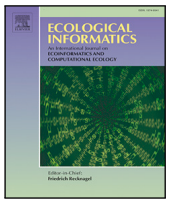
Ecological Informatics 92 (2025) 103500

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Comparing machine learning algorithms for simultaneous prediction of tree diameter distribution percentiles

Albert Ciceu [a,*], Hasan Aksoy [b,c], Ovidiu Badea [d,e], Bronson P. Bullock [f], Jacinta Ukamaka Ezenwenyi [g], Jose Javier Gorgoso-Varela [h], Ştefan Leca [d], Thomas Ledermann [a], Harri Mäkinen [i], Friday N. Ogana [j], Sheng-I Yang [f], Lauri Mehtätalo [c]

[a] Austrian Research Center for Forests (BFW), Seckendorff-Gudent-Weg 8, 1130 Vienna, Austria
[b] Sinop University, Vocational School of Ayancik, Department of Forestry, Program of Forestry and Forest Products, Sinop, Turkiye
[c] Natural Resources Institute Finland (Luke), Yliopistokatu 6, 80100, Joensuu, Finland
[d] National Institute for Research and Development in Forestry "Marin Drăcea" - INCDS, Blvd. Eroilor, 128, Voluntari 077190, Ilfov, Romania
[e] Transilvania University, Faculty of Silviculture and Forest Engineering, 1, Ludwig van Beethoven Street, Braşov 500123, Romania
[f] Warnell School of Forestry and Natural Resources, University of Georgia, 180 E Green Street, Athens, GA 30602, United States of America
[g] Department of Forestry and Wildlife, Nnamdi Azikiwe University, Awka, Nigeria
[h] Campus de Lugo, University of Santiago de Compostela, 27002 Lugo, Spain
[i] Natural Resources Institute Finland (Luke), Latokartanonkaari 9, 00790 Helsinki, Finland
[j] Virginia Polytechnic Institute and State University, Department of Forest Resources and Environmental Conservation, 310 West Campus Dr., 24061 Blacksburg, United States of America

ARTICLE INFO

Dataset link: https://github.com/AlbertCiceu/Multi-Output-DL-RF-for-Diameter-Percentile-Prediction.git

Keywords:
Multi-output regression
Machine learning
Model comparison
Tree diameter distribution
Prediction
Percentile prediction

ABSTRACT

Accurate predictions of tree diameter distributions are important for assessing forest structure, quantifying biodiversity, and estimating carbon sequestration. Percentile-based approaches are among the most effective methods for reconstructing diameter distributions from stand-level variables. In this study, we compared three modelling approaches, generalised least squares (GLS), Multi-Output Random Forest (MORF), and a multi-output deep learning-based model (MODL), across nine datasets representing different forest types and management regimes, aiming to predict simultaneously six diameter distribution percentiles. Our results show that MODL consistently outperformed both GLS and MORF in predictive accuracy across all nine training subsets and five out of nine test subsets, demonstrating strong generalisation across diverse forest types. MODL was particularly effective in achieving high accuracy while preserving the standard deviation of the response variables. While GLS performed slightly better in predicting the 100th percentile, MODL showed superior performance at the lower percentiles in most datasets. Interestingly, although MORF was generally the least accurate, it was the only method that consistently maintained the monotonicity of the predicted percentiles, a desirable property not inherently ensured by GLS or MODL, especially in the case of narrow diameter distributions. These findings underscore the strong potential of deep learning models for predicting diameter distribution percentiles and position MODL as a promising alternative to traditional parametric approaches.

Abbreviations: GLS, generalised least squares; MORF, Multi-Output Random Forest; MODL, multi-output regression; ML, machine learning; DD, diameter distribution; DBH, diameter at breast height; k-NN, k-nearest neighbours; SUR, Seemingly Unrelated Regression; OLS, ordinary least squares; RF, random Forests; BRT, Boosted Regression Trees; SVM, Support Vector Machines; ANN, artificial neural networks; RO-NF-MX, natural temperate mixed forest stands in Romania; CU, cluster units; SP, sampling subplots; NG-NF-MX, natural tropical mixed rain forest in Nigeria; ES-PF-PR, plantation forests of Monterey pine in Spain; ES-PF-EG, plantation forests of Tasmanian blue gum in Spain; ES-PF-EG, permanent sampling plots; US-DE-PT, density experiments of Loblolly pine in the United States; CPCD, Coastal Plain Culture x Density study; AT-TE-PA, Norway spruce thinning experiments in Austria; FI-TE-PS, Scots pine thinning experiments in Finland; FI-TE-PA, Norway spruce thinning experiments in Finland; TR-NF-PS, natural Scots pine forests in Türkiye; P, percentiles; Z, thinning status; T, stand age; D_g, quadratic mean diameter; *DDOM*, dominant diameter; N, number of trees per hectare; DL, deep learning; MSE, mean squared error; RMSE, root mean squared error; MAE, mean absolute error; ME, mean error; cRMSE, centred RMSE

* Corresponding author.
E-mail addresses: albert.ciceu@bfw.gv.at (A. Ciceu), haksoy@sinop.edu.tr, ext.hasan.aksoy@luke.fi (H. Aksoy), obadea@icas.ro (O. Badea), BronsonBullock@uga.edu (B.P. Bullock), uj.ezenwenyi@unizik.edu.ng (J.U. Ezenwenyi), josejavier.gorgoso@usc.es (J.J. Gorgoso-Varela), stefan.leca@icas.ro (Ş. Leca), thomas.ledermann@bfw.gv.at (T. Ledermann), harri.makinen@luke.fi (H. Mäkinen), fnogana23@vt.edu (F.N. Ogana), syang23@uga.edu (S.-I. Yang), lauri.mehtatalo@luke.fi (L. Mehtätalo).

https://doi.org/10.1016/j.ecoinf.2025.103500
Received 7 July 2025; Received in revised form 28 October 2025; Accepted 28 October 2025
Available online 31 October 2025

1. Introduction

Diameter distributions (DD) are widely used in forestry to summarise and describe data from inventoried areas. Once the trees within a plot have been recorded, including their species and dimensions, most commonly diameter at breast height (DBH), the data are typically visualised using a histogram. The histogram groups trees into fixed-diameter classes, or "bins", with each bar showing the number of trees within a particular size range (Garcıa, 1992). This provides a discrete frequency distribution that has been used to describe forest stands in terms of development stage (Feldmann et al., 2018; Chivulescu et al., 2020), age structure (even-aged or uneven-aged) (Pach and Podlaski, 2015), degree of naturalness (Alessandrini et al., 2011; Myllymäki et al., 2024), stand structure in managed and unmanaged forests (Janowiak et al., 2008), and stand structure relationship with biodiversity and productivity (Bohn and Huth, 2017).

In forestry, DD are commonly used to estimate the volume or biomass of a forest stand, providing essential information to evaluate the production of various wood products (Knowe et al., 1997; Yen et al., 2010; Yang et al., 2022). In addition to supporting volume and yield evaluations, DD are used to generate synthetic tree lists based on stand-level predictions, which serve as input for simulations in individual tree growth models (Weiskittel et al., 2011).

Since empirical DD are derived from tree lists obtained through forest inventories, they are time-consuming and costly to construct, as their accuracy depends on the number of trees measured. A practical alternative is to estimate DD from easily measured stand variables, typically by fitting theoretical distributions to empirical data. Two common approaches are used to reconstruct a DD: the parameter prediction and parameter recovery methods (Hyink and Moser, 1983). Parameter prediction estimates distribution parameters from models fitted with one or more stand-level predictors (Cao, 2004), whereas parameter recovery derives parameters directly from known stand characteristics, such as moments (e.g., mean, quadratic mean), percentiles (e.g., median), or other derived metrics (Bankston et al., 2021). For general relationships between distribution parameters and stand characteristics, see Siipilehto and Mehtätalo (2013) for the two-parameter Weibull function and Mehtätalo (2004) for percentile-based distributions.

Various theoretical distributions have been used to describe tree DD, with the Weibull distribution (Bailey and Dell, 1973) and its variants (Ciceu et al., 2021; Ogana, 2022) among the most widely applied. The Weibull distribution is widely regarded as one of the most versatile probability functions for modelling forest stand structures: its two- or three-parameter forms can assume a variety of shapes, accommodating diverse stand patterns (Merganič and Sterba, 2006; Palahí et al., 2007), although more flexible four-parameter models also exist (Wang and Rennolls, 2005). Despite this flexibility, the Weibull and other theoretical distributions often struggle to capture complex, multilayered stands (Maltamo et al., 2000; Kangas and Maltamo, 2000). To address this, multi-component approaches have been proposed, such as combining several Weibull functions (Zhang et al., 2001) or using Gaussian mixture models (Horodnic and Roibu, 2018).

As an alternative, nonparametric methods have been developed to overcome these limitations. Unlike parametric approaches, nonparametric methods do not rely on predefined mathematical functions and can adapt to more irregular or diverse distribution shapes. Examples of such methods include the k-nearest neighbours (k-NN) approach (Haara et al., 1997) and percentile-based techniques (Borders et al., 1987).

Percentile-based DD have been shown to produce more accurate estimates than theoretical distributions such as the Weibull (Maltamo et al., 2000; Kangas and Maltamo, 2000; Pogoda et al., 2019), largely because they reconstruct the empirical distribution based on a set of percentiles, typically providing more reference points than the limited number of parameters used by theoretical distributions.

The method presented by Borders et al. (1987) involves predicting 13 percentiles of the DD using stand-level predictors. This approach is implemented in two steps: first, a predefined set of percentiles is computed from inventory data; second, these percentiles are modelled as functions of stand attributes.

An alternative approach is to bypass the first step entirely and fit the percentile models directly to the observed diameter data using quantile regression (Mehtätalo et al., 2008). However, in this study, we focus on the original two-step approach proposed by Borders et al. (1987).

To estimate the model parameters, Borders et al. (1987) employed Seemingly Unrelated Regression (SUR) (Zellner, 1962), a method also adopted in subsequent studies using this framework (Kangas and Maltamo, 2000; Stankova and Zlatanov, 2010; Pogoda et al., 2019). In SUR, a system of equations is fitted jointly to account for the correlations among the error terms of the individual equations, as the residuals are correlated due to the interrelated nature of the percentiles being predicted.

The fitting process begins by estimating the coefficients separately for each equation using ordinary least squares (OLS). Then, the correlations among the residuals are estimated and incorporated into a second step, where a final, more efficient set of coefficients is computed (Mehtätalo and Lappi, 2020) using GLS. Although GLS provides unbiased estimates for each equation individually, modelling efficiency can be improved by accounting for the correlations between the errors, particularly when the independent variables differ across equations. By leveraging the correlations among equations, SUR enables information sharing across models, resulting in improved estimation efficiency. However, if the same independent variables are used in all equations, the regression coefficient estimates obtained through SUR will be identical to those obtained by independently fitting each model using GLS.

