Article

Reducing the Uncertainty of Radiata Pine Site Index Maps Using an Spatial Ensemble of Machine Learning Models

Gonzalo Gavilán-Acuña[1]*, Guillermo Federico Olmedo [1], Pablo Mena-Quijada [1], Mario Guevara[2], Beatriz Barría-Knopf[1] and Michael S. Watt[3]

[1] Investigaciones Forestales Bioforest S.A., Camino a Coronel, Km. 15, Concepción, 403 0000, Chile
[2] Department of Plant and Soil Sciences, University of Delaware, Newark, DE 19716, USA
[3] Scion, 10 Kyle St, Christchurch 8011, New Zealand

Abstract: Site Index has been widely used as an age normalised metric to account for variation in forest height at a range of spatial scales. Although previous research has used a range of modelling methods to describe regional variation in Site Index little research has examined gains that can be achieved through use of regression kriging or spatial ensemble methods. In this study an extensive set of environmental surfaces were used as covariates to predict Site Index measurements covering the environmental range of *Pinus radiata* D. Don plantations in Chile. Using this dataset, the objectives of this research were to (i) compare predictive precision of a range of geostatistical, parametric and non-parametric models, (ii) determine if significant gains in precision can be attained through use of regression kriging, (iii) evaluate the precision of a spatial ensemble model that utilises predictions from the five most precise models, through using the model prediction with lowest error for a given pixel and (iv) produce a map of Site Index across the study area. The five most precise models were all geostatistical and included ordinary kriging and four regression kriging models that were based on partial least squares or random forests. A spatial ensemble model constructed from these five models was the most precise of those developed (RMSE = 1.851 m, RMSE% = 6.38%) and had relatively little bias. Climatic and edaphic variables were the strongest determinants of Site Index and in particular, variables related to soil water balance were well represented within the most precise predictive models. These results highlight the utility of predicting Site Index using a range of approaches, as these can be used to construct a spatial ensemble that may be more precise than predictions from the constituent models.

Keywords: ensemble of models; site productivity; machine learning; precision silviculture

1. Introduction

Pinus radiata D. Don (radiata pine) is the most widely established plantation species within the Southern Hemisphere, and this species constitutes a large proportion of plantations in Chile, New Zealand and Australia [1]. This species is very responsive to environment and as a consequence productivity has been found to range widely across the environments over which it is grown [2,3]. A number of process-based models, such as 3PG [4], CenW [5] and CABALA [6] have been developed to describe how the environment influences growth of plantation species such as *P. radiata* (e.g. Kirschbaum and Watt [7]). However empirical or hybrid models are still most widely used for predictions of plantation productivity as these models are simpler to parameterise and can provide more precise estimates of growth than process-based approaches [8].

Stand productivity is modelled by empirical models as a function of stand age, using non-linear functional forms. Variation in productivity between stands is accounted for by standardised measurements of productivity at a given age that are used to adjust both the trajectory and the asymptote of predictions of productivity over time. Site Index, which expresses the height of dominant or co-dominant trees at a reference age [9], has been most widely used to account for this inter-stand variation as this metric is correlated with productivity [10,11] and the height of dominant trees is relatively invariant to stand density [12–14].

Environmental surfaces have been widely used through a range of modelling approaches to develop maps of Site Index for *P. radiata* [2,15] and many other coniferous tree species [16–21]. Compared to direct measurements of Site Index made using plot data, which are typically averaged to the stand level, predictions of Site Index from environmental surfaces open up a range of applications that are not available from traditional inventory. The resulting spatial description of Site Index provides insight into the key environmental drivers of productivity and allows managers to understand how productivity is likely to vary across the landscape and where the optimal productivity will occur at range of resolutions from the intra-stand to the regional level [2,15]. In contrast to spatial predictions of Site Index from remotely sensed data, such as LIDAR, [22], surfaces of productivity, created from environmental surfaces can also be used to estimate productivity for unplanted areas, providing managers with insight into the potential value of land that they intend to purchase.

The use of Site Index surfaces to parameterise empirical growth models incorporates elements of process-based modelling as Site Index integrates the most important determinants of tree growth including topography, soil characteristics and climate [23]. Consequently, spatial predictions of Site Index provide a means of generating stand growth curves that are sensitive to fine and coarser scale landscape level changes in climatic and edaphic conditions [24]. These estimates of stand development allow managers to spatially optimise the timing of a range of silvicultural operations including thinning and pruning, across their estate [25,26]. Site Index surfaces can also be used as input to models used for key management decisions such as the optimisation of final crop stand density (S_{opt}) and development of surfaces showing spatial variation in S_{opt} [27].

A large number of modelling methods with varying levels of complexity have been used to predict Site Index for a wide range of forest species growing in Europe, North America and New Zealand. These methods range from relatively simple approaches such as multiple linear regression [3,17–21,28–39] to more complex parametric methods such as Partial Least Squares, Lasso, Elastic Net, Least Angle Regression and Infinitesimal Forward Stagewise Regression [40]. A wide range of non-parametric methodologies have also been used to model Site Index including Random Forests [41,42], Boosted Trees [28,29], Classification and Regression Trees [28,29], Neural Networks [29], Generalised Additive Models [28,29,43], and Multivariate Adaptive Regression Splines [40].

Parametric methods that utilise the spatial correlation between the underlying plot data describing Site Index have been less frequently used to develop models and surfaces of Site Index. Amongst these geostatistical methods the most commonly used techniques are ordinary kriging and regression kriging [2,15]. As predictions are made by ordinary kriging through interpolating values between measured plots this method is most precise when plots are located in relatively close proximity [2]. Regression kriging is less reliant on a dense plot network than ordinary kriging as this method fits an underlying regression model and then geospatially refines these estimates through kriging the model residual variation across the area of interest [2].

The recent emergence of advanced machine learning methods allows greater utilisation of the increasing amount of information in geospatial surfaces as these models can often accommodate collinearity between closely correlated environmental variables. Despite this advantage, few studies have compared predictive precision of these methods with more traditional approaches. For forest species located in Belgium and Turkey, Site Index was more precisely predicted using non-parametric methods than multiple linear regression, and amongst non-parametric methods, artificial neural networks had the highest predictive performance [28]. Site Index of plantation grown *Pinus taeda* in the United States was more precisely predicted using the non-parametric Random Forests than parametric non-linear regression, but it was noted that Random Forests had the most potential for erroneous predictions when extrapolating the model beyond the fitted range [41].

Comparative studies of model performance undertaken in *P. radiata* plantations have highlighted the precision of regression kriging and more advanced non-parametric models, but as with other forest species, have not included a comprehensive comparison of models. Within New Zealand plantations, regression kriging was found to be marginally more precise than ordinary kriging, which in turn was more precise than Partial Least Squares [2,15], with the best Regression Kriging model accounting for 82% of the variance in Site Index [15]. Using a regional New Zealand dataset multiple regression models of Site Index were found to have a slightly superior precision to those created using Random Forests [22]. A comparison of seven modelling methods using data collected from northwest Spain found the non-parametric Multivariate Adaptive Regression Splines (MARS) most precisely predicted Site Index, which was closely followed by the parametric methods of stepwise regression and PLS. The best MARS model accounted for 50% of the variation in the data using 13 predictors [40].

As each modelling method has its own limitations and advantages [44] an alternative approach for improving the overall model precision is to combine predictions from each model [45,46]. This method, which is known as Ensemble Modelling, is a well known methodology that can improve prediction through integrating knowledge from many sources [46]. Although this technique has been used for prediction of many soil attributes [46,47] and class prediction studies [45], we are unaware of any studies that use spatial Ensemble Models for the prediction of Site Index.

P. radiata is the predominant plantation species within Chile and there is considerable interest within the forest sector in the accurate prediction of Site Index for this species [48]. Although different Site Index curves have been developed for each region and geographic area, these local predictions are relatively inaccurate and there is little understanding of how Site Index responds to topography, climatic and edaphic conditions[49]. Given the wide diversity of environmental conditions within the region over which plantations are grown we assumed that more than one modelling method would be required to best predict Site Index across south-central Chile. Consequently, the objectives of this study were to compare the precision of a wide range of modelling algorithms and determine if the combination of multiple algorithms (e.g., by the means of spatial ensemble learning) could more precisely predict Site Index than the best performing single modelling algorithm.

2. Materials and methods

2.1. Data and covariates description

Stand level data describing Site Index of *P. radiata* was extracted from 20 year stands. Site index for *P. radiata* is defined as the mean top height at age 20 years old, where mean top height is defined as the mean height of the 100 largest diameter trees [50]. As stands used in this study were planted between 1987 - 1997 Site Index could be estimated at 20 years of age rather than being projected forward to 20 years as is commonly done [51]. In total, there were 64,190 observations of Site Index available for modelling that were dispersed from Región del Maule (latitude 35°14′) to Región de los Ríos (latitude 40°6′) and covered a Site Index range of 14.2 - 42 m with a mean of 29.0 m. These observations were randomly split, with 75 % used for the fitting dataset, 12.5% for the calibration dataset (used for the ensemble methodology only) and 12.5% for the validation dataset. All three datasets covered a similar geographic range that was representative of the location of *P. radiata* plantations through Chile (Figure 1). Site Index and enviromental conditions were very similar between the three data sets, and are summarized in Table 1.

Table 1. Site variation in climatic and Site Index for the fitting, calibration and validation data sets. Values shown represent the mean, followed in brackets by the range.

