climate

MDPI

Article

Data-Driven Analysis of Forest–Climate Interactions in the Conterminous United States

Olga Rumyantseva and Nikolay Strigul *

Department of Mathematics and Statistics, Washington State University, 14204 NE Salmon Creek Avenue, Vancouver, WA 98686, USA; olga.rumyantseva@wsu.edu
* Correspondence: nick.strigul@wsu.edu

Abstract: A predictive understanding of interactions between vegetation and climate has been a grand challenge in terrestrial ecology for over 200 years. Developed in recent decades, continental-scale monitoring of climate and forest dynamics enables quantitative examination of vegetation–climate relationships through a data-driven paradigm. Here, we apply a data-intensive approach to investigate forest–climate interactions across the conterminous USA. We apply multivariate statistical methods (stepwise regression, principal component analysis) including machine learning to infer significant climatic drivers of standing forest basal area. We focus our analysis on the ecoregional scale. For most ecoregions analyzed, both stepwise regression and random forests indicate that factors related to precipitation are the most significant predictors of forest basal area. In almost half of US ecoregions, precipitation of the coldest quarter is the single most important driver of basal area. The demonstrated data-driven approach may be used to inform forest-climate envelope modeling and the forecasting of large-scale forest dynamics under climate change scenarios. These results have important implications for climate, biodiversity, industrial forestry, and indigenous communities in a changing world.

Keywords: climate–vegetation interactions; data-intensive modeling; dimensionality reduction; forest inventories; multivariate statistics; machine learning; random forests; stepwise regression

Citation: Rumyantseva, O.; Strigul, N. Data-Driven Analysis of Forest–Climate Interactions in the Conterminous United States. *Climate* **2021**, *9*, 108. https://doi.org/10.3390/cli9070108

Academic Editors: Marcello Vitale and Alessio Collalti

Received: 27 April 2021
Accepted: 21 June 2021
Published: 30 June 2021

1. Introduction

Understanding the interactions between vegetation and climate is a central question in ecology and biogeography [1,2]. The seminal work by von Humboldt and Bonpland [3] built a foundation for quantitative analysis of climatic factors controlling vegetation at a large scale. Climate–vegetation observations collected over the 19th century resulted in the Köppen climate classification, developed by Köppen in 1884–1936 [4–6]. The Köppen classification was critically evaluated, developed, modified, and improved over the next century [7–10]. Several of its extensions are known as Köppen and Geiger [10,11] and Trewartha classifications [12,13]. The Köppen classification is also employed in the delineation of Bailey's US ecoregions [14]. Another bioclimatic scheme, called the Holdridge life zone system, was proposed by Holdridge in 1947 [15,16], initially for tropical regions and later extended to boreal and temperate zones [17]. These bioclimatic classification schemes served as a scientific basis and inspiration for modern developments, including climate envelope models (CEMs), species distribution models (SDMs), and dynamic global vegetation models (DGVMs), within Earth system models (ESMs), focusing on both the understanding of vegetation formations and the prediction of climate effects across various temporal and spatial scales [18,19].

Forested and atmospheric systems are complex adaptive systems, as they operate at multiple scales, preserve structural unity, and demonstrate self-organizational patterns [20–23]. The determination of macroscopic characteristics capable of capturing complex system dynamics is a challenging problem due to their multidimensional and multi-scale nature [24–27]. While this complexity is broadly acknowledged, atmospheric systems

in vegetation modeling are often characterized by macroscopic mean-field approximations [18]. In particular, average temperature and precipitation are widely used as climatic factors determining the distribution and abundance of plant communities, gradients in the vegetation continuum, and primary productivity [1]. In a seminal work, von Humboldt and Bonpland [3] used temperature as the primary factor and air pressure as the secondary factor to quantify altitude-related climatic controls on vegetation. The first Köppen classification scheme was based on the mean temperature, as precipitation was added as another factor in an updated classification [4]. Yet, mean temperature and precipitation patterns were insufficient for adequate climate–vegetation classification, and only the inclusion of interseasonal changes allowed differentiation between climatic zones [5,6]. The Holdrige system [15,16] employs three primary variables: (1) precipitation, (2) biotemperature or the annual mean temperature adjusted to the duration of the vegetation period, and (3) potential evapotranspiration ratio. Due to the complexity of atmospheric and vegetative systems, traditional reasoning often dictated the selection of mean temperature and precipitation as primary climatic parameters in large-scale climate-envelope and plant-species distribution models [19].

A data-intensive approach provides an opportunity to quantitatively evaluate relationships within and across complex ecological systems through data mining and data-driven modeling [28–30]. In particular, currently available large-scale spatially explicit climatic data sets and continental-level individual-based forest inventories allow a data-intensive analysis of climate–vegetation systems and data mining for possible connections between different quantitative characteristics of climate and vegetation [19,24,26]. In this work, we employ a data-intensive paradigm to examine the continental-scale effects of climate on forested ecosystems. Specifically, we investigate the relationship between forest basal area and climatic factors in the conterminous United States, for each ecological domain (Humid, Dry, and Humid Tropical Domains) and for each ecoregion as per Bailey's classification scheme [14,31,32]. Our goal is to rank climatic factors by their effect on forest basal areas across these ecoregions.

We employ a combination of two distinct approaches: (1) traditional multivariate statistical methods (correlation analysis, principal component analysis, and stepwise regression; Section 2.2.1) and (2) a recently developed machine learning approach (random forests; Section 2.2.2). Multiple regression models in combination with multivariate statistics is a commonly used statistical methodology for exploring climatic drivers of forested ecosystems. For example, the link between forest growth rate and temperature and precipitation patterns was explored in [33] using regression analysis. Stepwise linear regression was used in [34] to describe climate effects on a tree species. Various linear regressions were implemented to model stem basal area dynamics related to climatic factors in [35]. Random forests is an ensemble machine learning method [36], averaging over multiple decision trees to derive a robust classifier or regressor (Section 2.2.2). In this work, we perform an intercomparison between stepwise multiple regression and random forests through a data-intensive paradigm by applying both methods to infer forest–climate interactions. While the outcomes of both methods were similar, our intercomparison reveals that random forests is an efficient tool for data mining and modeling (see Section 3.3).

2. Materials and Methods

2.1. Data Mining

We mined two continental-scale databases: the USDA Forest Inventory and Analysis (FIA) database and the WorldClim climatic database (https://www.fia.fs.fed.us/ and https://www.worldclim.org/, last time accessed November 2020). These data sets cover three ecological domains and 36 ecological regions across continental USA (Appendix A). The WorldClim data set contains normals for different climatic variables computed with a spatial resolution of approximately 1 km since 1950 [37]. The FIA data set includes information on 211,949 forest plots observed over the 1968–2013 period. Figure A1 visualizes the spatial distribution of FIA forest plots across ecoregions. Figure 1 summarizes the data used in our

study. In this study, we characterize forested ecosystems using basal area (see [24,38,39] for details on forest basal area computation with the FIA data set). After mining the FIA data set, we obtained 409,868 observations of basal area across the conterminous USA. This includes different forested plots observed at irregular time intervals due to the sampling rotation rule implemented by the US Forest Service. To normalize the observations and to avoid duplicate records, we selected the data snapshot for the year 2000. Specifically, we used only one observation of every FIA forest inventory plot acquired near the year 2000. These forest basal area measurements were linked with 19 climatic variables of three types (temperature, precipitation, variability) using the WorldClim data set (see Figure 1).