Machine learning algorithms have introduced flexible, data-driven alternatives to traditional parametric approaches (Breiman, 2001b). Models such as Random Forests (RFs), Boosted Regression Trees (BRTs), Support Vector Machines (SVMs), and artificial neural networks (ANNs) are particularly effective at capturing complex, nonlinear relationships between predictors and response variables, and are increasingly applied across various scientific fields (Pichler and Hartig, 2023; Ciceu et al., 2023). Some of these algorithms natively support multi-output predictions (Borchani et al., 2015), making them especially valuable for forestry modelling tasks where correlations exist between dependent variables, such as in the prediction of DD percentiles. Suppose these approaches can effectively account for the correlations between the errors of the predicted percentiles. In that case, they may better capture the underlying relationships within a unified framework, leading to gains in predictive accuracy and more realistic model behaviour. While machine learning algorithms have shown strong performance in various forestry contexts (Ciceu et al., 2024; Huy et al., 2024; Vázquez-Veloso et al., 2025; Hobiger et al., 2025; Yang et al., 2025), their application to multivariate regression problems, particularly for predicting DD percentiles, remains largely unexplored.

This study aims to test whether multi-output tree-based ensemble models and deep learning models can provide improved solutions for predicting tree DD percentiles compared to the classical parametric approach. To do so, we applied these algorithms to a diverse set of datasets representing different forest types and management regimes, including unmanaged tropical forests, planted monocultures, thinning and density experiments, and naturally regenerated stands from several countries such as Romania, Finland, Türkiye, Spain, Nigeria, and the United States. Specifically, we address the following research questions:

1. Do multi-output machine learning models outperform classical parametric models in predicting DD percentiles across different forest datasets?

2. Can machine learning models maintain important statistical properties, such as the standard deviation and monotonicity across percentiles, more effectively than parametric models?

3. Are machine learning models consistently superior across a range of forest types, stand structures, and ecological conditions?

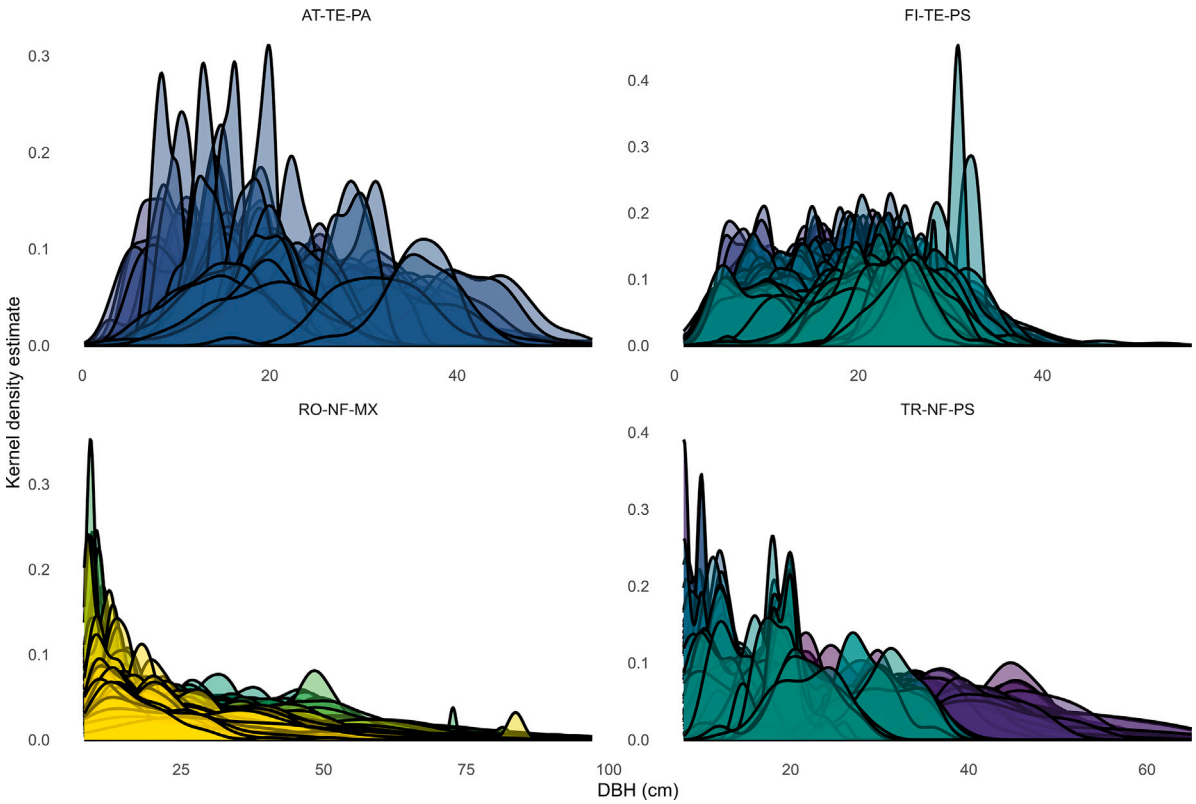


Fig. 1. Structural heterogeneity of four of the nine available datasets. Kernel density estimates of DBH are shown for individual plots across four datasets. Each coloured curve represents the smoothed DBH distribution within a single plot.

2. Methods

2.1. Data

In this study, we used nine datasets spanning diverse climatic zones and forest ecosystems, including boreal coniferous, temperate broadleaf, warm temperate coniferous, and tropical forests. Table 1 provides a detailed description of each dataset. For clarity, datasets are referenced using a standardised abbreviation scheme consisting of a country code, forest type code (NF = natural forest, PF = plantation forest, MX = mixed species stands, TE = thinning experiments, DN = density experiments), and, when applicable, a two-letter species code derived from the genus and species epithet (e.g., PS = *Pinus sylvestris*).

2.1.1. Natural temperate mixed forest stands in Romania (RO-NF-MX)

The RO-NF-MX dataset comes from a cluster-based sampling network established in Romania, covering 14262.7 ha. The network includes 783 cluster units (CUs), each with two circular sampling subplots (SPs) spaced 60 m apart (Ciceu et al., 2021). Subplot size depends on the largest tree: 200 m^2 for trees with DBH < 28 cm and 500 m^2 for larger trees. If a subplot was outside forest cover or inaccessible, only one was sampled. All trees with DBH > 8 cm were measured, totalling 23250 trees across 35 species (Ciceu et al., 2025).

The stands are predominantly naturally regenerated, resulting in high structural diversity. European beech (*Fagus sylvatica* L.) dominates, with Norway spruce (*Picea abies* (L.) Karst.) and silver fir (*Abies alba* Mill.) forming mixed-species stands. Management is close to nature with long regeneration cycles. For this study, we selected CUs with at least 15 individuals per species, totalling 390 CUs. Diameter quantiles were calculated at the CU level by aggregating all species to predict stand-level DD. Summary statistics are provided in Table 1 (see Fig. 1).

2.1.2. Natural tropical mixed rain forest in Nigeria (NG-NF-MX)

Data from Nigeria were collected randomly from various mixed-species stands in the tropical rainforest vegetation zone, which is characterised by high species diversity, an indeterminate age structure, and a wide range of tree sizes. Two datasets were combined, one of which was originally gathered as part of a REDD+ research project funded by the African Forest Forum in collaboration with the Swiss Agency for Development and Cooperation. Together, the datasets include measurements of the DBH of 4244 individual trees representing more than 150 species in 84 sample plots (Ogana and Ercanli, 2022). The plot size was 0.25 ha, and only trees with a DBH greater than 10 cm were recorded.

These datasets are characterised by substantial variability, a common trait of unmanaged tropical rainforests. The stands display wide and often multimodal DD due to the high species richness, mixed age classes, and complex forest dynamics inherent to these ecosystems.

2.1.3. Pine monoculture plantations in the Mediterranean region of Spain

Two datasets from monoculture stands in the Mediterranean region of Spain were also included in this study. These datasets represent Monterey pine (*Pinus radiata* D. Don) (ES-PF-PR) and Tasmanian blue gum (*Eucalyptus globulus* Labill) (ES-PF-EG). Pine and eucalyptus species are the most commonly used in production stands in this region, accounting for approximately 50% of the country's timber harvest (Gorgoso-Varela et al., 2015). Monoculture stands of Monterey pine occupy 96177 ha and 25385 ha, respectively. Tasmanian blue gum stands extend over 320774 ha in the Galicia region (Gorgoso-Varela et al., 2019).

A total of 96 permanent sampling plots (PSPs) for Monterey pine and 308 PSPs for Tasmanian blue gum were available for this study. The plot sizes ranged from 375 to 900 m^2; to achieve a minimum of 30 trees per plot. DBH was measured using a calliper to the nearest 0.1 cm. In total, 12985 trees were measured for Monterey pine, with a minimum DBH of 1.3 cm, and 17641 trees for Tasmanian blue gum,

with a minimum DBH of 0.1 cm. Compared to the RO-NF-MX and NG-NF-MX datasets, these datasets are characterised by high stand homogeneity, low dimensional variability, and unimodal, narrow DD.

2.1.4. Loblolly pine (*Pinus taeda* L.) density experiments in the southeastern United States (US-DE-PT)

Data used in the analyses were collected from the Coastal Plain Culture x Density study (CPCD), established and maintained by the Plantation Management Research Cooperative at the University of Georgia. In 1995–1996, a series of experimental installations (locations) were established in the lower Coastal Plain of Georgia, Florida, and South Carolina. In CPCD, all installations were planted with loblolly pine from a first-generation, open-pollinated family. Each installation is composed of six planting densities: 741, 1483, 2224, 2965, 3706 and 4448 trees per hectare (300, 600, 900, 1200, 1500, and 1800 trees per acre). Measurements, including DBH and total tree height, were taken for every live tree that exceeded breast height (1.37 m above ground) on a plot. Each plot was measured every two years until age 12 and every three years afterwards. Non-thinned operational plots from sixteen installations were selected and used in this study, with stand ages ranging from 6 to 27 years. The dataset comprises a total of 62504 trees, with a minimum diameter of 0.5 cm.

2.1.5. Norway spruce thinning experiments in Austria (AT-TE-PA)

The AT-TE-PA dataset comprises over 11000 tree measurements collected from 84 plots across 9 different experiments conducted in pure stands of Norway spruce. These experiments include thinning trials with varying levels of thinning intensity, spacing trials with different initial planting densities, and unmanaged control plots. All research sites were established and are maintained by the Austrian Research Centre for Forests and are located in eastern Austria, spanning the provinces of Lower and Upper Austria.

The diameter DBH of all trees taller than 1.3 m is measured at five-year intervals. Plot sizes range from 420 to 2590 m^2, stand ages from 16 to 57 years, and DBH values from 0.2 to 54.4 cm. One trial originated from natural regeneration, while the remaining stands were planted, some on former agricultural land. Initial stand densities at the time of first measurement ranged from 12000 to 321 trees per hectare, while final stand densities at the most recent measurement ranged from approximately 7600 to 220 trees per hectare.