Variable	Fitting data set	Calibration data set	Validation data set
Site index (m)	28.9 (14.2-42)	29.3 (17.1-40.4)	29.0 (14.3 -38.9)
Mean annual Temperature (°C)	12.4 (9.7-14.9)	12.4 (10.1 -14.6)	12.5 (10-14.7)
Max temperature of warmest month (°C)	25.1 (17.5-31.7)	25.3 (18.3-31.5)	25.3 (17.7-31.5)
Min temperature of coldest month (°C)	3.8 (0.7-7.9)	3.7 (1.4-7.1)	3.8 (1.3 - 7.9)
Total (annual) precipitation (mm)	1385 (548-2807)	1374 (562-2467)	1364 (591-2569)
Precipitation of wettest month (mm)	278 (133-552)	276 (137-496)	273 (145-545)
Precipitation of driest month (mm)	22.6 (0.9-70.3)	23.2 (0.9-66.6)	22.3 (1.6 - 66.4)

The 64 environmental factors or covariates, listed in Appendix B, were extracted for each of the plot locations at a 90 x 90 m resolution from spatial layers. These spatial layers described topography, vegetation index, soil properties and climate. Topography was characterised from a Digital Elevation Model (DEM) created using LiDAR. Automated Geoscientific Analyses (SAGA) was used to extract the topographical variables listed in Appendix B from this DEM. Enhanced vegetation index (EVI) was used to characterise the vegetation. Values for EVI were derived from MODIS images collected between 1987-2017 period that were reclassified to 90 x 90 m. Values of EVI describing the mean, range and standard deviation were extracted from this imagery (Appendix B). The soil morphology was determined from CIREN Chile. Soil properties were determined by interpolating data from ten thousand soil pits distributed across south-central Chile that belong to Arauco. The soil surfaces available from this dataset included soil depth, clay content, nutrient content (C:N ratio, N content) and physical properties (available soil water, bulk density, hydraulic conductivity). Long term monthly air temperature, rainfall, evapotranspiration and water balance was obtained from CR2 (Center for Climate and resilience research [52]).

In addition to long term averages, mean climatic data that was linked to the time period of each plot was also used within analyses. These covariates were developed for the model prediction using the climate that was experienced by each forest stand (observations), from the plantation establishment until an age of 20 years. Where these variables were significant they were used to construct the models. The map development used raster surfaces of these variables using the average values over the last 30 years for air temperature, water balance, rainfall and evapotranspiration.

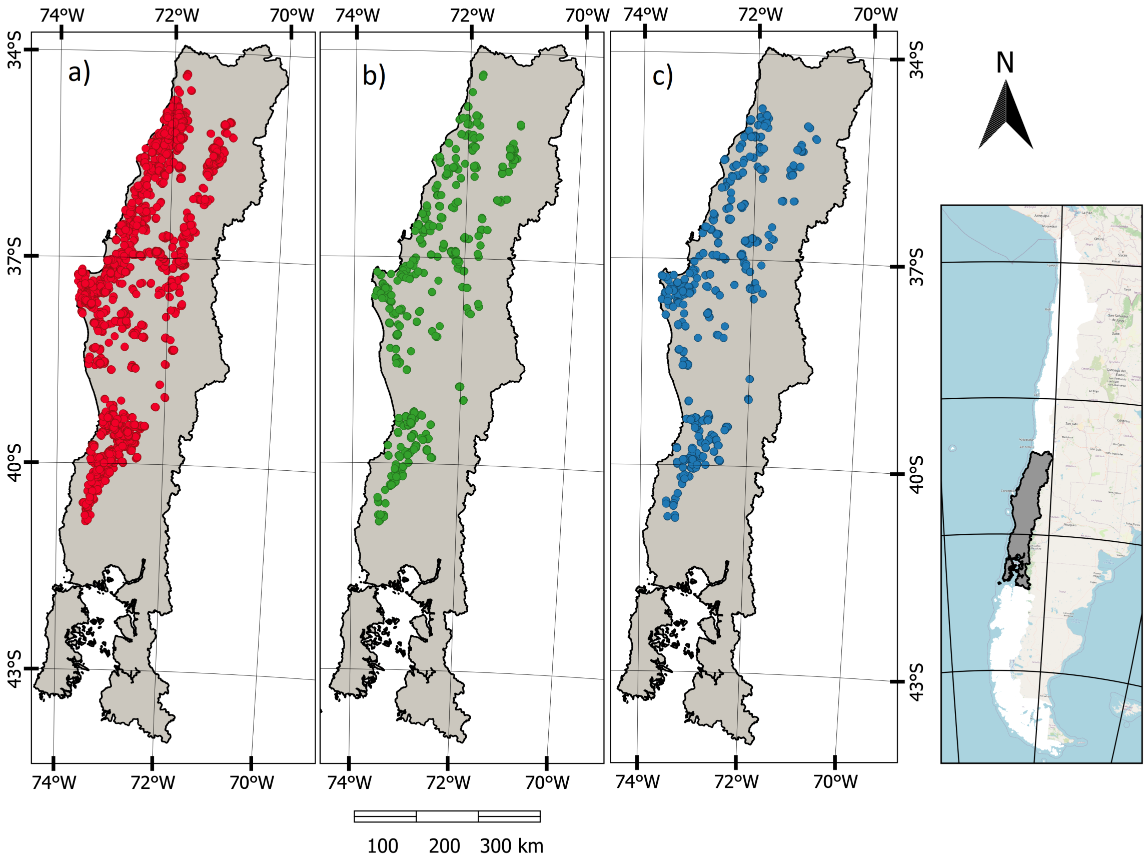

Figure 1. Map of spatial distribution of observations, showing the location of **a** Fitting (n = 48,117), **b** Calibration (n = 8,036) and **c** Validation (n = 8,037) data sets.

2.2. *Subsetting of covariates for the modelling*

A pre-processing method was used to extract a subset of relevant features from the available 64 covariates. According to Weston *et al.* [53], this plays an important role in improving prediction performance and can reduce overfitting. Two types of redundant features selection were made. The first was for non-parametric models (NPM) which included random forest (RF), support vector machines (SVM), neural network (ANN), eXtreme Gradient Boosting (XGBoost) and Multivariate Adaptive Regression Splines (MARS). The second method was used for parametric models (PM) including multiple linear regression (MLR), partial least squares (PLS), and elastic net (EN). Using these methods a total of 18 variables were selected for PM while 20 were selected for NPM. From these selected 'optimal' variables the top 5 were identified. Models of Site Index were developed using PM and NPM methods based on both the optimal selection and the top five variables.

2.2.1. Covariate selection for non-parametric models

Recursive Feature Elimination (RFE) implemented via caret [54] was used to subset variables for the NPM. This method seeks to improve generalization performance through removing the features that have the least effect on training errors [55]. RFE is basically a backward selection of the predictors, which selects features by recursively considering smaller and smaller sub sets of features, and then builds a model using the remaining attributes and calculates model accuracy with an internal cross-validation [56].

During the reduction of input selection with RFE the importance of each variable is measured based on the mean reduction in accuracy (MDA). This process assesses how much the prediction accuracy drops by randomly permuting the values of each input variable (one at a time). A higher importance of the input variable under consideration corresponds to larger reductions in the prediction accuracy [57]. The top five covariates were selected based on RFE importance level.

2.2.2. Covariate selection for parametric models

A penalized regression was used to reduce the number of predictors for PM as this method has been shown to produce more parsimonious models [58]. A least absolute shrinkage and selection operator (LASSO) was used to penalize the parameter estimates to avoid overfitting. LASSO finds the best variables and coefficients by minimizing the residual sum of squares and adding penalties that are useful for fitting a wide variety of linear models [59]. This technique requires the selection of a tuning parameter λ that determines the amount of shrinkage.

This method was followed by a complementary collinearity test using the variance inflation factor (VIF). VIF can be used to identify correlated variables, that can inflate coefficient values, and is determined from the following formulation [60],

$$VIF_i = \frac{1}{1 - R_i^2} \tag{1}$$

Where R2 is the coefficient of determination.

The variables identified as most important by penalized regression were reduced to a set of 18 predictors using a procedure that sequentially eliminated correlated variables with a VIF that exceeded four [60].

After LASSO and VIF were completed a ranking of the variables was obtained based on the strength of the regression between the variable and Site Index, using the Pearson correlation coefficient. This ranking was used to select the top five covariates.

2.3. Modelling approach

2.3.1. Overview

Five types of modelling methods were used to predict Site Index. Each of them used various machine learning and regression algorithm methods, that included: (1) geostatistical models (ordinary kriging); (2) parametric models (multiple linear regression, partial least squares, elastic net); (3) non-parametric models (random forest, support vector machines, neural network, XGBoost and MARS); (4) hybrid models (random forest-kriging and PLS-kriging) and; (5) a modelling ensemble approach using the best five models from the previous four categories. The strategy for fitting the models is outlined below. The parametric, non-parametric models and ordinary kriging were fitted to Site Index using the fitting dataset. Regression kriging was then used to create a range of hybrid models from the fitting dataset, through kriging the residuals of the partial least squares and random forest models. Predictions from all of these models were then made on the calibration dataset. The five most precise models were selected from this process and the difference between actual and predicted Site Index (residuals) for these five models were determined. The residuals for these five most precise models were kriged using OK. For all pixels the model with the lowest residual was selected and this combination of predictions from all five models was termed the model ensemble. The final precision of all of the models that were developed, including the ensemble, was determined through predictions that were made on the validation dataset

2.3.2. Model description and fitting procedure

A brief explanation of all the parametric and non-parametric models is provided in Appendix A. Each PM and NPM machine learning models was fitted using the caret package within R [54], using a five fold cross validation. Hyperparameters for the models were optimized using a grid search that started with a wide search radius and narrowed down to the final values over a series of iterations. A summary of the best tune hyperparameters can be observed in Table 2.