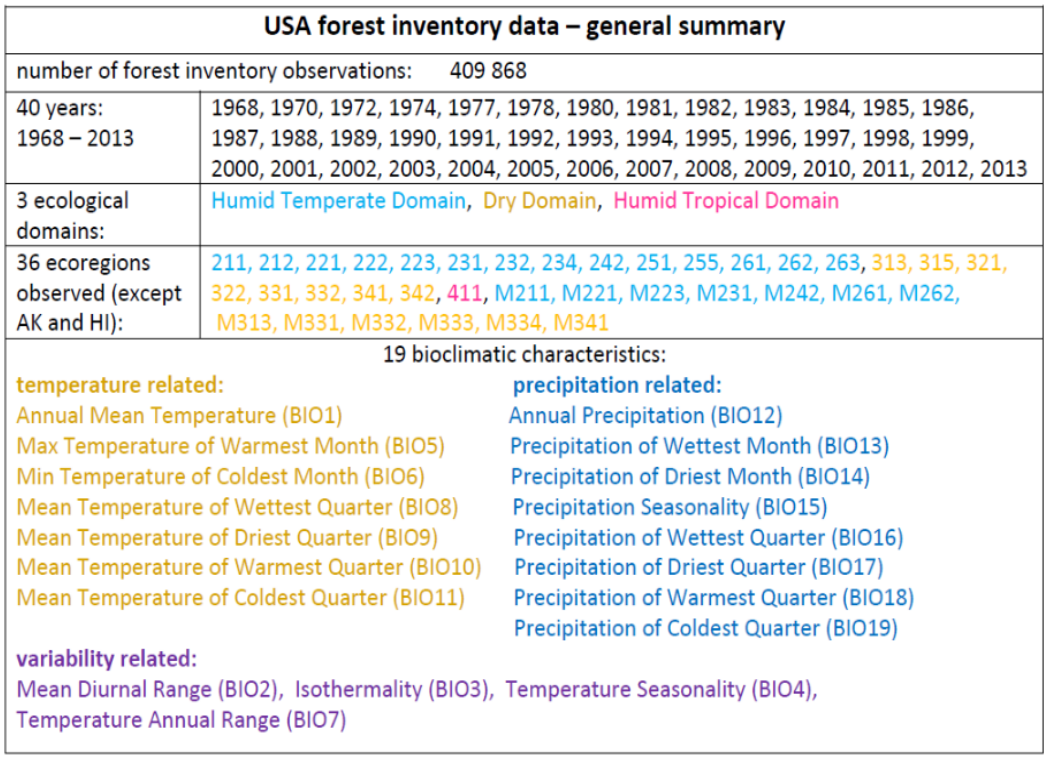

Figure 1. General data overview: FIA and WorldClim data sets. Ecoregions are colored depending on the domain: blue—Humid Domain, yellow—Dry Domain, pink—Tropical Domain. Bioclimatic characterisics colors: yellow—temperature related, blue—precipitation related, violet—variability related.

2.2. Data Analysis and Software

Our primary goal was to reveal large-scale climatic drivers of forest basal area dynamics. We employed stepwise linear regression and a machine learning approach known as random forests (Section 3.2, [36]) with climatic factors as the independent variables. In concert with stepwise regression, we applied traditional multivariate statistical techniques, correlation analysis (Section 3.1.1), and principal component analysis (Section 3.1.2). We employed multivariate traditional statistical and recent machine learning approaches to analyze the (a) conterminous USA, (b) Humid, Dry, and Humid Tropical Domains, and (c) 36 US ecoregions (Figure 1 and Appendix A). We used the open-source R language for statistical computing (www.r-project.org), packages *FactoMineR* (http://factominer.free.fr/) and *factoextra* (https://rpkgs.datanovia.com/factoextra/index.html) to conduct and visualize the principal component analysis (PCA), and the popular Python library *scitkit-learn* (scikit-learn.org) to run the random forests analysis. Original software used in this study is freely available at the following `GitHub` repository: `Olia8848/US-forest-dynamics-vs-climate`.

2.2.1. Stepwise Linear Regression

We employed stepwise multiple regression to develop predictive models and to rank climatic factors with respect to their linear connection with forest basal area. We built stepwise regression models sequentially by adding climatic factors one-by-one and re-evaluating model parameters at every step.

Step 1. Simple linear regression

At the first step, we built a collection of simple linear regression models for all climatic variables in a focal ecoregion and chose the 'best' model with the highest coefficient of determination R^2.

Let us consider an ecoregion with n forest inventory plots. Let $\vec{y} = (y_1, y_2, \ldots, y_n)$ be the vector of basal area observations at n forest plot locations. We compute WorldClim values for all $k = 19$ climatic variables at these n locations. Let $\vec{c^j} = (c_1^j, c_2^j, ..., c_n^j)$ be the vector of values of a climatic characteristic $c^j, j = 1 \ldots 19$. We build a simple linear regression model of the following type: **basal area** $\sim$ **single climatic characteristic** for each $c^j, j = 1 \ldots 19$:

$$\vec{y} = \alpha + \beta \cdot \vec{c^j} + \vec{\varepsilon}, \quad (1)$$

where $\vec{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ is a random error vector, and α and β are the intercept and slope, respectively.

We then select the 'best' model with the highest coefficient of determination R^2 among these 19 linear regression models. We consider the chosen model as the 'best' model per R^2 and the Euclidean distance of data points from the regression line ($0 \leq R^2 \leq 1$, and the statement $R^2 = 1$ means that all data points belong to the line). We determine the slope and intercept of the simple linear regression model (1), and, in this case, the coefficient of determination R^2 is equal to the square of the sample Pearson correlation coefficient, r_{y,c^j}. Therefore, at the first step, the 'best' model shows which climatic characteristic has the highest correlation with forest basal area in the given ecoregion.

Step 2. Regression with two climatic factors

At the second step, we construct 18 multiple regression models with two climatic factors, one of which is determined by the 'best' model at the previous step: **basal area** $\sim$ **2 climatic characteristics**:

$$\vec{y} = \alpha + \beta^1 \cdot \vec{c^1} + \beta^2 \cdot \vec{c^2} + \vec{\varepsilon}, \quad (2)$$

where $\vec{y} = (y_1, y_2, \ldots, y_n)$ is the basal area vector, c^1 is the climatic variable determined by the 'best' model selected in the 1st step, $\vec{c^1} = (c_1^1, c_2^1, \ldots, c_n^1)$ is the vector of c^1 values, $\vec{c^j} = (c_1^j, c_2^j, \ldots, c_n^j)$ is the vector of values of other climatic characteristics $c^j, j = 2 \ldots 19$, and $\vec{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ is the random error vector. The model parameters are: α—intercept term and β^1, β^2—regression coefficients.

As in the first step, we choose the 'best' model or that having the highest coefficient of determination R^2 among 18 multiple regression models. The coefficient of determination R^2 is equal to the square of the multiple correlation coefficient. This 'best' model defines two climatic variables that we will use at the next regression step.

Step k. Regression with k climatic factors

Theoretically, we can continue adding new climatic variables until all 19 variables are incorporated in the linear regression model. In the presented work, we limited the process to the first five steps $k = 5$, as adding more than 5 climatic variables did not substantially improve the goodness of fit (measured by R^2). This was in agreement with the results of the principal component analysis of our data set (see Section 3.1.2). In general, at the k^{th} step, we construct $19 - k + 1$ multiple regression models, **basal area** $\sim$ k **climatic characteristics**:

$$\vec{y} = \alpha + \beta^1 \cdot \vec{c^1} + \beta^2 \cdot \vec{c^2} + \ldots + \beta^k \cdot \vec{c^k} + \vec{\varepsilon}, \quad (3)$$

where $\vec{y} = (y_1, y_2, \ldots, y_n)$ is the basal area vector, $\vec{c^1} = (c_1^1, c_2^1, \ldots, c_n^1)$ is the the 1st climatic variable vector, $\vec{c^2} = (c_1^2, c_2^2, \ldots, c_n^2)$ is the the 2nd climatic variable vector, ...,

$\vec{c^k} = (c_1^k, c_2^k, \ldots, c_n^k)$ is the the kth climatic variable vector, and $\vec{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$ is the random error vector. The model parameters are: α—intercept term and β^1, β^2, …, β^k—regression parameters.

In all $19 - k + 1$ regression models (3), the first $k - 1$ climatic characteristics, $c^1, c^2, \ldots, c^{k-1}$, were chosen at the previous $k - 1$ steps (as climatic characteristics with the maximal R^2 at their particular steps). At the current k step, we again determined the 'best' model (3) per the highest coefficient of determination R^2 from $19 - k + 1$ models. This process determines the 'best' c^k climatic factor. At the theoretical last step, $k = 19$, we will have only one variable left to include in the model.

As a form of model optimization, the statistical methodology described above allows us to reveal which climatic characteristics explain the majority of basal area variation. Implementing the abovementioned series of regression models, we obtain the ordered (by its influence on basal area) set of climatic factors important for basal area.

2.2.2. Random Forests

We apply the random forests algorithm [36] to determine the climatic factors driving basal area dynamics in the contiguous US. Random forests is a machine learning model based on the concept of regularization through bootstrap aggregation (i.e., bagging) with out-of-bag generalization error estimation. This allows it to average over several decision tree models learned by training on data subsets and obtain a robust ensemble model. The method has been widely used previously in investigating climate impact on vegetation [40–47]. For example, in [41], the authors investigate the effects of climate change on conifer tree species distributions. A similar analysis was performed in [40], where the impact of temperature-related bioclimatic factors on wine regions was investigated.

First, we perform data preparation. We correct for missing points and remove outliers. Next, we apply a low-variance filter. We remove climatic characteristics having low variance as noninformative. We also apply a high-correlation filter. We delete highly correlated climatic factors. For example, looking at the correlation plot for the conterminous USA (Figure A1), we see that some climatic variables are highly correlated. Dropping one of two highly correlated variables, we filter data by reducing the number of climatic parameters.

