Forestry: An International Journal of Forest Research, 2024, **97**, 512–530
https://doi.org/10.1093/forestry/cpad061
Advance access publication date 4 December 2023
Original Article

Assessing the potential of synthetic and *ex situ* airborne laser scanning and ground plot data to train forest biomass models

Jannika Schäfer[1,*], Lukas Winiwarter[2,3,4], Hannah Weiser[2], Jan Novotný[5], Bernhard Höfle[2,6], Sebastian Schmidtlein[1], Hans Henniger[7], Grzegorz Krok[8], Krzysztof Stereńczak[8] and Fabian Ewald Fassnacht[9]

[1]Institute of Geography and Geoecology (IFGG), Karlsruhe Institute of Technology (KIT), Kaiserstraße 12, 76131 Karlsruhe, Germany
[2]3DGeo Research Group, Institute of Geography, Heidelberg University, Im Neuenheimer Feld 368, 69120 Heidelberg, Germany
[3]Department of Forest Resources Management, Faculty of Forestry, University of British Columbia, 2424 Main Mall, Vancouver, BC V6T 1Z4, Canada
[4]Photogrammetry Research Area, Department of Geodesy and Geoinformation, TU Wien, Wiedner Hauptstraße 8-10, 1040 Wien, Austria
[5]Global Change Research Institute of the Czech Academy of Sciences, Bělidla 986/4a, 603 00 Brno, Czech Republic
[6]Interdisciplinary Center for Scientific Computing (IWR), Heidelberg University, Im Neuenheimer Feld 205, 69120 Heidelberg, Germany
[7]Department of Ecosystem Analysis, Helmholtz Centre for Environmental Research (UFZ), Permoserstraße 15, 04318 Leipzig, Germany
[8]Department of Geomatics, Forest Research Institute, Sekocin Stary, 3 Braci Leśnej St., 05-090 Raszyn, Poland
[9]Remote Sensing and Geoinformatics, Freie Universität Berlin, Malteserstraße 74-100, 12249 Berlin, Germany
[*]Corresponding author: jannika.schaefer@kit.edu

Abstract

Airborne laser scanning data are increasingly used to predict forest biomass over large areas. Biomass information cannot be derived directly from airborne laser scanning data; therefore, field measurements of forest plots are required to build regression models. We tested whether simulated laser scanning data of virtual forest plots could be used to train biomass models and thereby reduce the amount of field measurements required. We compared the performance of models that were trained with (i) simulated data only, (ii) a combination of simulated and real data, (iii) real data collected from different study sites, and (iv) real data collected from the same study site the model was applied to. We additionally investigated whether using a subset of the simulated data instead of using all simulated data improved model performance. The best matching subset of the simulated data was sampled by selecting the simulated forest plot with the highest correlation of the return height distribution profile for each real forest plot. For comparison, a randomly selected subset was evaluated. Models were tested on four forest sites located in Poland, the Czech Republic, and Canada. Model performance was assessed by root mean squared error (RMSE), squared Pearson correlation coefficient (r^2), and mean error (ME) of observed and predicted biomass. We found that models trained solely with simulated data did not achieve the accuracy of models trained with real data (RMSE increase of 52–122 %, r^2 decrease of 4–18 %). However, model performance improved when only a subset of the simulated data was used (RMSE increase of 21–118 %, r^2 decrease of 5–14 % compared to the real data model), albeit differences in model performance when using the best matching subset compared to using a randomly selected subset were small. Using simulated data for model training always resulted in a strong underprediction of biomass. Extending sparse real training datasets with simulated data decreased RMSE and increased r^2, as long as no more than 12–346 real training samples were available, depending on the study site. For three of the four study sites, models trained with real data collected from other sites outperformed models trained with simulated data and RMSE and r^2 were similar to models trained with data from the respective sites. Our results indicate that simulated data cannot yet replace real data but they can be helpful in some sites to extend training datasets when only a limited amount of real data is available.

Introduction

The accurate estimation of forest biomass is essential for quantifying carbon stocks and fluxes at local to global scales (Dixon *et al.*, 1994). One data source for predicting aboveground biomass across larger areas is airborne laser scanning (ALS) (McRoberts *et al.*, 2015). ALS is increasingly used for forest inventories (Achim *et al.*, 2022), including biomass inventories, because it allows the collection of forest structure information in large areas (Moudrý *et al.*, 2023). ALS cannot measure biomass directly, but metrics derived from ALS point clouds can be used as predictors in empirical models with biomass as response. Accordingly, additional biomass reference data are required to train often applied supervised machine-learning models (Hawbaker *et al.*, 2009). In the area-based approach (ABA), plot-based field measurements of biomass are linked to metrics derived from ALS point clouds extracted from the same plots to build a model that can then be used to predict biomass of the entire area covered by ALS data, resulting in a wall-to-wall map of biomass predictions (White *et al.*, 2013). The number, size, shape, and geolocation accuracy of the field plots affect the accuracy of the biomass predictions. According to earlier studies, the accuracy increases with a greater number of field plots, larger plot sizes, plot shapes with a smaller perimeter-to-area ratio, and smaller geolocation errors (Gobakken & Næsset, 2008, Frazer *et al.*, 2011, Lisańczuk *et al.*, 2020, Packalen *et al.*, 2023). However, field measurements are time consuming and costly, especially when field plots are remote or difficult

Handling editor: Dr. Tzeng Yih Lam
Received 7 May 2023. **Revised** 7 November 2023. **Accepted** 9 November 2023

to access (Hawbaker *et al.*, 2009, Rana *et al.*, 2016), and dense canopy and complex topography can reduce Global Navigation Satellite System positioning accuracy (Næsset & Gjevestad, 2008, Dalponte *et al.*, 2011). For cost reasons, it is hence of interest to keep the number of field plots small (Gobakken & Næsset, 2009, Dalponte *et al.*, 2011, Stereńczak *et al.*, 2018). At the same time, it is important to ensure that the field plots represent the full range of biomass values and corresponding ALS metrics of the study area to minimise extrapolation (Dalponte *et al.*, 2011, Maltamo *et al.*, 2011, Fekety *et al.*, 2015).

There are several approaches to optimising the number of field plots and the associated workload of field measurements. One option is the re-use of field and ALS data that have been collected at an earlier (or later) time or in a different location. If both field and ALS data from one time are available for the study area, these can be used to build a model that can be applied to the ALS data acquired at another time, provided the distribution of the extracted metrics is the same. Alternatively, if field and ALS data are available from different times, growth models can be used to project the field data to the year of the ALS data (Domingo *et al.*, 2019, Lera Garrido *et al.*, 2020). Temporal transferability of ALS-based biomass models has been demonstrated in several studies (Fekety *et al.*, 2015, Zhao *et al.*, 2018, Domingo *et al.*, 2019, de Lera Garrido *et al.*, 2020). However, the temporally transferred models often performed worse than models trained with field and ALS data acquired at the same time (Domingo *et al.*, 2019, de Lera Garrido *et al.*, 2020).

Spatial model transfer requires similar forest conditions in the region where the model was trained and in the region where it is to be applied, as the relationship between ALS-derived metrics and biomass may differ between regions (Næsset & Gobakken, 2008, Tompalski *et al.*, 2019). Studies evaluating the performance of models trained with data collected over a larger area (e.g. national models) to predict local forest parameters have found that additional model calibration with a small set of local data improves model performance for local predictions (Breidenbach *et al.*, 2008, Kotivuori *et al.*, 2016, van Ewijk *et al.*, 2020) and that even models trained with only 50 local training sample plots can perform better than models trained with many more training data collected from other areas (Suvanto & Maltamo, 2010).

A major drawback of the spatial and the temporal model transfers is that, although they reduce the number of new field observations, they still require data of a forest with a similar structure, or of the same area at different times. Such data may not always be available. Another promising approach to minimise field work is to reduce the number of field plots by using a stratified sampling method to select plots (Goodbody *et al.*, 2023). When comparing stratified sampling with random sampling, several studies have shown that stratified sampling based on ALS-derived metrics gives more stable results and higher model accuracy than random sampling of field plot locations (Hawbaker *et al.*, 2009, Dalponte *et al.*, 2011, Maltamo *et al.*, 2011). Stratified sampling can be used to find the minimum number of field plots that still cover the full range of forest structural variability that can be inferred from ALS data. For example, Dalponte *et al.* (2011) obtained almost the same accuracy for the prediction of stem volume when using 53 field plots that had been selected based on the mean height of the ALS returns, compared to a model using all available 534 field plots.

The easiest way around the need for field measurements would be to generate training data simply by computer simulation. For simulating ALS data of forests, a laser scanning simulation approach can be applied to a 3D model of a virtual forest. Existing methods differ in terms of the complexity of both the forest representation and the laser scanning simulation approach. Trees can either be represented by simple geometric objects, such as cones, spheres, and cylinders (e.g. Nelson, 1997, Frazer *et al.*, 2011, Palace *et al.*, 2015, Knapp *et al.*, 2018), by more detailed, realistically rendered tree models as created by the OnyxTREE software (https://www.onyxtree.com, e.g. Disney *et al.*, 2010), or by tree models extracted from high-resolution real laser scanning data (e.g. Fassnacht *et al.*, 2018, Schäfer *et al.*, 2023). Laser scanning can be simulated using simplified statistical models (e.g. Nelson, 1997, Wang *et al.*, 2013, Palace *et al.*, 2015, Spriggs *et al.*, 2015, Knapp *et al.*, 2018), or using computationally more intensive approaches (e.g. Holmgren *et al.*, 2003, Disney *et al.*, 2010, Roberts *et al.*, 2020, Zhu *et al.*, 2020, Schäfer *et al.*, 2023) that allow the simulation of the scanning process itself and thus the effects of different laser scanning acquisition settings. The latter include for example the Discrete Anisotropic Radiative Transfer (DART) model (Gastellu-Etchegorry *et al.*, 2016, Yin *et al.*, 2016) and the Heidelberg LiDAR Operations Simulator [HELIOS++, Winiwarter *et al.*, 2022). Computer simulations are a time- and cost-efficient way to generate large amounts of laser scanning data and associated field data. They allow control of both the laser scanning acquisition settings and the forest composition (Frazer *et al.*, 2011). In addition, the location and properties of each tree in the virtual forest are known. These data offer therefore plenty of opportunities for sensitivity analyses as well as method development that are much more difficult to perform with real data (Disney *et al.*, 2010). Accordingly, simulated ALS data of forests have been used for a variety of applications, e.g. to assess the influence of laser scanning acquisition settings on ALS-derived structural parameters, such as canopy height and canopy closure (Holmgren *et al.*, 2003, Disney *et al.*, 2010, Roberts *et al.*, 2020), or to analyse the influence of field plot size and co-registration error on ALS-based biomass predictions (Frazer *et al.*, 2011, Fassnacht *et al.*, 2018). Simulated ALS data have also been used to validate methods for tree delineation (Wang *et al.*, 2013) and effective leaf area index estimation (Zhu *et al.*, 2020), and to find the best ALS-derived metrics for biomass predictions (Knapp *et al.*, 2018). Some studies also explored the potential of simulated data to derive predictive equations or look-up tables for relating forest structural parameters to ALS data (Nelson *et al.*, 1997, Palace *et al.*, 2015, Spriggs *et al.*, 2015).

Schäfer *et al.* (2023) demonstrated that HELIOS++ laser scanning simulations of virtual stands composed of real laser scanning tree point clouds can produce data that are sufficiently realistic for training biomass models, even if the prediction accuracy was lower than for models trained with real data. They used real forest inventory data to generate the virtual stands, which strongly limits the number of the synthetic forest plots. In this study, we overcome this limitation by creating the virtual stands based on simulated forest compositions. Our main aim was to explore the potential of such synthetic ALS and forest inventory datasets to reduce the amount of field reference data required for the laser scanning-based prediction of forest aboveground biomass. We conducted three experiments. In the first two experiments, we trained biomass models with (i) simulated data only, and (ii) mixed sets of simulated and real data. In the third experiment, we tested a spatial model transfer and trained biomass models with real *ex situ* data, i.e. real ALS and field data collected from other sites. Model performance was always evaluated on real data, and compared with models trained with real *in situ* data that were excluded from the evaluation. Our objective was to answer the following research questions using datasets obtained

from study sites located in Poland, the Czech Republic, and Canada:

1. How accurately can random forest models that have been trained with simulated forest inventory and virtual laser scanning data predict biomass of real forest sites compared to models that have been trained with real data collected at the same site (Experiment 1) or at different sites (Experiment 3)?
2. When there are little real training data available, can model accuracy be improved by extending real training datasets with synthetic data? If so, up to what number of real training samples does a model trained with additional synthetic data outperform a model trained with real data only (Experiment 2)?

