Using Synthetic Datasets to predict Forest Aboveground Biomass from Airborne Laser Scanning Data

Zur Erlangung des akademischen Grades einer

DOKTORIN DER NATURWISSENSCHAFTEN
(Dr. rer. nat)

von der KIT-Fakultät für
Bauingenieur-, Geo- und Umweltwissenschaften des
Karlsruher Instituts für Technologie (KIT)

genehmigte

DISSERTATION

von

Jannika Sophie Schäfer
aus Aachen

Tag der mündlichen Prüfung: 03. Juni 2024
Referent: Prof. Dr. Sebastian Schmidtlein
Korreferenten: Prof. Dr. Fabian Ewald Faßnacht,
Prof. Dr.-Ing. Boris Jutzi

Karlsruhe, 2024

Danksagung

Ich möchte mich bei all den Menschen bedanken, die mich in der Zeit meiner Promotion begleitet und unterstützt haben.

Mein größter Dank gilt Fabian Faßnacht für die Idee zum SYSSIFOSS-Projekt und die weltbeste Betreuung. Du warst immer für mich erreichbar und hast in meine Fähigkeiten vertraut, auch wenn ich mal wieder am Verzweifeln war, und mich auch dann nicht verstoßen, als ich mit Kartoffelkochen fast dein Haus abgebrannt hätte.

Ich möchte mich bei Sebastian Schmidtlein und Bernhard Höfle für ihre Unterstützung und ihr wertvolles Feedback bedanken, sowie meiner gesamten Promotionskommission für die Zeit und Mühe bei der Bewertung meiner Arbeit.

Ich bin meinen Co-Autoren und meiner Co-Autorin dankbar für die extrem konstruktive Kritik, auf die ich mich beim Schreiben immer verlassen konnte. Ich danke außerdem den studentischen Hilfskräften, die bei der Datenaufnahme geholfen haben und den Kollegen und Kolleginnen aus Polen, Tschechien und Kanada, die mir ihre Daten zur Verfügung gestellt haben.

Hannah Weiser, ich bin extrem froh, dich als Kollegin gehabt zu haben. Ohne dich hätte das SYSSIFOSS-Projekt nicht funktioniert.

Lukas Winiwarter, ich danke dir für die Hilfe mit Deep Learning und die Unterstützung, wenn ich mal wieder Schreibhemmungen hatte, sowie für das, was du mir über CNNs hinaus sonst noch so gezeigt hast. Ich bin sehr froh, dass du mir damals vorgeschlagen hattest, meinen Forschungsaufenthalt in Wien zu machen. Ich habe es kein einziges Mal bereut.

Ich danke meinen Kolleginnen und Kollegen am IFGG für die gemeinsamen Mittagessen und Kaffeepausen, es war eine wunderbare Zeit mit euch. Insbesondere möchte ich dem antizyklischen Pistensportverein bestehend aus Felix Schiefer, Pia Labenski, Johannes Senn und Michael Ewald danken, sowie Anne Lewerentz und Elham Shafeian. Danke für eure Geduld, euren Witz, und das Erklären der Witze. Pia, ich bin unglaublich dankbar, in dir eine Freundin gefunden zu haben, mit der ich über alles sprechen kann. Felix, ich bin froh darüber, dass wir all die Jahre das Büro geteilt haben. Unsere Freitagnachmittage im 7. Stock werden mir immer in guter Erinnerung bleiben. Eli, ich danke dir dafür, dass du mit mir Kashk-e-Bademjan gekocht hast, obwohl du gegen Auberginen allergisch bist.

Ich danke GRACE für die finanzielle Unterstützung, die mir die Teilnahme an Konferenzen und meinen Forschungsaufenthalt ermöglicht hat.

Ich habe durch meine Promotion viele tolle Menschen kennen gelernt, deren Bekanntschaft dazu geführt hat, dass ich mich auf Konferenzen nicht alleine gefühlt habe, oder die mein

Leben anderweitig bereichert und/oder meine Arbeit unterstützt haben (auch wenn ich manche davon nie persönlich getroffen habe). Dazu gehören Moritz Bruggisser, Aline Bornand, Anna Iglseder, die INSANE-Gruppe aus Karlsruhe und Ljubljana, Hans Henniger und Grzegorz Krok. Ich werde es vermissen, nicht mehr teil der Waldfernerkundungs-Community zu sein.