We run the algorithm for both the contiguous USA and for each ecoregion using a random forests regressor. Using SciKit-Learn, we import the `RandomForestRegressor` class from the `sklearn.ensemble` module to instantiate a regressor object. For each tree, we use a default model fitness criterion of mean squared error (MSE); we will describe this parameter later) in our objective function. We select the number of trees to generate in our random forests using the parameter *n_estimators* in `RandomForestRegressor`. Suppose that we start with three trees: n_estimators = 3.

Denote $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ is the vector of basal area observations at n forest plot locations, and $\mathbf{C}^1, \mathbf{C}^2, \ldots, \mathbf{C}^{19}$ are the vectors of our 19 climatic characteristics (features). We fit the random forest regressor: **basal area** $\sim$ **climatic factors**:

$$\mathbf{y} \sim \mathbf{C}^1, \mathbf{C}^2, ..., \mathbf{C}^{19} \tag{4}$$

For each tree, we take a set of m ($1 \leq m \leq 19$, where m is a hyperparameter) randomly chosen (repetitions are allowed) climatic factors: $(\mathbf{C}^{j_1}, \mathbf{C}^{j_2}, \ldots, \mathbf{C}^{j_m})$ For notational convenience, we use $(\mathbf{C}^1, \mathbf{C}^2, \ldots, \mathbf{C}^m)$.

In order to make our explanations more transparent, suppose $m = 2$. Then, $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, $\mathbf{C}^1 = (C_1^1, C_2^1, \ldots, C_n^1)$ and $\mathbf{C}^2 = (C_1^2, C_2^2, \ldots, C_n^2)$. Suppose R is the set of the points (y_i, C_i^1, C_i^2), $i = \overline{1, n}$. We see that R contains n points, and it is a subset of $\mathbb{R}^{m+1} = \mathbb{R}^3$. R is our bootstrap sample.

Now, we can 'grow' a regression tree. Growing a regression tree within a random forest is equivalent to approximating the dependence $\mathbf{y} = f(\mathbf{C}^1, \mathbf{C}^2)$ by a piecewise continuous function. We use a greedy (i.e., locally optimal) algorithm in order to approximate the relationship $\mathbf{y} = f(\mathbf{C}^1, \mathbf{C}^2)$ (`RandomForestRegressor` uses a greedy algorithm by default).

The **first step** in the greedy algorithm (the first tree node) is the following. In the bootstrap sample R, we choose a feature C (C is either C^1 or C^2, but we simply use notation C for convenience) and a split s of the values of C such that the error is minimized:

$$\min_{\hat{y}} \sum_{i:\, C_i > s} (\hat{y} - y_i)^2 + \min_{\hat{y}} \sum_{i:\, C_i \leq s} (\hat{y} - y_i)^2, \tag{5}$$

where the constant $\hat{y}$ is an estimate of y_i (y_i is i-th coordinate of basal area $\mathbf{y}$). Because we have a finite number of data points, we only need to consider finitely many splits s. At the first step, we split our data using the split s. This way, R is divided into two subsets: R_1 and R_2 (suppose R_1 has more points). By construction, this division produces the most separation between the data points.

At the **second step** (consecutive tree nodes), we perform the same procedure for subset R_1. We look for feature $\widetilde{C}$ and such a split $\widetilde{s}$ of R_1, which together minimize the error (produce the most data separation) defined in our objective function:

$$\min_{\hat{\hat{y}}} \sum_{j:\, \widetilde{C}_j > \widetilde{s}} (\hat{\hat{y}} - \widetilde{y}_j)^2 + \min_{\hat{\hat{y}}} \sum_{j:\, \widetilde{C}_j \leq \widetilde{s}} (\hat{\hat{y}} - \widetilde{y}_j)^2. \tag{6}$$

In practice, minimizing the errors (5) and (6), and the consecutive ones, we build a stepwise function that approximates the basal area $\mathbf{y}$. To build such a stepwise function, we cut R into smaller and smaller subsets (choosing feature C and its split s at every step). We perform cuts until we reach the case when we have only one point (or some small finite number of points) in a subset R_k. This case corresponds to a leaf of the tree (a node with no offspring).

The random forests algorithm naturally splits data into training and testing sets through bagging and out-of-bag error estimation, similar to k-fold cross-validation. However, we explicitly constructed training and testing data sets to get a better notion of the generalization error. We used 5% of observations as the test set with the rest of the data serving as the training set. First, the model runs for a training set. Then, we used the testing set in order to evaluate model generalization (accuracy, overfitting). This gives a stronger overall picture of real-world model performance.

Random forests can optionally estimate the feature importance scores, determined by randomly permuting each feature and calculating the corresponding change to the out-of-bag error. The feature importance score of a climatic variable (feature) can be intuitively understood as a measure of its significance in explaining basal area. Consider a tree in a random forest. In an internal node of the tree, the algorithm chooses the feature that reduces the variance of basal area. In other words, it looks for the feature that decreases the impurity of the split. After we average a feature's importance scores over all of the trees, we obtain the final feature importance score. This is the percent reduction in classification accuracy compared to an out-of-bag error with all variables left intact [36].

3. Results and Discussion

3.1. Stepwise Regression and Multivariate Statistical Analysis

We applied multivariate statistical methods (correlation and principal component analyses), including stepwise regression, to reveal linear relationships between climatic variables and forest basal area. The correlation analysis and principal component analysis (PCA) allowed us to evaluate the intrinsic dimensionality of our data set and provided necessary ground information for stepwise regression modeling.

3.1.1. Correlation Analysis

We computed the Pearson correlation coefficient, r, between basal area and each of the 19 climatic variables for each ecoregion. In all ecoregions, basal area was weakly correlated with climatic variables, $|r| < 0.3$. At the same time, some climatic variables strongly covaried. A correlation matrix computed for the conterminous US is shown in Figure A2 in Appendix B. In most ecoregions, a large number of observations showed correlation coefficients statistically significant with p-values smaller than 0.05. Ecoregion 262 (California Dry Steppe Province) had only five observations; hence, we omitted results for this territory. Regarding Ecoregion 261 (California Dry Steppe Province), the only climatic factors showing significant correlation with basal area are Annual Precipitation (BIO12) and Precipitation of Wettest Month (BIO13).

Weak correlations between basal area and climatic factors are explained by high variation in local topography, disturbance, and land-use history within ecoregions. Natural and anthropogenic disturbances of varying magnitudes create complex patchwork mosaic patterns across forested landscapes [24,26]. This mosaic includes a large number of forest stands (i.e., patches) in different successional stages, leading to variation in forest basal area and biomass values. Forest inventory plots are designed to uniformly sample the landscape, and, therefore, they reflect some aspects of this patch mosaic [24,26,38]. In addition, many ecoregions include several forest and soil types, with different correlations between basal area and climatic factors for each. In the present analysis, we investigate correlations between climatic factors and forest basal area at the ecoregion level without controlling for the patchwork mosaic, forest type, or local topography. That low correlations are often shown between forest basal area and climatic factors is not surprising given the importance of local topography and land-use history, motivating future studies along these lines.

3.1.2. Principal Component Analysis

Principal component analysis (PCA) was performed on the set of 19 climatic factors together with forest basal area. The first five principal components explained most of the variation for all ecoregions (Figure 2), while the first three principal components retained much of this variation. On average, for all ecoregions, three principal components retained around 84% of the variation. On the other hand, climatic factors show low contributions or loadings of the principal components. Maximally contributing to Principal Component 1, climatic factors have on average contribution scores around 6%. In this way, principal axes describe much of the data variation, while each climatic factor does not contribute substantially to the primary principal component. PCA results are summarized in Figure A3.

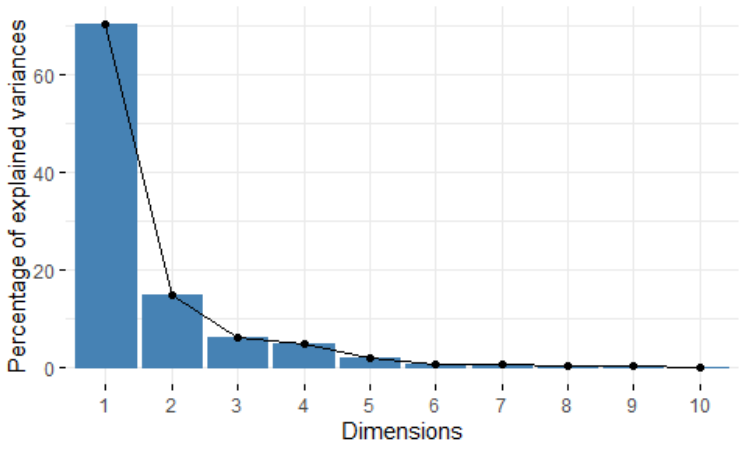

Figure 2. Percentages of explained variance in Ecoregion 332 (Great Plains Steppe Province).