Materials and methods

Study sites

We tested our approach using a total of four real datasets obtained from the Milicz forest district in Poland, the Silesian Beskids (Těšínské Beskydy) forest in the Czech Republic, the DendroNET sites in the Czech Republic, and the Petawawa Research Forest in Canada.

The Milicz forest district is located in the south-west of Poland. The dominant tree species is Scots pine (*Pinus sylvestris* L.), accompanied by European beech (*Fagus sylvatica* L.) and oaks (*Quercus* spp. L.). Approximately 70 % of the forest stands are pure pine stands (Stereńczak *et al.*, 2018). Field reference data were collected in summer 2015 for 500 circular plots (Stereńczak *et al.*, 2018). ALS data were acquired at the same time under leaf-on conditions.

The Silesian Beskids are a mountain range in southern Poland and eastern Czech Republic. Data were collected in the Czech part. The forest there is dominated by Norway spruce (*Picea abies* (L.) H. Karst) and European beech. ALS data and field data were collected in July 2019 for 130 plots. Study site and data have been described in more detail by Brovkina *et al.* (2022).

The DendroNET (http://dendronet.cz) is a network of small forest sites located across the Czech Republic. 47 plots were used in this study, 22 of them are located in spruce forest, 10 in pine forest, 12 in beech forest, and 3 in mixed forest. Field data were collected for each tree within a 30 m × 30 m square. ALS data were acquired in October 2021.

The Petawawa Research Forest is located in the Great Lakes–St. Lawrence Forest region in southern Ontario, Canada. The most frequent tree species are white pine (*Pinus strobus* L.), trembling aspen (*Populus tremuloides* Michx.), red oak (*Quercus rubra* L.), red pine (*Pinus resinosa* Ait.), white birch (*Betula papyrifera* Marsh), maple (*Acer* spp.), and white spruce (*Picea glauca* (Moench) Voss) (Wetzel *et al.*, 2011). Several remotely sensed and ancillary datasets are available for this remote sensing supersite (https://opendata.nfis.org/mapserver/PRF.html). A summary of the open-access datasets can be found in White *et al.* (2019). Here, we used the ALS data of 2012. Field measurements were conducted in 2014 in 223 circular plots (White *et al.*, 2019). The field data collection is described in the Field Procedures Manual which is provided with the data. Table 1 gives an overview of the laser scanning acquisition settings and resulting mean pulse densities and mean planar point densities of all study sites.

Simulated data

Simulated data were generated by applying the HELIOS++ laser scanning simulator to simulated forest stands. For simulating forest compositions, we used Forest Factory 2.0, a forest generator based on the forest gap model FORMIND (Bohn & Huth, 2017, Henniger *et al.*, 2023). Forest Factory generates virtual forest stands with different species composition and structure, without taking into account the forest development over time. This reduces the computational time compared to forest growth simulators such as FORMIND (Fischer *et al.*, 2016) or SILVA (Pretzsch *et al.*, 2002). Forest Factory is initialised with a region-specific parameterisation, i.e. species pool and productivity. Additionally, an initial minimum and maximum tree height and an initial maximum stand density is set. Each forest stand is created tree by tree. First, a stand-specific height range and species pool is randomly selected from the initial height range and species pool. Then, for each tree that is to be placed in the forest stand, tree species and height are randomly sampled from the stand-specific species pool and height range. Trees are added until no tree with a positive annual productivity (photosynthetic production is higher than respiration) can be placed, or until there is no canopy space for the tree that is to be placed. Stands are limited to a size of 20 m × 20 m (Henniger *et al.*, 2023). In our study, we used Forest Factory to simulate 2500 plots of 400 m^2 each with different compositions of pine, spruce, beech, and oaks. Forest Factory has also been calibrated for more tree species and plant functional types (Bohn & Huth, 2017, Bruening *et al.*, 2021, Henniger *et al.*, 2023), but we excluded these from forest simulations because there were only few or no tree point clouds of these species available (see next paragraph). The maximum tree height of each species was set to be 5 m higher than the maximum height of the available tree point clouds of this species.

Virtual 3D representations of the Forest Factory stands were created by making use of individual tree point clouds that had been extracted from laser scanning data acquired by an uncrewed aerial vehicle (UAV) under leaf-on conditions in temperate forests in southwestern Germany. These individual tree point clouds have been published by Weiser *et al.* (2022b), they can be downloaded from https://pytreedb.geog.uni-heidelberg.de/. A description of the tree point cloud dataset can be found in Weiser *et al.* (2022a). Here, we only used tree point clouds having a high to medium segmentation quality (q1–q3). The segmentation quality score ranges from high (q1) to low (q6). It is a subjective measure of the probability of segmentation and extraction errors that was assigned by the person who manually extracted the tree point cloud. After applying the filtering criteria, there were 102 tree point clouds for pine, 191 tree point clouds for spruce, 345 tree point clouds for beech, and 154 tree point clouds for oaks available.

For each tree in the Forest Factory stands, a point cloud of a tree was selected randomly from all tree point clouds of that species with a height ± 4 m the height specified by Forest Factory. If there was no point cloud of a tree of matching height available, the point cloud of the tree with the smallest height difference was selected. This filtering procedure is similar to the one applied in Schäfer *et al.* (2023) except that we omit the crown diameter filter here. The tree point cloud was scaled along the Z-axis so that the height of the point cloud matched the height of the Forest Factory tree. It was randomly rotated around the Z-axis and placed at the location of the tree in the Forest Factory stand. For simulating laser scanning, the 20 m × 20 m Forest Factory stands were arranged to larger scenes of 100 m × 100 m (1 ha). To prevent border effects of tree crowns reaching into neighbouring stands, the composite of the tree point clouds of a stand was clipped to the stand boundaries. The resulting forest point clouds were converted into opaque voxels with 3 cm side length to create 3D

Table 1. Laser scanning acquisition settings and resulting mean pulse density and mean planar point density

	Milicz Forest	Silesian Beskids	DendroNET	Petawawa Research Forest
Sensor	*RIEGL* LMS-Q680i	*RIEGL* LMS-Q780	*RIEGL* LMS-Q780	*RIEGL* LMS-Q680i
Laser beam divergence[a]	0.5 mrad	0.25 mrad	0.25 mrad	0.5 mrad
Pulse repetition frequency	360 kHz[b] [300 kHz]	400 kHz	400 kHz	150 kHz
Scan frequency	? [140 lines/s]	125 lines/s	160 lines/s	76.67 lines/s
Scan angle off nadir	±30°	±30°	±30°	±20°
Altitude above ground	480–620 m [480 m]	819 m	515 m	750 m
Flight speed	54 m/s	56 m/s	56 m/s	? [54 m/s]
Flight line distance	? [≈ 296 m]	440 m	–	250 m [≈ 242 m]
Flight pattern	Parallel[c]	Parallel[c]	Perpendicular[c]	Parallel[c]
Mean pulse density	9.4 pulses/m^2 [12.4 pulses/m^2]	7.0 pulses/m^2 [9.7 pulses/m^2]	13.2 pulses/m^2 [18.0 pulses/m^2]	5.4 pulses/m^2 [4.3 pulses/m^2]
Mean planar point density	19.9 points/m^2 [24.4 points/m^2]	12.8 points/m^2 [17.8 points/m^2]	23.9 points/m^2 [31.7 points/m^2]	11.7 points/m^2 [8.3 points/m^2]

Numbers in square brackets indicate values of the simulations differing from reported values of the real acquisitions.[a]Measured at the 1/e^2 points.[b]The reported pulse repetition frequency was 360 kHz, but simulations with a pulse repetition frequency of 300 kHz matched better to the real point clouds.[c]The simulations were performed with flight strips that were not perfectly parallel/perpendicular to reflect deviations from the flight pattern in the real data.

voxel scenes as input for the simulations. A horizontal plane was added as ground layer.

HELIOS++ has been validated with DART (Winiwarter *et al.*, 2022) and was already successfully applied to simulate ALS data of synthetic forest stands composed of real ULS tree point clouds (Schäfer *et al.*, 2023), which is why we used HELIOS++ for the laser scanning simulations. HELIOS++ allows simulating full waveform and discrete return laser scanning. Beam divergence is modelled by subrays of different base power. The returned waveforms of all subrays are binned and summed up to generate the full waveform. On this waveform, a local maximum filter is employed to detect return points. The simulations are configured by the scene to be scanned, the scanner parameters and the position and movement of the platform on which the virtual scanner is mounted. Additional parameters such as the temporal window size for echo detection, and the number of generated subrays can be defined (Winiwarter *et al.*, 2022). We conducted laser scanning simulations with the same acquisition settings as in the real acquisitions (Table 1), resulting in four different simulated laser scanning datasets. In case of unknown acquisition settings, different values were tested, and the best approximation was selected based on comparisons of the resulting point patterns of simulated and real point clouds. The simulations were performed with a temporal window size of 1 ns in the local maximum filter, for the number of subrays the default value was used (beamSampleQuality = 3). As full waveform data from the real study sites were not available, we simulated only discrete, albeit multiple return point clouds and not full waveform data for the virtual forest stands.

Extraction of biomass reference data and predictor variables

Biomass reference data and ALS metrics were extracted from the real-world datasets and the Forest Factory datasets. The available data from the four study sites and the simulated data differed in plot shape and size, sampling strategy (measurement of all trees or sampling based on the diameter at breast height, $D_{1.3}$), and whether individual tree positions were included in the data. Therefore, data were extracted in ways specific to the study sites (Fig. 1).

Individual tree biomass values of the Forest Factory stands were calculated based on $D_{1.3}$ and height using species-specific allometric equations that were developed for the German National Forest Inventory, available in the R package "rBDAT" (Vonderach *et al.*, 2021).

The Milicz Forest field data included information on species, $D_{1.3}$, height, and location of every tree within a radius of 12.62 m from the plot centre. Biomass values of the individual trees were also calculated using the allometric equations of the German National Forest Inventory, assuming that the allometry of trees in Poland and Germany does not differ significantly. As the Forest Factory plots are squares of 20 m side length, the Milicz Forest field plots and the Forest Factory plots do not overlap perfectly. Therefore, simulated and real data were extracted from the largest square area that fits into both plot shapes (17.8 m × 17.8 m). The biomass of all trees located within this subplot was summed and divided by the area to derive estimates in t/ha.

In the Silesian Beskids, field data had been collected in nested plots with a maximum radius of 12.62 m. Due to the $D_{1.3}$-dependent sampling design, data had not been recorded for all trees within the plots. Accordingly, only plot-based estimates of biomass were available. Because of the smaller plot size of the simulated data, these could not be cropped to the plot shape of the real data. For simplicity, it was assumed that the provided biomass estimates (in t/ha) of the 500 m^2 circular Silesian Beskids plots were also representative for 20 m × 20 m square plots. For these plots, ALS metrics were extracted from the real and simulated data, and biomass reference values were derived from the Forest Factory stand information.

The DendroNET dataset included individual tree locations and biomass values derived from species-specific allometric equations. The data were cropped to 20 m × 20 m square plots. This allowed biomass reference values and ALS metrics to be derived from plots of the same shape and size for both the real and simulated data.

For the Petawawa Research Forest, individual tree biomass predictions were available, but not tree locations, so biomass could only be calculated for the entire 625 m^2 plots. As for the Silesian Beskids dataset, we assumed these area-based biomass estimates to be representative for the forest stands at the plot locations and extracted ALS metrics and the Forest Factory biomass reference values from 20 m × 20 m square plots.