Ich hatte eine wahnsinnig schöne Zeit bei meinem Forschungsaufenthalt in Wien was insbesondere an den Menschen lag, die ich dort kennen lernen durfte und dir mir das Gefühl gegeben haben, genau so richtig zu sein, wie ich bin. Ich habe mich an der TU Wien dank der unglaublich netten Kolleginnen und Kollegen extrem wohl gefühlt. Ein ganz herzlicher Dank geht daher an die Photo-Gruppe rund um Norbert Pfeifer. Es wäre mir in Wien nicht so gut gegangen, wenn es nicht Kai und die Doppelkopfrunde gegeben hätte. Durch euch habe ich Karlsruhe keine Sekunde vermisst.

Die letzten Jahre hätte ich nicht ohne die Pandemiegruppe (die schon vor Corona so hieß) überstanden. Die Spieleabende, -wochenenden und -urlaube waren immer ein großes Vergnügen. Sara und Valentin, ich danke euch dafür, dass ihr immer ein offenes Ohr für mich habt. Benni, danke, dass du immer freundlich bleibst, wenn wir dir das Lesen und Verstehen der Spieleanleitungen komplett alleine überlassen. Feli und Steven, euch danke ich für die unglaubliche Hilfsbereitschaft und Pia und Reiner für die unübertreffliche Gastfreundschaft. Johannes, ich bin dankbar, in dir einen Wanderpartner mit dem perfekten Schritttempo und der gleichen Vorliebe für lange Strecken gefunden zu haben. Mit dir gehen nie die Diskussionsthemen aus.

Alex, ich danke dir für den Beistand in den letzten Jahren. Auch wenn wir uns durchgehend in den Haaren haben, war es oft auch sehr schön mit dir. Und nur durch dich habe ich den Corona-Lockdown als eine tolle Zeit in Erinnerung.

Thorben, ich bin sehr froh, dass wir das monatliche Seebaden seit so vielen Jahren gemeinsam durchziehen. Es tut immer wieder gut, den inneren Schweinehund zu überwinden und festzustellen, dass es im Nachhinein gar nicht so schlimm war.

Jannik, vielen Dank für die Kletterausflüge. Du stellst immer die richtigen Fragen, wenn ich meine Probleme schildere (sowohl in Bezug auf die Promotion als auch aufs Leben im Allgemeinen).

Judith und Bella, es gibt nichts, was mir vor euch peinlich ist. Danke, dass ihr aus blöden Ereignissen witzige macht, allein dadurch, dass ich sie euch erzählen kann.

Zu guter Letzt möchte ich mich bei meiner Familie bedanken. Mama und Papa, danke dafür, dass ihr mir absolut vertraut und mir vollkommen freie Wahl in meiner Lebensgestaltung lasst. Laurin und Maora, danke für eure zahlreichen Besuche in Karlsruhe und für eure Beratung bei Design-Fragen und sonstigen Problemen. Valentin und Büşra, ich bin froh, dass es euch gibt. Darüber hinaus gilt mein Dank auch der ganzen Großfamilie rund um Oma Wisi und Opa Bruno. Durch diese unsere Familie haben sich meine Naturverbundenheit und mein Umweltbewusstsein entwickelt, die mich dazu bewogen haben, Geoökologie zu studieren, was letztendlich zu dieser Promotion geführt hat.

Danke an alle, die mir zugehört haben, wenn ich mal wieder Redebedarf hatte (also immer).

Abstract

Airborne laser scanning (ALS) data are increasingly being used to estimate forest aboveground biomass (AGB). However, AGB information cannot be directly derived from ALS data. Instead, regression models are built to link metrics describing the ALS point clouds to field-measured AGB estimates. Therefore, the use of ALS data for estimating AGB is limited by the availability of field data, which are time-consuming and costly to collect. One potential solution for overcoming this limitation is the use of synthetic data generated by computer simulations.

This thesis aims to evaluate the potential of synthetic data for training AGB models that can subsequently be applied to real ALS data. The workflow for generating the synthetic data involves a forest simulator, realistic tree models, and a laser scanning simulator. Three-dimensional forest scenes are created based on forest stand information generated using the forest simulator Forest Factory 2.0 and point clouds of individual trees extracted from real laser scanning data acquired by an unoccupied aerial vehicle (UAV). The lidar simulator HELIOS++ is then used to simulate airborne laser scanning of these forest scenes. Based on three studies, this thesis evaluates the synthetic data in terms of their performance as training data for AGB models.