2.1.6. Scots pine (*Pinus sylvestris* L.) and Norway spruce thinning experiments in Finland

The datasets included both thinned and unthinned plots from two Finnish growth and yield experiment series: ARPVANHA and HARKAS. The ARPVANHA experiments, the older of the two, were established in the 1920s and 1930s (Ilvessalo, 1932). The HARKAS experiments were primarily established in the 1970s, with some dating back to the 1960s and a few more recent trials initiated in the 1990s (Mäkinen and Isomäki, 2004b,a).

The stands consisted of pure or nearly pure Scots pine (FI-TE-PS; 92 experiments, 562 plots, 82903 trees sampled) and Norway spruce (FI-TE-PA; 26 experiments, 138 plots, 16793 trees sampled), all growing on mineral soils. Plot sizes averaged 0.13 ha (ranging from 0.01 ha to 0.25 ha), and the minimum measured DBH was 1 cm. Treatment regimes included unthinned control plots and plots thinned from below, with thinning intensities ranging from low (6%–20% basal area removal) to as high as 54%. The experiments followed a randomised block design, typically with one to three replicates per thinning intensity.

Some findings from the older experiments have been previously published (Nyyssonen, 1950; Vuokila, 1962; Mielikäinen, 1979). For the present study, one measurement occasion was randomly selected from each experiment. The average measurement period was nine years, with durations ranging from 3 to 22 years.

2.1.7. Natural scots pine forests in Türkiye (TR-NF-PS)

Another dataset used in this study was obtained from pure Scots pine stands naturally regenerated in northern Türkiye.

Field measurements for this dataset were conducted on 126 sample plots, with plot sizes ranging from 400 m^2 (radius = 11.28 m) to 800 m^2 (radius = 16 m), depending on stand crown closure. Sample plots were selected to represent each development stage, crown closure level, and site index class, maintaining a minimum distance of 300 m between plots. Within each plot, all trees with a DBH of 8 cm or greater were measured. In total, 5018 individual tree DBH measurements were collected and used to construct the dataset.

The stands exhibited high structural variability due to a mix of young and old trees. Regeneration occurred several years ago in multiple plots, which now contain a high number of small-diameter trees.

2.2. Percentile-based tree diameter distribution

We focused on six of the thirteen percentiles (P) originally proposed by Borders et al. (1987), specifically the 0th, 20th, 40th, 60th, 80th, and 100th percentiles for each plot. The 0th and 100th percentiles were included to evaluate each method's ability to accurately capture the full range of the DD, while the intermediate percentiles provided a representative summary of its internal structure. These percentiles were modelled as functions of stand-level predictors that are both practical and efficient to measure in the field.

To investigate how different levels of information affect model performance, we used four sets of predictor variables, as shown in Eq. (1)–(4).

For the RO-NF-MX, NG-NF-MX, and TR-NF-PS datasets, Eq. (1) was applied, as stand age (T) was unavailable, as these stands originate from natural regeneration where precise age determination is not possible. In the thinning trial datasets FI-TE-PS and FI-TE-PA, Eq. (2) was used, incorporating thinning status (Z). For the ES-PF-EG, US-DE-PT, and ES-PF-PR datasets, Eq. (3) was employed, adding stand age (T) to the basic stand variables. Finally, the AT-TE-PA dataset, which contained the most comprehensive set of stand-level information and was also part of a thinning trial, was modelled using Eq. (4).

$$P_k = f(D_g, DDOM, N) \tag{1}$$

$$P_k = f(D_g, DDOM, N, Z) \tag{2}$$

$$P_k = f(D_g, DDOM, T, N) \tag{3}$$

$$P_k = f(D_g, DDOM, T, N, Z) \tag{4}$$

where P_k is the kth percentile of the DD, with $k \in \{0, 20, 40, 60, 80, 100\}$; Dg is the quadratic mean diameter, $DDOM$ is the dominant DBH, defined as the DBH of the 100 largest trees per hectare, N is the number of trees per hectare, T is the stand age, and Z is a one-hot encoded dummy variable, which takes the value 1 if thinning was applied and 0 if the stand is unthinned.

2.3. Models

2.3.1. Generalised least squares (GLS)

For each dataset, we constructed a system of six equations to describe the selected percentiles based on stand-level predictors. We applied SUR, as introduced by Zellner (1962), to model the relationships between the percentiles and the predictors. SUR estimates parameters using the GLS method, which accounts for possible correlations among the error terms of the equations. To fairly assess the performance of the three modelling frameworks, we used the same set of predictors across all fitting methods and equations, resulting in parameter estimates equivalent to those obtained by GLS. The statistical significance of individual coefficients was considered secondary; the primary objective

Table 1
Summary statistics for stand-level predictors and DD percentiles across the nine datasets. Values in parentheses represent standard deviations. N denotes the number of trees per hectare, Dg is the quadratic mean diameter, and $DDOM$ is the dominant diameter, defined as the diameter of the 100 largest trees per hectare. T refers to stand age. Percentiles from P_0 to P_{100} represent the distribution of tree diameters within each plot.

Dataset	N	Dg	$DDOM$	T	P_0	P_{20}	P_{40}	P_{60}	P_{80}	P_{100}
RO-NF-MX	635 (461)	32.2 (10.5)	48.8 (12.3)	–	10.2 (4.2)	17.2 (8)	23.5 (10.6)	30.5 (12.4)	39.9 (14.3)	63.3 (18.1)
NG-NF-MX	207 (75)	34.1 (6.4)	44.4 (8.6)	–	11.2 (1.9)	16.2 (3.3)	21.4 (4.4)	27.6 (6.1)	39 (8.8)	92.8 (32.3)
ES-PF-EG	1149 (342)	13.4 (4.1)	20.4 (6.9)	10.4 (4.8)	3.3 (1.8)	8.4 (3.2)	11.3 (3.7)	13.9 (4.3)	16.5 (5.2)	22.5 (8.2)
ES-PF-PR	1343 (542)	17.8 (3.2)	27.1 (4.3)	17.6 (4)	5.3 (2.5)	11.8 (2.8)	15.2 (3.1)	18.5 (3.5)	21.9 (4)	30.7 (5.4)
US-DE-PT	1937 (1075)	16.4 (5.9)	20.5 (6.4)	14.4 (6.7)	8.1 (4.7)	13.5 (5.3)	15.3 (5.6)	16.9 (6)	18.7 (6.5)	23.1 (7.9)
AT-TE-PA	1459 (1402)	22.3 (8.3)	28.7 (8.9)	34.3 (11.2)	11.8 (7.8)	18.5 (7.8)	21 (8.2)	23 (8.5)	25.3 (9)	31.7 (10.3)
FI-TE-PA	1069 (558)	23.3 (6.1)	30.3 (5.8)	–	13.1 (7.1)	19.2 (6.2)	21.7 (6)	23.9 (6.1)	26.4 (6.3)	34.4 (7.1)
FI-TE-PS	1158 (769)	18.7 (4.7)	24.7 (4.2)	–	10.3 (5.4)	15.1 (5.2)	17.2 (5)	19.1 (4.8)	21.3 (4.8)	27.8 (5.1)
TR-NF-PS	598 (229)	26.7 (10.8)	33.6 (12.4)	–	15.2 (7.3)	22.2 (10)	25.1 (10.9)	27.5 (11.6)	30.3 (12.4)	36.8 (12.9)

was to maintain consistency in the predictor sets to ensure a fair and unbiased comparison between parametric and nonparametric modelling approaches.

Because the relationships between the percentiles and predictors were inherently nonlinear, we linearised them by applying logarithmic transformations to both the percentiles and the predictors. Predictions were then back-transformed, with half of the residual variance added to correct for the bias introduced by the transformation. The system of equations was fitted using R (Team, 2025) and *systemfit* function from systemfit package (Henningsen and Hamann, 2007).

2.3.2. Multi-output random forest (MORF)

The first machine learning algorithm we used was MORF, an adaptation of Random Forest (RF) designed to predict multiple output variables simultaneously. RF is one of the few machine learning algorithms that adapted to natively support multi-output regression (Breiman, 2001a; De'Ath, 2002). Unlike single-output models, where each regression tree optimises a single target variable, MORF optimises all outputs collectively, capturing interactions among them. The splitting criterion at each node is based on the average reduction in impurities across all m percentiles, ensuring that the model considers all outputs when making splits. In MORF, the impurity of a node is redefined by summing the univariate impurity measures across the multivariate response. Specifically, the splitting criterion selected was the mean squared error (MSE), which corresponds to variance reduction and minimises the average squared difference between the predicted and observed values of the multivariate response.

In our case, we used six percentiles as output variables, with stand-level predictors as explanatory variables. The model formulation is a straightforward extension of the single-output RF: instead of delivering a one-column array (single prediction) per set of observations, the model receives an array of six output values, corresponding to the six percentiles. The training, tuning, and evaluation of MORF were performed using Keras (Chollet et al., 2015) within the Python environment (Van Rossum and Drake Jr., 1995).

2.3.3. Multi-output deep learning (MODL)

Deep learning (DL) is a subfield of machine learning that employs artificial neural networks to learn patterns from data. Originally inspired by human cognition (McCulloch and Pitts, 1943), DL models are structured as multilayer networks consisting of an input layer, one or more hidden layers, and an output layer (LeCun et al., 2015).

In single-output architectures, the output layer has one neuron, while in multi-output setups, it includes multiple neurons to enable simultaneous predictions. Each neuron receives weighted inputs, applies an activation function, and transmits the result to the next layer. During training, these weights are iteratively adjusted to minimise a single loss function that jointly considers all output dimensions.

In our multi-output regression task, we used MSE loss computed across all outputs. This loss encourages the model to simultaneously minimise errors for all targets, enabling it to learn both how inputs affect each output and how the outputs co-vary. The gradients from this joint loss are backpropagated through shared hidden layers, so all outputs influence the same internal feature space.

The term "deep learning" refers to the depth of these networks, which can range from just a few layers to thousands. In this study, we employed a shallow architecture with only two hidden layers, chosen to reflect the limited sample size. The feedforward architecture comprised an input layer aligned with the number of predictors, two hidden layers using ReLU activations, and a final linear output layer with six nodes to estimate the target percentiles. The models were compiled using the Adam optimiser, with MSE as both the loss function and evaluation metric.

Because the predictors and response variables were on different scales, we first standardised both sets of values, except the one-hot encoded dummy variable Z. Once the number of hidden layers and output neurons was defined, model performance depended on several hyperparameters, including batch size, learning rate, number of neurons and the number of training epochs. These hyperparameters were tuned individually for each dataset using a random search within a predefined search space (see Table 2).

To prevent overfitting and ensure optimal model performance, we implemented an early stopping strategy. The training process was monitored based on the loss function and automatically halted if no improvement was observed for 250 consecutive epochs. Upon stopping, the model weights from the epoch with the lowest loss were restored.

All feedforward DL models were built, trained, tuned, and evaluated using Keras (Chollet et al., 2015) with TensorFlow (Abadi et al., 2015) as the backend, via the Sequential API in Python (Van Rossum and Drake Jr., 1995).

2.4. Hyperparameter tuning

To evaluate model performance on unseen data, we first split each dataset into training and testing subsets. This separation allows us to train the model on one portion of the data while assessing its predictive accuracy on a different, unused subset, ensuring a more realistic estimate of how the model will perform in practice.