ALS metrics were calculated for each plot using the R package "lidR" (Roussel *et al.*, 2020, Roussel & Auty, 2021). The point clouds were cropped to the plot extents and the height values were normalised using the *normalize_height()* function included in the

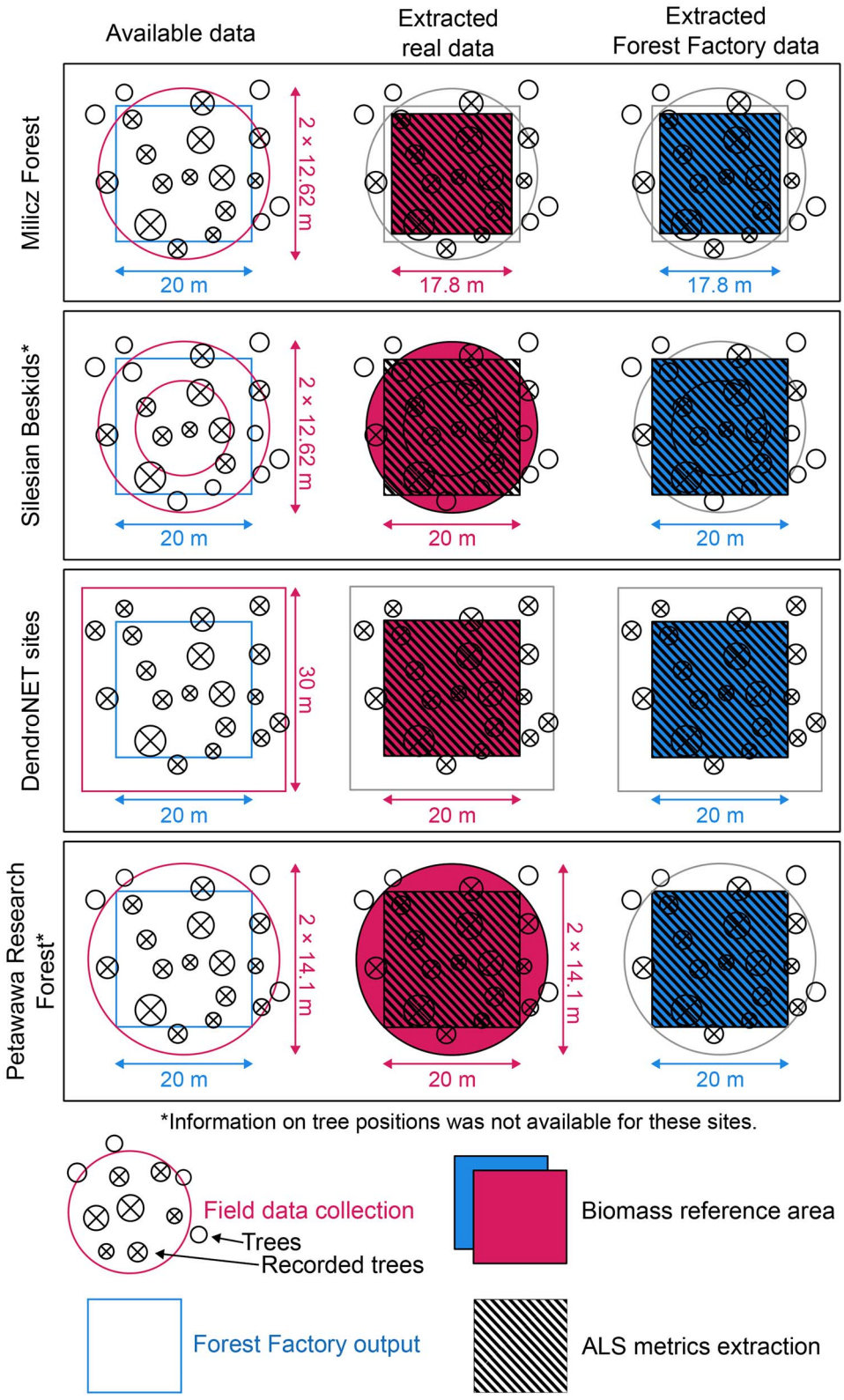

Figure 1. Schematic overview of field plot sizes and shapes of the four study sites in comparison to the Forest Factory plots (left). Areas from which biomass reference data were collected and areas from which ALS metrics were extracted are highlighted by colour and stripe pattern, respectively, for the real-world data (centre) and the Forest Factory data (right).

"lidR" package. For both all returns and first returns, the following metrics were calculated from the return heights: the maximum height, the mean height, the standard deviation, the skewness, the kurtosis, and the entropy of the height distribution, the percentage of returns above the mean height, the percentage of returns above 2 m, the 5^{th}–95^{th} (in steps of 5) height percentiles, and the cumulative percentage of returns in the 1^{st}–9^{th} layer. In addition, the number of returns, the percentage

of the first to fifth returns, the percentage of ground returns, the sum of intensities of the returns, the maximum intensity, the standard deviation of intensity, the skewness and kurtosis of the intensity distribution, the percentage of intensity of ground returns, and the percentage of intensity returned below the 10^{th}–90^{th} (in steps of 10) percentile of height were computed. Because the number of returns and the intensity-related metrics differed significantly between simulated and real data, these metrics were excluded from models trained with simulated data. To ensure comparability of the models, we did not perform any feature selection (e.g. based on feature importance) during modelling.

Subsampling of simulated training data

The Forest Factory stands covered a wide range of forest structures, with stem numbers ranging from 1 to 585 trees per 20 m × 20 m plot and biomass values ranging from 0.26 to 1323.81 t/ha, including plots with low tree numbers and low biomass, low tree numbers and high biomass, high tree numbers and low biomass, and high tree numbers and high biomass. The use of all 2500 Forest Factory plots as training data may hence result in less appropriate biomass models for study sites with less structural diversity. In addition, using all data leads to increased computing times. Therefore, we tested a sampling approach to reduce the amount of simulated training data while maintaining or even improving model performance when applied to real data.

For this, in each study site, the simulated data were filtered for the Forest Factory plots that were best matching the real data, assuming that only ALS-derived information were available. For both the real plots and the simulated plots, the relative number of ALS returns per 1 m height bin for the height of 0–50 m above ground was calculated to derive height distribution profiles of the returns. Adapting a waveform matching approach that was developed to compare simulated and real full-waveform light detection and ranging (LiDAR) data (Blair & Hofton, 1999, Hancock *et al.*, 2019, Lang *et al.*, 2022), the Pearson correlation coefficients between all simulated and all real height distribution profiles were calculated. For each real plot, the Forest Factory plot with the highest Pearson correlation coefficient was selected. This resulted in a subset of simulated data equal in size to the number of real plots, or smaller if a simulated plot was the best match for several real plots. Figure 2 shows two examples of the return height distributions of real forest plots and the corresponding Forest Factory plots.

Biomass prediction models

The random forest algorithm (Breiman, 2001) as implemented in the R package "randomForest" (Liaw & Wiener, 2002) was used to build regression models for the prediction of biomass from ALS metrics. We conducted three experiments in which we compared how different training datasets affect the model performance. In all experiments, the performance of the models was assessed by comparing the predicted and observed biomass values of the real forest plots. For this, 30 % of the real data were randomly sampled for model testing. The experiments were performed separately for each study site. Each experiment was repeated 500 times, i.e. the model performance was assessed for 500 (partially overlapping) test datasets per study site. The RMSE, the squared Pearson correlation coefficient (r^2), and – as a measure of bias – the ME between observed and predicted biomass values were calculated.

The first experiment tested whether simulated data alone could be used to train biomass models. For this purpose, models were trained using (i) all simulated data, (ii) a randomly selected subset of the simulated data, and (iii) the best matching subset of the simulated data filtered by the waveform matching approach. As a benchmark, models were also trained with the remaining 70 % of the real data that were not used for model testing. The size of the randomly selected subset (ii) was chosen to be equal to the size of the real training dataset.

In the second experiment, we tested whether simulated data can be used to extend the training dataset if only limited real training data are available. Again, the real dataset was randomly divided into 30 % test data and 70 % training data in each of the 500 runs. We trained models with mixed sets of simulated and real data, gradually increasing the number of real samples from two plots to the maximum number of training data available, to investigate whether and up to which number of real training data the model benefits from supplementary simulated data. To reduce computation time, only the randomly selected subset (b) and the best matching subset of the simulated data (c) were used in this experiment. In case of the random selection, the number of randomly selected data was adjusted such that the total number of training data (simulated plus real data) was always equal to the maximum number of real training data. In contrast, the best matching subset was always used in its entirety, i.e. in this case, the total number of training data was equal to the number of best-matching simulated data plus the number of real data used in each increment.

In the first two experiments, we investigated the potential of simulated training data in comparison to real data collected from the same study site that the biomass models were being applied to. In the third experiment, we tested a spatial model transfer. We assumed that only real data from the other real sites were available and evaluated how biomass models perform when trained with these data. For each study site, we trained biomass models using all available data from the three other sites, irrespective of the fact that plot sizes and ALS acquisition settings differed between the datasets.

The data were processed, analysed, and visualised in R version 4.0.4 (R Core Team, 2021) within the RStudio interface (RStudio Team, 2016) making use of the packages "data.table" (Dowle & Srinivasan, 2021), "rgdal" (Bivand et al., 2022), "DescTools" (Signorell, 2021), "ggplot2" (Wickham, 2016), "viridis" (Garnier *et al.*, 2021), "ggpubr" (Kassambara, 2020), and their dependencies.

Results

Comparison of real and simulated data

The biomass values of the real plots ranged from 0.98 t/ha to 583.20 t/ha. This range was completely covered and exceeded by the Forest Factory plots (0.26–1323.81 t/ha for 20 m × 20 m plots, 0.00–1488.57 t/ha for 17.8 m × 17.8 m plots). The mean biomass was highest for the DendroNET sites (268.38 t/ha) and lowest for the Petawawa Research Forest (157.70 t/ha). The mean biomass of the simulated forest plots was significantly lower (136.16 t/ha for 20 m × 20 m plots, 137.67 t/ha for 17.8 m × 17.8 m plots). The stand density ranged from 0 trees/ha to 14897 trees/ha in the Forest Factory plots, with a mean value of 499 trees/ha. The Petawawa Research Forest plots had a similar range of stand density (32–13 024 trees/ha), but the mean was significantly higher (2500 trees/ha). The DendroNet sites had the smallest mean stand density (846 trees/ha) and also the smallest range (89–1600 trees/ha). The stand density of the Milicz Forest plots ranged from 32 trees/ha to 4261 trees/ha, with a mean of 951 trees/ha. Information on the stand density of the Silesian Beskids