The first study investigates whether applying HELIOS++ on forest scenes composed of real tree point clouds can generate realistic ALS data. The scenes are created based on the stand composition of real forest plots located in southwestern Germany and laser scanning is simulated with the same parameters as in the real acquisitions. This allows for comparison between the simulated and real ALS data from the same plots. In addition to the real tree point clouds, simplified tree point clouds with cylindrical stems and spheroidal crowns are used as tree models to assess the influence of tree model complexity on the usability of the synthetic data. The analysis reveals that using simplified tree models results in more accurate values for canopy cover, while the height distribution of the returns is better represented by the simulated data based on real tree models for most study sites. Training AGB models on point cloud metrics derived from the simulated data shows that using real tree models results in an overprediction of AGB when applied to real ALS data, whereas using simplified tree models results in an underprediction of AGB. In both cases, the prediction accuracy in terms of root mean squared error (RMSE) is slightly worse than for models trained on metrics derived from real data.

While the first study demonstrates that AGB models perform better when trained on real data than on synthetic data, the question remains as to whether synthetic data could still be valuable in situations where little to no real data are available. Consequently, the second study aims to determine if the prediction performance of AGB models, trained on a limited sample of real data, can be enhanced by extending the training dataset with synthetic data. Additionally, the study compares models trained solely on synthetic data to models trained on real *ex situ* data, i.e. data collected from different study sites than the one where the model is applied. Four real datasets from study sites in Poland, the Czech Republic (2 $\times$), and Canada are used for the experiments. Synthetic datasets comprising of AGB estimates and ALS point cloud metrics for 2 500 plots are created using Forest Factory 2.0 and HELIOS++ with the identical ALS settings as the real acquisitions. The analysis demonstrates that extending the training dataset with synthetic data can improve the accuracy of AGB

predictions in terms of RMSE and squared Pearson correlation coefficient (r^2), as long as a limited number of real training samples is available (12 to 346, depending on the study site). However, using synthetic data for model training consistently results in a strong underprediction of AGB. For three of the four study sites, training models on real *ex situ* data yields prediction accuracies comparable to models trained on data from the respective sites. AGB models trained on synthetic data only outperform models trained on real *ex situ* data in case of the Polish dataset.
The second study confirms that using real data to train AGB models results in higher prediction accuracies compared to using synthetic data and reveals that this is in most cases true even when the real data are collected from different sites. The advantage of using synthetic data only becomes apparent when there is a very limited amount of real training data available. This rises the question of whether applications with an extremely high demand for training data, such as deep learning methods, can particularly benefit from the utilization of synthetic data. A novel approach is developed for predicting AGB from ALS point clouds: three-dimensional convolutional neural networks (CNNs) are trained to predict AGB based on sequences of images derived from cross sections of the point clouds. The initial weights of the CNNs are transferred from a 2D CNN pre-trained on ImageNet data. The CNNs are compared to random forest regression models based on typical point cloud metrics, using the same real datasets as in the second study. Further, the study investigates whether the predictive performance can be improved by additionally pre-training the CNNs on cross section images derived from synthetic data. As in the previous studies, the results vary across study sites and depend on the number of real samples available for model training. In most cases, the CNNs perform slightly worse than the random forest models. Pre-training on synthetic data only improves the CNN predictions when the number of real training samples is extremely limited.

This thesis demonstrates that synthetic data cannot yet substitute real data in AGB models without a noticeable decrease in prediction accuracy. It appears that the synthetic data differ from real data in a way that affects their suitability as training data. The generation of synthetic data needs to be improved in terms of the realism of laser scanning simulations, forest stand composition, tree models, and the placement of tree models. Furthermore, this work shows that making real datasets publicly available could be a more promising solution for addressing the issue of limited training data in ALS-based predictions, rather than using synthetic data. Nevertheless, synthetic data still hold potential in applications that necessitate precise information on individual trees and the ability to test different laser scanning acquisition settings and field sampling methods, such as sensitivity analyses or the development of methods at the individual tree level.