Each dataset was treated independently. To maintain representativeness across the full range of stand structures, we stratified the training and testing splits based on quantile bins of the unscaled Dg values, allocating 80% of the data for training and 20% for testing.

Hyperparameter tuning for both machine learning algorithms was performed using a random search strategy on the training set. To ensure robust performance estimation and reduce the risk of overfitting, we implemented a 5-fold cross-validation scheme with two repetitions for all models. This approach allowed us to assess model generalisability and verify that the selected hyperparameters produced stable performance across different data splits.

For each algorithm, 500 hyperparameter combinations were randomly sampled and evaluated using repeated 5-fold cross-validation, resulting in a total of 5000 runs per model. The tuning objective was to minimise the MSE. The full list of hyperparameters and their respective search spaces is provided in Table 2.

Table 2
Hyperparameter search space for MORF and MODL.

Algorithm	Parameters	Min	Max	Step
MORF	Number of trees	50	500	10
	Max features	sqrt	log2	–
	Max depth	None	10	1
	Min sample split	5	25	1
	Min sample leaf	3	10	1
	Bootstrap	FALSE	TRUE	–
MODL	Neurons layer 1	32	512	5
	Neurons layer 2	32	512	5
	Learning rate	0.0001	0.01	0.001
	Batch size	32	64	2
	Epchos	500	5000	500

In the final stage, after selecting the best-performing hyperparameter set, the models were retrained using the entire training set and evaluated on the test set.

2.5. Model evaluation

To identify the best-performing algorithm, we employed two complementary approaches. First, the three modelling approaches were evaluated using both the training and test datasets. Predictive performance was assessed using standard error metrics: mean squared error (MSE, Eq. (5)), root mean squared error (RMSE, Eq. (6)), mean absolute error (MAE, Eq. (7)), and mean error (ME, Eq. (8))

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

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

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

$$\mathrm{ME} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \tag{8}$$

where $\hat{y}_i$ is the predicted value for the ith observation, y_i the observed value for the ith observation, $\bar{y}$ the mean of the observed values, and n the number of observations.

To enable a consistent comparison across multiple error metrics and percentile levels, a composite ranking method was applied. For each combination of dataset, data subset (training or test), and percentile, models were ranked individually for each performance metric based on a relative ranking function. These metric-specific ranks were then summed to generate a total rank score for each model within that specific context.

To derive an overall assessment of model performance, the total ranks were aggregated across all percentiles for each model. The resulting cumulative scores were then used to assign final ranks within each dataset and data subset, enabling an evaluation of model performance across all error metrics and percentile levels.

$$R_m = 1 + \frac{(n-1)(S_\mathrm{m} - S_\mathrm{min})}{S_\mathrm{max} - S_\mathrm{min}} \tag{9}$$

where R_m is the relative rank, with m taking values between 1 and 3, corresponding to the three modelling frameworks tested. S_m represents the MSE, RMSE, MAE, or ME values, while S_min and S_max denote the minimum and maximum values of S_m, respectively.

In addition to these conventional metrics, we employed the Taylor diagram (Taylor, 2001) as a graphical summary tool to further compare model performance. For this analysis, we used only the test dataset, which was excluded entirely from the model training process to ensure an unbiased assessment of predictive performance. The Taylor diagram offers an integrated and intuitive visual representation of model performance, displaying three complementary statistics: the correlation coefficient, the standard deviation, and the centred RMSE (Eq. (10) - cRMSE).

$$\mathrm{cRMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[(\hat{y}_i - \bar{y}) - (y_i - \bar{y})\right]^2} \tag{10}$$

Taylor diagram allows for the simultaneous visualisation of how well each model reproduces the spatial pattern (through correlation), the amplitude of the variability (through standard deviation), and the overall deviation from the observations (through cRMSE). By plotting all models on the same diagram relative to a reference point representing the observed data, the Taylor diagram highlights both the relative strengths and weaknesses of each approach and facilitates quick identification of models that strike the best balance between accuracy and robustness.

3. Results

3.1. Parameters and hyperparameters of the tested models

For the GLS models, Dg and DDOM emerged as the most important stand-level predictors across multiple datasets and percentiles (Table A.1). As expected, DDOM was consistently significant for the upper percentiles, particularly for P_{100}, where it was significant in all nine datasets. Dg showed significance across most percentiles and datasets, with only a few exceptions. At least one variable was significant for each percentile in every dataset, except for P_0 in the NG-NF-MX dataset. However, for the other percentiles within this dataset, most variables were statistically significant. A similar pattern was observed in the models developed for the TR-NF-PS and US-DE-PT datasets, where the vast majority of variables were found to be significant.

In a model developed for practical applications, non-significant variables would typically be excluded, and the use of SUR would be an appropriate approach for further model development. However, in our case, to enable meaningful comparisons across models, we retained all variables, including those not statistically significant, to assess overall predictive performance.

The best-performing hyperparameter configurations varied notably across datasets, reflecting differences in both data complexity and learning behaviour.

The MODL models, trained separately on each of the nine datasets, exhibited distinct learning dynamics and could be categorised as either "fast learners" or "slow learners" based on their hyperparameters. The "slow learners", trained on the NG-NF-MX, RO-NF-MX, and TR-NF-PS datasets, required substantially more training epochs, with optimal performance achieved at 2500, 3000, and 5000 epochs, respectively. Notably, the model for TR-NF-PS was the only one that responded best to a relatively high learning rate (0.008), whereas all other datasets consistently favoured lower learning rates (0.001) (Table A.2).

Architecturally, five out of nine MODL models adopted a funnel-shaped configuration, where the number of neurons in the first layer (*neurons*_1) was slightly smaller than in the second layer (*neurons*_2). The "slow learners" tended to favour simpler architectures with fewer neurons overall, particularly in the second layer. Among all models, the one developed for FI-TE-PS yielded optimal results with the smallest network size.

Deeper network structures, with both *neurons*_1 and *neurons*_2 exceeding 450 units, were optimal for datasets such as AT-TE-PA, ES-MF-EG, and ES-MF-PR. In contrast, models trained on FI-TE-PS, RO-NF-MX, and NG-NF-MX performed best with more compact configurations (Table A.2).

Smaller batch sizes (34–44) were most effective for the majority of datasets. Only a few, including ES-MF-PR and TR-NF-PS, showed improved performance with larger batch sizes (54–64).

For the MORF models, most achieved optimal results with a moderate number of estimators, typically around 150 or 370. However, the "slow learners" (trained on TR-NF-PS and RO-NF-MX) benefited from significantly larger ensembles, with up to 490 estimators.

Most MORF models favoured smaller values for *min_samples_leaf* (typically 3) but moderate values for *min_samples_split* (6–7), resulting in shallower trees. An exception was the MORF trained on NG-NF-MX, where deeper trees improved performance, driven by a lower threshold for node splitting.

The *max_features* parameter was predominantly set to `log2`, although a few MORF models trained on FI-TE-PS and US-DE-PT performed better with `sqrt`. Tree depth (*max_depth*) was generally constrained to shallow values (e.g., 7 or 10), except for the MORF model trained on ES-MF-EG, where no depth limit was imposed, allowing trees to grow until other stopping criteria were met. A depth of 10 was also typical among the "slow learners" (Table A.2).

Finally, bootstrapping was generally disabled across datasets. Exceptions included the models trained on FI-TE-PS, US-DE-PT, and TR-NF-PS, where enabling bootstrapping improved performance.

3.2. Model performance across individual datasets and percentiles

Taylor diagrams indicate strong correlations between observed and predicted values, with all models effectively capturing variability and providing estimates suitable for operational use. Overall, MODL showed superior performance at lower percentiles (P_0 to P_{40}), outperforming the other models in seven out of nine datasets for P_0, and in five datasets for both P_{20} and P_{40}. At higher percentiles, GLS outperformed both machine learning approaches in six datasets for P_{60} and in eight datasets for P_{100}. MODL was the best-performing model in three datasets at P_{60}, one at P_{100}, and notably outperformed GLS and MORF at P_{80} in six out of nine datasets.

At the dataset level, no single model consistently outperformed all others across every percentile, except for the GLS model developed for the NG-NF-MX dataset (Fig. 2). However, none of the models succeeds in preserving the variability observed in the data, instead producing homogenised estimates. This shortcoming is further reflected in the weak correlations between predicted and observed values, the lowest among the nine datasets. In terms of similarity to the actual standard deviation of the measured values, all three modelling approaches performed the worst on this dataset. Moreover, the cRMSE for the highest percentile exceeds 25 cm for MODL and 15 cm for GLS, despite GLS delivering the best results. Such levels of error limit the practical applicability of these models, especially for predicting the largest percentiles, calling for reconsideration either of the model structure or the input variables. These results obtained for the NG-NF-MX dataset obtained highlight the high structural variability of tropical forests, where maximum tree diameters vary widely between sites due to past disturbances, species composition, and absence of management. In contrast, the models developed for RO-NF-MX dataset exhibit a divergent results, with the MODL approach consistently outperforming the other models based on centred RMSE, alignment with the observed standard deviation, and correlation coefficients across all percentiles except at P_{100} (Fig. 3).

Although the differences among the three modelling approaches were visually striking in the RO-NF-MX and NG-NF-MX datasets, the ES-PF-EG, (Fig. A.1), FI-TE-PA (Fig. 4) and ES-PF-PR (Fig. A.2) datasets present a different pattern. Only MORF stands out across multiple percentiles, whereas GLS and MODL produce similar predictions. Among these, MODL performs slightly better, consistently delivering accurate estimates across percentiles while preserving the variability in the predictions. As observed in the RO-NF-MX dataset, MODL stands out particularly at the first percentile, providing more accurate estimates compared to the other two approaches. In contrast, for the upper percentiles, all three models produce similar results, with GLS providing slightly better results.

A similar pattern is observed for the AT-TE-PA (Fig. A.5), FI-TE-PS (Fig. A.3) and TR-NF-PS (Fig. A.4 datasets, where the GLS and MODL models consistently alternate for the top position across percentiles, yielding very similar predictions (Fig. A.2). The MORF algorithm consistently ranks third; however, cRMSE values remain below 4 cm across all six percentiles, indicating a good overall fit.

For the US-DE-PT dataset, the three models performed equally well in all cases except P_0, where MODL and MORF showed slightly better performance (Fig. 5). In all other cases, the differences among the models' predictions were negligible, with all three typically producing predictions with errors of less than one cm. Notably, US-DE-PT and FI-TE-PS are the only datasets in which MORF produced predictions comparable to the other two modelling approaches across all percentiles.

3.3. Overall model performance

The relative ranking of models revealed that machine learning approaches generally outperformed the parametric model on the training data (Fig. 6). MODL and MORF consistently occupied the top two positions in seven out of nine datasets, with MODL ranking first across all nine training subsets. The performance gap between these two leading algorithms varied considerably, ranging from 1.1 to 2.9. Although in many datasets GLS ranked third on the training data, it surpassed MORF on the test subsets, highlighting the limited generalisability of the MORF algorithm. In the test subsets, MODL remained the best-performing model, ranking first in five out of nine datasets.