Table 2. Optimal values Hyperparameters used per model.The term"opt" or "5" is added to specify if the model was fitted using the top five or the optimum number of covariates.

Model	Hyperparameter	Best Tune (5)	Best Tune (opt)
PLS	Components (ncomp)	4	12
Elastic net	Mixing Percentage (alpha)	0.02020202	0.03030303
	Regularization Parameter (lambda)	0.0102	0.0102
Random Forest	Ramdomly selected parameter (mtry)	2	11
	Splitting rule	extratrees	extratrees
	Minimum node size	5	5
SVM	Cost	1	0.25
	Lost function	L2	L2
Neural network	Hidden Units (size)	7	7
	Weight decay	0.1	0.1
XGBoost	Boosting Iterations (nrounds)	200	200
	Max Tree Depth	2	2
	Shrinkage (eta)	0.3	0.3
	Minimum Loss Reduction (gamma)	0	0
	Subsample Ratio of Columns (colsample_bytree)	0.8	0.8
	Minimum Sum of Instance	1	1
	Subsample Percentage	0.7	1
MARS	Terms (nprune)	23	34
	Product Degree	2	2

Geostatistical models have been widely used for predicting spatially continuous variables based only on geospatial locations. Ordinary kriging (OK) is one of the most widely used geostatistical methods [2,15,61–63], in which the value for an unsampled point is estimated based on the weighted average of observed neighbouring points within a given area [63]. The neighborhood was restricted to include only the 100 nearest neighbours. Ordinary Kriging was used to predict Site Index from,

$$\hat{Z}(S_0) = \sum_{i=1}^{n} \lambda_i Z(S_i) \tag{2}$$

where $\hat{Z}(S_0)$ is the predicted value of an unvisited location S_0, $Z(S_1), .., Z(S_n)$ are the measured values and their location, and λ_i are the weights that depend on the spatial autocorrelation of the variable as defined by the semi-variogram (see below for description).

Regression kriging is a hybrid modelling technique that combines model prediction of the dependant variables with ordinary kriging of the model residuals [64,65]. This method was used to predict Site Index, at an unsampled site, from,

$$\hat{Z}(S_0) = \hat{m}(S_0) + \hat{\epsilon}(S_0) \tag{3}$$

where the drift $\hat{m}$ refers to the predictions made by the modelling method (described above) and the residuals from these models, $\hat{\epsilon}$, are interpolated using ordinary kriging. In this study the model drift was estimated using random forests and partial least squares.

Semi-variance is used within ordinary kriging and regression kriging to describe spatial autocorrelation of measured values between locations. The plot of semi-variance against distance is a semi-variagram [65]. In this study the semi-variogram was fitted using the autofitVariogram from the R package automap, which tests various models (spherical model, exponential model, gaussian model, Stein's parameterization) and automatically fits the most precise to the data [66].

2.3.3. Model evaluation

Each model was evaluated by comparing the performance of our predicted Site Index with validation data set, following the procedure proposed in Guevara and Olmedo [62]. Using modStats from the openair library from R, the statistics of different models were compared [67]. These statistics included root mean squared error (RMSE), Mean Bias (MB), the Pearson correlation coefficient (r) and Index of Agreement based on Willmontt (IOA), which were defined as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y} - y_1)^2}{n}} \tag{4}$$

$$MB = \frac{1}{n}\sum_{i=1}^{n}(y_1 - \hat{y}) \tag{5}$$

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i-1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i-1}^{n}(y_i - \overline{y})^2}} \tag{6}$$

$$IOA = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(|\hat{y}_i - \overline{y}| + |y_i - \overline{y}|)^2} \tag{7}$$

where $\hat{y}$ is the predicted value in i, $\overline{y}$ is the average of the observed values (analogously for $\overline{x}$), y_i is the observed value (analogously for x), and n is the number of plots.

3. Results

3.1. Covariate selection

Using the covariate selection process described above the number of independant variables used in the modelling was reduced from sixty-four to twenty for NPM and eighteen for PM (Table 3). These were further subsetted to the most precise five variables. The 18 key variables selected for PM were predominantly related to topography, with the remainder evenly distributed across the climate, soil properties and vegetation categories. Eight of the 20 variables selected for NPM were related to climate with the remainder evenly distributed within the other three categories. The top five variables from the selected covariates related mainly to soil properties for PM and included soil hydraulic conductivity, available soil water, C:N ratio, and soil depth. From NPM these top five variables were all related to climate and included growing degree days, rainfall, and two variables that accounted for seasonality in rainfall and air temperature.

Table 3. Selected environmental covariates for parametric models (PM) and non-parametric models (NPM). A more detailed description of each variable is given in Appendix B.

Covariate	Category	PM Ranking Selection	NPM Ranking Selection
AGDD5	Climate		Top 5
ET	Climate	Optimum	Optimum
Rain1	Climate	Optimum	Top 5
Rain2	Climate		Optimum
Rain4	Climate		Top 5
biovar.3	Climate	Optimum	Optimum
biovar.4	Climate	Optimum	Top 5
biovar.15	Climate		Top 5
Total nitrogen content	Soil Properties		Optimum
Soil hydraulic conductivity	Soil Properties	Top 5	Optimum
Clay content	Soil Properties		Optimum
Available soil water	Soil Properties	Top 5	
Carbon to Nitrogen ratio	Soil Properties	Top 5	
Soil Depth	Soil Properties	Top 5	Optimum
Elevation	Topography	Optimum	

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2020 doi:10.20944/preprints202012.0285.v1

Table 3 continued from previous page

Covariate	Category	PM Ranking Selection	NPM Ranking Selectio
Aspect	Topography		Optimum
Chanel Network Base Level	Topography		Optimum
Channel Network Distance	Topography	Top 5	
Slope Length and Steepness Factor	Topography	Optimum	
Tangential Curvature	Topography	Optimum	
Terrain Surface Convexity	Topography	Optimum	Optimum
Profile Curvature	Topography	Optimum	
Valley Depth	Topography	Optimum	Optimum
EVI mean	Vegetation	Optimum	Optimum
EVI min	Vegetation	Optimum	Optimum
EVI range	Vegetation		Optimum
EVI sd	Vegetation	Optimum	Optimum

3.2. Creation of the ensemble model

Predictions made on the calibration data showed that the five most precise models were either geostatistical or hybrid models. Four of these models were developed through regression kriging using PLS and random forests with either the optimum set of covariates or the top five covariates while the fifth model used OK. Within the calibration dataset residuals were extracted from these five models and OK was used to spatially interpolate values throughout the study area. An ensemble model was then created from these five models by selecting from each pixel the model with the lowest residual (Figure 2).

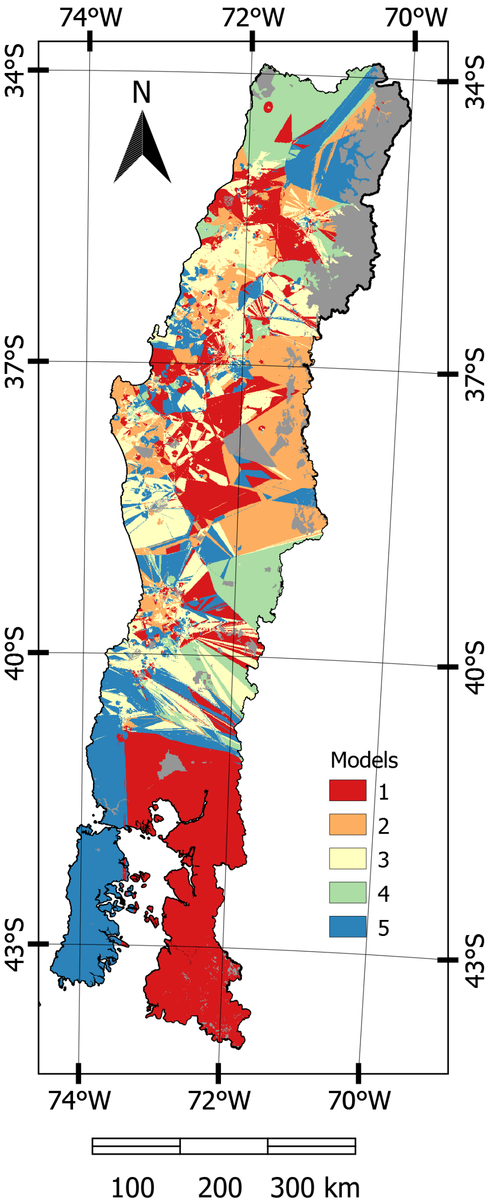

Figure 2. Map showing the allocation of the most precise model per pixel, represented by 1) PLS-kriging (opt); 2) PLS-kriging (5); 3) Ordinary Kriging; 4) Random Forest-Kriging (5); and 5) Random Forest-Kriging (opt).

3.3. Validation of the models

All of the created models were fitted to the validation dataset and the model statistics are displayed for the top 5 models (Table 4) and all models (Appendix C). The model ensemble was the least biased (mean bias = -0.0227 m) with the highest precision (RMSE = 1.8505 m, r = 0.8103 and IOA = 0.7228). The five models that were used to create the ensemble were also the next most precise, with RMSE ranging from 1.8888 – 2.0373 m (Table 4) and mean bias ranging from 0.0309 – 0.5537 m. The number of variables included in the model did not markedly affect model precision as these five models included regression kriged PLS and RF models that had either five variables or the entire set of covariates (18 for PLS and 20 for RF).