We also performed principal component analysis (PCA) on the set of 19 climatic factors for every ecoregion without a basal area. The results were similar to the PCA with the basal area variable. In this case, principal components retain much of the data variation (about 90% for three main principal axes), while individual climatic variables show only a small contribution to each principal component (about 10%). We interpret this as a reflection of the *gestalt* of the climate system or that the whole is greater than the mean-field approximation of its parts (an analogy to the *gestalt* principle for ecosystems [14,31]).

Our PCA reveals that the system can be considered five-dimensional, while even a three-dimensional model would provide a relatively good approximation. At the same time, none of the climatic variables provide a large single contribution to the principal axes. This is consistent with the results of the correlation analysis, where none of the climatic factors demonstrated a strong correlation with the basal area. Theoretically, based on the PCA, five synthetic variables can be extracted as principal axes and employed in regression modeling, while unexplained variance can be considered as random noise. We considered this research direction; however, we decided to build stepwise regression models in five steps using the original climatic variables as independent variables. While the stepwise regression models using original climatic variables have less predictive power than models based on the synthetic variables determined by PCA, these models allow us to rank climatic factors using multiple correlation coefficients as a selection criterion.

The PCA results raise questions regarding the complexity of the climate–vegetation system. Our set of 19 climatic variables, including seasonal and interannual characteristics (Figure 1), taken as a whole characterizes forest basal area well even in the linear modeling framework (Figure 2). At the same time, none of these variables can explain more than 10% of the variation in any given ecoregion, which are distinct with respect to vegetation and climate. Most of the traditional climate–vegetation theories and models are based on easily measurable climatic factors, most notably average temperature and precipitation. However, the PCA results indicate that none of these quantities are particularly important. Thus, there is a possibility that we often employ climatic variables that are convenient to measure and compute but that do not accurately represent essential climatic characteristics for forested ecosystems.

3.1.3. Stepwise Regression

We implemented the linear stepwise regression analysis for every ecoregion with a five-step depth. In line with the results of the correlation analysis, we observed low R^2 values (less than 15%) in 28 out of 35 ecoregions with a substantial number of observations for statistical analysis. Therefore, in these ecoregions, the stepwise regression did not reveal substantial linear relationships between forest basal area and climatic variables based on five regression steps. This is in the agreement with the PCA results, indicating that the first principal axes are not closely correlated with any particular climatic variables.

Figure A4 summarizes our stepwise regression results for 11 ecoregions, with a significantly high coefficient of determination after five steps ($R^2 > 13\%$): 242, 261, M242, M261, M262, 313, 315, 321, 322, M331, and M341. For Ecoregion 242, we obtain that linear regression **basal area** $\sim$ **Mean Temperature of Warmest Quarter** gives $R^2 = 12.3\%$, and multiple regression **basal area** $\sim$ **BIO10, BIO11, BIO4, BIO6, BIO15** explains 13.1% of basal area variation. Using the division of 19 climatic characteristics into three groups (precipitation, temperature, variability; Figure 1), we visualize the results in Figure 3. The map contains colored ecoregions, where the color depends on the climatic characteristic that maximally contributes to basal area variation in the given ecoregion. The gray areas are the locations where regressions do not capture essential basal area climate dependence (low R^2).

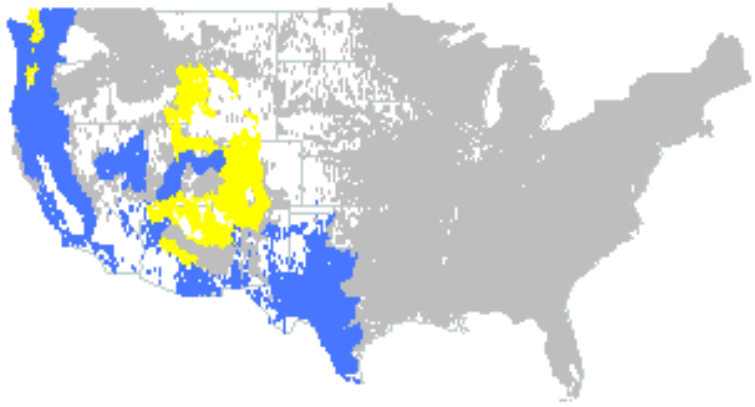

Figure 3. Ecoregions 242, 261, M242, M261, M262, 313, 315, 321, 322, M331, and M341, colored depending on which climatic characteristic maximally contributes to basal area variation in the ecoregion (yellow—temperature-related characteristics and blue—precipitation related). Gray areas are locations where regressions have low R^2 values.

It is noticeable that the majority of these 11 ecoregions are located in desert and arid provinces or in mountainous areas in the Pacific Northwest and Rocky Mountain regions (Figure A4). It is not surprising that we observe the highest R^2 values in Ecoregion 322 (American Semidesert and Desert Province), where three of the five most significant climatic factors are precipitation related: Precipitation of Driest Month (BIO14), Precipitation of Wettest Month (BIO13), and Precipitation of Wettest Quarter (BIO16). These results clearly indicate that precipitation is a crucial factor for this territory.

The first line in Figure A4 contains climatic factors maximally contributing to basal area variation. In Ecoregions 261, M242, M261, M262, 315, 321, 322, and M341, it is precipitation of the extreme (driest, wettest, coldest, or warmest) quarter or month. In Ecoregions 242 and M331, temperature of the warmest period is the most significant factor, while in Ecoregion 313 Annual Mean Temperature (BIO1), most contributes to basal area variation. There is a noticeable overlap between these 11 ecoregions and ecoregions where forest gap dynamics are not a primary driver of stand succession [39].

3.2. Random Forests

As an alternative approach to the linear multivariate and regression analysis, we applied the random forests (RF) algorithm [36] to infer climatic factors linked to forest basal area. The application of RF to the conterminous United States reveals that Precipitation of Coldest Quarter (BIO19) has the highest importance score, while Mean Temperature of Coldest Quarter (BIO11) has secondary importance (see Figure 4 for the top five climatic factors and their importance scores).

Figures A5 and 5 summarize the results of our RF application at the ecoregion level. Figure A5 contains the list of main climatic characteristics together with their feature importance scores for various ecoregions (Figure A5). Figure 5 is the visualization of these results. We see that in Ecoregions 212, 234, 255, 261, 313, 315, 321, M221, M223, M242, M261, M262, M313, M331, and M333, the leading climatic factor is Precipitation of Coldest Quarter (BIO19). This correlates with results that we obtained when running RF for all of the US. There are few ecoregions (221, Dry Steppe 331, and the southern ones 231, 232, 322, and M231) where the most important factor is temperature related. In the other group of ecoregions (222, 223, 242, 263, 342, M211, M332, and M341), the leading factor is variability related (Temperature Seasonality (BIO4) or Temperature Annual Range (BIO7)). We found that in the majority of ecoregions (22 out of 36), precipitation-related climatic factors are most strongly correlated with forest basal area.

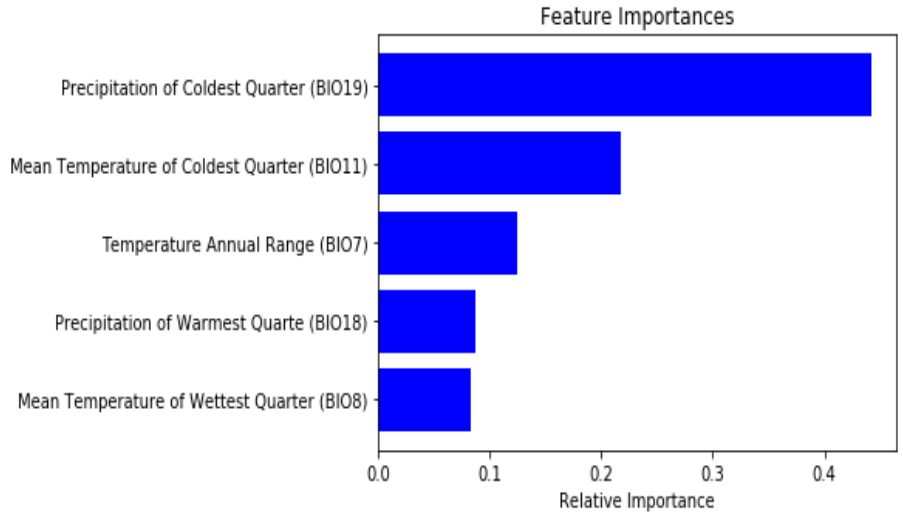

Figure 4. Top five importance scores of climatic characteristics (features) based on random forests for the conterminous USA.