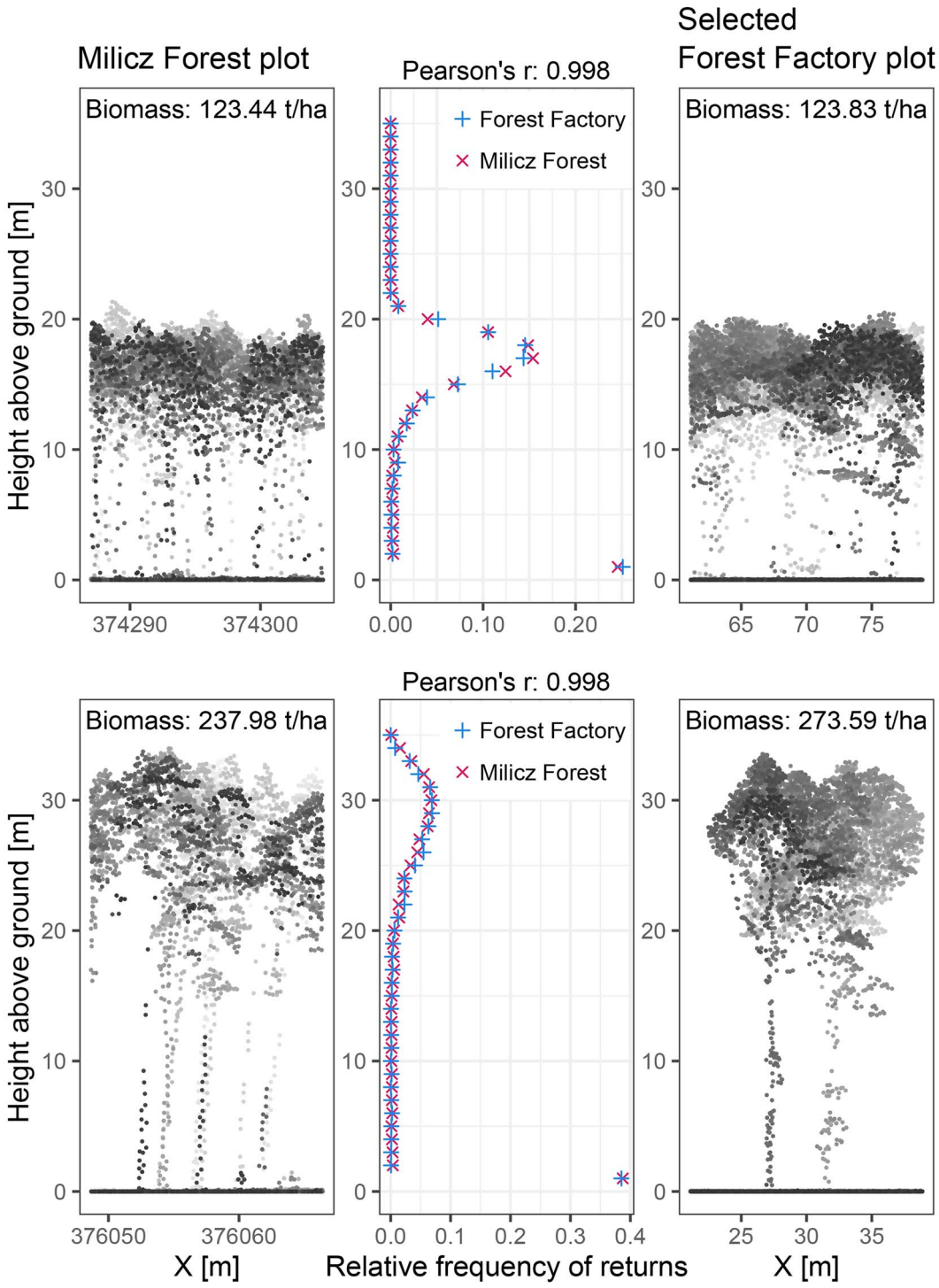

Figure 2. Two examples of the applied waveform matching approach for selecting the best matching Forest Factory plot for each real plot. The left images show vertical sections of real ALS point clouds of plots located in the Milicz Forest, the right images show the simulated ALS point clouds of the best matching Forest Factory plots, and the centre images show the derived return height distribution ("waveform") profiles of both. While the Pearson correlation coefficient of the height distribution profiles is very high (r = 0.998) for both examples, the biomass of the real and the selected simulated plot are very similar for the upper example (123.44 and 123.83 t/ha), but have a higher difference for the lower example (237.98 and 273.59 t/ha). Points are coloured according to their position in Y-direction.

plots was not available. Figure 3 shows histograms of biomass, stand density, and the maximum and mean height of the real and simulated forest plots. The maximum and mean height were calculated from the ALS point clouds, because information on individual tree heights was only available for the Milicz Forest and the Petawawa Research Forest. The simulated data contained proportionally more plots with a maximum height ≥ 30 m than the real-world data, especially when compared to the Milicz Forest and the DendroNET sites. The mean height of returns was on average higher for the real plots than for the simulated plots (except for the Petawawa Research Forest). The best matching subsets of the simulated data fit slightly better to the real data

Figure 3. Relative distribution of biomass, stand density, maximum height of returns (Hmax), and mean height of returns (Hmean) for the real forest plots and the simulated Forest Factory plots. Information on stand density was not available for the Silesian Beskids.

than all simulated data, but there are still large differences in the distribution of biomass, stand density and maximum and mean height of returns.

The mean height of returns (Hmean) was highly correlated with plot biomass (Pearson correlation coefficient > 0.86 for the real data and > 0.73 for the simulated data). Scatter plots of biomass and Hmean show that the simulated data covered a wider range of structural diversity as expressed by these two metrics (Fig. 4). Compared to the real data, the simulated data show lower biomass values in relation to Hmean (Fig. 4, left column). This trend is less pronounced but still visible in the best matching subset of the simulated data (Fig. 4, centre column). Especially for the DendroNET sites, there are large deviations in the ratio of biomass and Hmean between simulated and real data. In contrast, the real data from the different sites have more similar ranges (Fig. 4, right column).

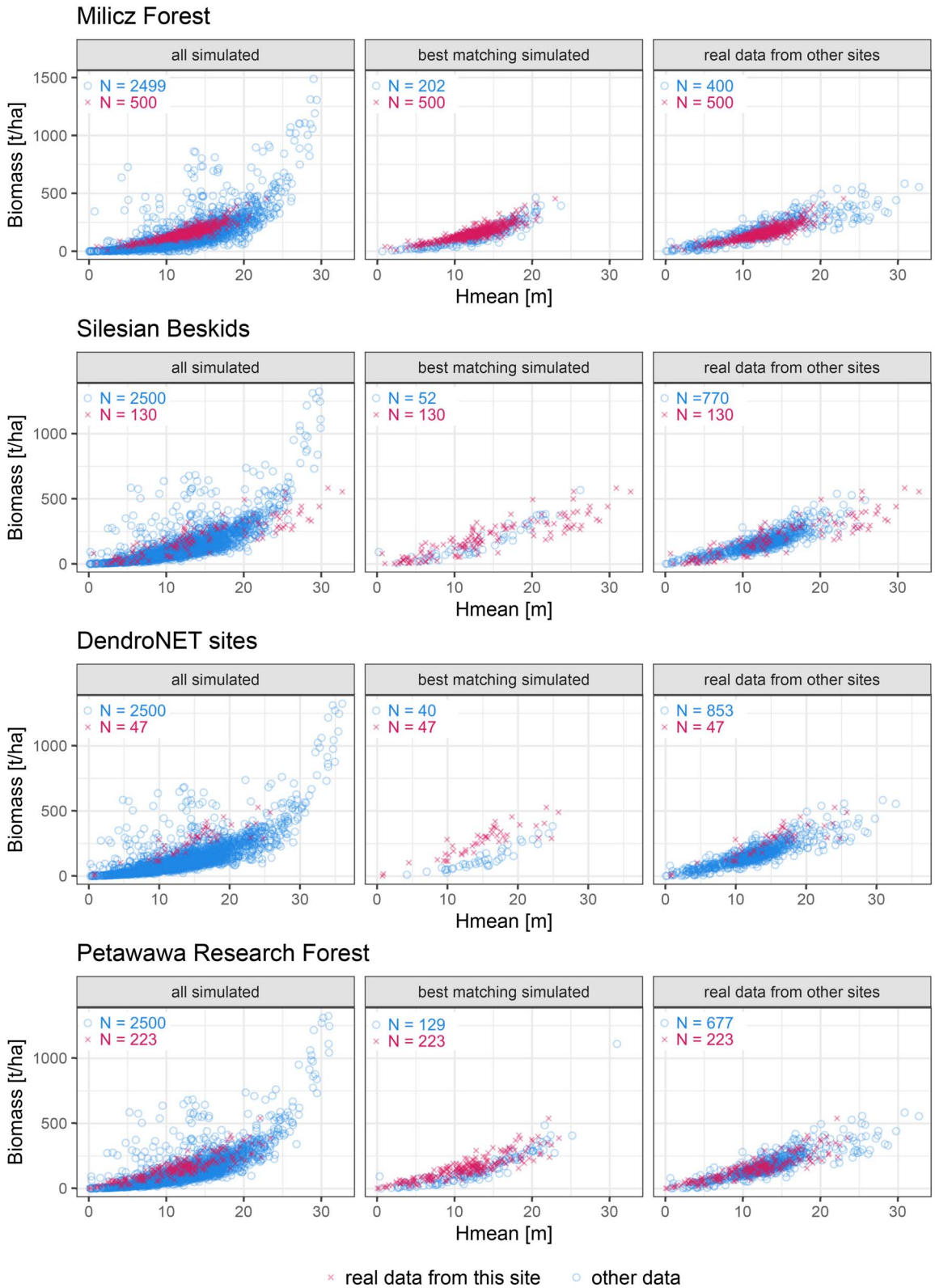

Figure 4. Biomass and mean height of returns (Hmean) of real and simulated forest plots. The real data collected from a site are compared to all simulated data (left), the best matching subset of the simulated data (centre), and the real data collected from the other sites (right).

Biomass models

Experiment 1 (using only simulated data for model training)

In the first experiment, we tested how biomass models perform when trained with simulated data compared to models trained with real data from the same study site the model was applied to. Figure 5 shows scatter plots of the predicted and observed biomass values of all field plots. The predicted values were calculated as the mean of all predictions for one field plot. Since the data were randomly divided into training and test data in the

500 model iterations, the number of predictions varied slightly per field plot, depending on how often this plot was sampled for the test dataset. Model performance metrics (RMSE, ME, r^2) were calculated based on the mean predicted and the observed values. The "all real" models that were trained with all data from the respective study site (excluding the test data) served as benchmark for evaluating the performance of models trained with other data. The first experiment revealed that for all study sites, models trained with real *in situ* data outperformed models trained with simulated data. The difference in model performance was most clearly expressed by the ME, which was negligible for models trained with real data and significantly higher (6.22–118.90 t/ha) for models trained with all simulated data (Fig. 5, second column). Accordingly, the models trained with simulated data underpredicted the biomass values of the real plots. With regard to RMSE, differences in model performance were most pronounced for the DendroNET sites (DN) and the Petawawa Research Forest (PRF). Here, the RMSE of the models trained with all simulated data was about twice as high as for models trained with real data (136.15 t/ha vs. 61.36 t/ha for DN, 73.65 t/ha vs. 37.16 t/ha for PRF). For the Milicz Forest (MF) and the Silesian Beskids (SB), the relative difference in RMSE was slightly smaller (40.86 t/ha vs. 26.94 t/ha for MF, 101.08 t/ha vs. 63.88 t/ha for SB). The difference in r^2 was highest for the Petawawa Research Forest (simulated data: 0.68, real data: 0.83), and smallest for the DendroNET sites (simulated data: 0.69, real data: 0.72).

Using a randomly selected subset of the simulated data instead of using all simulated data for model training increased the model performance in most cases (Fig. 5, third column). An exception were the predictions for the Silesian Beskids, where the models trained with all simulated data strongly overpredicted the biomass of three plots (cf. outliers in Fig. 5, second row, second column). This led to a very low ME between observed and predicted biomass values compared to the models trained with a randomly selected subset of the simulated data (all simulated: 6.22 t/ha, randomly selected simulated: 32.37 t/ha), while the RMSE was much higher (all simulated: 101.08 t/ha, randomly selected simulated: 77.17 t/ha) and the r^2 was much lower (all simulated: 0.67, randomly selected simulated: 0.74).

The performance of models trained with the best matching subset of the simulated data was similar to the performance of models trained with a randomly selected subset (Fig. 5, fourth column), but differed between study sites. While the difference in RMSE for the Milicz Forest and the Silesian Beskids was < 1 t/ha, for the DendroNET sites and the Petawawa Research Forest, the RMSE was lower for the models trained with randomly selected simulated data than for models trained with the best matching data (127.15 t/ha vs. 133.58 t/ha for DN, 67.83 t/ha vs. 72.92 t/ha for PRF). The difference in the ME was negligible for the Milicz Forest (16.48 t/ha vs. 17.38 t/ha). For the Silesian Beskids, the absolute ME was higher when the models were trained with a randomly selected subset than when they were trained with the best matching subset (32.37 t/ha vs. 18.01 t/ha). In contrast, for the DendroNET sites and the Petawawa Research Forest, using a randomly selected subset for model training resulted in a slightly lower absolute ME than using the best matching subset (103.63 t/ha vs. 114.85 t/ha for DN, 49.19 t/ha vs. 53.74 t/ha for PRF). The absolute difference in r^2 was 0.2 for all study sites, with a higher r^2 for the randomly selected subset for the Silesian Beskids and the Petawawa Research Forest, and a higher r^2 for the best matching subset for the Milicz Forest and the DendroNET sites.

Figure 6 shows the mean performance metrics calculated from each of the 500 model iterations for each study site and training data type, including the results for models that were trained with different numbers of real training samples. Because of the random sampling of field plots for the test datasets, the model performance metrics differ slightly from the values presented in Fig. 5. Models that were trained with simulated data resulted, in most cases, in higher prediction accuracies, as expressed by RMSE and r^2, than models that were trained with real data when the number of real training samples was very low. Table 2 and the dashed vertical lines in Fig. 6 show up to which number of real training samples models that were trained with simulated data only, or with real data collected from other sites (Experiment 3), performed better than models that were trained with real data collected from the same site the models were applied to.

Experiment 2 (extending the real training dataset with simulated data)

In the second experiment, we tested whether the accuracy of biomass models could be increased by extending a small real dataset with additional simulated data for model training. For all study sites, model accuracy in terms of RMSE and r^2 improved by adding simulated training data to a small number of real training data (Fig. 6). However, as the amount of real training data increased, the positive effect of additional training data decreased and eventually disappeared. In contrast to RMSE and r^2, the absolute ME of the models was always lowest when only real data were used for model training. As the number of real training samples increased, the accuracies of models trained with real data only and those trained with additional randomly selected simulated data converged because the number of additional simulated samples decreased when more real samples were used. The numbers of real training samples up to which the addition of simulated data resulted in higher model accuracies in terms of RMSE and r^2 are given in Table 2.