A plot of predicted against actual and residual values for the ensemble showed little apparent bias (Figure 3). This bias was smallest for the high density observations and residuals for these observations largely did not exceed 2 m. However the model did slightly overpredict at low values and underpredict at high values of Site Index but this bias was generally constrained to outlying points that occurred at low density (Figure 3).

Table 4. Model validation to different statistical estimators. Mean Bias (MB), Root mean square error (RMSE), Pearson correlation coefficient (r) and Index of Agreement based on Willmontt (IOA). The term "opt" or "5" is added to specify if the model uses the top five or the optimum number of covariates.

Model	MB	RMSE	r	IOA
Model Ensemble	-0.0227	1.8505	0.8103	0.7228
PLS-kriging(opt)	0.2513	1.8888	0.8072	0.7155
Ordinary Kriging	0.1603	1.9332	0.7920	0.7081
PLS-kriging(5)	0.0309	1.9632	0.7893	0.7000
Random Forest-Kriging(5)	0.1718	1.9974	0.7839	0.7025
Random Forest-Kriging(opt)	0.5537	2.0373	0.7853	0.6916

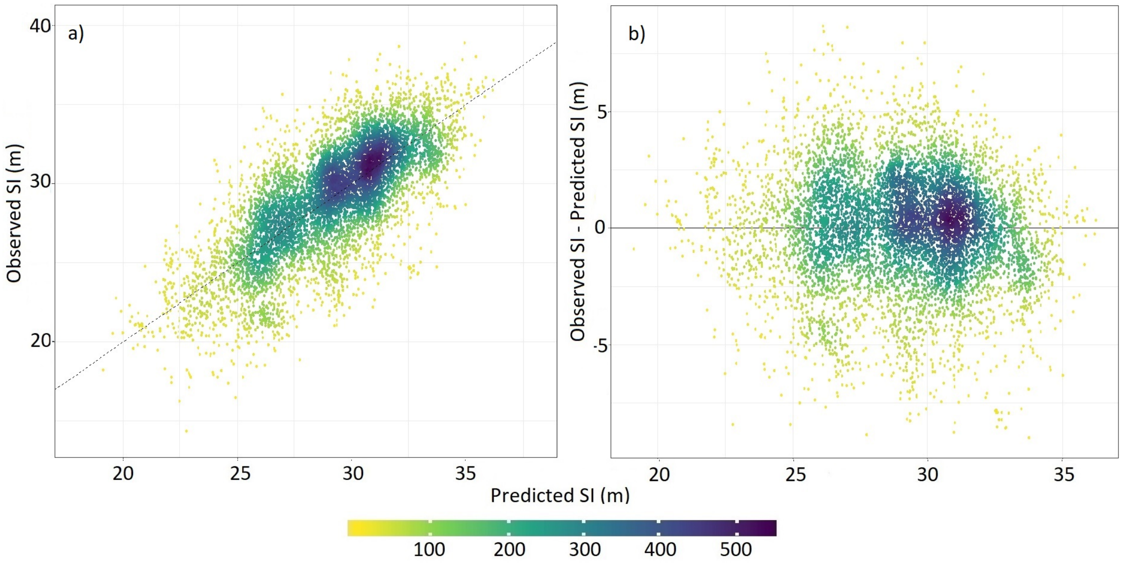

Figure 3. Relationship between Site Index predicted by the ensemble model and **a** observed Site Index and **b** residual Site Index.

3.4. Predictions of the models

The SI prediction of the five most precise model along with the ensemble approach is illustrated in the following Figure 4. All five models that were used to produce the ensemble show the highest values of Site Index occurred at a latitude of ca. 36 – 38°S within coastal and some parts of the inland areas. With the exception of OK the lowest Site Index was predicted to occur in northern regions and in eastern regions close to the Andes. OK did not predict low values in eastern regions as there were few plots in this area but did predict low values in the north where plot density was higher. Predictions from the ensemble largely reflected the values from the five constituent models. This model predicted moderate to high values of Site Index in coastal areas and within the central valley with the lowest predicted values occurring in the eastern and northern regions.

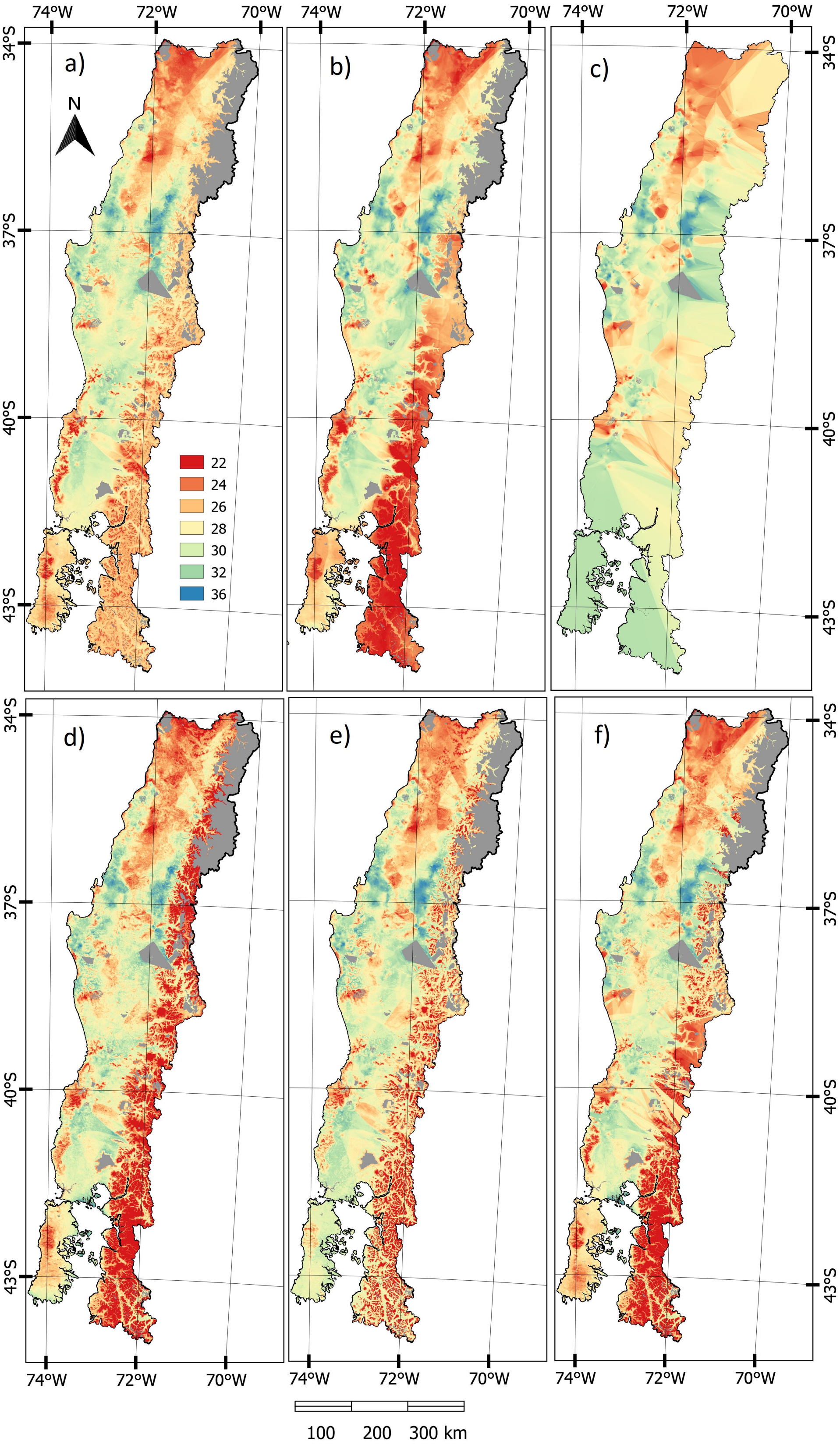

Figure 4. Site Index maps predictions of the five most precise models and the ensemble approach; **a** Random Forest-Kriging(opt); **b** Random Forest-Kriging(5); **c** Ordinary Kriging; **d** PLS-Kriging(opt); **e** PLS-Kriging(5); and **f** Ensemble map

4. Discussion

This study clearly demonstrated the utility of geostatistical methods for predicting Site Index of *P. radiata*. Using a novel spatial ensemble approach the most precise model for each pixel was combined to produce an overall model with a more precise prediction that the constituent models. The final model had an RMSE of 1.88 m which compared favourably with previous predictions for both *P. radiata* and other coniferous species [2,15,28,29]. The variable reduction process highlighted the sensitivity of Site Index to climatic and edaphic factors.

The geostatistical models of ordinary kriging or regression kriging provided the most precise predictions of Site Index among the compared methods. Previous studies have focussed primarily on comparisons of precision between models of Site Index, that do not include a geostatistical component, demonstrating that non-parametric generally outperform parametric models [28,40,41]. Our results extend this research through demonstrating that addition of a spatial component to both of these model types outperform models without this component. Gains through regression kriging over the base model were particularly marked for the most precise model, that utilised PLS (r = 0.807 vs 0.705) demonstrating the utility of this approach. Although regression kriging has been widely used for prediction in other domains [68–70] with few exceptions Kimberley *et al.* [15] very little research has used regression kriging for prediction of productivity indices. As noted by Samuel-Rosa *et al.* [71] there was generally a consistent but small reduction in precision when the number of variables in the models was reduced from between 18-20 to five.