Kurzfassung

Airborne (luftgestütztes) Laserscanning (ALS) wird zunehmend für die Schätzung der oberirdischen Biomasse (AGB) von Wäldern verwendet. AGB-Informationen lassen sich jedoch nicht direkt aus ALS-Daten ableiten. Daher werden Regressionsmodelle erstellt, die von den ALS-Punktwolken abgeleitete Metriken mit im Feld erhobenen AGB-Schätzungen verknüpfen. Dementsprechend ist die Verwendung von ALS-Daten zur Schätzung der AGB limitiert durch die Verfügbarkeit der Felddaten, deren Erhebung zeitaufwändig und kostspielig ist. Ein Ansatz, diese Limitierung zu umgehen, ist die Verwendung von durch Computersimulationen erzeugten synthetischen Daten. Ziel dieser Arbeit ist es, das Potenzial synthetischer Daten für das Training von AGB-Modellen, die auf realen Daten angewendet werden, zu bewerten. Der Ansatz zur Erzeugung der synthetischen Daten umfasst den Waldsimulator Forest Factory 2.0, realistische Baummodelle und den Laserscanning-Simulator HELIOS++. Dreidimensionale Waldszenen werden auf der Grundlage von simulierten Waldbestandesinformationen und Punktwolken einzelner Bäume, die aus realen Laserscanning-Daten extrahiert wurden, erstellt. ALS dieser Waldszenen wird mit HELIOS++ simuliert. Auf Grundlage von drei Studien evaluiert diese Arbeit die synthetischen Daten hinsichtlich ihrer Eignung als Trainingsdaten für AGB-Modelle.

In der ersten Studie wird untersucht, ob die Anwendung von HELIOS++ auf Waldszenen, die aus echten Baumpunktwolken bestehen, realistische ALS-Daten erzeugen kann. Die Waldszenen werden entsprechend der Zusammensetzung echter Waldbestände in Südwestdeutschland erstellt und das Laserscanning wird mit den selben Einstellungen wie in den echten Aufnahmen simuliert. Dies ermöglicht einen Vergleich zwischen simulierten und echten ALS-Daten derselben Flächen. Neben den realen Baumpunktwolken (RTM) werden vereinfachte Baumpunktwolken (STM) mit zylindrischen Stämmen und Kronen in Form von Rotationsellipsoiden als Baummodelle verwendet, um den Einfluss der Komplexität der Baummodelle auf das Anwendungspotential der synthetischen Daten zu erfassen. Die Analyse ergibt, dass mit den STM ein realistischerer Kronenschlussgrad erzeugt werden kann, wohingegen die Höhenverteilung der Reflexionen für die meisten Untersuchungsgebiete besser über die RTM reproduziert werden kann. Wenn AGB-Modelle mit von simulierten ALS-Punktwolken abgeleiteten Metriken trainiert und danach auf echten Daten angewendet werden zeigt sich, dass die Verwendung der RTM zu einer Überschätzung der AGB führt, während die Verwendung der STM zu einer Unterschätzung führt. In beiden Fällen ist die über den Root Mean Squared Error (RMSE) erfasste Modellgüte etwas schlechter als bei Modellen, die auf echten Daten trainiert wurden.

Während die erste Studie zeigt, dass AGB-Modelle eine höhere Modellgüte aufweisen, wenn sie mit echten Daten trainiert werden, blieb die Frage offen, ob synthetische Daten einen Mehrwert liefern können, wenn wenige oder gar keine realen Daten verfügbar sind. Daher wird in der zweiten Studie untersucht, ob die Modellgüte von AGB-Modellen, die auf einer begrenzten Anzahl echter Daten trainiert wurden, durch die Ergänzung des Trainingsdatensatzes mit synthetischen Daten verbessert werden kann. Darüber hinaus werden Modelle, die ausschließlich auf synthetischen Daten trainiert wurden, mit solchen verglichen, die auf *Ex-situ*-Daten trainiert wurden. Letztere umfassen Daten, die in anderen Untersuchungsgebieten erhoben wurden als dem, in dem das Modell angewendet wurde. Für die Experimente werden vier reale Datensätze aus Untersuchungsgebieten in