Experiment 3 (using real data from other study sites for model training)

Using data collected from other sites (*ex situ* data) for training biomass models resulted in high model accuracies for all sites but Milicz Forest (Fig. 5, last column). Compared to models that were trained with real data collected from the Silesian Beskids and the Petawawa Research Forest, respectively, the RMSE increased by only 0.54 t/ha and 2.70 t/ha when models were trained with data from the other sites. The increase in RMSE was slightly higher for the DendroNET sites (17.79 t/ha), but still much lower than when models were trained with simulated data (increase in RMSE: ≥ 65.79 t/ha). The ME indicated an overprediction of biomass for the Silesian Beskids (-17.95 t/ha), and an underprediction of biomass for the DendroNET sites (54.61 t/ha) and the Petawawa Research Forest (5.20 t/ha). The r^2 value of models trained with real data collected from other sites decreased by 0.03 for the Petawawa Research Forest, and even increased by 0.02 and 0.04 for the Silesian Beskids and the DendroNET sites, compared to models trained with real data collected from the respective sites. For three of the study sites, the Silesian Beskids, the DendroNET sites, and the Petawawa Research Forest, using real data collected from other sites for model training resulted in significantly higher model accuracies than using simulated data, regardless of whether all simulated data or a subset were used. In contrast, for the Milicz Forest, the accuracy was much higher for models trained with simulated data, especially when only a subset of the

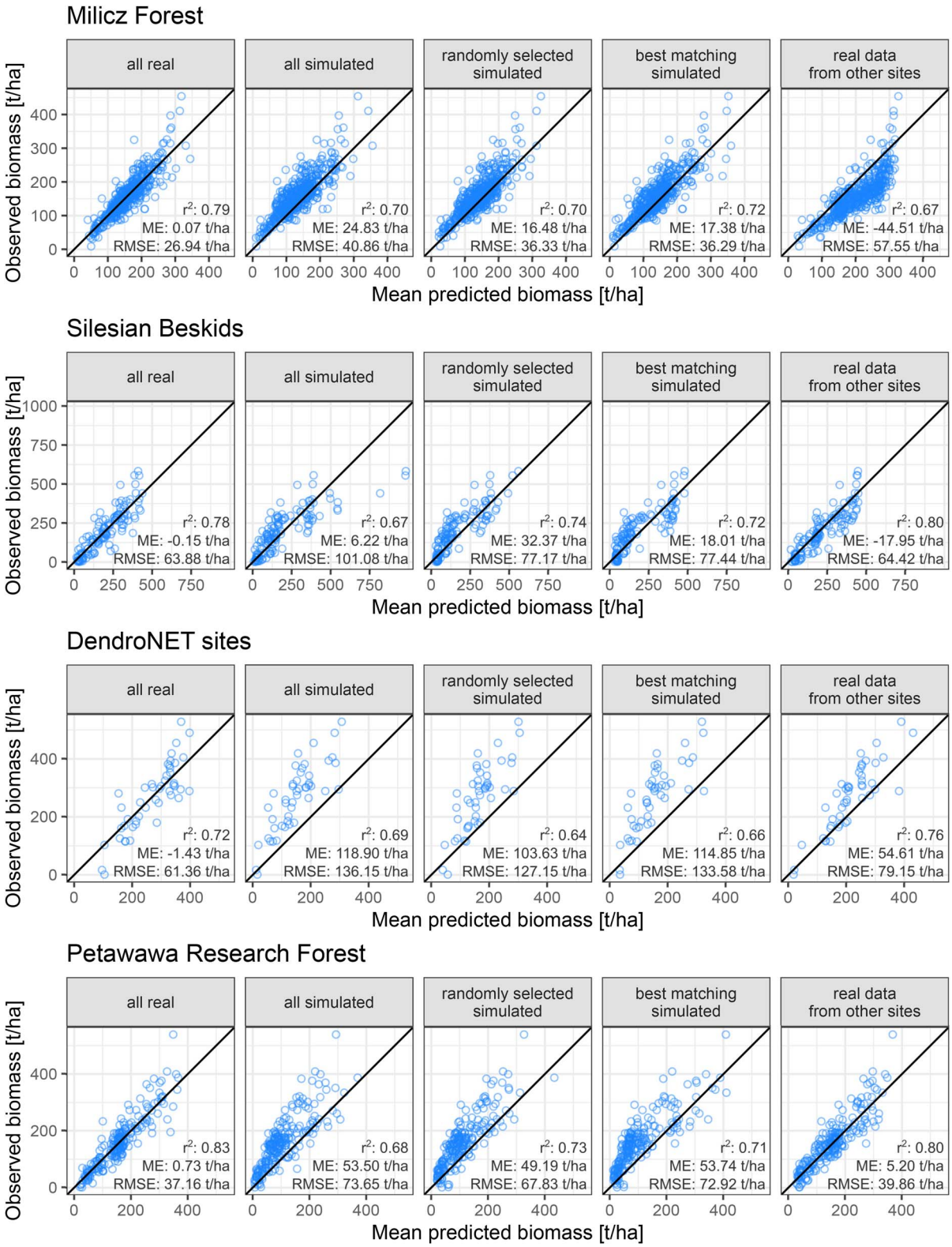

Figure 5. Mean predicted biomass and observed biomass of all field plots by study site and training data type. Model building and predictions were repeated 500 times for each training data type. In each of the 500 iterations, the real data were randomly split into 30 % test data and 70 % training data, i.e. each plot was included in the test and training data several times. The mean predicted biomass was calculated as the mean of all predictions for one field plot. The squared Pearson correlation coefficient (r^2), the ME, and the RMSE are given.

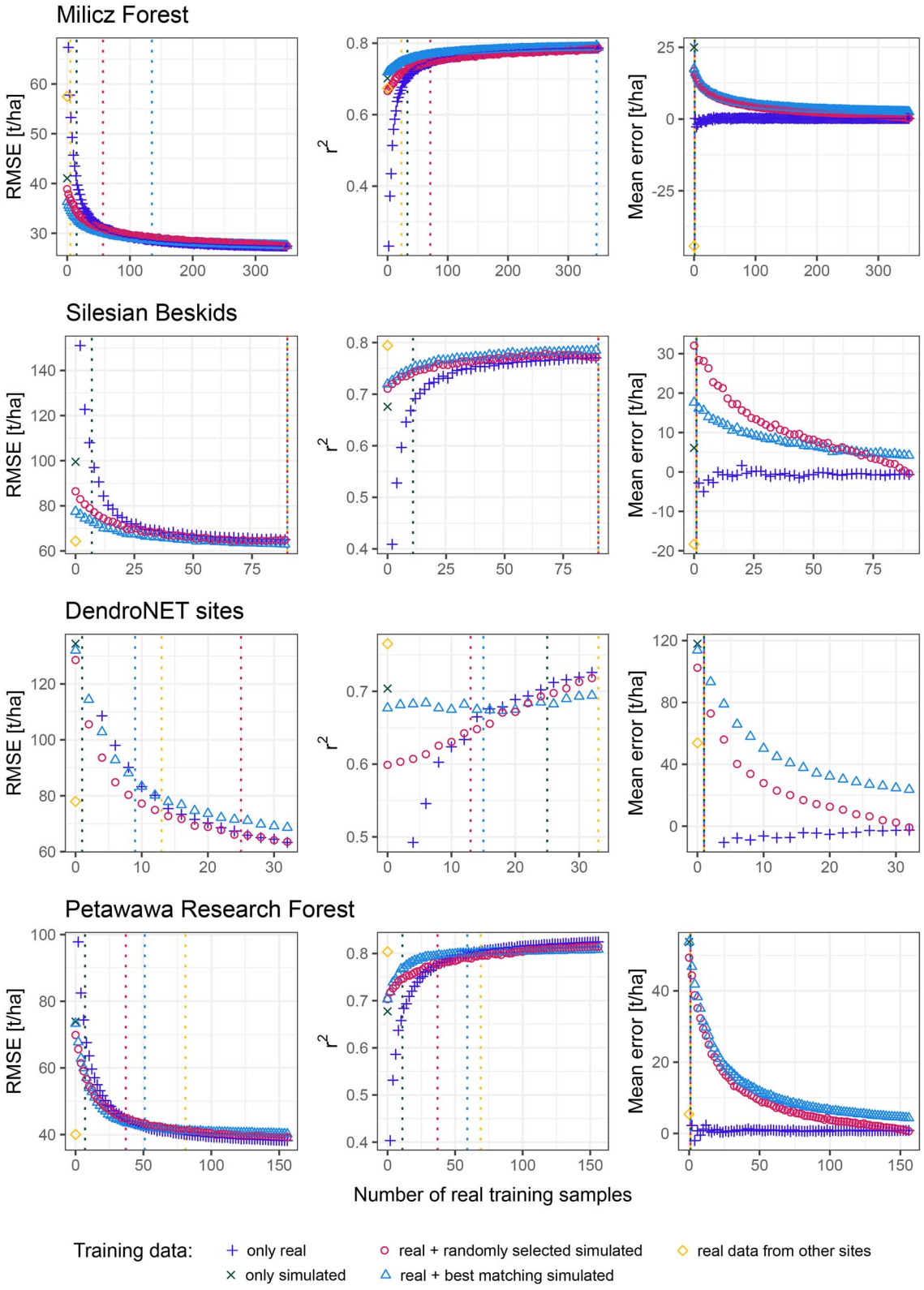

Figure 6. Mean RMSE, mean squared Pearson correlation coefficient (r^2), and ME of the biomass predictions for different training data types. Model building and predictions were repeated 500 times for each training data type and each number of real training samples. Dashed vertical lines show at which number of real training samples the "only real" model performed better than the other models, as indicated by colour.

Table 2. Number of training samples up to which models trained with real *in situ* data performed worse than models trained with other data (as specified in column *Training data type*).

Study Site	Training data type	Number of real training samples measured by	
		RMSE	r^2
Milicz Forest	All simulated	≤ 14	≤ 32
	Randomly selected simulated	≤ 20	≤ 10
	Best matching simulated	≤ 22	≤ 40
	Real data from other sites	≤ 4	≤ 22
	Real + randomly selected simulated	≤ 56	≤ 70
	Real + best matching simulated	≤ 134	≤ 346
Silesian Beskids	All simulated	≤ 6	≤ 10
	Randomly selected simulated	≤ 10	≤ 16
	Best matching simulated	≤ 14	≤ 16
	Real data from other sites	> 90	> 90
	Real + randomly selected simulated	> 90	> 90
	Real + best matching simulated	> 90	> 90
DendroNET sites	All simulated	≤ 2	≤ 24
	Randomly selected simulated	≤ 2	≤ 6
	Best matching simulated	≤ 2	≤ 6
	Real data from other sites	≤ 12	>32
	Real + randomly selected simulated	≤ 24	≤ 12
	Real + best matching simulated	≤ 8	≤ 14
Petawawa Research Forest	All simulated	≤ 6	≤ 10
	Randomly selected simulated	≤ 6	≤ 14
	Best matching simulated	≤ 6	≤ 14
	Real data from other sites	≤ 80	≤ 68
	Real + randomly selected simulated	≤ 36	≤ 36
	Real + best matching simulated	≤ 50	≤ 60

simulated data was used. Compared to the models trained with the best matching subset, the RMSE of models trained with real data from other sites increased by 21.26 t/ha, and r^2 decreased by 0.05. Training models with real data collected from other sites resulted in a significant overprediction of biomass for the Milicz Forest, with an ME of -44.51 t/ha.

