct a tree height–diameter model for the main dominant tree species in northern Ontario, Canada. After adding random effects, the fitting results and prediction accuracy of the models for all tree species were improved [23]. In forestry modeling, compared to methods that only used fixed effects, the mixed-effects method could incorporate more variable information, such as plot location information, into the model. And plot information contains many useful factors that cannot be included in fixed effects models. Therefore, compared with traditional regression methods, tree height–diameter models with random effects usually perform better in terms of model fitting performance and prediction accuracy.

4.2. The Influence of Temperature and Competition on the H-D Relationship

The results of the significance test of the model parameters showed that although the *p*-value of *MCMT* was at the 0.1 level, the rest of the coefficients were all significant at the 0.05 level. The coefficient of *rSDI* was positive, indicating that individual trees under high-density pressure will invest more resources in tree height growth to obtain more sunlight. Many studies have found that competition will have a significant impact on the allometric growth of trees. The greater the intensity of competition is, the thinner the individual trees in the stand will grow [24–28]. The coefficient of *MCMT* was positive, which meant that the higher the average temperature in the coldest month is, the faster the increase in the tree height-to-diameter ratio of individual trees in the study area is. Previous research results have shown that for the growth of individual trees, there was an optimal temperature for photosynthesis and growth, and the influence of the temperature on tree height growth was greater than that on diameter growth [26]. The study areas in this paper all belong to northern China, and the annual average temperature is relatively cold and has not reached the turning point of the optimal temperature. Therefore, for individual trees in the study area, an increase in temperature would make the tree height increase more rapidly.

4.3. Machine Learning Algorithms and Forestry Modeling

Machine learning methods have gradually become a hotspot in the field of forest modeling in recent years, because they often have better predictive capabilities than traditional regression methods. These methods can contain more variables. Ogana et al. [29] used three methods, the DLA (Deep Learning Algorithm), NLS, and NLME, to study the H-D relationship of complex tropical rainforest trees. It was found that the DLA model was superior to the NLS and NLME models. Compared with NLS and NLME, the error of estimating aboveground biomass by tree height predicted by using the DLA model was reduced by more than 30%. The deep learning network model can be regarded as an alternative to traditional nonlinear regression techniques. Qin et al. used the DLA method and the NLME method to construct a natural mixed forest crown model. It was found that the best DLA model can explain 69% of the crown variation. When all 22 input variables were used for modeling, the DLA model performed better than the NLME model [30].

However, Dantas et al. employed the ANN (Artificial Neural Network), SVM (Support Vector Machine), and NLME methods to construct a volume model for *eucalyptus* plantations in Minas Gerais, Brazil. They found that the nonlinear mixed-effects model performed the best [31]. A possible reason for this result was that the research object of the article was *eucalyptus* plantations, which had a relatively simple structure. The variables affecting the volume of *eucalyptus* were more explicit. Therefore, using the nonlinear mixed-effects method could already accurately express the variations in *eucalyptus* volume. The two machine learning methods, ANN and SVM, required more hyperparameters to be adjusted. It was more difficult to find the optimal parameters in the parameter space to express the variations in *eucalyptus* volume. Consequently, in this study, the nonlinear mixed-effects model showed the best performance.

Our research results showed that compared with the nonlinear mixed-effects method, in the performance of the fitting set of the Random Forest model, R^2 was increased by 33.83%, and the *MAE* and *RMSE* were decreased by 67.74% and 66.44%, respectively; in

the performance of the validation set, R^2 was increased by 9.88%, and the *MAE* and *RMSE* were decreased by 14.38% and 12.05%, respectively. The Random Forest method showed excellent performance, which was similar to the results of previous studies. Yu Yang et al. used beta regression and the Random Forest algorithm to develop a crown ratio (CR) model. Their research results showed that the CR model developed based on the RF algorithm was superior to the model developed by beta regression. The algorithm idea based on the integrated system can improve the accuracy by itself. Coupled with the interpretation ability of multiple variables, the Random Forest algorithm has a stronger prediction ability [32].