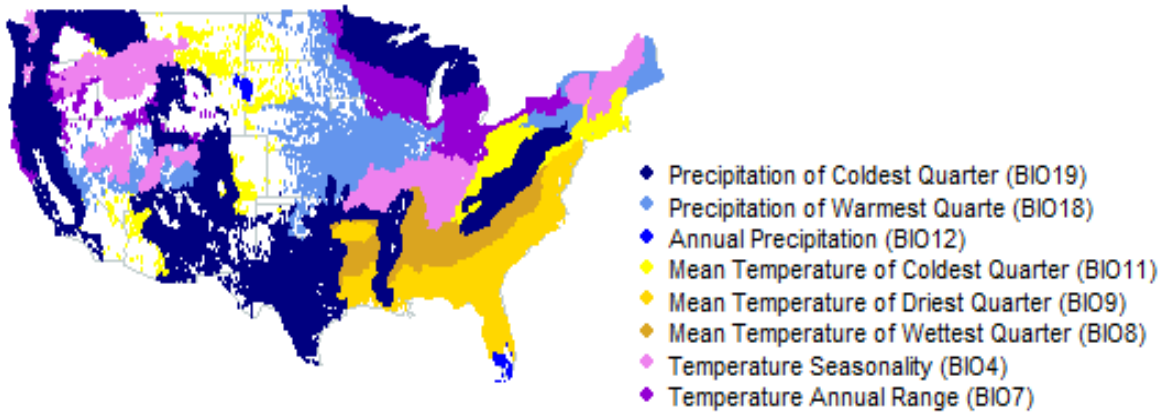

Figure 5. Climatic characteristics having the highest feature importance scores according to the random forests algorithm. Temperature, precipitation, and variability related (yellow, blue, and violet hues, correspondingly).

3.3. Stepwise Linear Regression Versus Random Forests

In this work, we applied to the same general problem two different data-intensive methodologies: stepwise linear regression and random forests. Linear regression is able to capture only linear relationships between variables, while random forests can theoretically capture both linear and nonlinear relationships. The multilinear regression approach gave us nontrivial results only for several USA ecological regions 242, 261, M242, M261, M262, 313, 315, 321, 322, M331, and M341 (Figure A4). Therefore, the intercomparison of these two methods is restricted to these ecoregions. In Ecoregions M242, M261, and 321, both the regression analysis and random forest reveals Precipitation of Coldest Quarter (BIO19) as the leading bioclimatic characteristics (Figures A4 and A5). With respect to Ecoregions 242, 261, M262, and 315, both methods suggest a key climatic factor belonging to the same group (temperature and precipitation groups), while the specific factors were different (Figures A4 and A5). Finally, the random forests algorithm and the regression approach give completely different key climatic characteristics in Ecoregions 313, 322, M331, and M341 (Figures A4 and A5).

From this analysis, we find that both methods give similar results. Random forests appears to be a more advanced and efficient method, yet it essentially remains a 'black-box' approach. Random forests appears to be an effective and powerful dimension reduction tool for large multidimensional data sets (dimension of 19 climatic characteristics in our case). As we see in Figure A5, over all ecoregions, we have only 8 out of 19 climatic factors as the most essential factors. At the same time, the regression approach applied to the climatic characteristics did not substantially reduce dimensionality that can be observed in the comparison of R^2 values as PCA results (Figures A4 and 2). Overall, we conclude that random forests is the preferred method where possible.

3.4. Summary

In the presented research, we investigate how climatic changes may affect forest basal area in the USA. We apply various statistical and machine learning techniques in order to infer which bioclimatic factors mostly influence basal area in US forests. We built a series of multiple regression models based on linear regression. For some ecoregions, the models gave good results and we found the most influential bioclimatic factors for these areas. Precipitation-related factors turn out to be crucial for estimating basal area in the USA. However, there are many ecoregions where this linear analysis is inconclusive.

Principal component analysis gave us an interesting result: in all ecoregions, three main principal components described more than 80% of data variation, while each climatic factor's contribution to the main principal axis is low. We also used an advanced machine learning technique—the random forest algorithm. We generated a map of the US with ecological regions, which shows the climatic factor that most influences basal area. Comparing multiple regression with random forests, we see that the latter is a more suitable tool for data analysis of climate–forest relationships. However, random forests and multiple linear regression give similar results: precipitation-related factors are the most important climatic factors controlling basal area. In particular, Precipitation of Coldest Quarter (BIO19) is shown as a key factor in the majority of US ecoregions.

3.5. Future Research

We anticipate that the application of deep learning algorithmscombined with additional data mining may be productive. In particular, various neural-network-based dimensionality reduction techniques could be used. One could also apply backward feature elimination or forward feature selection for each ecoregion. This may be a good strategy, as we have relatively small data sets within an ecoregion. The other direction would be investigating data for nonlinearities and applying projection-based dimensionality reduction techniques.

4. Conclusions

We examined relationships between climate and forested ecosystems in the contiguous United States using a data-driven paradigm. Data mining of the fForest Inventory and Analysis and WorldClim data sets allowed us to link quantitative measurements of forested ecosystems and climate. We employed two data analysis approaches, multivariate statistics and machine learning to examine the multidimensional structure of the climate–forest complex system and to reveal linkages between variables. PCA reveals that our system can be considered five-dimensional; however, none of the climatic variables are tightly correlated with principal axes. Stepwise linear regression revealed leading climatic characteristics for 11 ecoregions, located in desert and semidesert provinces, the Pacific Northwest, and the southern Rocky Mountains. In these areas, the results obtained with the random forests algorithm and stepwise regression were similar. However, random forests appears to be a more versatile and powerful approach than traditional regression analysis. We anticipate that the application of more advanced deep learning methods with additional data mining may be useful for understanding and predicting forest–climate relationships.

Author Contributions: Conceptualization, N.S.; methodology, N.S., O.R.; software, O.R.; statistical analysis, O.R.; writing—original draft preparation, O.R.; writing—review and editing, N.S.; visualization, O.R.; supervision, N.S.; project administration, N.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by a grant from the Simons Foundation (Grant No. 283770 to N.S.).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: We would like to thank Andrey Sarantsev (University of Nevada) for his advice and help with manuscript editing and Adam Erickson (NASA) for help with the manuscript proofreading.

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

Abbreviations

The following abbreviations are used in this manuscript:

USDA	United States Department of Agriculture
FIA	USDA Forest Service Forest Inventory and Analysis Program
GIS	Geographic Information Systems
PCA	Principal component analysis
RF	Random forests

Appendix A. USA Ecological Subdivisions

Ecological regions in the conterminous United States according to the Bailey's classification:

- 211: Northeastern Mixed Forest Province
- 212: Laurentian Mixed Forest Province
- 221: Eastern Broadleaf Forest Province
- 222: Midwest Broadleaf Forest Province
- 223: Central Interior Broadleaf Forest Province
- 231: Southeastern Mixed Forest Province
- 232: Outer Coastal Plain Mixed Forest Province
- 234: Lower Mississippi Riverine Forest Province
- 242: Pacific Lowland Mixed Forest Province
- 251: Prairie Parkland (Temperate) Province
- 255: Prairie Parkland (Subtropical) Province
- 261: California Coastal Chaparral Forest and Shrub Province
- 262: California Dry Steppe Province
- 263: California Coastal Steppe, Mixed Forest, and Redwood Forest Province
- 313: Colorado Plateau Semidesert Province
- 315: Southwest Plateau and Plains Dry Steppe and Shrub Province
- 321: Chihuahuan Semidesert Province
- 322: American Semidesert and Desert Province
- 331: Great Plains Palouse Dry Steppe Province
- 332: Great Plains Steppe Province
- 341: Intermountain semidesert and Desert Province
- 342: Intermountain semidesert Province
- 411: Everglades Province
- M211: Adirondack New England Mixed Forest and Coniferous Forest, Alpine Meadow Province
- M221: Central Appalachian Broadleaf Forest Coniferous Forest Meadow Province
- M223: Ozark Broadleaf Forest Meadow Province
- M231: Ouachita Mixed Forest Meadow Province

- M242: Cascade Mixed Forest and Coniferous Forest Alpine Meadow Province
- M261: Sierran Steppe Mixed Forest and Coniferous Forest Alpine Meadow Province
- M262: California Coastal Range Open Woodland and Shrub Coniferous Forest Meadow Province
- M313: Arizona-New Mexico Mountains Semidesert and Open Woodland Coniferous Forest Alpine Meadow Province
- M331: Southern Rocky Mountain Steppe and Open Woodland Coniferous Forest Alpine Meadow Province
- M332: Middle Rocky Mountain Steppe and Coniferous Forest Alpine Meadow Province
- M333: Northern Rocky Mountain Forest and Steppe Coniferous Forest Alpine Meadow Province
- M334: Black Hills Coniferous Forest Province
- M341: Nevada-Utah Mountains Semidesert and Coniferous Forest Alpine Meadow Province