Regression kriging and ordinary kriging have been found to be most precise when applied to high density datasets. The accuracy of OK is most influenced by the spatial point pattern (random, aggregated, or regular), sampling density (high or low), autocorrelation, data distribution (normal and skewed), and heterogeneity of the data [72]. The high density of observations in this study favoured the use of ordinary kriging and regression kriging which is consistent with a model of *P. radiata* Site Index developed in New Zealand [15]. Regression kriging is less sensitive to the spatial distribution of the sample plots than ordinary kriging as this method also includes an underlying model based on environmental variables. As a result regression kriging can outperform OK when datasets include a range of plot densities across the area of interest (e.g. [15]).

A novel ensemble approach was used here to spatially combine predictions from the five most precise prediction models. Although ensemble methods have been widely used in other disciplines [45,47,73,74] this method has not previously been used to predict Site Index. Most ensemble approaches combine predictions from all models across the entire study area using a range of approaches to weight the individual model predictions [73,75]. Our approach differs in that a single model was used to predict Site Index within each pixel which improved the overall predictive precision over any of the five constituent models.

One of the advantages of our ensemble approach, is that this method highlighted regions in which each model performed best, which may be useful if predictions need to be made within a sub-set of the study area. The RF-Kriging method was found to be well suited to northern regions with sparse observations and southern parts of the study area without observations, which highlights the utility of this approach where observation density is low. OK had less error in regions where there were both sparse and denser observations. In high altitude eastern areas where the OK model was not selected predictions from this model are likely be higher than actual values as there were not any points to interpolate within this region. In the central part of the study area, as well as southern parts of the Andes, PLS-Kriging had the lowest prediction error and this region generally included a dense concentration of observations.

The five environmental variables that were the most important determinants of Site Index were all climatic variables for the RF-Kriging model and almost all edaphic variables for the PLS-Kriging model. Growing degree days, accumulated rainfall during years 1 and 4 and variables describing the rainfall and temperature seasonality were the most important climatic determinants of Site Index. Growing degree days has a sound physiological basis as a predictive variable as *P. radiata* height extension is strongly regulated by air temperature [76,77] and consequently this variable controls the length of the growing season. The sensitivity of Site Index to rainfall has been well established [3,7] and our study clearly demonstrates the importance of adequate rainfall during the years immediately after establishment. The seasonality of air temperature and rainfall were also important regulators of Site Index within Chile where both variables exhibit a marked seasonal variance [78].

Important soil properties included C:N ratio, soil hydraulic conductivity, available soil water and soil depth. Soil C:N ratio has been found to be a key determinant of conifer productivity [79] and is a more precise proxy of soil nutrient availability than N as C:N ratio accounts for the positive relationship between carbon content and nitrogen immobilisation [80–84]. Both available soil water and soil depth control the amount of water available to trees which are key attributes within the study area where rainfall is often sparse and highly seasonal [78]. Similarly, soil hydraulic conductivity is also related to water availability and reflects the soils ability to transmit water when subjected to a hydraulic gradient, therefore controlling the partitioning of precipitation between surface runoff and groundwater recharge [85].

Although enhanced vegetation index was not included in the top five variables, variables describing the mean and dispersion of EVI consistently featured among the optimum variables for both types of model. Previous research has found EVI to be a useful predictor of forest canopy structure [86]. There is a strong physiological link between this variable and growth rate as EVI has also been found to be strongly related to leaf area index (LAI) [87,88] and EVI can also be used to identify the start of the growing season [89].

Topographic variables that were well represented for prediction of Site Index in both types of models, included terrain surface convexity (TSC) and valley depth. These covariates influence local microclimate and soil-forming processes, and are consequently associated with the soil type [38]. Both TSC and valley depth are also associated with multiple environmental variables such as water drainage and water availability, accumulation of clay and other soil particles. Valley depth is also likely to be a proxy for local exposure to the wind. As higher windspeeds result in reduced tree height and increased diameter [90–93] *P. radiata* located in deep valleys with little wind exposure has been found to have significantly greater height than trees located on ridges or more exposed areas [94,95].

The direct estimation of Site Index from height data collected at the index age of 20 years reduced error from extrapolation. According to Burkhart and Tomé [96] estimates of Site Index using measurements that coincide with 20 years are rare. As a result most SI studies use equations to extrapolate height to the required index age, which is an approach that is potentially biased [97,98]. More recent methods such as the generalised algebraic difference approach (GADA), which include polymorphic models, provide a more accurate estimation method but still include uncertainties in the final prediction [97]. An additional advantage of using an older dataset was that site specific climatic conditions could be estimated over a uniform period of the rotation length from establishment to 20 years of age.

In conclusion, we found geostatistical models of Site Index to outperform a range of parametric and non-parametric models without a geo-spatial component. These five geostatistical models were successfully combined into an ensemble model that was more precise than the constituent models. Climatic and edaphic variables were most strongly related to Site Index although EVI and many topographic variables were also widely used within the five most precise models. Variables related to soil water balance, such as rainfall, soil depth and water holding capacity were well represented in the top five models reflecting the importance of water limitations in regulating growth across the study area.

Future research could improve the spatial resolution on this product by incorporating high point density lidar measurements as calibration data. And in this way, improving both the spatial resolution and height measurement precision.

Author Contributions: Conceptualization, Gonzalo Gavilán-Acuña and Guillermo Federico Olmedo; Data curation, Pablo Mena-Quijada; Formal analysis, Gonzalo Gavilán-Acuña, Guillermo Federico Olmedo and Michael S. Watt; Funding acquisition, Beatriz Barria-Knopf; Methodology, Gonzalo Gavilán-Acuña, Guillermo Federico

Olmedo, Pablo Mena-Quijada, Mario Guevara and Michael S. Watt; Validation, Michael S. Watt; Visualization, Beatriz Barria-Knopf; Writing – original draft, Gonzalo Gavilán-Acuña, Guillermo Federico Olmedo and Michael S. Watt; Writing – review & editing, Pablo Mena-Quijada, Mario Guevara and Beatriz Barria-Knopf.

Acknowledgments: The authors would like to thank Forestal Arauco, for their role in providing the study site observations, and financial support to this study. Additionally, we want to acknowledge Rodrigo Sobarzo, Fernando Bustamante, Juan Jose Rivera and Eledemar Jaskiu from Forestal Arauco for validating and reviewing the content of this document. Special thanks to Rodrigo Ahumada and Sebastian Fernandez for their support in the publication of this manuscript.

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

5. Appendix A

5.1. Parametric models

5.1.1. Multiple linear regression

Multiple linear regression is a statistical method that predicts the response variable from more than one independent variable. Use of this method assumes that independent variables are not too highly correlated and that residuals from the final model are normally distributed [99].

5.1.2. Partial least squares

Partial least squares (PLS) condenses the most useful information from a large number of predictors to a reduced set of uncorrelated components. These components are then used within a regression to predict the dependant variable [2,100]. The main advantage of this method is that provides lower risk of chance correlation and higher predictive accuracy than multiple regression particularly when there are a large number of correlated predictors in the dataset [101]. The number of components within the model was optimised.

5.1.3. Elastic net

Elastic net (EN) is a form of regularised regression. This method incorporates penalities from both lasso and ridge regression that constrain the coefficient size within the model. EN performs automatic variable selection like Lasso, while the penalization from the Ridge term stabilizes the solution paths which improves the prediction accuracy. The model was fitted using the 'glmnet' method, which was used to optimise the two hyperparameters alpha (mixing percentage) and lambda (regularization parameter).

5.2. Non-parametric models

5.2.1. Random Forest

Random Forest (RF) consists of a combination of many binary decision trees built using several bootstrap samples from a supervised machine learning algorithm, which randomly chooses at each node a subset of explanatory variables [102]. RF belongs to the family of ensemble methods, in where the final prediction comprises the the average predictions from individual trees [103].

Using the caret package [54], a more memory efficient implementation of RF, using the 'ranger' method was applied [104].For this methodology the following parameters were optimised: (i) the number of randomly selected predictors (mtry), (ii) the minimum node size and (iii) the splitting rule.

5.2.2. Support vector machines

Support vector machine (SVM) applies a simple linear technique to the data but in a high-dimensional feature space that is non-linearly related to the input space [105]. The regularized SVM Machine (dual) with Linear Kernel was fitted to the data [54]. The covariates were standardized

doi:10.20944/preprints202012.0285.v1

using the pre-processing of the predictor data ("center, and "scale"). The loss function and cost parameters were optimised within the model.

5.2.3. Neural network

Neural networks or Artificial neural networks (ANN) consist of processing units called neurons or nodes, whose functionality is loosely based on the structure and behaviour of the natural neuron. This technique uses mathematical models that learn nonlinear relationship between the data set, and response variable for both prediction and classification [106]. ANN was fitted using "nnet" method [54] and the hidden units and weight decay parameters were optimised.

5.2.4. XGBoost

The method eXtreme Gradient Boosting (XGBoost) is a scalable implementation of gradient boosting framework developed by Friedman [107]. XGBoost is a supervised machine learning, using decision-tree algorithms. This method builds models from adding individual so called "weak learners" from a gradient descent algorithm over an objective function. XGBoost used the caret package with the "xgbTree" methodology [54]. The hyperparameters that were tuned for this method, included : (i) boosting iterations, (ii) maximum tree depth, (iii) shrinkage, (iv) minimum loss reduction, (v) subsample ratio of columns, (vi) minimum sum of instance weight and (vii) subsample percentage.

5.2.5. Multivariate Adaptive Regression Splines

Multivariate Adaptive Regression Splines (MARS) is a regression technique that captures the nonlinear relationships in the data by assessing cutpoints (knots), which identifies regions where the relationship between the predictor variable and the response changes [108]. The main advantage of MARS is that it enhances the interpretability of complex interactions between the predictor and the response variables. This model was fitted using the "earth" method in the caret package[54], which tuned the number of terms (the maximum number of knots) and the number of variable interactions.