4. Discussion

4.1. Machine learning approaches for tree diameter distribution modelling

When modelling DD, a common approach is to apply machine learning algorithms within a parameter prediction framework, where the objective is to estimate the parameters of a theoretical probability distribution. For instance, Abbasi et al. (2008) and Diamantopoulou et al. (2015) employed ANNs to estimate the parameters of the Weibull distribution and reported that ANNs outperformed other algorithms evaluated in their studies.

A similar methodology has also been applied in other fields. For example, Qamar et al. (2025) used ANNs to predict the parameters of a Weibull distribution modelling sunspot activity. Likewise, Alrashidi (2023) demonstrated that neural networks can be used to estimate theoretical distribution parameters, although maximum likelihood achieved superior performance.

In our study, we found that the MODL approach outperformed the standard parametric method typically used to estimate the parameters of the Weibull distribution. Similarly, Huy et al. (2024) reported that MODL provided more accurate estimates than weighted nonlinear SUR when predicting both above- and below-ground tree biomass components using the same set of predictors. MODL's performance across datasets and forest types highlights its robustness and suitability for DD modelling, supporting the broader reliability of deep learning approaches in forestry applications (Hamedianfar et al., 2022).

An alternative to parameter prediction is using machine learning to directly model the empirical probability density function without assuming any theoretical distribution. This approach, explored by Leduc (2001) and Bolat and Ercanlı (2016), reconstructs DD by directly estimating the PDF. However, it requires predefined diameter ranges and class intervals, which limit flexibility. For example, Leduc (2001) divided the diameter range into 20 fixed one-inch classes (1–20 inches). Once set, these classes cannot be easily modified, restricting predictions to the specified intervals.

Our analysis, which incorporates datasets from multiple regions worldwide, shows that DD ranges vary substantially depending on forest management practices and the degree of naturalness. When the

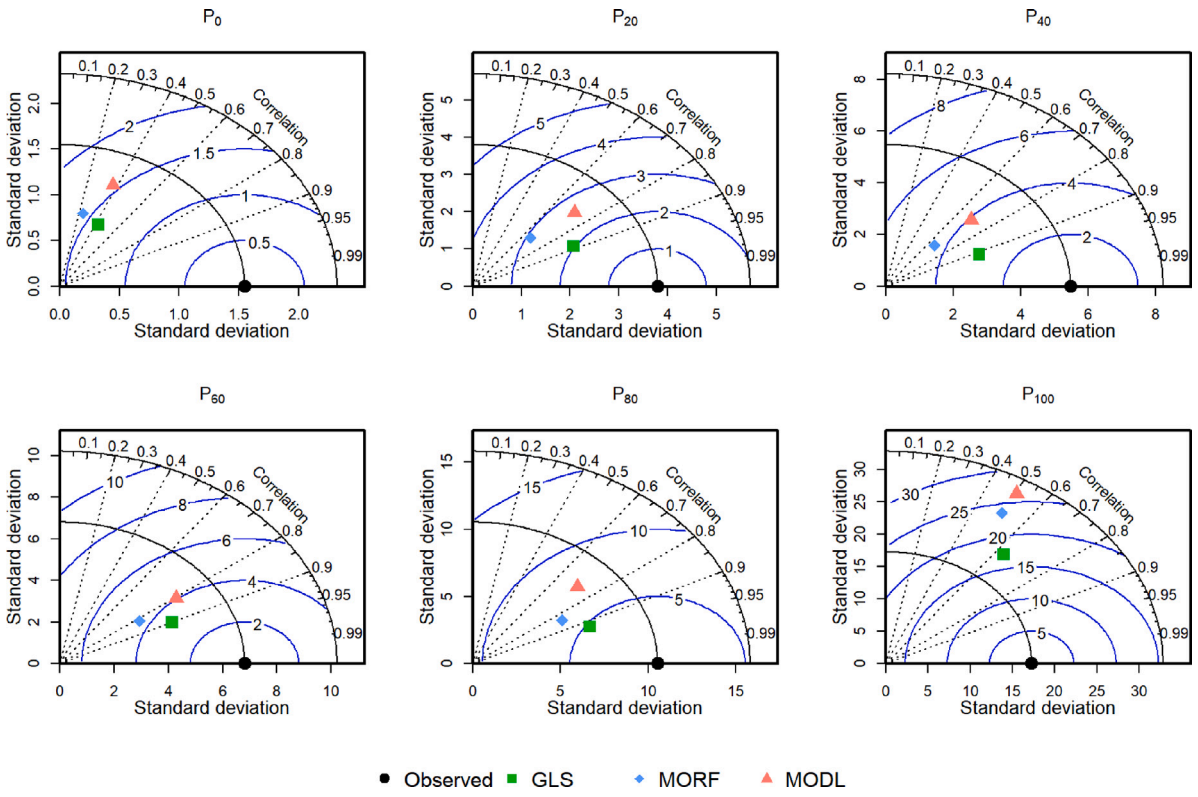


Fig. 2. Taylor diagram comparing the performance of the three modelling approaches for NG-NF-MX. Each subfigure is labelled using the percentile P for which the statistics are reported, with P_0 representing the 0th percentile and subsequent labels following in order. MODL refers to the multi-output deep learning model, MORF represents the multi-output random forest, and SUR corresponds to the seemingly unrelated regression. The diagram illustrates each model's ability to reproduce the standard deviation of the observed data, the correlation between predicted and observed values, and the cRMSE (blue arcs). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

diameter range is predefined using a limited number of plots that do not capture the full variability of possible DDs for a species, substantial errors can arise in volume and per-hectare calculations. This issue is critical because large-diameter trees contribute disproportionately to total stand volume. Underestimating their frequency can therefore lead to significant biases in overall volume estimates.

The approach employed in this study, following the methodology of Borders et al. (1987), reconstructs the DD based on multiple percentiles. By using multiple percentiles, our model allows for predictions of tree counts across any desired diameter class, without being constrained by a fixed class structure.

Only a few machine learning algorithms have been tested for DD modelling, most commonly ANNs (Leduc, 2001; Abbasi et al., 2008; Cai et al., 2012; Diamantopoulou et al., 2015; Bolat and Ercanlı, 2016; Ercanli and Bolat, 2017), SVMs (Bolat and Ercanlı, 2016), MaxEnt (Chen et al., 2019), and k-NN (Haara et al., 1997; Maltamo and Kangas, 1998). Our study is the first to adopt a comprehensive approach by testing multiple algorithms across a wide range of datasets covering different forest types and management practices.

Furthermore, to our knowledge, MORF have not been previously applied in this context. In our study, MORF showed the weakest ability to generalise, which is consistent with the findings of Vázquez-Veloso et al. (2025), who reported that univariate RF performed best on training data but worst under cross-validation when predicting tree mortality, compared to Decision Trees, Naive Bayes, k-NN, and SVMs.

Despite its limitations, MORF has been applied in forestry for tasks such as ensemble vegetation index mapping (Kocev et al., 2009), while RF is already widely used for various forest modelling tasks (Silva et al., 2017; Yang et al., 2022; Liu et al., 2024). In our study, MORF ranked third overall, with predictive performance only slightly lower than MODL and GLS across most percentiles and datasets.

A key advantage of MORF is its ability to ensure monotonicity, which is essential when reconstructing DDs from predicted percentiles. To be valid, percentiles must satisfy the condition: $P_0 < P_{10} < \cdots < P_{100}$.

Among the three approaches tested, only MORF consistently produced monotonic predictions across all datasets and plots. In contrast, MODL and GLS generated non-monotonic predictions in several cases, five plots for MODL and thirteen for GLS.

MORF's ability to guarantee monotonicity makes it particularly useful in stands with narrow DD structures, where its predictive performance is comparable to MODL and GLS. Moreover, it provides a practical solution when time or data for model development are limited. For MODL and GLS, monotonicity can still be achieved by algebraically transforming target percentiles before model training (Borders et al., 1987; Kangas et al., 2007; Stankova and Zlatanov, 2010).

A common critique of multi-output machine learning algorithms is that they may distribute accuracy across outputs, potentially reducing optimal performance for individual targets and, in some cases, resulting in higher accuracy for single-output models (Tran et al., 2024). However, this trade-off is not universally observed (Borchani et al., 2015). In fact, several studies have shown that multi-output models can achieve superior overall predictive performance compared to single-output approaches, primarily because they reduce overfitting and capture dependencies among outputs (Kocev et al., 2009; Claveria et al., 2015; Xi et al., 2018), even when the outputs are not linearly correlated (Mastelini et al., 2018). Furthermore, fitting a joint model ensures logical consistency among predicted variables, which is particularly important in ecological modelling where multiple outputs are often interdependent.

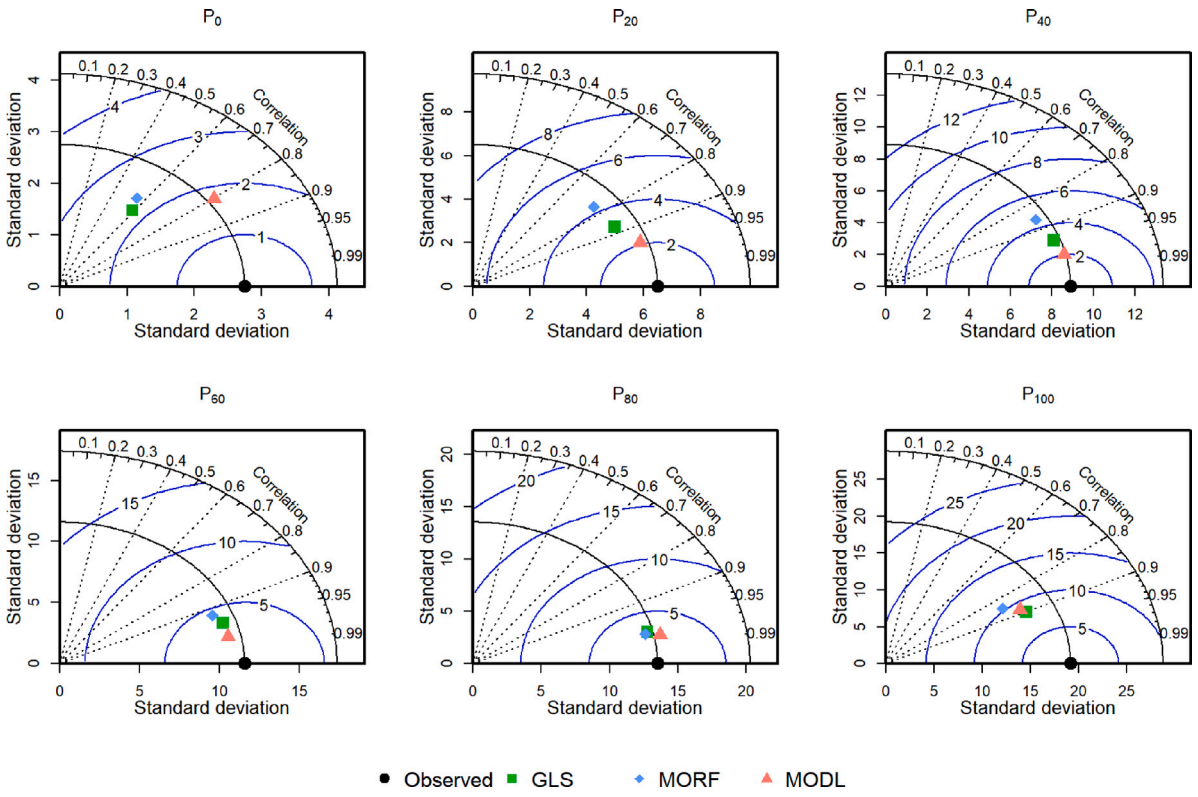


Fig. 3. Taylor diagram comparing the performance of the three modelling approaches for RO-NF-MX. Each subfigure is labelled using the percentile P for which the statistics are reported, with P_0 representing the 0th percentile and subsequent labels following in order. MODL refers to the multi-output deep learning model, MORF represents the multi-output random forest, and SUR corresponds to the seemingly unrelated regression. The diagram illustrates each model's ability to reproduce the standard deviation of the observed data, the correlation between predicted and observed values, and cRMSE (blue arcs). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.2. Statistical and algorithmic modelling comparison