Discussion

Experiment 1

The results of the first experiment suggest that models trained only with simulated data do not reach the performance of models trained with real data, as long as a sufficient amount of real data is available. The gap in model accuracy when simulated data were used for model training instead of real data differed between study sites. Regardless of the study site, all models trained exclusively or additionally (see Experiment 2) with simulated data significantly underpredicted the biomass of the real plots, whereas models trained with real data (collected from the same site the model was applied to) did not show any bias (Fig. 6). The underprediction was highest for the DendroNET sites (ME 103.63–118.90 t/ha) and the Petawawa Research Forest (ME 49.19–53.74 t/ha). We suppose that there are several reasons why the difference in model performance between models trained with simulated data and models trained with real data was highest for these two study sites (also with regard to RMSE and r^2). Most of the DendroNET sites are located in single species forest with only one layer. The plots there have high biomass values but low maximum and mean heights of return compared to the Forest Factory plots, but also compared to the other real forest plots (Figs 3 and 4). As the stand density at the DendroNET sites is also rather low, the high plot biomass is probably the result of a specific silvicultural strategy which is not captured by the other real datasets and by the growth simulator. In contrast to the DendroNET sites, the Petawawa Research Forest has a high structural diversity resulting from diverse species compositions and complex management histories (White *et al.*, 2021), and the stand density is comparatively high. Multiple forest layers and the occurrence of undergrowth shift the relative return height metrics toward the lower heights, resulting in an underprediction of biomass when models are trained with less complex data. In addition, forest plots with a high stand density may have a similar return height distribution as plots with a lower stand density but similar tree sizes, resulting in a much higher plot biomass. It is therefore also possible that the high stand density is a reason for the underprediction of biomass by models trained with simulated data. An in-depth analysis of why the models trained with the simulated data performed differently for the four study sites was not possible, as this would have required detailed information on individual trees (e.g. tree height and location), which was only available for the Milicz Forest.

Differences in the feature space can cause problems in model transferability (Meyer & Pebesma, 2021). If simulated data are to replace real data in model training, it needs to be ensured that they cover all the features of real-world data. Even if the simulated data covered the whole range of biomass values of the real forest plots, they did not cover the complete range of all predictors and the relation between LiDAR metrics and biomass was not the same. For example, the extremely high overprediction of biomass for three of the Silesian Beskids plots when using all simulated data for model training (Fig. 5, second row, second column) can be explained by the fact that the Silesian Beskids plots with

higher mean return heights have much lower biomass values than the Forest Factory plots (Fig. 4, second row, first column). The observed differences between the simulated and the real data can be caused by several factors, that are related to either (i) the simulation of the forest stand composition or (ii) the simulation of the laser scanning. First of all, the forest composition differs between simulated and real forests. Forest Factory uses the region-specific parameterisations implemented in the FORMIND model, which do not yet include all tree species that occur at our study sites (Henniger *et al.*, 2023). In addition, the availability of tree point clouds that are needed to create the 3D representations of the Forest Factory stands further limits the number and size range of tree species included in the Forest Factory simulations. Therefore, the simulated forest stands only include four tree species. They also lack understorey elements such as shrubs and small trees, resulting in a lower structural complexity than real forests (Bruening *et al.*, 2021). The goal of Forest Factory is to generate as many potential forest states as possible. The resulting large variability of the generated stands, as reflected by the wide range of biomass values in relation to the mean height of returns, indicates that the underlying stem size distributions of the simulations include extremes that do not occur in our study sites. However, there will also be real forest compositions that are not captured by the Forest Factory simulations. A drawback of using Forest Factory to simulate forest stand composition is that tree positions are randomly assigned within the plot area. As a result, trees may be placed unrealistically close together. Compared to Forest Factory 2.0, other forest simulators such as SILVA (Pretzsch *et al.*, 2002) have the disadvantage that parameterisation and simulations take much longer, which means that significantly fewer stands can be generated in a reasonable time. On the other hand, SILVA has the ability to apply different management strategies and takes into account competition from neighbouring trees. It is therefore likely that the stands that are generated will be more similar to real forest stands. Further research should investigate how the use of different forest simulators affects the quality (in terms of usability) of the simulated data.

Another factor that can lead to differences between the simulated and real data is the simulation of laser scanning. Both the pulse densities and the resulting planar point densities differed between simulated and real datasets (Table 1). Because some of the acquisition settings of the real laser scanning campaigns were unknown, it was not possible to exactly reproduce the acquisitions. Furthermore, the laser scanning simulation process implemented in HELIOS++ is sensitive to other parameters, such as the size of the voxels that are used to convert the forest point clouds into scannable objects, the point density of the tree point clouds, and the temporal window size for echo detection (Winiwarter *et al.*, 2022). Consequently, the parameterisation of HELIOS++ and the implemented laser scanning simulation approach should be further optimised, also with regard to more realistic intensity values and numbers of returns, so that metrics related to these point cloud characteristics could also be derived from the simulated data. However, when models were trained with the real *in situ* data, we did not observe significant differences when these metrics were included or excluded from model training.

Including understorey elements in the simulated forest stands and choosing a different voxel size and temporal window size could contribute to shifting the distribution of simulated returns to lower heights, and thus better fit the relation between biomass and mean return height of the simulated data to the real data. Furthermore, this relation could also be affected by the method for calculating the individual tree biomass values. Allometric equations are commonly applied for predicting biomass from stem diameters and in some cases tree height, and different equations can result in different biomass predictions for the same tree (Zianis *et al.*, 2005, Ameztegui et al., 2022). In cases where no allometries are available for a specific location and its site conditions, existing equations developed for a similar site are used. However, these equations were not necessarily developed with trees that fully match the range of diameters present in the studied site or in few cases even several matching equations may available that, however, differ in their predictions. The resulting uncertainty in the biomass reference values additionally affects the model performance. To exclude potential effects of allometric equations, it would be best to use the same equations for predicting biomass for all study sites. Here, we only used the same set of equations for calculating biomass of the Forest Factory trees and the trees in the Milicz Forest, for the other sites, the provided biomass estimates were used.

One disadvantage of the Forest Factory simulations is that the forest stands are limited to a fixed size of 20 m × 20 m. In case of larger field plots in the real data acquisitions, it was therefore not possible to extract data from plots of the same size and shape in the Forest Factory stands. We decided to keep the plot size and shape the same for the extraction of laser scanning data from the simulated and from the real forest stands. For the Silesian Beskids and the Petawawa Research Forest, no information on individual tree positions was available, and thus the biomass could not be calculated for the same plot that was used for the laser scanning data extraction and the extraction of both biomass information and simulated laser scanning data from the Forest Factory stands. Hayashi *et al.* (2015) analysed how the plot radius for the LiDAR metrics extraction affects biomass prediction in the Acadian Forest using biomass reference data obtained from nested circular and variable radius plots, and found little influence on model performance. Zhao *et al.* (2009) found no differences in parametric regression performance between models trained with squared or circular plots, but notable differences when comparing models trained with plot sizes from 0.01 to 1 ha (for the extraction of both LiDAR metrics and biomass reference data). However, their results also indicate that model performances are very comparable if differences between plot sizes are as small as in our study. We therefore expect that the differences in plot size had only a minor impact on our results. Especially since in the simulated data models that performed worst, i.e. those for the DendroNET sites, biomass values and ALS metrics were collected and extracted from the same sized plots for both the simulated training data and the real test data. Nevertheless, we acknowledge that with the experimental setup presented, it is not possible to fully disentangle the effects of the simulated data from the effects of different plot shapes and sizes, especially since random forest has also been found to decrease its performance when applied to datasets that are not fully comparable to the training data (Hayashi *et al.*, 2015).

One explanation for the slight increase in model performance when a randomly selected subset of the simulated data was used for model training instead of all data could be that the random sample is less likely to include the extreme values in the simulated data that are out of the range of the real data, resulting in better fitting models. Surprisingly, the best matching samples did not provide the expected additional benefit compared to random samples. This is probably because the shift in the relation between biomass and return heights that was observed between simulated and real datasets is still present in the best matching samples (Fig. 4). Unlike stratified sampling approaches relying on one or more ALS-derived metrics, our sampling was

based on the overall height distribution. Bruening et al. (2021) applied a similar approach based on the relative overlap of the height distribution profiles to match simulated GEDI waveforms of Forest Factory stands and real forest stands. In contrast to our results, they found a good fit in the biomass distribution of the selected simulated data and the real data. Since they eliminated potential effects of different allometry models, differences in real and simulated LiDAR data, and influences of understorey elements in their study, this may be an indicator that the observed shift in our simulated datasets compared to the real datasets might be related to one or more of these factors. Apart from that, the study by Bruening et al. (2021) provides another possible explanation why the best matching approach for the training data selection did not significantly improve our model accuracies. They explored the non-uniqueness of LiDAR signals, which was also described by Zolkos et al. (2013), and showed that forest stands of different composition can produce similar LiDAR waveforms but have a different biomass, and vice versa. Accordingly, one LiDAR waveform should be associated with a range of biomass values. These findings can be transferred to discrete return laser scanning data. This might explain why Forest Factory plots with the same mean height of returns have a wide range of biomass values and also why the best matching simulated data did not greatly improve the model accuracy compared to randomly selected simulated data. A high correlation of the height distribution profiles of two plots is not necessarily related to similar biomass values of these plots (see Fig. 2) and it is imaginable that the biomass value of the second best matching plot would fit much better. It should be investigated whether a sampling approach based on other ALS metrics would result in a better match between the biomass values of the selected simulated and real forest plots. Auxiliary data, such as information on the forest structure (e.g. tree density), could also be helpful for solving this issue.

Experiment 2

The second experiment showed that simulated data can be used to extend sparse real training datasets. However, the positive effect of additional training data on model accuracy decreased as the number of real training samples increased, and even with relatively low quantities of real training data, the increase in model accuracy was small. Our findings with respect to up to which number of real training samples the model accuracy increased when additional simulated data were used differed between the study sites, making a generalised statement difficult. The models for the DendroNET sites benefited the least from the additional training data, which is probably because the simulated data did not fit well to the real data of these sites. Future works could expand the presented analysis with more datasets to better understand which combination of plots of different forest structures benefits in which way from the additional simulated data.

Models that were trained with mixed datasets composed of real and the best matching simulated data performed slightly better in most cases than models that were trained with mixed datasets composed of real and randomly selected simulated data, but this could also be an effect of the different compositions and sizes of the training datasets (fixed number of simulated data in case of the best matching subset, varying number of simulated data in case of the randomly selected subset).

Stereńczak et al. (2018) analysed how many field samples are required for the accurate prediction of growing stock volume in the Milicz Forest using an ordinary least square multiple regression and found that model performance did not change much when at least 200 samples were used for model training, except for relative bias, which was lowest when at least 500 samples were used. Using synthetic forest data, Fassnacht et al. (2018) observed an increase in the accuracy of random forest models for biomass prediction with increasing sample size, particularly for small sample sizes, using 50–500 samples. Nevertheless, according to a review by Fassnacht et al. (2014), 73 % of the reviewed studies on remote sensing-based forest biomass predictions had sample sizes smaller than 100, and 53 % had sample sizes smaller than 50. Synthetic data could therefore be of great value if they could improve model performance when limited real-world training data are available. While RMSE and r^2 did indeed improve by training models on mixed datasets of real and simulated data (up to 12–346 real training samples depending on the study site), the increase in bias whenever simulated data were included in the training dataset is a major concern. In this study, we used the random forest algorithm as prediction method because it has been shown to outperform other commonly used methods for ALS-based forest biomass prediction (Fassnacht et al. (2014). However, Yang et al. (2019) found that compared to other prediction methods, random forest models resulted in a high overprediction of forest volume when combined with variable probability selection methods. Hayashi et al. (2015) tested a spatial transfer of biomass models and found that the performance of random forest models decreased when applied to an *ex situ* dataset, while the performance of non-linear mixed effects models did not change when applied to *in situ* or *ex situ* data. Accordingly, the random forest algorithm might not be the best choice for our study. In addition, random forests are not designed to handle multiple training datasets and treat them differently, e.g. by giving them different weights. Instead of simply merging simulated and real data into one training dataset, as we did for the random forest models, one could also use the simulated data to pre-train a model and then use the real data to fine-tune it. This transfer learning approach is often used in deep learning, where large amounts of labelled training data are required (Hamedianfar et al., 2022). Transfer learning has also been implemented for linear regression under covariate shift, reducing the amount of required target data (Wu et al., 2022). Future work should explore whether model accuracies could further be improved by using transfer learning methods.