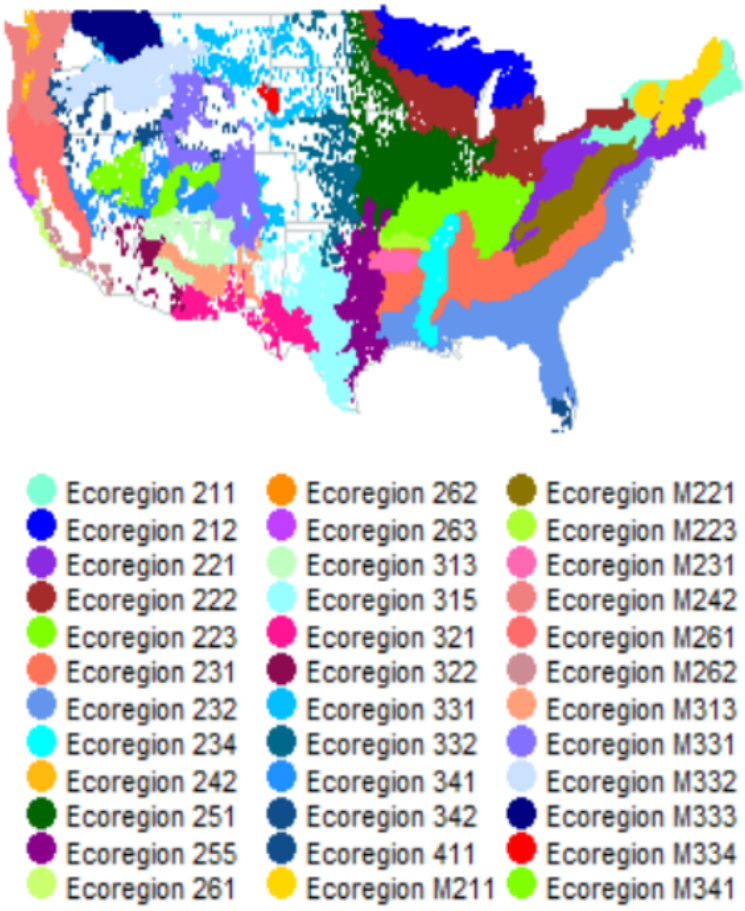

Figure A1. Distribution of FIA forest inventory plots in different ecoregions.

Appendix B. Supplementary Statistical Results

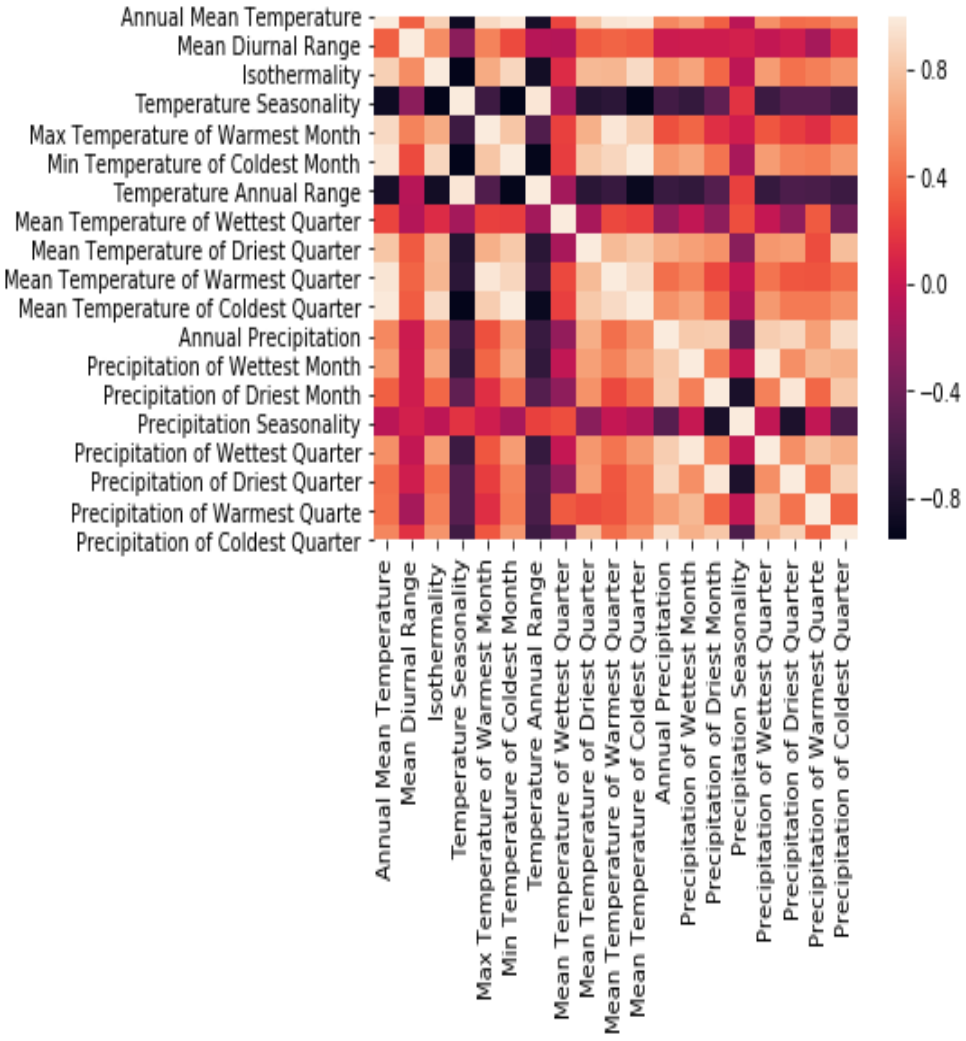

Figure A2. Correlations between various climatic characteristics in the USA.

eco	variation retained by PC1	variation retained by PC1,PC2	variation retained by PC1-PC3	climatic factor closest to PC1/ contribution		climatic factor closest to PC2/ contribution		climatic factor closest to PC3/ contribution	
211	34.66	65.46	76.71	BIO19	12.98	BIO11	13.35	BIO18	28.44
212	61.78	77.05	85.68	BIO15	7.85	BIO5	29.95	BIO13	23.97
221	48.62	72.03	80.65	BIO4	8.78	BIO8	15.42	BIO2	26.21
222	59.07	78.98	87.03	BIO6	8.25	BIO16	21.53	BIO8	43.07
223	53.50	73.06	81.52	BIO12	8.93	BIO5	24.31	BIO8	22.60
231	45.46	67.63	78.87	BIO11	8.70	BIO14	17.41	BIO3	18.42
232	51.91	69.27	83.75	BIO11	9.37	BIO19	26.48	BIO2	15.27
234	59.99	74.79	83.16	BIO4	8.18	BIO5	27.23	BIO15	39.78
242	47.69	79.28	87.08	BIO1	9.25	BIO16	14.07	BIO17	19.47
251	61.05	80.11	87.26	BIO6	7.99	BIO15	16.05	BIO2	34.71
255	51.98	74.80	86.47	BIO11	9.01	BIO12	20.18	BIO2	20.55
261	37.86	65.05	85.82	BIO12	11.76	BIO4	17.04	BIO1	15.74
263	48.97	81.04	88.60	BIO18	9.55	BIO4	14.70	BIO13	12.75
313	37.72	70.84	82.54	BIO10	11.39	BIO4	11.85	BIO15	22.70
315	51.19	76.60	86.39	BIO2	8.57	BIO8	11.52	BIO5	26.33
321	42.11	70.72	85.51	BIO18	10.62	BIO2	14.70	BIO6	16.62
322	49.01	75.99	86.55	BIO14	8.70	BIO15	13.99	BIO7	43.21
331	41.58	66.87	82.58	BIO17	10.00	BIO1	15.19	BIO16	24.63
332	70.28	85.14	91.19	BIO1	6.82	BIO2	26.71	BIO18	29.96
341	40.45	65.84	78.01	BIO12	9.90	BIO4	12.89	BIO2	21.78
342	37.42	65.05	78.74	BIO6	11.51	BIO12	16.21	BIO10	26.35
411	53.64	73.34	84.64	BIO6	8.89	BIO17	24.69	BIO8	23.28
M211	50.05	76.61	84.49	BIO4	8.88	BIO16	14.53	BIO8	24.50
M221	57.04	79.01	86.94	BIO19	7.91	BIO5	19.66	BIO15	22.35
M223	36.59	68.62	78.88	BIO18	10.15	BIO15	11.74	BIO6	27.54
M231	45.30	67.56	78.58	BIO17	9.91	BIO6	20.08	BIO13	26.79
M242	51.29	81.81	87.78	BIO7	8.91	BIO10	12.84	BIO15	36.19
M261	47.25	79.15	86.32	BIO11	10.02	BIO4	11.81	BIO2	33.84
M262	46.82	71.49	81.37	BIO1	9.47	BIO7	15.21	BIO3	23.42
M313	45.55	64.75	78.76	BIO5	9.94	BIO15	18.94	BIO7	18.62
M331	42.51	68.44	80.15	BIO10	10.05	BIO18	15.35	BIO6	22.95
M332	32.70	58.71	75.12	BIO11	14.02	BIO12	15.77	BIO18	18.94
M333	38.04	67.62	79.93	BIO1	12.41	BIO12	16.33	BIO3	36.99
M334	66.42	80.58	88.80	BIO12	7.17	BIO18	16.45	BIO6	16.76
M341	49.02	66.74	76.26	BIO12	9.03	BIO4	23.32	BIO15	42.25

Figure A3. PCA results. Variations retained by principal components and climatic characteristics closest to principal components.