At first glance, a comparison between statistical and algorithmic modelling may appear futile, given the fundamental differences between the two modelling paradigms (Ij, 2018). Statistical models rely on explicit assumptions about data generation and predefined functional forms, often informed by prior subject-matter knowledge about the process that generated the data (McCullagh, 2002). In our case, a logarithmic transformation was applied to linearise relationships and facilitate non-biased GLS parameter estimation, which can impose constraints on model flexibility.

By contrast, algorithmic approaches such as machine learning make no assumptions about the form of the systematic component (Breiman, 2001b). Instead, they rely on data-driven algorithms and optimisation criteria to learn the underlying structure directly from the data, with a primary focus on achieving the best predictive fit. This allows for greater flexibility and the ability to capture complex, non-linear patterns. However, even comparisons between machine learning models with different architectures, as is the case in our study, are not entirely fair. One could say, no model comparison is ever perfectly balanced.

Even when the same model fitting criterion is used across all methods, as we have done here, the architectural differences lead to substantial disparities in model complexity, making direct comparisons problematic. The statistical models in our study estimate between three and five parameters, depending on the dataset. In contrast, the MORF models included 150 to 490 trees, each with depths ranging from 7 to 10. The MODL architecture used between 224 and 959 neurons across two hidden layers, resulting in up to 234466 weights in the most complex configuration.

From the perspective of model parsimony, the statistical model outperforms its algorithmic counterparts. Furthermore, statistical models offer greater control over model behaviour through explicit definition of the systematic component, which can enhance interpretability, transferability, and predictive robustness under extrapolation (Vázquez-Veloso et al., 2025). Computational efficiency is another key advantage (Makridakis et al., 2018): all statistical models converged in under one minute, while the MODL model for the RO-NF-MX dataset alone required approximately 17 h to train on a standard machine. In addition, statistical models facilitate straightforward estimation of prediction uncertainty through the variance–covariance matrix, which enables analysis of prediction uncertainty and model calibration (Mehtätalo, 2005). Another important advantage of statistical models is their suitability for small datasets. Unlike machine learning models, which generally require large volumes of data to learn and achieve generalisation (Rajput et al., 2023), statistical models can be reliably fitted even with relatively limited data.

However, in the era of fast computing, some advantages of statistical modelling may be less critical. Statistical models are computationally efficient but require substantial prior knowledge to define the systematic component. In contrast, machine learning models demand more training time but can be developed with minimal assumptions. It is therefore essential to define the objective of the comparison clearly from the outset.

When predictive accuracy is the primary objective, computational resources are not a constraint, and data is abundant, comparing the two modelling paradigms becomes not only valuable but necessary, as statistical models serve as a baseline that machine learning models are expected to outperform. In such cases, fairness in comparison becomes secondary to assessing how well each method performs on the task at hand.

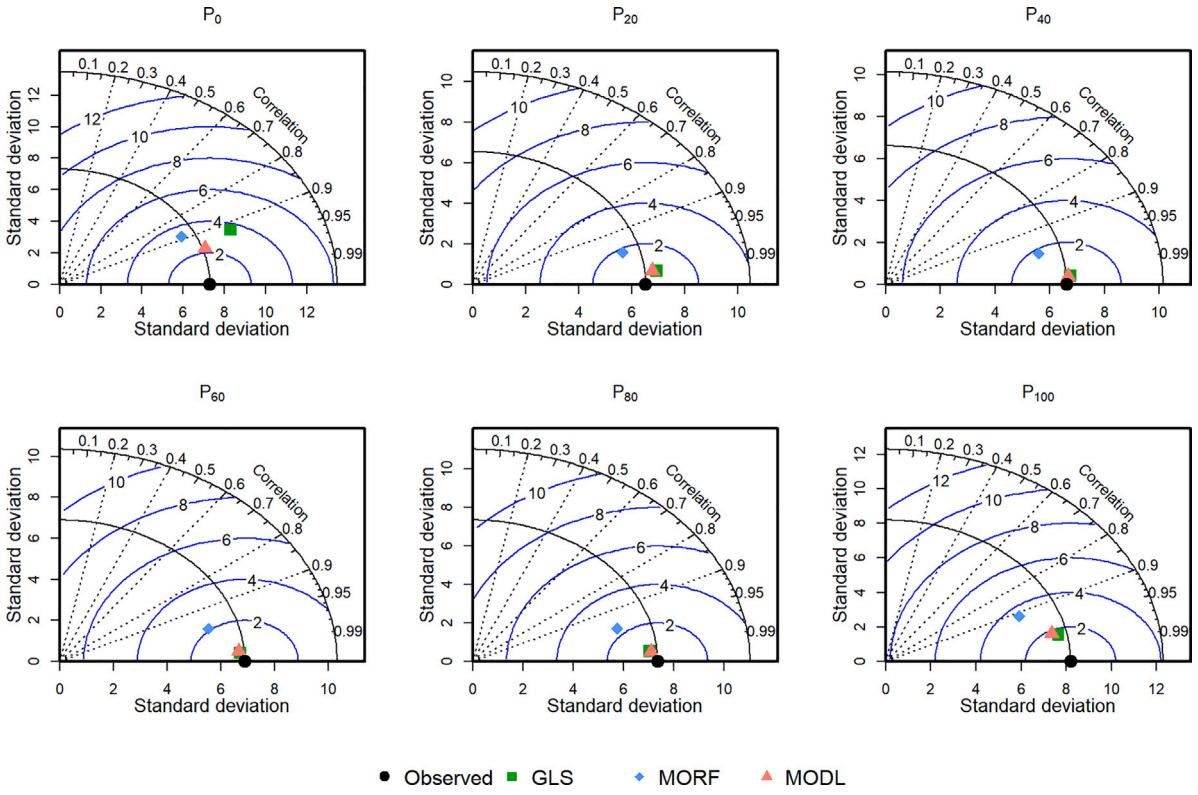


Fig. 4. Taylor diagram comparing the performance of the three modelling approaches for the FI-TE-PA dataset. Each subfigure is labelled using the percentile P for which the statistics are reported, with P_0 representing the 0th percentile and subsequent labels following in order. MODL refers to the multi-output deep learning model, MORF represents the multi-output random forest, and SUR corresponds to the seemingly unrelated regression. The diagram illustrates each model's ability to reproduce the standard deviation of the observed data, the correlation between predicted and observed values, and the cRMSE (blue arcs). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5. Conclusion

In this study, we demonstrate that multi-output machine learning models, particularly MODL, provide a powerful and flexible approach for predicting DD percentiles across diverse forest types. MODL delivered the best overall predictive performance and generalisation, while MORF, although slightly less accurate, consistently preserved monotonicity, making it a practical choice in time or resource-constrained settings. The parametric statistical model, though simpler and highly efficient, achieved accuracy comparable to MODL, highlighting their value when computational resources or training data are limited.

From a methodological perspective, future research should compare reconstructing DD from percentile-based MODL predictions with directly fitting MODL to the empirical probability density function. Evaluating these approaches across multiple datasets would provide a robust assessment of the most effective method for DD reconstruction. A key finding of this study is that MODL performs consistently well across different datasets, providing strong evidence that it can be reliably applied without extensive benchmarking against other algorithms.

From a forest management perspective, the improved DD predictions provided by MODL enable more precise assessment of carbon stocks and better planning of thinning, harvesting, and other silvicultural interventions. This supports sustainable resource use and enhances long-term stand stability.

Ecologically, modelling DD improves our understanding of forest structure, dynamics, and biodiversity conservation. As DD shape is widely used as an indicator of stand stability and biodiversity, these findings underscore the importance of continuous monitoring and further research to develop fast, accurate methods for reconstructing DD, ultimately supporting resilient and biodiverse forest ecosystems.

CRediT authorship contribution statement

Albert Ciceu: Writing – original draft, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Hasan Aksoy:** Writing – review & editing, Resources. **Ovidiu Badea:** Writing – review & editing, Resources, Conceptualization. **Bronson P. Bullock:** Writing – review & editing, Resources. **Jacinta Ukamaka Ezenwenyi:** Writing – review & editing, Resources. **Jose Javier Gorgoso-Varela:** Writing – review & editing, Resources. **Ştefan Leca:** Writing – review & editing, Resources. **Thomas Ledermann:** Writing – review & editing, Resources. **Harri Mäkinen:** Writing – review & editing, Resources. **Friday N. Ogana:** Writing – review & editing, Resources. **Sheng-I Yang:** Writing – review & editing, Resources. **Lauri Mehtätalo:** Writing – review & editing, Supervision, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the European Union and the Romanian Government through the Competitiveness Operational Programme 2014–2020, under the project "Increasing the economic competitiveness of the forestry sector and the quality of life through knowledge transfer, technology and CDI skills" (CRESFORLIFE), ID P 40 380/105506, subsidiary contract no. 17/2020 and partially by the FORCLIMSOC Nucleu Programme (Contract 12N/2023), project PN