Experiment 3

Training biomass models with data that were collected from other study sites (*ex situ*) resulted in surprisingly high prediction accuracies for the Silesian Beskids, the DendroNET sites and the Petawawa Research Forest. Although the study sites had different species compositions (and in case of the Petawawa Research Forest even completely different tree species), different allometric equations were used for calculating the biomass, the ALS point clouds characteristics such as point density differed, and the data were extracted from differently shaped and sized plots, the merged datasets were well suited for model training, and the spatially transferred models resulted in RMSE values and squared Pearson correlation coefficients similar to models that were trained with real data collected from the same site the model was applied to. These results indicate that the aforementioned factors are less likely the reason for the decrease in model accuracies when simulated data were used for model training. Suvanto & Maltamo (2010) compared models for predicting forest characteristics that were trained with local data only to models that were trained with a mixed dataset of local and additional *ex situ* data, and found that the local model outperformed the mixed model already at sample sizes below 50. In our study, we

could only partially confirm these observations. Even if we did not use mixed datasets but only data collected from other sites, for the Silesian Beskids, model training with data from other sites resulted in higher prediction accuracies in terms of RMSE and r^2 than the model training with local data, even when the maximum of 90 training samples was used. Our observations that the absolute ME of models trained with other simulated or real data was always higher than of models trained with real data from the site the model is applied to are in line with the observations made by Suvanto & Maltamo (2010).

Regarding RMSE and r^2, using training data collected from other sites worked best for the Silesian Beskids and the Petawawa Research Forest. For these sites, the ranges in biomass and mean height of returns of the on-site data and of the other data fit very well, apart from the higher ranges of the mean height of returns in the Silesian Beskids and the lower ranges of the maximum height of returns in the Petawawa Research Forest that were not covered by the data from the other sites. Compared to the DendroNET sites, the other real field plots had lower biomass values at the same mean return heights, which probably led to the strong underprediction when these data were used to predict biomass of the DendroNET sites.

The Milicz Forest was the only study site for which models that were trained with real data collected from the other sites performed worse than models trained with simulated data. From the information that was available for the study sites, we are unable to explain why the results for the Milicz Forest differ from the results for the other study sites. We assume that the low performance of the spatial model transfer is related to the fact that the range of associated biomass values in relation to the mean height of returns (and other ALS metrics) is much wider for the other sites (1.0–583.2 t/ha biomass at mean return heights of 0.0–32.7 m) than for the Milicz Forest (8.9–454.3 t/ha biomass at mean return heights of 1.0–23.0 m) but we do not know which characteristics of the study sites lead to these differences.

As the second experiment showed that prediction accuracies can be improved when sparse training datasets are extended by additional simulated data, it should be tested if similar results can be observed when real data collected from other sites are used to extend the training datasets. Taking into account that for three of the four study sites, models trained with with real data collected from other sites performed better than models trained with simulated data, we would expect even better results from mixing the real local and non-local datasets than from mixing simulated and real data. This would also be in line with findings of Breidenbach *et al.* (2008), Kotivuori *et al.* (2016), and van Ewijk *et al.* (2020) who demonstrated that calibrating models with a small local dataset in combination with a larger dataset collected from other sites improves prediction accuracies compared to models that were only calibrated based on the larger dataset.

Conclusions

This study investigated the potential of simulated data for training biomass models for real forest plots. Our experiments revealed that simulated data generated by applying the HELIOS++ laser scanning simulator to Forest Factory 2.0 forest plots cannot yet compete with real data. Models can be trained using simulated data only, but we observed a strong underprediction of biomass for three of the four study sites, and the model performance generally improves when real data are included in the training dataset. However, when only a limited number of real training samples is available, simulated data can be used to extend the training dataset. It depends on the study site and the measure of model performance up to which number of real training samples the model accuracy can be increased by the additional simulated data. While the prediction accuracy of models trained with simulated data may be satisfactory for various applications, the significant underprediction of biomass presents a challenge. Therefore, the workflow for generating simulated data needs improvement in order to achieve a better match between the simulated data and the real data in terms of the relation of biomass to ALS metrics. In addition, future research should explore alternative methods for selecting the samples of the simulated data that best match the real data (using only ALS-derived information) and investigate transfer learning methods.

Our experiments also demonstrate that real data collected from different locations can be very suitable for training biomass models, even if the laser scanning acquisition settings, the plot design, the forest composition, and the method to calculate biomass values differ from the site the model is applied to. It would therefore be beneficial for the research community, but also for forest practitioners, if reference data were made more widely available to others. These data may still be useful even if they do not include ALS data, as laser scanning point clouds could be generated with our simulation approach, at least if information on all trees in the field plots is provided, and not only summarised information on a plot level.

Acknowledgements

We thank Joanne White and the Canadian Forest Service for providing the data for the Petawawa Research Forest, and the technicians from the Institute of Forest Ecosystem Research and the Global Change Research Institute for field data collection at the Silesian Beskids sites and within the DendroNetwork. We would also like to thank Florian Hartig for thoughtful discussions on the analysis. We are very grateful to the two anonymous reviewers and the editors whose constructive comments helped us to significantly improve the paper.

CRediT statement

Jannika Schäfer: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing—Original Draft, Writing—Review & Editing, Visualization. **Lukas Winiwarter:** Investigation, Methodology, Writing—Review & Editing. **Hannah Weiser:** Investigation, Data curation, Writing—Review & Editing. **Jan Novotný:** Data curation, Writing—Original Draft, Writing—Review & Editing. **Bernhard Höfle:** Conceptualization, Data curation, Resources, Writing—Review & Editing, Supervision, Funding acquisition. **Sebastian Schmidtlein:** Resources, Writing—Review & Editing, Supervision. **Hans Henniger:** Software, Writing—Review & Editing. **Grzegorz Krok:** Data curation, Writing—Review & Editing. **Krzystzof Stereńczak:** Investigation, Data curation, Writing—Review & Editing. **Fabian Ewald Fassnacht:** Conceptualization, Methodology, Software, Resources, Writing—Review & Editing, Supervision, Funding acquisition.

Conflict of Interest statement: None declared.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the frame of the project SYSSIFOSS - 411263134 / 2019-2022; by the Polish State Forests National Forest Holding in the frame of the project "Development of the method of forest inventory using the results of the REMBIOFOR project" (Project No. 500463, agreement No. EO.271.3.12.2019, signed on 14.10.2019); and by the National Centre for Research and Development (Poland) in the frame of the REMBIOFOR project "Remote sensing-based assessment of woody biomass and carbon storage in forests" as part of the BIOSTRATEG programme (Agreement No. BIOSTRATEG1/267755/4/NCBR/2015).

Data availability statement

The tree point clouds are available at PANGAEA, at https://doi.pangaea.de/10.1594/PANGAEA.942856. R code for the creation of 3D models of the Forest Factory stands is available on GitHub, at https://github.com/JannikaSchaefer/Syssifoss.

References

Achim A, Moreau G, Coops NC, *et al.* The changing culture of silviculture. *Forestry* 2022; **95**:143–52. https://doi.org/10.1093/forestry/cpab047.

Ameztegui A, Rodrigues M, Granda V. Uncertainty of biomass stocks in Spanish forests: a comprehensive comparison of allometric equations. *Eur J For Res* 2022; **141**:395–407. https://doi.org/10.1007/s10342-022-01444-w.

Signorell A, *et al.* DescTools: Tools for descriptive statistics. R package version 0.99.44, 2021. https://cran.r-project.org/package=DescTools.

Bivand R, Keitt T, Rowlingson B. Rgdal: Bindings for the 'geospatial' data abstraction library. R package version 1 , 2022. https://CRAN.R-project.org/package=rgdal.

Blair J, Hofton M. Modeling laser altimeter return waveforms over complex vegetation using high-resolution elevation data. *Geophys Res Lett* 1999; **26**:2509–12. https://doi.org/10.1029/1999GL010484.

Bohn FJ, Huth A. The importance of forest structure to biodiversity–productivity relationships. *R Soc Open Sci* 2017; **4**:160521. https://doi.org/10.1098/rsos.160521.

Breidenbach J, Kublin E, McGaughey R, *et al.* Mixed-effects models for estimating stand volume by means of small footprint airborne laser scanner data. *Photogramm J Finland* 2008; **21**:4–15.

Breiman L. Random forests. *Mach Learn* 2001; **45**:5–32. https://doi.org/10.1023/A:1010933404324.

Brovkina O, Navrátilová B, Novotný J, *et al.* Influences of vegetation, model, and data parameters on forest aboveground biomass assessment using an area-based approach. *Eco Inform* 2022; **70**:101754. https://doi.org/10.1016/j.ecoinf.2022.101754.

Bruening JM, Fischer R, Bohn FJ, *et al.* Challenges to aboveground biomass prediction from waveform LiDAR. *Environ Res Lett* 2021; **16**:125013. https://doi.org/10.1088/1748-9326/ac3cec.

Dalponte M, Martinez C, Rodeghiero M, *et al.* The role of ground reference data collection in the prediction of stem volume with LiDAR data in mountain areas. *ISPRS J Photogramm Remote Sens* 2011; **66**:787–97. https://doi.org/10.1016/j.isprsjprs.2011.09.003.

de Lera Garrido A, Gobakken T, Ørka HO, *et al.* Reuse of field data in ALS-assisted forest inventory. *Silva Fennica* 2020; **54**(5):1–18. https://www.silvafennica.fi/article/10272.

Disney MI, Kalogirou V, Lewis P, *et al.* Simulating the impact of discrete-return LiDAR system and survey characteristics over young conifer and broadleaf forests. *Remote Sens Environ* 2010; **114**:1546–60. https://doi.org/10.1016/j.rse.2010.02.009.

Dixon RK, Solomon AM, Brown S, *et al.* Carbon pools and flux of global Forest ecosystems. *Science* 1994; **263**:185–90, https://doi.org/10.1126/science.263.5144.185.

Domingo D, Alonso R, Lamelas MT, *et al.* Temporal transferability of pine Forest attributes modeling using low-density airborne laser scanning data. *Remote Sens (Basel)* 2019; **11**:261. https://doi.org/10.3390/rs11030261.

Dowle M, Srinivasan A. data.table: Extension of 'data.frame'. R package version 1.14.0, 2021. https://CRAN.R-project.org/package=data.table.

Fassnacht FE, Hartig F, Latifi H, *et al.* Importance of sample size, data type and prediction method for remote sensing-based estimations of aboveground forest biomass. *Remote Sens Environ* 2014; **154**:102–14. https://doi.org/10.1016/j.rse.2014.07.028.

Fassnacht FE, Latifi H, Hartig F. Using synthetic data to evaluate the benefits of large field plots for forest biomass estimation with LiDAR. *Remote Sens Environ* 2018; **213**:115–28. https://doi.org/10.1016/j.rse.2018.05.007.

Fekety PA, Falkowski MJ, Hudak AT. Temporal transferability of LiDAR-based imputation of forest inventory attributes. *Can J For Res* 2015; **45**:422–35. https://doi.org/10.1139/cjfr-2014-0405.

Fischer R, Bohn F, Dantas de Paula M, *et al.* Lessons learned from applying a forest gap model to understand ecosystem and carbon dynamics of complex tropical forests. *Ecol Model* 2016; **326**:124–33. https://doi.org/10.1016/j.ecolmodel.2015.11.018.

Frazer GW, Magnussen S, Wulder MA, *et al.* Simulated impact of sample plot size and co-registration error on the accuracy and uncertainty of LiDAR-derived estimates of forest stand biomass. *Remote Sens Environ* 2011; **115**:636–49. https://doi.org/10.1016/j.rse.2010.10.008.

Garnier S, Ross N, Rudis R, *et al.* Viridis - colorblind-friendly color maps for R. R package version 0.6.2, 2021. https://sjmgarnier.github.io/viridis/.

Gastellu-Etchegorry J-P, Yin T, Lauret N, *et al.* Simulation of satellite, airborne and terrestrial LiDAR with DART (I): waveform simulation with quasi-Monte Carlo ray tracing. *Remote Sens Environ* 2016; **184**:418–35. https://doi.org/10.1016/j.rse.2016.07.010.

Gobakken T, Næsset E. Assessing effects of laser point density, ground sampling intensity, and field sample plot size on biophysical stand properties derived from airborne laser scanner data. *Can J For Res* 2008; **38**:1095–109. https://doi.org/10.1139/X07-219.

Gobakken T, Næsset E. Assessing effects of positioning errors and sample plot size on biophysical stand properties derived from airborne laser scanner data. *Can J For Res* 2009; **39**:1036–52. https://doi.org/10.1139/X09-025.

Goodbody TRH, Coops NC, Queinnec M, *et al.* sgsR: a structurally guided sampling toolbox for LiDAR-based forest inventories. *Forestry* 2023; **96**:411–24. https://doi.org/10.1093/forestry/cpac055.

Hamedianfar, A., Mohamedou, C., Kangas, A., Vauhkonen, J. Deep learning for forest inventory and planning: a critical review on the remote sensing approaches so far and prospects for further applications. *Forestry* 2022; **95**:1–15. https://doi.org/10.1093/forestry/cpac002.