242		261	
Mean Temperature of Warmest Quarter (BIO10)	12.3	Precipitation of Driest Quarter (BIO17)	15.1
Mean Temperature of Coldest Quarter (BIO11)	12.4	Precipitation of Warmest Quarter (BIO18)	17.7
Temperature Seasonality (BIO4)	12.7	Precipitation Seasonality (BIO15)	21.1
Min Temperature of Coldest Month (BIO6)	13	Mean Diurnal Range (BIO2)	22.3
Precipitation Seasonality (BIO15)	13.1	Isothermality (BIO3)	22.5

M242		M261	
Precipitation of Coldest Quarter (BIO19)	8.7	Precipitation of Coldest Quarter (BIO19)	8.7
Mean Temperature of Wettest Quarter (BIO8)	13.9	Mean Temperature of Wettest Quarter (BIO8)	13.9
Precipitation Seasonality (BIO15)	14.7	Precipitation Seasonality (BIO15)	14.7
Min Temperature of Coldest Month (BIO6)	14.9	Min Temperature of Coldest Month (BIO6)	14.9
Isothermality (BIO3)	14.6	Isothermality (BIO3)	14.6

M262		313	
Precipitation of Warmest Quarter (BIO18)	15.6	Annual Mean Temperature (BIO1)	18.3
Mean Temperature of Driest Quarter (BIO9)	16.7	Precipitation of Driest Quarter (BIO17)	20.7
Mean Diurnal Range (BIO2)	16.6	Mean Diurnal Range (BIO2)	22.7
Temperature Seasonality (BIO4)	16.7	Precipitation of Coldest Quarter (BIO19)	23.4
Isothermality (BIO3)	15	Precipitation Seasonality (BIO15)	25.7

315		321	
Precipitation of Driest Month (BIO14)	9.9	Precipitation of Coldest Quarter (BIO19)	22.7
Max Temperature of Warmest Month (BIO5)	12.1	Precipitation of Warmest Quarter (BIO18)	24.8
Precipitation of Warmest Quarter (BIO18)	12.8	Mean Temperature of Driest Quarter (BIO9)	25.4
Annual Precipitation (BIO12)	13.1	Mean Temperature of Coldest Quarter (BIO11)	25.5
Precipitation of Wettest Quarter (BIO16)	13.3	Precipitation of Wettest Month (BIO13)	25.5

322		M331	
Precipitation of Driest Month (BIO14)	36.1	Max Temperature of Warmest Month (BIO5)	25.8
Temperature Seasonality (BIO4)	41	Precipitation of Wettest Month (BIO13)	27.8
Precipitation of Wettest Month (BIO13)	41.3	Min Temperature of Coldest Month (BIO6)	27.8
Precipitation of Wettest Quarter (BIO16)	41.3	Temperature Annual Range (BIO7)	27.9
Mean Temperature of Wettest Quarter (BIO8)	44.8	Mean Temperature of Driest Quarter (BIO9)	27.7

M341	
Precipitation of Wettest Quarter (BIO16)	18.3
Precipitation of Driest Month (BIO14)	19.5
Precipitation of Driest Quarter (BIO17)	21.1
Mean Temperature of Driest Quarter (BIO9)	22.6
Precipitation of Warmest Quarter (BIO18)	22.9

Figure A4. R^2 values (in percentage) of multiple regression models for various ecoregions. Regarding Ecoregion 242, linear regression **basal area** $\sim$ **BIO10** has $R^2 = 12.3\%$, and multiple regression **basal area** $\sim$ **BIO10, BIO11, BIO04, BIO06 and BIO15** has $R^2 = 13.1\%$.

eco	climatic characteristic	feature importance score
211	Precipitation of Warmest Quarter (BIO18)	0.24
212	Precipitation of Coldest Quarter (BIO19)	0.69
221	Mean Temperature of Coldest Quarter (BIO11)	0.27
222	Temperature Annual Range (BIO7)	0.38
223	Temperature Seasonality (BIO4)	0.43
231	Mean Temperature of Wettest Quarter (BIO8)	0.23
232	Mean Temperature of Driest Quarter (BIO9)	0.31
234	Precipitation of Coldest Quarter (BIO19)	0.32
242	Temperature Seasonality (BIO4)	0.4
251	Precipitation of Warmest Quarter (BIO18)	0.35
255	Precipitation of Coldest Quarter (BIO19)	0.33
261	Precipitation of Coldest Quarter (BIO19)	0.77
263	Temperature Annual Range (BIO7)	0.53
313	Precipitation of Coldest Quarter (BIO19)	0.25
315	Precipitation of Coldest Quarter (BIO19)	0.44
321	Precipitation of Coldest Quarter (BIO19)	0.6
322	Mean Temperature of Coldest Quarter (BIO11)	0.5
331	Mean Temperature of Coldest Quarter (BIO11)	0.21
332	Precipitation of Warmest Quarter (BIO18)	0.57
341	Precipitation of Warmest Quarter (BIO18)	0.2
342	Temperature Annual Range (BIO7)	0.31
411	Annual Precipitation (BIO12)	0.51
M211	Temperature Seasonality (BIO4)	0.59
M221	Precipitation of Coldest Quarter (BIO19)	0.48
M223	Precipitation of Coldest Quarter (BIO19)	0.5
M231	Mean Temperature of Driest Quarter (BIO9)	0.39
M242	Precipitation of Coldest Quarter (BIO19)	0.43
M261	Precipitation of Coldest Quarter (BIO19)	0.46
M262	Precipitation of Coldest Quarter (BIO19)	0.35
M313	Precipitation of Coldest Quarter (BIO19)	0.4
M331	Precipitation of Coldest Quarter (BIO19)	0.23
M332	Temperature Seasonality (BIO4)	0.24
M333	Precipitation of Coldest Quarter (BIO19)	0.34
M334	Annual Precipitation (BIO12)	0.57
M341	Temperature Seasonality (BIO4)	0.3

Figure A5. Feature importance scores for various ecoregions. Scores are based on the random forests analysis and vary from 0 to 1.