6. Appendix B

Table 5. Environmental raster covariates list

Code	Category	Description
AGDD5	Climate	Growing degrees day base 5 (°C)
AGDD10	Climate	Growing degrees day base 10 (°C)
AGDD15	Climate	Growing degrees day base 15 (°C)
Rain	Climate	accumulative rainfall per stand (mm)
DH	Climate	accumulative water deficit index per stand (mm)
DH4	Climate	accumulative water deficit index per stand first 4 years (mm)
ET	Climate	accumulative evapotranspiration per stand (mm)
Rain1	Climate	accumulative rainfall per stand first year (mm)
Rain2	Climate	accumulative rainfall per stand first 2 years (mm)
Rain4	Climate	accumulative rainfall per stand first 4 years (mm)
ColdD	Climate	accumulative cold days per stand (days)
biovar.1	Climate	bio1 = Mean annual temperature (°C)
biovar.2	Climate	bio2 = Mean diurnal range (mean of max temp - min temp) (°C)
biovar.3	Climate	bio3 = Isothermality (bio2/bio7) (* 100) (%)
biovar.4	Climate	bio4 = Temperature seasonality (standard deviation *100) (%)
biovar.5	Climate	bio5 = Max temperature of warmest month (°C)
biovar.6	Climate	bio6 = Min temperature of coldest month (°C)
biovar.7	Climate	bio7 = Temperature annual range (bio5-bio6) (°C)

Table 5 continued from previous page

Code	Category	Description
biovar.8	Climate	bio8 = Mean temperature of the wettest quarter (°C)
biovar.9	Climate	bio9 = Mean temperature of driest quarter (°C)
biovar.10	Climate	bio10 = Mean temperature of warmest quarter (°C)
biovar.11	Climate	bio11 = Mean temperature of coldest quarter (°C)
biovar.12	Climate	bio12 = Total (annual) precipitation (mm)
biovar.13	Climate	bio13 = Precipitation of wettest month (mm)
biovar.14	Climate	bio14 = Precipitation of driest month (mm)
biovar.15	Climate	bio15 = Precipitation seasonality (coefficient of variation) (%)
biovar.16	Climate	bio16 = Precipitation of wettest quarter (mm)
biovar.17	Climate	bio17 = Precipitation of driest quarter (mm)
biovar.18	Climate	bio18 = Precipitation of warmest quarter (mm)
biovar.19	Climate	bio19= Precipitation of coldest quarter (mm)
Frost	Climate	average number of cold Day (days)
PRM	Soil Morphology	Parent rock material
BLD	Soil Properties	Bulk density
SHC_0_30	Soil Properties	Soil hydraulic conductivity (0-30 cm)
CLAY	Soil Properties	Clay content (%)
ASW	Soil Properties	Available soil water
N_0_60	Soil Properties	Total Nitrogen content from 0 to 60 cm of soil depth
C_N	Soil Properties	Carbon to Nitrogen ratio
SoilDepth	Soil Properties	Soil Depth
Elevation	Topography	LIDAR + SRTM elevation (m.a.s.l)
Aspect	Topography	Aspect Degree (%)
Slope	Topography	Slope Degree (%)
CNBL	Topography	Channel Network Base Level (m.a.s.l)
CND	Topography	Channel Network Distance
LC	Topography	Longitudinal Curvature
CI	Topography	Convergence Index
LS_factor	Topography	Slope Length and Steepness Factor
Max_Curv	Topography	Maximal Curvature
Min_Curv	Topography	Minimal Curvature
Prof_Curv	Topography	Profile Curvature
Tang_Curv	Topography	Tangential Curvature
TSC	Topography	Terrain Surface Convexity
TRI	Topography	Terrain ruggedness index
TPI	Topography	Topographic Position Index
TWI	Topography	Topographic Wetness Index
ValDepth	Topography	Valley Depth
EVI_mean	Vegetation	Average Enhance vegetation index from last 20 years
EVI_min	Vegetation	Minimum Enhance vegetation index from last 20 years
EVI_peak	Vegetation	Maximum Enhance vegetation index from last 20 years
EVI_range	Vegetation	Enhanced vegetation index range from last 20 years
EVI_sd	Vegetation	Standard deviation for EVI from last 20 years
PWU	Water balance	Potential water use
WDI	Water balance	Water deficit index
WS	Water balance	Water surplus

7. Appendix C

Spatial model validation, including the ensemble approach with the validation data set. Elastic net and Support vector machine(SVM) using five best covariates were excluded from the validation table, as in both cases the RMSE exceeded 4 meters.

Table 6. Model validation to different statistical estimators. Mean Bias (MB), Root mean square error (RMSE), Pearson correlation coefficient (r) and Index of Agreement based on Willmontt (IOA). The term "opt" or "5" is added to specify if the model use the top five or the optimum number of covariates.

model	MB	RMSE	r	IOA
Model Ensemble	-0.0228	1.8505	0.8103	0.7229
PLS-kriging(opt)	0.2609	1.8884	0.8074	0.7154
Ordinary Kriging	0.1813	1.9477	0.7888	0.7059
PLS-kriging(5)	0.0368	1.9587	0.7901	0.7009
Random Forest-Kriging(5)	0.1767	1.9927	0.7849	0.7035
Random Forest-Kriging(opt)	0.5590	2.0340	0.7864	0.6921
Random Forest (opt)	0.5334	2.0496	0.7793	0.6844
XGBoost (opt)	0.5996	2.1420	0.7583	0.6714
Random Forest(5)	0.6542	2.1795	0.7521	0.6708
Neural network(opt)	0.2924	2.3034	0.7010	0.6524
XGBoost(5)	0.8710	2.3129	0.7344	0.6354
MARS(opt)	0.6446	2.4419	0.6934	0.6106
PLS(opt)	1.2366	2.5921	0.7052	0.5924
Multiple linear regression(opt)	1.2368	2.5922	0.7052	0.5924
Elastic net(opt)	1.2429	2.6019	0.7030	0.5907
Neural network(5)	1.0331	2.6561	0.6380	0.5924
MARS(5)	1.6199	2.8209	0.6929	0.5499
Support vector machine(opt)	1.1034	3.1136	0.4028	0.5114
Multiple linear regression(5)	1.7278	3.3910	0.4025	0.4514
PLS(5)	1.7298	3.3930	0.4036	0.4513
Support vector machine(5)	3.1245	4.5882	-0.1750	0.2521

References

1. Lewis, N.B.; Ferguson, I.S.; Sutton, W.; Donald, D.; Lisboa, H.; others. *Management of radiata pine.*; Inkata Press Pty Ltd/Butterworth-Heinemann, 1993.
2. Palmer, D.J.; Höck, B.; Kimberley, M.O.; Watt, M.S.; Lowe, D.J.; Payn, T.W. Comparison of spatial prediction techniques for developing Pinus radiata productivity surfaces across New Zealand. *Forest Ecology and Management* **2009**, *258*, 2046–2055.
3. Watt, M.S.; Palmer, D.J.; Kimberley, M.O.; Höck, B.K.; Payn, T.W.; Lowe, D.J. Development of models to predict Pinus radiata productivity throughout New Zealand. *Canadian Journal of Forest Research* **2010**, *40*, 488–499.
4. Landsberg, J.; Waring, R. A generalised model of forest productivity using simplified concepts of radiation-use efficiency, carbon balance and partitioning. *Forest ecology and management* **1997**, *95*, 209–228.
5. Kirschbaum, M.U. CenW, a forest growth model with linked carbon, energy, nutrient and water cycles. *Ecological Modelling* **1999**, *118*, 17–59.
6. Battaglia, M.; Sands, P.; White, D.; Mummery, D. CABALA: a linked carbon, water and nitrogen model of forest growth for silvicultural decision support. *Forest Ecology and Management* **2004**, *193*, 251–282.
7. Kirschbaum, M.U.; Watt, M.S. Use of a process-based model to describe spatial variation in Pinus radiata productivity in New Zealand. *Forest Ecology and Management* **2011**, *262*, 1008–1019.
8. Pinjuv, G.; Mason, E.G.; Watt, M. Quantitative validation and comparison of a range of forest growth model types. *Forest Ecology and Management* **2006**, *236*, 37–46.
9. Skovsgaard, J.P.; Vanclay, J.K. Forest site productivity: A review of the evolution of dendrometric concepts for even-aged stands. *Forestry* **2008**, *81*, 13–31. doi:10.1093/forestry/cpm041.
10. Bontemps, J.D.; Bouriaud, O. Predictive approaches to forest site productivity: recent trends, challenges and future perspectives. *Forestry* **2014**, *87*, 109–128.