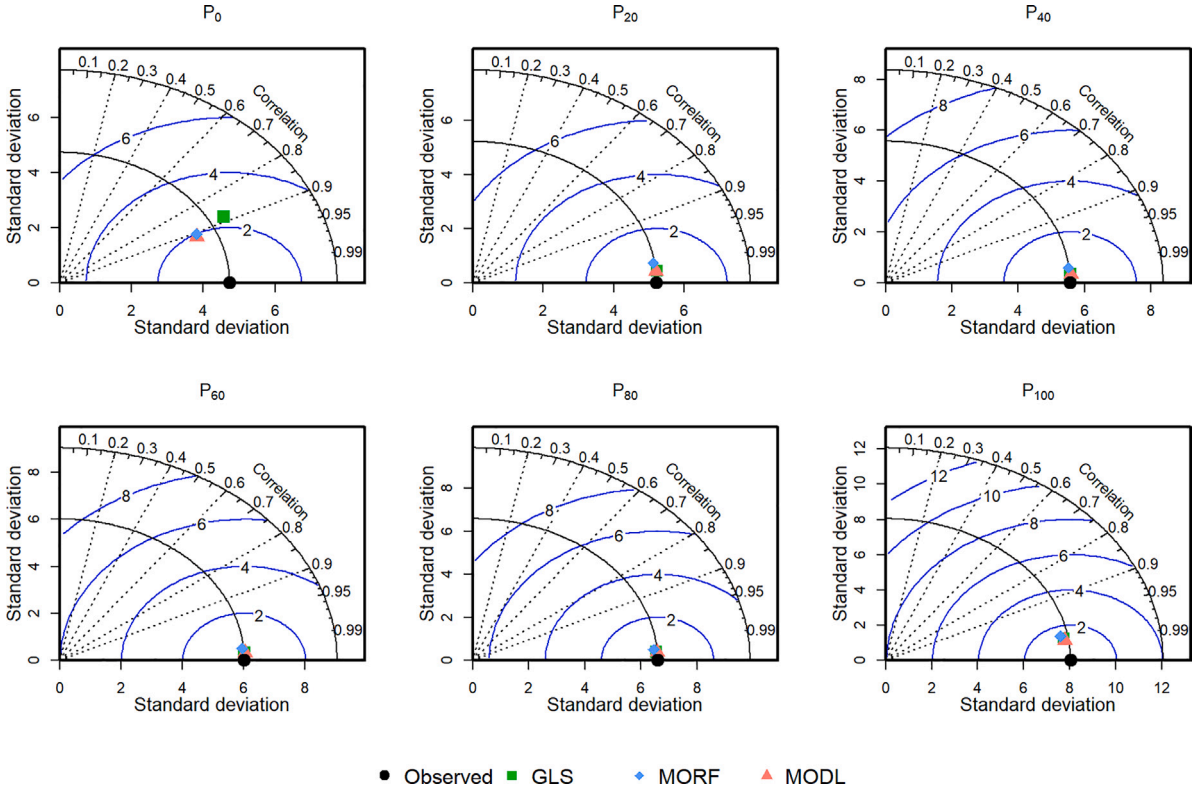


Fig. 5. Taylor diagram comparing the performance of the three modelling approaches for the US-DE-PT dataset. Each subfigure is labelled using the percentile P for which the statistics are reported, with P_0 representing the 0th percentile and subsequent labels following in order. MODL refers to the multi-output deep learning model, MORF represents the multi-output random forest, and SUR corresponds to the seemingly unrelated regression. The diagram illustrates each model's ability to reproduce the standard deviation of the observed data, the correlation between predicted and observed values, and the cRMSE (blue arcs). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

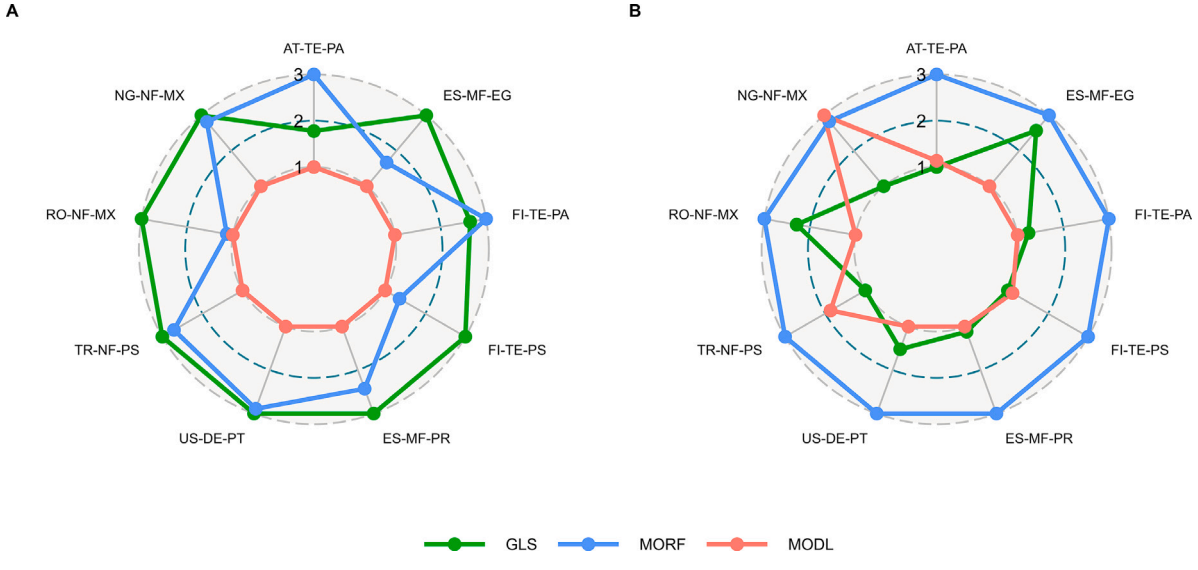


Fig. 6. Overall relative ranking of the models across nine datasets and data subsets. The numbers displayed on the axes represent the rankings, where lower values indicate better performance. Panel A shows the training subset, and Panel B shows the test subset.

23090101 and CresPerfInst project (Contract 34PFE/30.12.2021) "Increasing the institutional capacity and performance of INCDS 'Marin Drăcea' in RDI activities - CresPerfInst". We are very grateful to the General Directorate of Forestry, Department of Forest Management and Planning team for the Türkiye forest inventory data.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ecoinf.2025.103500.

Data availability

The R and Python scripts used to train the three models are publicly available at: https://github.com/AlbertCiceu/Multi-Output-DL-RF-for-Diameter-Percentile-Prediction.git.

The RO-NF-MX dataset is publicly available and can be accessed at the link above. The other datasets are part of ongoing projects and cannot be shared with this study; however, they may be obtained from the corresponding author upon reasonable request.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. URL https://www.tensorflow.org/.

Abbasi, B., Rabelo, L., Hosseinkouchack, M., 2008. Estimating parameters of the three-parameter Weibull distribution using a neural network. Eur. J. Ind. Eng. 2 (4), 428–445.

Alessandrini, A., Biondi, F., Di Filippo, A., Ziaco, E., Piovesan, G., 2011. Tree size distribution at increasing spatial scales converges to the rotated sigmoid curve in two old-growth beech stands of the Italian apennines. Forest Ecol. Manag. 262 (11), 1950–1962.

Alrashidi, M., 2023. Estimation of weibull distribution parameters for wind speed characteristics using neural network algorithm. Comput. Mater. Contin. 75 (1), 1073–1088.

Bailey, R.L., Dell, T., 1973. Quantifying diameter distributions with the Weibull function. For. Sci. 19 (2), 97–104.

Bankston, J.B., Sabatia, C.O., Poudel, K.P., 2021. Effects of sample plot size and prediction models on diameter distribution recovery. For. Sci. 67 (3), 245–255.

Bohn, F.J., Huth, A., 2017. The importance of forest structure to biodiversity–productivity relationships. R. Soc. Open Sci. 4 (1), 160521.

Bolat, F., Ercanlı, İ., 2016. Using artificial neural network in describing diameter distribution in an even-aged forest. In: International Forestry Symposium (IFS 2016) Proceedings. pp. 07–10.

Borchani, H., Varando, G., Bielza, C., Larranaga, P., 2015. A survey on multi-output regression. Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 5 (5), 216–233.

Borders, B., Souter, R., Bailey, R., Ware, K., 1987. Percentile-based distributions characterize forest stand tables. For. Sci. 33 (2), 570–576.

Breiman, L., 2001a. Random forests. Mach. Learn. 45, 5–32.

Breiman, L., 2001b. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statist. Sci. 16 (3), 199–231.

Cai, S., Kang, X., Zhang, L., 2012. Simulation of trunk diameter distribution for three broadleaved trees based on artificial neural networks. Int. J. Adv. Comput. Technol. 4, 520–527.

Cao, Q.V., 2004. Predicting parameters of a Weibull function for modeling diameter distribution. For. Sci. 50 (5), 682–685.

Chen, Y., Wu, B., Min, Z., 2019. Stand diameter distribution modeling and prediction based on maximum entropy principle. Forests 10 (10), 859.

Chivulescu, S., Ciceu, A., Leca, S., Apostol, B., Popescu, O., Badea, O., 2020. Development phases and structural characteristics of the penteleu-viforâta virgin forest in the curvature carpathians. IForest-Biogeosciences For. 13 (5), 389.

Chollet, F., et al., 2015. Keras. https://keras.io.

Ciceu, A., Bălăcenoiu, F., De Groot, M., Chakraborty, D., Avtzis, D., Barta, M., Blaser, S., Bracalini, M., Castagneyrol, B., Chernova, U.A., et al., 2024. The ongoing range expansion of the invasive oak lace bug across europe: current occurrence and potential distribution under climate change. Sci. Total Environ. 949, 174950.

Ciceu, A., Chakraborty, D., Ledermann, T., 2023. Examining the transferability of height–diameter model calibration strategies across studies. For.: An Int. J. For. Res. cpad063.

Ciceu, A., Leca, Ş., Badea, O., Mehtätalo, L., 2025. Nonlinear multilevel seemingly unrelated height-diameter and crown length mixed-effects models for the southern transylvanian forests, Romania. For. Ecosyst..

Ciceu, A., Pitar, D., Badea, O., 2021. Modeling the diameter distribution of mixed uneven-aged stands in the south western carpathians in Romania. Forests 12 (7), 958.

Claveria, O., Monte, E., Torra, S., et al., 2015. Multiple-input multiple-output vs. single-input single-output neural network forecasting. Res. Inst. Appl. Econ. 1–29.

De'Ath, G., 2002. Multivariate regression trees: a new technique for modeling species–environment relationships. Ecology 83 (4), 1105–1117.

Diamantopoulou, M.J., Özçelik, R., Crecente-Campo, F., Eler, Ü., 2015. Estimation of Weibull function parameters for modelling tree diameter distribution using least squares and artificial neural networks methods. Biosyst. Eng. 133, 33–45.

Ercanli, Í., Bolat, F., 2017. Diameter distribution modeling based on artificial neural networks for kunduz forests. In: International Symposium on New Horizons in Forestry, Isparta, Turkey. pp. 18–20.

Feldmann, E., Glatthorn, J., Hauck, M., Leuschner, C., 2018. A novel empirical approach for determining the extension of forest development stages in temperate old-growth forests. Eur. J. For. Res. 137, 321–335.

Garcıa, O., 1992. What is a diameter distribution. In: Proceedings of the Symposium on Integrated Forest Management Information Systems. International Union of Forest Research Organizations Tsukuba, Japan, pp. 11–29.

Gorgoso-Varela, J., García-Villabrille, J., Rojo-Alboreca, A., 2015. Modeling extreme values for height distributions in pinus pinaster, pinus radiata and eucalyptus globulus stands in northwestern Spain. iforest-biogeosci forest 9: 23-29.