Hancock S, Armston J, Hofton M, *et al.* The GEDI simulator: a large-footprint waveform Lidar simulator for calibration and validation of Spaceborne missions. *Earth Space Sci* 2019; **6**:294–310. https://doi.org/10.1029/2018EA000506.

Hawbaker TJ, Keuler NS, Lesak AA, *et al.* Improved estimates of forest vegetation structure and biomass with a LiDAR-optimized

sampling design: LiDAR-optimized sampling. *J Geophys Res Biogeo* 2009; **114**:1–11. https://doi.org/10.1029/2008JG000870.

Hayashi R, Kershaw JA, Weiskittel A, *et al.* Evaluation of alternative methods for using LiDAR to predict aboveground biomass in mixed species and structurally complex forests in northeastern north america. *Math Comput For Nat Resour Sci* 2015; **7**:49–65.

Henniger H, Huth A, Frank K, *et al.* Creating virtual forests around the globe and analysing their state space. *Ecol Model* 2023; **483**:110404. https://doi.org/10.1016/j.ecolmodel.2023.110404.

Holmgren J, Nilsson M, Olsson H. Simulating the effects of LiDAR scanning angle for estimation of mean tree height and canopy closure. *Can J Remote Sens* 2003; **29**:623–32. https://doi.org/10.5589/m03-030.

Kassambara A. ggpubr: 'ggplot2' based publication ready plots. R package version 0.4.0, 2020. https://CRAN.R-project.org/package=ggpubr.

Knapp N, Fischer R, Huth A. Linking LiDAR and forest modeling to assess biomass estimation across scales and disturbance states. *Remote Sens Environ* 2018; **205**:199–209. https://doi.org/10.1016/j.rse.2017.11.018.

Kotivuori E, Korhonen L, Packalen P. Nationwide airborne laser scanning based models for volume, biomass and dominant height in Finland. *Silva Fennica* 2016; **50**(4):1–28. https://doi.org/10.14214/sf.1567. https://www.silvafennica.fi/article/1567.

Lang N, Kalischek N, Armston J, *et al.* Global canopy height regression and uncertainty estimation from GEDI LIDAR waveforms with deep ensembles. *Remote Sens Environ* 2022; **268**:112760. https://doi.org/10.1016/j.rse.2021.112760.

Liaw A, Wiener M. Classification and regression by randomForest. *R News* 2002; **2**:18–22 https://cran.r-project.org/web/packages/randomForest.

Lisańczuk M, Mitelsztedt K, Parkitna K, *et al.* Influence of sampling intensity on performance of two-phase forest inventory using airborne laser scanning. *For Ecosyst* 2020; **7**:65. https://doi.org/10.1186/s40663-020-00277-6.

Maltamo M, Bollandsås OM, Næsset E, *et al.* Different plot selection strategies for field training data in ALS-assisted forest inventory. *Forestry* 2011; **84**:23–31. https://doi.org/10.1093/forestry/cpq039.

McRoberts RE, Næsset E, Gobakken T, *et al.* Indirect and direct estimation of forest biomass change using forest inventory and airborne laser scanning data. *Remote Sens Environ* 2015; **164**:36–42. https://doi.org/10.1016/j.rse.2015.02.018.

Meyer H, Pebesma E. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods Ecol Evol* 2021; **12**:1620–33. https://doi.org/10.1111/2041-210X.13650.

Moudrý V, Cord AF, Gábor L, *et al.* Vegetation structure derived from airborne laser scanning to assess species distribution and habitat suitability: the way forward. *Divers Distrib* 2023; **29**:39–50. https://doi.org/10.1111/ddi.13644.

Nelson R. Modeling forest canopy heights: the effects of canopy shape. *Remote Sens Environ* 1997; **60**:327–34. https://doi.org/10.1016/S0034-4257(96)00214-3.

Nelson R, Oderwald R, Gregoire TG. Separating the ground and airborne laser sampling phases to estimate tropical forest basal area, volume, and biomass. *Remote Sens Environ* 1997; **60**:311–26. https://doi.org/10.1016/S0034-4257(96)00213-1.

Næsset E, Gjevestad JG. Performance of GPS precise point positioning under conifer Forest canopies. *Photogramm Eng Remote Sens* 2008; **74**:661–8. https://doi.org/10.14358/PERS.74.5.661.

Næsset E, Gobakken T. Estimation of above- and below-ground biomass across regions of the boreal forest zone using airborne laser. *Remote Sens Environ* 2008; **112**:3079–90. https://doi.org/10.1016/j.rse.2008.03.004.

Packalen P, Strunk J, Maltamo M, *et al.* Circular or square plots in ALS-based forest inventories—does it matter? *Forestry* 2023; **96**:49–61. https://doi.org/10.1093/forestry/cpac032.

Palace MW, Sullivan FB, Ducey MJ, *et al.* Estimating forest structure in a tropical forest using field measurements, a synthetic model and discrete return LiDAR data. *Remote Sens Environ* 2015; **161**:1–11. https://doi.org/10.1016/j.rse.2015.01.020.

Pretzsch H, Biber P, Ďurský J. The single tree-based stand simulator SILVA: construction, application and evaluation. *For Ecol Manage* 2002; **162**:3–21. https://doi.org/10.1016/S0378-1127(02)00047-6.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2021. https://www.R-project.org/.

Rana P, Gautam B, Tokola T. Optimizing the number of training areas for modeling above-ground biomass with ALS and multispectral remote sensing in subtropical Nepal. *Int J Appl Earth Obs Geoinform* 2016; **49**:52–62. https://doi.org/10.1016/j.jag.2016.01.006.

Roberts O, Bunting P, Hardy A, *et al.* Sensitivity analysis of the DART model for Forest mensuration with airborne laser scanning. *Remote Sens* 2020; **12**:247. https://doi.org/10.3390/rs12020247.

Roussel, J.-R., Auty, D.. Airborne LiDAR data manipulation and visualization for forestry applications. R package version 3.2.3, 2021. https://cran.r-project.org/package=lidR.

Roussel J-R, Auty D, Coops NC, *et al.* lidR: An R package for analysis of airborne laser scanning (ALS) data. *Remote Sens Environ* 2020; **251**:112061. https://doi.org/10.1016/j.rse.2020.112061.

RStudio Team. RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA, 2016. http://www.rstudio.com/.

Schäfer J, Weiser H, Winiwarter L, *et al.* Generating synthetic laser scanning data of forests by combining forest inventory information, a tree point cloud database and an open-source laser scanning simulator. *Forestry* 2023; **cpad006**:1–19. https://doi.org/10.1093/forestry/cpad006.

Spriggs RA, Vanderwel MC, Jones TA, *et al.* A simple area-based model for predicting airborne LiDAR first returns from stem diameter distributions: an example study in an uneven-aged, mixed temperate forest. *Can J For Res* 2015; **45**:1338–50. https://doi.org/10.1139/cjfr-2015-0018.

Stereńczak K, Lisańczuk M, Parkitna K, *et al.* The influence of number and size of sample plots on modelling growing stock volume based on airborne laser scanning. *Drewno Prace Naukowe Doniesienia Komunikaty* 2018; **61**:5–22.

Suvanto A, Maltamo M. Using mixed estimation for combining airborne laser scanning data in two different forest areas. *Silva Fennica* 2010; **44(1)**:91–107. https://doi.org/10.14214/sf.164.

Tompalski P, White JC, Coops NC, *et al.* Demonstrating the transferability of forest inventory attribute models derived using airborne laser scanning data. *Remote Sens Environ* 2019; **227**:110–24. https://doi.org/10.1016/j.rse.2019.04.006.

van Ewijk K, Tompalski P, Treitz P, *et al.* Transferability of ALS-derived Forest resource inventory attributes between an eastern and western Canadian boreal Forest Mixedwood site. *Can J Remote Sens* 2020; **46**:214–36. https://doi.org/10.1080/07038992.2020.1769470.

Vonderach C, Kublin E, Bösch B, *et al.* rBDAT: Implementation of BDAT tree taper Fortran functions. R package version 0.9.8, 2021. https://CRAN.R-project.org/package=rBDAT.

Wang L, Birt AG, Lafon CW, *et al.* Computer-based synthetic data to assess the tree delineation algorithm from airborne LiDAR survey. *GeoInformatica* 2013; **17**:35–61. https://doi.org/10.1007/s10707-011-0148-1.

Weiser H, Schäfer J, Winiwarter L, *et al.* Individual tree point clouds and tree measurements from multi-platform laser scanning in German forests. *Earth System Science Data* 2022a; **14**:2989–3012. https://doi.org/10.5194/essd-14-2989-2022.

Weiser H, Schäfer J, Winiwarter L, *et al.* Terrestrial, UAV-borne, and airborne laser scanning point clouds of central European forest plots, Germany, with extracted individual trees and manual forest inventory measurements. PANGAEA [data set] 2022b. https://doi.org/10.1594/PANGAEA.942856.

Wetzel S, Swift DE, Burgess D, *et al.* Research in Canada's National Research Forests—past, present and future. *For Ecol Manage* 2011; **261**:893–9. https://doi.org/10.1016/j.foreco.2010.03.020.

White JC, Chen H, Woods ME, *et al.* The Petawawa research Forest: establishment of a remote sensing supersite. *For Chron* 2019; **95**: 149–56. https://doi.org/10.5558/tfc2019-024.

White JC, Penner M, Woods M. Assessing single photon LiDAR for operational implementation of an enhanced forest inventory in diverse mixedwood forests. *For Chron* 2021; **97**:78–96. https://doi.org/10.5558/tfc2021-009.

White JC, Wulder MA, Varhola A, *et al.* A best practices guide for generating forest inventory attributes from airborne laser scanning data using an area-based approach. *For Chron* 2013; **89**:722–3. https://doi.org/10.5558/tfc2013-132.

Wickham, H.. *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York, 2016. https://ggplot2.tidyverse.org, https://doi.org/10.1007/978-3-319-24277-4.

Winiwarter L, Esmorís Pena AM, Weiser H, *et al.* Virtual laser scanning with HELIOS++: a novel take on ray tracing-based simulation of topographic full-waveform 3D laser scanning. *Remote Sens Environ* 2022; **269**:112772. https://doi.org/10.1016/j.rse.2021.112772.

Wu J, Zou D, Braverman V, *et al.* The power and limitation of pretraining-finetuning for linear regression under covariate shift. Advances in Neural Information Processing Systems 2022; **35**:33041–33053. http://arxiv.org/abs/2208.01857.

Yang T-R, Kershaw JA Jr, Weiskittel AR, *et al.* Influence of sample selection method and estimation technique on sample size requirements for wall-to-wall estimation of volume using airborne LiDAR. *Forestry: An International Journal of Forest Research* 2019; **92**:311–23. https://doi.org/10.1093/forestry/cpz014.

Yin T, Lauret N, Gastellu-Etchegorry J-P. Simulation of satellite, airborne and terrestrial LiDAR with DART (II): ALS and TLS multi-pulse acquisitions, photon counting, and solar noise. *Remote Sens Environ* 2016; **184**:454–68. https://doi.org/10.1016/j.rse.2016.07.009.

Zhao K, Popescu S, Nelson R. Lidar remote sensing of forest biomass: a scale-invariant estimation approach using airborne lasers. *Remote Sens Environ* 2009; **113**:182–96. https://doi.org/10.1016/j.rse.2008.09.009.

Zhao K, Suarez JC, Garcia M, *et al.* Utility of multitemporal LiDAR for forest and carbon monitoring: tree growth, biomass dynamics, and carbon flux. *Remote Sens Environ* 2018; **204**:883–97. https://doi.org/10.1016/j.rse.2017.09.007.

Zhu X, Liu J, Skidmore AK, *et al.* A voxel matching method for effective leaf area index estimation in temperate deciduous forests from leaf-on and leaf-off airborne LiDAR data. *Remote Sens Environ* 2020; **240**:111696. https://doi.org/10.1016/j.rse.2020.111696.

Zianis D, Muukkonen P, Mäkipää R, *et al.* Biomass and stem volume equations for tree species in Europe. *Silva Fennica Monographs* 2005; **2005**:1–63. https://doi.org/10.14214/sf.sfm4.

Zolkos, S. G., Goetz, S. J., Dubayah, R.. A meta-analysis of terrestrial aboveground biomass estimation using LiDAR remote sensing. *Remote Sens Environ* 2013; **128**, 289–98, https://doi.org/10.1016/j.rse.2012.10.017.