11. Eichhorn, F. Beziehungen zwischen bestandshöhe und bestandsmasse. *Allgemeine Forst-und Jagdzeitung* **1904**, *80*, 45–49.
12. Lanner, R.M. On the insensitivity of height growth to spacing. *Forest Ecology and Management* **1985**, *13*, 143–148.
13. Maclaren, J.; Grace, J.; Kimberley, M.; Knowles, R.; West, G. Height growth of Pinus radiata as affected by stocking. *NZJ For. Sci* **1995**, *25*, 73–90.
14. Pienaar, L.V.; Shiver, B.D. The effect of planting density on dominant height in unthinned slash pine plantations. *Forest science* **1984**, *30*, 1059–1066.
15. Kimberley, M.O.; Watt, M.S.; Harrison, D. Characterising prediction error as a function of scale in spatial surfaces of tree productivity. *New Zealand Journal of Forestry Science* **2017**, *47*, 1–11.
16. Fontes, L.; Tomé, M.; Thompson, F.; Yeomans, A.; Luis, J.S.; Savill, P. Modelling the Douglas-fir (Pseudotsuga menziesii (Mirb.) Franco) site index from site factors in Portugal. *Forestry* **2003**, *76*, 491–507.
17. Monserud, R.A.; Huang, S.; Yang, Y. Predicting lodgepole pine site index from climatic parameters in Alberta. *The Forestry Chronicle* **2006**, *82*, 562–571.
18. Palmer, D.J.; Watt, M.S.; Kimberley, M.O.; Dungey, H.S.; others. Predicting the spatial distribution of Sequoia sempervirens productivity in New Zealand. *New Zealand Journal of Forestry Science* **2012**, *42*, 81–89.
19. Seynave, I.; Gégout, J.C.; Hervé, J.C.; Dhôte, J.F.; Drapier, J.; Bruno, É.; Dumé, G. Picea abies site index prediction by environmental factors and understorey vegetation: a two-scale approach based on survey databases. *Canadian Journal of Forest Research* **2005**, *35*, 1669–1678.
20. Wang, G.G.; Huang, S.; Monserud, R.A.; Klos, R.J. Lodgepole pine site index in relation to synoptic measures of climate, soil moisture and soil nutrients. *The Forestry Chronicle* **2004**, *80*, 678–686.
21. Watt, M.S.; Palmer, D.J.; Dungey, H.; Kimberley, M.O. Predicting the spatial distribution of Cupressus lusitanica productivity in New Zealand. *Forest Ecology and Management* **2009**, *258*, 217–223.
22. Watt, M.S.; Dash, J.P.; Bhandari, S.; Watt, P. Comparing parametric and non-parametric methods of predicting Site Index for radiata pine using combinations of data derived from environmental surfaces, satellite imagery and airborne laser scanning. *Forest Ecology and Management* **2015**, *357*, 1–9.
23. Perron, J. Inventaire forestier. *Manuel de foresterie* **1996**, pp. 390–473.
24. McLeod, S.D.; Running, S.W. Comparing site quality indices and productivity in ponderosa pine stands of western Montana. *Canadian Journal of Forest Research* **1988**, *18*, 346–352.
25. Duncker, P.S.; Barreiro, S.M.; Hengeveld, G.M.; Lind, T.; Mason, W.L.; Ambrozy, S.; Spiecker, H. Classification of forest management approaches: a new conceptual framework and its applicability to European forestry. *Ecology and Society* **2012**, *17*.
26. Arano, K.G.; Munn, I.A. Evaluating forest management intensity: a comparison among major forest landowner types. *Forest Policy and Economics* **2006**, *9*, 237–248.
27. Watt, M.S.; Kimberley, M.O.; Dash, J.P.; Harrison, D. Spatial prediction of optimal final stand density for even-aged plantation forests using productivity indices. *Canadian Journal of Forest Research* **2017**, *47*, 527–535.
28. Aertsen, W.; Kint, V.; Van Orshoven, J.; Muys, B. Evaluation of modelling techniques for forest site productivity prediction in contrasting ecoregions using stochastic multicriteria acceptability analysis (SMAA). *Environmental Modelling & Software* **2011**, *26*, 929–937.
29. Aertsen, W.; Kint, V.; Van Orshoven, J.; Özkan, K.; Muys, B. Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests. *Ecological modelling* **2010**, *221*, 1119–1130.
30. Chen, H.Y.; Krestov, P.V.; Klinka, K. Trembling aspen site index in relation to environmental measures of site quality at two spatial scales. *Canadian journal of forest research* **2002**, *32*, 112–119.
31. CODILAN, A.L.; NAKAJIMA, T.; TATSUHARA, S.; SHIRAISHI, N.; others. Estimating site index from ecological factors for industrial tree plantation species in Mindanao, Philippines. *Bull. Univ. of Tokyo For* **2015**, *133*, 19–41.
32. Hamel, B.; Bélanger, N.; Paré, D. Productivity of black spruce and Jack pine stands in Quebec as related to climate, site biological features and soil properties. *Forest Ecology and Management* **2004**, *191*, 239–251.
33. Nigh, G.D.; Ying, C.C.; Qian, H. Climate and productivity of major conifer species in the interior of British Columbia, Canada. *Forest Science* **2004**, *50*, 659–671.

34. Pinno, B.D.; Paré, D.; Guindon, L.; Bélanger, N. Predicting productivity of trembling aspen in the Boreal Shield ecozone of Quebec using different sources of soil and site information. *Forest Ecology and Management* **2009**, *257*, 782–789.

35. Sánchez-Rodrıguez, F.; Rodrıguez-Soalleiro, R.; Español, E.; López, C.; Merino, A. Influence of edaphic factors and tree nutritive status on the productivity of Pinus radiata D. Don plantations in northwestern Spain. *Forest Ecology and Management* **2002**, *171*, 181–189.

36. Seynave, I.; Gégout, J.C.; Hervé, J.C.; Dhôte, J.F. Is the spatial distribution of European beech (Fagus sylvatica L.) limited by its potential height growth? *Journal of Biogeography* **2008**, *35*, 1851–1862.

37. Sharma, R.P.; Brunner, A.; Eid, T. Site index prediction from site and climate variables for Norway spruce and Scots pine in Norway. *Scandinavian Journal of Forest Research* **2012**, *27*, 619–636.

38. Socha, J. Effect of topography and geology on the site index of Picea abies in the West Carpathian, Poland. *Scandinavian Journal of Forest Research* **2008**, *23*, 203–213.

39. Wang, G.G. White spruce site index in relation to soil, understory vegetation, and foliar nutrients. *Canadian Journal of Forest Research* **1995**, *25*, 29–38.

40. González-Rodríguez, M.A.; Diéguez-Aranda, U. Exploring the use of learning techniques for relating the site index of radiata pine stands with climate, soil and physiography. *Forest Ecology and Management* **2020**, *458*, 117803. doi:10.1016/j.foreco.2019.117803.

41. Sabatia, C.O.; Burkhart, H.E. Predicting site index of plantation loblolly pine from biophysical variables. *Forest ecology and management* **2014**, *326*, 142–156.

42. Weiskittel, A.R.; Crookston, N.L.; Radtke, P.J. Linking climate, gross primary productivity, and site index across forests of the western United States. *Canadian Journal of Forest Research* **2011**, *41*, 1710–1721.

43. Shen, C.; Lei, X.; Liu, H.; Wang, L.; Liang, W. Potential impacts of regional climate change on site productivity of Larix olgensis plantations in northeast China. *iForest-Biogeosciences and Forestry* **2015**, *8*, 642.

44. Górecki, T.; Krzyśko, M. Regression methods for combining multiple classifiers. *Communications in Statistics-Simulation and Computation* **2015**, *44*, 739–755.

45. Taghizadeh-Mehrjardi, R.; Minasny, B.; Toomanian, N.; Zeraatpisheh, M.; Amirian-Chakan, A.; Triantafilis, J. Digital Mapping of Soil Classes Using Ensemble of Models in Isfahan Region, Iran. *Soil Systems* **2019**, *3*, 37.

46. Swiderski, B.; Osowski, S.; Kruk, M.; Barhoumi, W. Aggregation of classifiers ensemble using local discriminatory power and quantiles. *Expert Systems with Applications* **2016**, *46*, 316–323.

47. Dobarco, M.R.; Arrouays, D.; Lagacherie, P.; Ciampalini, R.; Saby, N.P. Prediction of topsoil texture for Region Centre (France) applying model ensemble methods. *Geoderma* **2017**, *298*, 67–77.

48. Salas, C.; Donoso, P.J.; Vargas, R.; Arriagada, C.A.; Pedraza, R.; Soto, D.P. The Forest Sector in Chile: An Overview and Current Challenges. *Journal of Forestry* **2016**, *114*, 562–571. doi:10.5849/jof.14-062.

49. García, O. *Indices de sitio para pino insigne en Chile. Instituto Forestal. Serie de Investigacion.*; Number 2 in January 1970, Instituto Forestal: Santiago, Chile, 1970; p. 30.

50. Goulding, C. Measurement of trees. *Forestry handbook* **2005**, *2005*, 145–148.

51. Palmer, D.J.; Kimberley, M.O.; Cown, D.J.; McKinley, R.B. Assessing prediction accuracy in a regression kriging surface of Pinus radiata outerwood density across New Zealand. *Forest Ecology and Management* **2013**, *308*, 9–16.

52. DE VISUALIZACIÓN, P.L.P. GUÍA DE REFERENCIA PARA LA PLATAFORMA DE VISUALIZACIÓN DE SIMULACIONES CLIMÁTICAS.

53. Weston, J.; Elisseeff, A.; Schölkopf, B.; Tipping, M. Use of the zero-norm with linear models and kernel methods. *Journal of machine learning research* **2003**, *3*, 1439–1461.

54. Kuhn, M. The caret package. *R Foundation for Statistical Computing, Vienna, Austria. URL https://cran. r-project. org/package= caret* **2012**.

55. Chen, X.w.; Jeong, J.C. Enhanced recursive feature elimination. Sixth International Conference on Machine Learning and Applications (ICMLA 2007). IEEE, 2007, pp. 429–435.

56. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Machine learning* **2002**, *46*, 389–422.

57. Bahl, A.; Hellack, B.; Balas, M.; Dinischiotu, A.; Wiemann, M.; Brinkmann, J.; Luch, A.; Renard, B.Y.; Haase, A. Recursive feature elimination in random forest classification supports nanomaterial grouping. *NanoImpact* **2019**, *15*, 100179.