Gorgoso-Varela, J.J., Ogana, F.N., Alonso-Ponce, R., 2019. Evaluation of direct and indirect methods for modelling the joint distribution of tree diameter and height data with the bivariate johnson's SBB function to forest stands. For. Syst. 28 (1), 1–10.

Haara, A., Maltamo, M., Tokola, T., 1997. The K-nearest-neighbour method for estimating basal-area diameter distribution. Scand. J. For. Res. 12 (2), 200–208.

Hamedianfar, A., Mohamedou, C., Kangas, A., Vauhkonen, J., 2022. Deep learning for forest inventory and planning: a critical review on the remote sensing approaches so far and prospects for further applications. Forestry 95 (4), 451–465.

Henningsen, A., Hamann, J.D., 2007. Systemfit: A package for estimating systems of simultaneous equations in r. J. Stat. Softw. 23 (4), 1–40, URL https://www.jstatsoft.org/v23/i04/.

Hobiger, J., Laa, U., Vospernik, S., 2025. Comparing traditional methods and modern statistical techniques for tree height prediction. Forests 16 (2), 271.

Horodnic, S.A., Roibu, C.C., 2018. A Gaussian multi-component model for the tree diameter distribution in old-growth forests. Eur. J. For. Res. 137 (2), 185–196.

Huy, B., Truong, N.Q., Poudel, K.P., Temesgen, H., Khiem, N.Q., 2024. Multi-output deep learning models for enhanced reliability of simultaneous tree above-and below-ground biomass predictions in tropical forests of Vietnam. Comput. Electron. Agric. 222, 109080.

Hyink, D.M., Moser, J.W., 1983. A generalized framework for projecting forest yield and stand structure using diameter distributions. For. Sci. 29 (1), 85–95.

Ij, H., 2018. Statistics versus machine learning. Nat Methods 15 (4), 233.

Ilvessalo, Y., 1932. The establishment and measurement of permanent sample plots in suomi (Finland).. Commun. Instituti For. Fenn. 17 (2), 1–39.

Janowiak, M.K., Nagel, L.M., Webster, C.R., 2008. Spatial scale and stand structure in northern hardwood forests: implications for quantifying diameter distributions. For. Sci. 54 (5), 497–506.

Kangas, A., Maltamo, M., 2000. Performance of percentile based diameter distribution prediction and Weibull method in independent data sets. Silva Fenn. 34 (4), 381–398.

Kangas, A., Mehtatalo, L., Maltamo, M., 2007. Modelling percentile based basal area weighted diameter distribution. Silva Fenn. 41 (3), 425.

Knowe, S.A., Ahrens, G.R., DeBell, D.S., 1997. Comparison of diameter-distribution-prediction, stand-table-projection, and individual-tree-growth modeling approaches for young red alder plantations. Forest Ecol. Manag. 98 (1), 49–60.

Kocev, D., Džeroski, S., White, M.D., Newell, G.R., Griffioen, P., 2009. Using single- and multi-target regression trees and ensembles to model a compound index of vegetation condition. Ecol. Model. 220 (8), 1159–1168.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

Leduc, D.J., 2001. Predicting diameter distributions of longleaf pine plantations: a comparison between artificial neural networks and other accepted methodologies, vol. 25, US Department of Agriculture, Forest Service, Southern Research Station.

Liu, S., Liu, Y., Xia, R., 2024. Using random forest to disentangle the effects of environmental conditions on height-to-diameter ratio of engelmann spruce. New For. 55 (2), 213–229.

Mäkinen, H., Isomäki, A., 2004a. Thinning intensity and growth of Norway spruce stands in Finland. Forestry 77 (4), 349–364.

Mäkinen, H., Isomäki, A., 2004b. Thinning intensity and growth of scots pine stands in Finland. Forest Ecol. Manag. 201 (2–3), 311–325.

Makridakis, S., Spiliotis, E., Assimakopoulos, V., 2018. Statistical and machine learning forecasting methods: Concerns and ways forward. PloS One 13 (3), e0194889.

Maltamo, M., Kangas, A., 1998. Methods based on k-nearest neighbor regression in the prediction of basal area diameter distribution. Can. J. Forest Res. 28 (8), 1107–1115.

Maltamo, M., Kangas, A., Uuttera, J., Torniainen, T., Saramäki, J., 2000. Comparison of percentile based prediction methods and the Weibull distribution in describing the diameter distribution of heterogeneous scots pine stands. Forest Ecol. Manag. 133 (3), 263–274.

Mastelini, S.M., Santana, E.J., da Costa, V.G.T., Barbon, S., 2018. Benchmarking multi-target regression methods. In: 2018 7th Brazilian Conference on Intelligent Systems. BRACIS, IEEE, pp. 396–401.

McCullagh, P., 2002. What is a statistical model? Ann. Statist. 30 (5), 1225–1310.

McCulloch, W.S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. 5, 115–133.

Mehtätalo, L., 2004. Predicting stand characteristics using limited measurements. (Dissertation). Finnish Forest Research Institute, Research Papers 929.

Mehtätalo, L., 2005. Localizing a predicted diameter distribution using sample information. For. Sci. 51 (4), 292–303.

Mehtätalo, L., Gregoire, T.G., Burkhart, H.E., 2008. Comparing strategies for modeling tree diameter percentiles from remeasured plots. Environmetrics: Off. J. Int. Environmetrics Soc. 19 (5), 529–548.

Mehtätalo, L., Lappi, J., 2020. Biometry for forestry and environmental data: With examples in R. Chapman and Hall/CRC.

Merganič, J., Sterba, H., 2006. Characterisation of diameter distribution using the Weibull function: method of moments. Eur. J. For. Res. 125, 427–439.

Mielikäinen, K., 1979. The influence of low thinnings on the wood production and value of a pine stand. Folia For. (Finland) 401.

Myllymäki, M., Tuominen, S., Kuronen, M., Packalen, P., Kangas, A., 2024. The relationship between forest structure and naturalness in the finnish national forest inventory. For.: Int. J. For. Res. 97 (3), 339–348.

Nyyssonen, A., 1950. Comparative observations on the structure and development of tended and natural pine stands.. Silva Fenn. 68, 1–48.

Ogana, F.N., 2022. Does the inclusion of truncation point in a finite mixture model improve diameter distribution estimation of degraded stand? J. Sustain. For. 41 (1), 77–91.

Ogana, F.N., Ercanli, I., 2022. Modelling height-diameter relationships in complex tropical rain forest ecosystems using deep learning algorithm. J. For. Res. 33 (3), 883–898.

Pach, M., Podlaski, R., 2015. Tree diameter structural diversity in central European forests with abies alba and fagus sylvatica: managed versus unmanaged forest stands. Ecol. Res. 30, 367–384.

Palahí, M., Pukkala, T., Blasco, E., Trasobares, A., 2007. Comparison of beta, johnson's SB, Weibull and truncated Weibull functions for modeling the diameter distribution of forest stands in catalonia (north-east of Spain). Eur. J. For. Res. 126, 563–571.

Pichler, M., Hartig, F., 2023. Machine learning and deep learning-a review for ecologists. Methods Ecol. Evol. 14 (4), 994–1016.

Pogoda, P., Ochał, W., Orzeł, S., 2019. Modeling diameter distribution of black alder (alnus glutinosa (l.) gaertn.) stands in Poland. Forests 10 (5), 412.

Qamar, W., Hussain, M., Zaheer, M.B., Akram, J., Sadiq, N., Uddin, Z., 2025. Prediction of sunspot numbers via Weibull distribution and deep learning. Astrophys. Space Sci. 370 (7), 1–9.

Rajput, D., Wang, W.-J., Chen, C.-C., 2023. Evaluation of a decided sample size in machine learning applications. BMC Bioinformatics 24 (1), 48.

Siipilehto, J., Mehtätalo, L., 2013. Parameter recovery vs. parameter prediction for the Weibull distribution validated for scots pine stands in Finland. Silva Fenn. 47 (4), 1–22.

Silva, C.A., Klauberg, C., Hudak, A.T., Vierling, L.A., Jaafar, W.S.W.M., Mohan, M., Garcia, M., Ferraz, A., Cardil, A., Saatchi, S., 2017. Predicting stem total and assortment volumes in an industrial pinus taeda l. forest plantation using airborne laser scanning data and random forest. Forests 8 (7), 254.

Stankova, T.V., Zlatanov, T.M., 2010. Modeling diameter distribution of Austrian black pine (pinus nigra arn.) plantations: a comparison of the Weibull frequency distribution function and percentile-based projection methods. Eur. J. For. Res. 129, 1169–1179.

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res.: Atmospheres 106 (D7), 7183–7192.

Team, R.C., 2025. R: A language and environment for statistical computing. vienna, austria. URL https://www.R-project.org/.

Tran, N.K., Kühle, L.C., Klau, G.W., 2024. A critical review of multi-output support vector regression. Pattern Recognit. Lett. 178, 69–75.

Van Rossum, G., Drake Jr., F.L., 1995. Python Reference Manual. Centrum voor Wiskunde en Informatica Amsterdam.

Vázquez-Veloso, A., Caicoya, A.T., Bravo, F., Biber, P., Uhl, E., Pretzsch, H., 2025. Does machine learning outperform logistic regression in predicting individual tree mortality? Ecol. Informatics 103140.

Vuokila, Y., 1962. The effect of thinnings on the yield of pine and birch stands.. Comm Inst Fenn 55, 12.

Wang, M., Rennolls, K., 2005. Tree diameter distribution modelling: introducing the logit logistic distribution. Can. J. Forest Res. 35 (6), 1305–1313.

Weiskittel, A.R., Hann, D.W., Kershaw Jr., J.A., Vanclay, J.K., 2011. Forest Growth and Yield Modeling. John Wiley & Sons.

Xi, X., Sheng, V.S., Sun, B., Wang, L., Hu, F., 2018. An empirical comparison on multi-target regression learning.. Comput. Mater. Contin. 56 (2).

Yang, S.-I., Brandeis, T.J., Helmer, E.H., Marcano-Vega, H., 2025. Predicting species-specific diameter growth rate for caribbean trees using mixed-effects extreme gradient boosting. Forest Ecol. Manag. 580, 122520.

Yang, S.-I., Brandeis, T.J., Helmer, E.H., Oatham, M.P., Heartsill-Scalley, T., Marcano-Vega, H., 2022. Characterizing height-diameter relationships for caribbean trees using mixed-effects random forest algorithm. Forest Ecol. Manag. 524, 120507.

Yen, T.-M., Ji, Y.-J., Lee, J.-S., 2010. Estimating biomass production and carbon storage for a fast-growing makino bamboo (phyllostachys makinoi) plant based on the diameter distribution model. Forest Ecol. Manag. 260 (3), 339–344.

Zellner, A., 1962. An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. J. Amer. Statist. Assoc. 57 (298), 348–368.

Zhang, L., Gove, J.H., Liu, C., Leak, W.B., 2001. A finite mixture of two Weibull distributions for modeling the diameter distributions of rotated-sigmoid, uneven-aged stands. Can. J. Forest Res. 31 (9), 1654–1659. http://dx.doi.org/10.1139/x01-086.