5. Conclusions

This paper used the nonlinear mixed-effects method and the Random Forest method to construct a climate-sensitive tree height–diameter model for natural coniferous and broad-leaved mixed forests in Jilin Province. The parameter estimation and test results of the nonlinear mixed-effects method showed that temperature and competition were the key variables affecting the allometric increase in individual tree height–diameter. For individual trees in natural mixed forests in Jilin Province, an increase in temperature would make the increase in individual tree height more rapid. The model test results showed that compared with the mixed-effects method, the model constructed by the Random Forest algorithm had a higher prediction accuracy. The Random Forest method showed a strong generalization ability when constructing the tree height–diameter model and had low requirements for modeling data. In application scenarios with high requirements for prediction accuracy, it had advantages over traditional models. Our research results can provide decision-making support for forest management in Jilin Province.

Author Contributions: Methodology, data curation, formal analysis, writing—original draft preparation, and writing—review and editing, Q.X.; formal analysis, software, and writing—review and editing, F.Y., S.H. and X.H.; funding acquisition, conceptualization, and writing—review and editing, Y.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China (2022YFE0112700), the National Key R&D Program of China (2023YFD2201705-6), and the Research on the Path of Baishanzu National Park to Assist Regional Carbon Neutrality (2022JBGS08).

Data Availability Statement: Data are contained within the article.

Acknowledgments: The authors are grateful to the State Key Laboratory of Efficient Production of Forest Resources, Key Laboratory of Forest Management and Growth Modelling, National Forestry and Grassland Administration, Institute of Forest Resource Information Techniques, and Chinese Academy of Forestry for providing the data used in this study. We would also like to thank the Editors and anonymous reviewers for their valuable comments.

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

References

1. Crecente-Campo, F.; Tomé, M.; Soares, P.; Diéguez-Aranda, U. A generalized nonlinear mixed-effects height–diameter model for Eucalyptus globulus L. in northwestern Spain. *For. Ecol. Manag.* **2010**, *259*, 943–952. [CrossRef]
2. Knoebel, B.R.; Burkhart, H.E. A bivariate distribution approach to modeling forest diameter distributions at two points in time. *Biometrics* **1991**, *47*, 241–253. [CrossRef]
3. Zucchini, W.; Schmidt, M.; von Gadow, K. A model for the diameter-height distribution in an uneven-aged beech forest and a method to assess the fit of such models. *Silva Fenn.* **2001**, *35*, 169–183. [CrossRef]
4. Sharma, R.P.; Vacek, Z.; Vacek, S.; Kučera, M. A nonlinear mixed-effects height-to-diameter ratio model for several tree species based on Czech national forest inventory data. *Forests* **2019**, *10*, 70. [CrossRef]
5. Chen, J.; Yang, H.; Man, R.; Wang, W.; Sharma, M.; Peng, C.; Parton, J.; Zhu, H.; Deng, Z. Using machine learning to synthesize spatiotemporal data for modelling DBH-height and DBH-height-age relationships in boreal forests. *For. Ecol. Manag.* **2020**, *466*, 118104. [CrossRef]
6. Zhang, X.; Duan, A.; Zhang, J.; Xiang, C. Estimating Tree Height-Diameter Models with the Bayesian Method. *Sci. World J.* **2014**, *2014*, 683691. [CrossRef]
7. Shen, J.; Hu, Z.; Sharma, R.P.; Wang, G.; Meng, X.; Wang, M.; Wang, Q.; Fu, L. Modeling height–diameter relationship for poplar plantations using combined-optimization multiple hidden layer back propagation neural network. *Forests* **2020**, *11*, 442. [CrossRef]