58. Bruce, P.; Bruce, A. *Practical statistics for data scientists: 50 essential concepts*; " O'Reilly Media, Inc.", 2017.
59. Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **1996**, *58*, 267–288.
60. Akinwande, M.O.; Dikko, H.G.; Samson, A.; others. Variance inflation factor: as a condition for the inclusion of suppressor variable (s) in regression analysis. *Open Journal of Statistics* **2015**, *5*, 754.
61. Hengl, T.; Heuvelink, G.B.; Rossiter, D.G. About regression-kriging: From equations to case studies. *Computers & geosciences* **2007**, *33*, 1301–1315.
62. Guevara, M.A.; Olmedo, G.F. Model evaluation in digital soil mapping. In *Soil Organic Carbon Mapping Cookbook*, 2 ed.; Yigini, Y.; Olmedo, G.F.; Reiter, S.; Baritz, R.; Viatkin, K.; Vargas, R.R., Eds.; FAO: Rome, Italy, 2018; chapter 8, pp. 133–143.
63. Farmer, W.H. Ordinary kriging as a tool to estimate historical daily streamflow records. *Hydrology and Earth System Sciences* **2016**, *20*, 2721.
64. Odeh, I.O.; McBratney, A.; Chittleborough, D. Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. *Geoderma* **1995**, *67*, 215–226.
65. Hengl, T.; Heuvelink, G.B.; Stein, A. A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma* **2004**, *120*, 75–93.
66. Hiemstra, P.; Hiemstra, M.P. Package 'automap'. *compare* **2013**, *105*, 10.
67. Carslaw, D.C.; Ropkins, K. Openair—an R package for air quality data analysis. *Environmental Modelling & Software* **2012**, *27*, 52–61.
68. Fox, E.W.; Ver Hoef, J.M.; Olsen, A.R. Comparing spatial regression to random forests for large environmental data sets. *PloS one* **2020**, *15*, e0229509.
69. Hengl, T.; Nussbaum, M.; Wright, M.N.; Heuvelink, G.B.; Gräler, B. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. *PeerJ* **2018**, *6*, e5518.
70. Li, J.; Heap, A.D.; Potter, A.; Huang, Z.; Daniell, J.J. Can we improve the spatial predictions of seabed sediments? A case study of spatial interpolation of mud content across the southwest Australian margin. *Continental Shelf Research* **2011**, *31*, 1365–1376.
71. Samuel-Rosa, A.; Heuvelink, G.; Vasques, G.; Anjos, L. Do more detailed environmental covariates deliver more accurate soil maps? *Geoderma* **2015**, *243*, 214–227.
72. Eldeiry, A.A.; Garcia, L.A. Evaluating the performance of ordinary kriging in mapping soil salinity. *Journal of irrigation and drainage engineering* **2012**, *138*, 1046–1059.
73. Diks, C.G.; Vrugt, J.A. Comparison of point forecast accuracy of model averaging methods in hydrologic applications. *Stochastic Environmental Research and Risk Assessment* **2010**, *24*, 809–820.
74. Padarian, J.; Minasny, B.; McBratney, A.; Dalgliesh, N. Predicting and mapping the soil available water capacity of Australian wheatbelt. *Geoderma Regional* **2014**, *2*, 110–118.
75. Nisbet, R.; Elder, J.; Miner, G. *Handbook of statistical analysis and data mining applications*; Academic Press, 2009.
76. Whitehead, D.; Kelliher, F.M.; Frampton, C.M.; Godfrey, M.J. Seasonal development of leaf area in a young, widely spaced Pinus radiata D. Don stand. *Tree Physiology* **1994**, *14*, 1019–1038.
77. Kimberley, M.O.; Richardson, B. Importance of seasonal growth patterns in modelling interactions between radiata pine and some common weed species. *Canadian Journal of Forest Research* **2004**, *34*, 184–194.
78. Watt, M.S.; Trincado, G. Modelling the influence of environment on basic density of the juvenile wood for Pinus radiata grown in Chile. *Forest Ecology and Management* **2019**, *448*, 112–118.
79. Watt, M.S.; Davis, M.R.; Clinton, P.W.; Coker, G.; Ross, C.; Dando, J.; Parfitt, R.L.; Simcock, R. Identification of key soil indicators influencing plantation productivity and sustainability across a national trial series in New Zealand. *Forest Ecology and Management* **2008**, *256*, 180–190.
80. Goodale, C.L.; Aber, J.D. The long-term effects of land-use history on nitrogen cycling in northern hardwood forests. *Ecological Applications* **2001**, *11*, 253–267.
81. Andersson, P.; Berggren, D.; Nilsson, I. Indices for nitrogen status and nitrate leaching from Norway spruce (Picea abies (L.) Karst.) stands in Sweden. *Forest Ecology and Management* **2002**, *157*, 39–53.
82. Ross, D.S.; Lawrence, G.B.; Fredriksen, G. Mineralization and nitrification patterns at eight northeastern USA forested research sites. *Forest Ecology and Management* **2004**, *188*, 317–335.
83. Parfitt, R.; Ross, D.; Coomes, D.; Richardson, S.; Smale, M.; Dahlgren, R. N and P in New Zealand soil chronosequences and relationships with foliar N and P. *Biogeochemistry* **2005**, *75*, 305–328.

84. Parfitt, R.; Yeates, G.; Ross, D.; Mackay, A.; Budding, P. Relationships between soil biota, nitrogen and phosphorus availability, and pasture growth under organic and conventional management. *Applied soil ecology* **2005**, *28*, 1–13.

85. Jarvis, N.; Koestel, J.; Messing, I.; Moeys, J.; Lindahl, A. Influence of soil, land use and climatic factors on the hydraulic conductivity of soil. *Hydrology & Earth System Sciences Discussions* **2013**, *10*.

86. Chen, J.; Sun, L. Using MODIS EVI to detect vegetation damage caused by the 2008 ice and snow storms in south China. *Journal of Geophysical Research: Biogeosciences* **2010**, *115*.

87. Alexandridis, T.K.; Ovakoglou, G.; Clevers, J.G. Relationship between MODIS EVI and LAI across time and space. *Geocarto International* **2019**, pp. 1–15.

88. Potithep, S.; Nagai, S.; Nasahara, K.N.; Muraoka, H.; Suzuki, R. Two separate periods of the LAI–VIs relationships using in situ measurements in a deciduous broadleaf forest. *Agricultural and forest meteorology* **2013**, *169*, 148–155.

89. Karkauskaite, P.; Tagesson, T.; Fensholt, R. Evaluation of the plant phenology index (PPI), NDVI and EVI for start-of-season trend analysis of the Northern Hemisphere boreal zone. *Remote Sensing* **2017**, *9*, 485.

90. Jacobs, M. The effect of wind sway on the form and development of Pinus radiata D. Don. *Australian Journal of Botany* **1954**, *2*, 35–51.

91. Telewski, F.W.; Jaffe, M.J. Thigmomorphogenesis: anatomical, morphological and mechanical analysis of genetically different sibs of Pinus taeda in response to mechanical perturbation. *Physiologia Plantarum* **1986**, *66*, 219–226.

92. Telewski, F.W.; Jaffe, M.J. Thigmomorphogenesis: field and laboratory studies of Abies fraseri in response to wind or mechanical perturbation. *Physiologia Plantarum* **1986**, *66*, 211–218.

93. Telewski, F. Structure and function of flexure wood in Abies fraseri. *Tree Physiology* **1989**, *5*, 113–121.

94. Watt, M.; Moore, J.; McKinlay, B. The influence of wind on branch characteristics of Pinus radiata. *Trees* **2005**, *19*, 58–65.

95. Watt, M.S.; Kirschbaum, M.U. Moving beyond simple linear allometric relationships between tree height and diameter. *Ecological Modelling* **2011**, *222*, 3910–3916.

96. Burkhart, H.E.; Tomé, M. *Modeling forest trees and stands*; Springer Science & Business Media, 2012.

97. Socha, J.; Pierzchalski, M.; Bałazy, R.; Ciesielski, M. Modelling top height growth and site index using repeated laser scanning data. *Forest Ecology and Management* **2017**, *406*, 307–317. doi:10.1016/j.foreco.2017.09.039.

98. Cieszewski, C.J.; Harrison, M.; Martin, S.W. Examples of practical methods for unbiased parameter estimation in self-referencing functions. Proceedings of the First International Conference on Measurements and Quantitative Methods and Management, 1999, pp. 17–18.

99. Seal, H.L. *The historical development of the Gauss linear model*; Yale University, 1968.

100. Wold, S.; Geladi, P.; Esbensen, K.; Öhman, J. Multi-way principal components-and PLS-analysis. *Journal of chemometrics* **1987**, *1*, 41–56.

101. Cramer, R.D. Partial least squares (PLS): its strengths and limitations. *Perspectives in Drug Discovery and Design* **1993**, *1*, 269–278.

102. Breiman, L.; Friedman, J.; Stone, C.J.; Olshen, R.A. *Classification and regression trees*; CRC press, 1984.

103. Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5–32.

104. Wright, M.N.; Ziegler, A. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *arXiv preprint arXiv:1508.04409* **2015**.

105. Karatzoglou, A.; Meyer, D.; Hornik, K. Support vector machines in R. *Journal of statistical software* **2006**, *15*, 1–28.

106. Haykin, S. *Neural networks: a comprehensive foundation, Second Edition*; Prentice Hall PTR, 1998.

107. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y. Xgboost: extreme gradient boosting. *R package version 0.4-2* **2015**, pp. 1–4.

108. Fridedman, J. Multivariate adaptive regression splines (with discussion). *Ann. Statist* **1991**, *19*, 79–141.