References

1. Whittaker, R. *Communities and Ecosystems*; Current Concepts in Biology; Macmillan: New York, NY, USA, 1970.
2. Woodward, F. *Climate and Plant Distribution*; Cambridge Studies in Ecology; Cambridge University Press: Cambridge, UK, 1987.
3. Von Humboldt, A.; Bonpland, A. *Essai sur la Géographie des Plantes*; Chez Levrault, Schoell et Compagnie: Paris, France, 1805.
4. Köppen, W. Die Wärmezonen der Erde, nach der Dauer der heissen, gemässigten und kalten Zeit und nach der Wirkung der Wärme auf die organische Welt betrachtet (The thermal zones of the Earth according to the duration of hot, moderate and cold periods and of the impact of heat on the organic world). *Meteorol. Z.* **1884**, *1*, 215–226.
5. Köppen, W. Versuch einer Klassifikation der Klimate, vorzugsweise nach ihren Beziehungen zur Pflanzenwelt. *Geogr. Z.* **1900**, *6*, 593–611.
6. Koppen, W. Klassifikation der Klima nach Temperatur, Niederschlag und Jahreslauf. *Petermanns Geogr. Mitteilungen* **1918**, *64*, 193–203.
7. Kottek, M.; Grieser, J.; Beck, C.; Rudolf, B.; Rubel, F. World map of the Köppen-Geiger climate classification updated. *Meteorol. Z.* **2006**, *15*, 259–263. [CrossRef]
8. Peel, M.C.; Finlayson, B.L.; Mcmahon, T.A. Updated world map of the Köppen-Geiger climate classification. *Hydrol. Earth Syst. Sci. Discuss.* **2007**, *4*, 439–473.
9. Rohli, R.V.; Joyner, T.A.; Reynolds, S.J.; Ballinger, T.J. Overlap of global Köppen–Geiger climates, biomes, and soil orders. *Phys. Geogr.* **2015**, *36*, 158–175. [CrossRef]
10. Rubel, F.; Kottek, M. Observed and projected climate shifts 1901–2100 depicted by world maps of the Köppen-Geiger climate classification. *Meteorol. Z.* **2010**, *19*, 135–141. [CrossRef]
11. Geiger, R.; Pohl, W. Eine neue Wandkarte der Klimagebiete der Erde nach W. Köppens Klassifikation (A New Wall Map of the Climatic Regions of the World According to W. Köppen's Classification). *Erdkunde* **1954**, *8*, 58–61.
12. Trewartha, G.; Horn, L. *An Introduction to Climate*, 5th ed.; McGraw-Hill Book Co.: New York, NY, USA, 1980.
13. Belda, M.; Holtanová, E.; Halenka, T.; Kalvová, J. Climate classification revisited: From Köppen to Trewartha. *Clim. Res.* **2014**, *59*, 1–13. [CrossRef]
14. Bailey, R.G. *Ecosystem Geography: From Ecoregions to Sites*; Springer Science & Business Media: Berlin, Germany, 2009.
15. Holdridge, L.R. Determination of world plant formations from simple climatic data. *Science* **1947**, *105*, 367–368. [CrossRef]
16. Holdridge, L.R. *Life Zone Ecology*; Tropical Science Center: Monteverde, Costa Rica, 1967.
17. Lugo, A.E.; Brown, S.L.; Dodson, R.; Smith, T.S.; Shugart, H.H. The Holdridge life zones of the conterminous United States in relation to ecosystem mapping. *J. Biogeogr.* **1999**, *26*, 1025–1038. [CrossRef]
18. Talluto, M.V.; Boulangeat, I.; Ameztegui, A.; Aubin, I.; Berteaux, D.; Butler, A.; Doyon, F.; Drever, C.R.; Fortin, M.J.; Franceschini, T.; et al. Cross-scale integration of knowledge for predicting species ranges: A metamodelling framework. *Glob. Ecol. Biogeogr.* **2016**, *25*, 238–249. [CrossRef]
19. Liénard, J.; Harrison, J.; Strigul, N. US forest response to projected climate-related stress: A tolerance perspective. *Glob. Chang. Biol.* **2016**, *22*, 2875–2886. [CrossRef] [PubMed]
20. Levin, S.A. Ecosystems and the Biosphere as Complex Adaptive Systems. *Ecosystems* **1998**, *1*, 431–436. [CrossRef]
21. Levin, S.A. Complex adaptive systems: Exploring the known, the unknown and the unknowable. *Am. Math. Soc.* **2003**, *40*, 3–19. [CrossRef]
22. Snyder, C.W.; Mastrandrea, M.D.; Schneider, S.H. The Complex Dynamics of the Climate System: Constraints on our Knowledge, Policy Implications and the Necessity of Systems Thinking. In *Philosophy of Complex Systems*; Handbook of the Philosophy of Science; Hooker, C., Ed.; North-Holland: Amsterdam, The Netherlands, 2011; Volume 10, pp. 467–505.
23. Mihailović, D.T.; Mimić, G.; Arsenić, I. Climate predictions: The chaos and complexity in climate models. *Adv. Meteorol.* **2014**, *2014*, 878249. [CrossRef]
24. Strigul, N.; Florescu, I.; Welden, A.R.; Michalczewski, F. Modelling of forest stand dynamics using Markov chains. *Environ. Model. Softw.* **2012**, *31*, 64–75. [CrossRef]
25. Strigul, N. Individual-based models and scaling methods for ecological forestry: Implications of tree phenotypic plasticity. In *Sustainable Forest Management*; Garcia, J., Casero, J., Eds.; InTech: Rijeka, Croatia, 2012; pp. 359–384. [CrossRef]
26. Lienard, J.F.; Gravel, D.; Strigul, N.S. Data-intensive modeling of forest dynamics. *Environ. Model. Softw.* **2015**, *67*, 138–148. [CrossRef]
27. Easterling, D.R.; Meehl, G.A.; Parmesan, C.; Changnon, S.A.; Karl, T.R.; Mearns, L.O. Climate extremes: Observations, modeling, and impacts. *Science* **2000**, *289*, 2068–2074. [CrossRef]
28. Kelling, S.; Hochachka, W.; Fink, D.; Riedewald, M.; Caruana, R.; Ballard, G.; Hooker, G. Data-intensive Science: A New Paradigm for Biodiversity Studies. *BioScience* **2009**, *59*, 613–620. [CrossRef]
29. Michener, W.K.; Jones, M.B. Ecoinformatics: Supporting ecology as a data-intensive science. *Trends Ecol. Evol.* **2012**, *27*, 85–93. [CrossRef] [PubMed]
30. Hargrove, W.W.; Hoffman, F.M. Potential of multivariate quantitative methods for delineation and visualization of ecoregions. *Environ. Manag.* **2004**, *34*, S39–S60. [CrossRef]
31. Bailey, R.G. Identifying Ecoregion Boundaries. *Environ. Manag.* **2004**, *34*, S14–S26. [CrossRef]
32. Bailey, R.G. *Description of the Ecoregions of the United States*, 2nd ed.; Number 1391; US Department of Agriculture, Forest Service: Washington, DC, USA, 1995.

33. Toledo, M.; Poorter, L.; Peña-Claros, M. Climate is a stronger driver of tree and forest growth rates than soil and disturbance. *J. Ecol.* **2011**, *99*, 254–264. [CrossRef]
34. Zhang, J.; Zhou, Y.; Zhou, G.; Xiao, C. Composition and Structure of Pinus koraiensis Mixed Forest Respond to Spatial Climatic Changes. *PLoS ONE* **2014**, *10*, e0097192. [CrossRef]
35. Khan, D.; Muneer, M.A.; Zaib-Un-Nisa. Effect of Climatic Factors on Stem Biomass and Carbon Stock of Larix gmelinii and Betula platyphylla in Daxing'anling Mountain of Inner Mongolia, China. *Adv. Meteorol.* **2019**, *2019*, 5692574. [CrossRef]
36. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
37. Hijmans, R.J.; Cameron, S.E.; Parra, J.L.; Jones, P.G.; Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol. A J. R. Meteorol. Soc.* **2005**, *25*, 1965–1978. [CrossRef]
38. Liénard, J.F.; Strigul, N.S. Modelling of hardwood forest in Quebec under dynamic disturbance regimes: A time-inhomogeneous Markov chain approach. *J. Ecol.* **2016**, *104*, 806–816. [CrossRef]
39. Liénard, J.; Florescu, I.; Strigul, N. An Appraisal of the Classic Forest Succession Paradigm with the Shade Tolerance Index. *PLoS ONE* **2015**, *10*, e0117138. [CrossRef]
40. Gaal, M.; Moriondo, M. Modelling the impact of climate change on the Hungarian wine regions using Random Forest. *Appl. Ecol. Environ. Res.* **2012**, *10*, 121–140. [CrossRef]
41. Garzón, M.B.; Sánchez de Dios, R. Effects of climate change on the distribution of Iberian tree species. *Appl. Veg. Sci.* **2008**, *11*, 169–178. [CrossRef]
42. Guo, F.T.; Guangyu, W. What drives forest fire in Fujian, China? Evidence from logistic regression and Random Forests. *Int. J. Wildland Fire* **2016**, *25*, 505–519. [CrossRef]
43. Evans, J.S.; Murphy, M.A. Modeling Species Distribution and Change Using Random Forest. In *Predictive Species and Habitat Modeling in Landscape Ecology*; Springer: Berlin, Germany, 2016; pp. 139–159.
44. Iverson, L.; Prasad, A. New machine learning tools for predictive vegetation mapping after climate change: Bagging and Random Forest perform better than Regression Tree Analysis. In *Landscape Ecology of Trees and Forests*; IALE: Manchester, UK, 2004; p. 317.
45. Hashimoto, H.; Wang, W.; Melton, F.S. High-resolution mapping of daily climate variables by aggregating multiple spatial data sets with the random forest algorithm over the conterminous United States. *Int. J. Climatol.* **2019**, *39*, 2964–2983. [CrossRef]
46. Mutanga, O.; Elhadi, A.; Azong Cho, M. High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *Int. J. Appl. Earth Obs. Geoinf.* **2012**, *18*, 399–406. [CrossRef]
47. Wang, L.; Zhou, X.; Zhu, X. Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. *Crop J.* **2016**, *4*, 212–219. [CrossRef]