8. Özçelik, R.; Diamantopoulou, M.J.; Crecente-Campo, F.; Eler, U. Estimating Crimean juniper tree height using nonlinear regression and artificial neural network models. *For. Ecol. Manag.* **2013**, *306*, 52–60. [CrossRef]
9. Breiman, L. Random Forest. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
10. Ou, Q.; Lei, X.; Shen, C. Individual tree diameter growth models of larch–spruce–fir mixed forests based on machine learning algorithms. *Forests* **2019**, *10*, 187. [CrossRef]
11. Jevšenak, J.; Skudnik, M. A random forest model for basal area increment predictions from national forest inventory data. *For. Ecol. Manag.* **2021**, *479*, 118601. [CrossRef]
12. Reineke, L.H. Perfecting a stand-density index for even-aged forests. *J. Agric. Res.* **1933**, *46*, 627–638.
13. Meng, X.; She, G.; Li, F.; Wang, X. *Forest Mensuration*; China Forestry Publishing House: Beijing, China, 2006.
14. Wang, T.; Wang, G.; Innes, J.L.; Seely, B.; Chen, B. ClimateAP: An application for dynamic local downscaling of historical and future climate data in Asia Pacific. *Front. Agric. Sci. Eng.* **2017**, *4*, 448–458. [CrossRef]
15. Scolforo, J.R.S.; Maestri, R.; Ferraz Filho, A.C.; de Mello, J.M.; de Oliveira, A.D.; de Assis, A.L. Dominant height model for site classification of Eucalyptus grandis incorporating climatic variables. *Int. J. For. Res.* **2013**, *2013*, 139236.
16. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [CrossRef]
17. Zang, H.; Lei, X.; Ma, W.; Zeng, W. Spatial heterogeneity of climate change effects on dominant height of larch plantations in northern and northeastern China. *Forests* **2016**, *7*, 151. [CrossRef]
18. Team, R.C. *R: A Language and Environment for Statistical Computing*; Foundation for Statistical Computing: Vienna, Austria, 2013; Available online: http://www.r-project.org (accessed on 12 January 2016).
19. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M. {TensorFlow}: A system for {Large-Scale} machine learning. In Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), Savannah, GA, USA, 2–4 November 2016; pp. 265–283.
20. Gulli, A.; Pal, S. *Deep Learning with Keras*; Packt Publishing Ltd.: Birmingham, UK, 2017.
21. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
22. Ciceu, A.; Garcia-Duro, J.; Seceleanu, I.; Badea, O. A generalized nonlinear mixed-effects height–diameter model for Norway spruce in mixed-uneven aged stands. *For. Ecol. Manag.* **2020**, *477*, 118507. [CrossRef]
23. Meng, S.X.; Huang, S.; Lieffers, V.J.; Nunifu, T.; Yang, Y. Wind speed and crown class influence the height–diameter relationship of lodgepole pine: Nonlinear mixed effects modeling. *For. Ecol. Manag.* **2008**, *256*, 570–577. [CrossRef]
24. Crecente-Campo, F.; Corral-Rivas, J.J.; Vargas-Larreta, B.; Wehenkel, C. Can random components explain differences in the height–diameter relationship in mixed uneven-aged stands? *Ann. For. Sci.* **2014**, *71*, 51–70. [CrossRef]
25. Forrester, D.I.; Benneter, A.; Bouriaud, O.; Bauhus, J. Diversity and competition influence tree allometric relationships–Developing functions for mixed-species forests. *J. Ecol.* **2017**, *105*, 761–774. [CrossRef]
26. Fortin, M.; Van Couwenberghe, R.; Perez, V.; Piedallu, C. Evidence of climate effects on the height-diameter relationships of tree species. *Ann. For. Sci.* **2019**, *76*, 1. [CrossRef]
27. Garber, S.M.; Temesgen, H.; Monleon, V.J.; Hann, D.W. Effects of height imputation strategies on stand volume estimation. *Can. J. For. Res.* **2009**, *39*, 681–690. [CrossRef]
28. Temesgen, H.; v Gadow, K. Generalized height–diameter models—An application for major tree species in complex stands of interior British Columbia. *Eur. J. For. Res.* **2004**, *123*, 45–51. [CrossRef]
29. Ogana, F.N.; Ercanli, I. Modelling height-diameter relationships in complex tropical rain forest ecosystems using deep learning algorithm. *J. For. Res.* **2022**, *33*, 883–898. [CrossRef]
30. Qin, Y.; Wu, B.; Lei, X.; Feng, L. Prediction of tree crown width in natural mixed forests using deep learning algorithm. *For. Ecosyst.* **2023**, *10*, 100109. [CrossRef]
31. Dantas, D.; Calegario, N.; Acerbi, F.W.; Carvalho, S.d.P.C.; Isaac, M.A.; Melo, E.d.A. Multilevel nonlinear mixed-effects model and machine learning for predicting the volume of Eucalyptus spp. trees. *Cerne* **2020**, *26*, 48–57. [CrossRef]
32. Yu, Y.; Zhou, Z.; Sharma, R.P.; Zhang, L.; Du, M.; Zhang, H. Comparing crown ratio models for spruce-fir broadleaved mixed forests using beta regression and random forest algorithm. *Comput. Electron. Agric.* **2024**, *225*, 109302. [CrossRef]