Polen, Tschechien (2 ×) und Kanada verwendet. Synthetische Datensätze bestehend aus AGB-Werten und ALS-Punktwolkenmetriken von 2 500 Plots werden mit Forest Factory 2.0 und HELIOS++ unter Verwendung derselben ALS-Aufnahmeparameter wie bei den realen Aufnahmen erstellt. Die Analyse zeigt, dass die Ergänzung des Trainingsdatensatzes mit synthetischen Daten die Modellgüte der AGB-Schätzungen in Bezug auf den RMSE und den quadrierten Pearson-Korrelationskoeffizienten (r^2) verbessern kann, solange wenig reale Trainingsdaten zur Verfügung stehen (12 bis 346 je nach Untersuchungsgebiet). Die Verwendung synthetischer Daten für das Modelltraining führt jedoch durchweg zu einer signifikanten Unterschätzung der AGB. Bei drei der vier Untersuchungsgebiete erreichen Modelle, die mit echten *Ex-situ*-Daten trainiert wurden, eine ähnliche Modellgüte wie mit *In-situ*-Daten trainierte Modelle. Nur für den polnischen Datensatz erzielen AGB-Modelle, die mit synthetischen Daten trainiert wurden, höhere Modellgüten als *Ex-situ*-Modelle.
Die zweite Studie bestätigt die Ergebnisse der ersten Studie: Die Verwendung realer Daten für das Training von AGB-Modellen führt zu einer höheren Modellgüte als die Verwendung synthetischer Daten, und das in den meisten Fällen auch dann, wenn die Daten in anderen Gebieten erhoben wurden. Ein Nutzen der synthetischer Daten zeigt sich nur, wenn nur sehr wenige reale Trainingsdaten verfügbar sind. Dadurch ergibt sich die Frage, ob Anwendungen mit einem extrem hohen Bedarf an Trainingsdaten, wie z. B. Deep-Learning-Methoden, insbesondere von der Verwendung synthetischer Daten profitieren könnten. Die dritte Studie befasst sich daher mit einem neuartigen Ansatz zur Schätzung von AGB aus ALS-Punktwolken, bei dem mit dreidimensionalen Convolutional Neural Networks (CNNs) AGB aus Bildsequenzen von Querschnitten der Punktwolken abgeleitet wird. Die initialen Gewichte der CNNs werden transferiert von einem mit dem ImageNet-Datensatz vortrainierten 2D-CNN. Die Schätzgüte der CNNs wird mit der eines auf Punktwolkenmetriken basierenden Random-Forest-Modells verglichen. Außerdem wird untersucht, ob die Modellgüte durch zusätzliches Vortraining der CNNs auf Querschnittsbildern von synthetischen Daten verbessert werden kann. Die Ergebnisse variieren zwischen den Untersuchungsgebieten und hängen von der Anzahl der verwendeten realen Trainingsdaten ab. Zumeist schneiden die CNNs etwas schlechter ab als die Random-Forest-Modelle. Das Vortraining auf synthetischen Daten verbessert die CNN-Schätzungen nur dann, wenn die Anzahl der realen Trainingsdaten sehr klein ist (10 bis 40).
Diese Arbeit zeigt, dass synthetische Daten reale Daten in AGB-Modellen noch nicht ersetzen können, ohne die Schätzgenauigkeit zu verringern. Die synthetischen Daten weisen im Vergleich zu realen Daten Diskrepanzen auf, die ihre Eignung als Trainingsdaten beeinträchtigen. Die Methode zur Erzeugung synthetischer Daten muss daher in Bezug auf die Realitätsnähe der Laserscanning-Simulationen, der Zusammensetzung des Waldbestands, der Baummodelle und der Platzierung der Baummodelle verbessert werden. Darüber hinaus hat diese Arbeit gezeigt, dass die Veröffentlichung realer Datensätze eine vielversprechendere Lösung für das Problem der begrenzten Trainingsdaten bei ALS-basierten AGB-Schätzungen ist als die Verwendung synthetischer Daten. Nichtsdestoweniger haben synthetische Daten großes Potenzial für Anwendungen, die genaue Informationen über einzelne Bäume und die Möglichkeit zum Testen verschiedener Laserscanner-Aufnahmeparameter und Feldaufnahmemethoden erfordern, wie beispielsweise Sensitivitätsanalysen oder die Entwicklung von Einzelbaum-basierten Methoden.

Contents

List of Publications

1. Jannika Schäfer, Hannah Weiser, Lukas Winiwarter, Bernhard Höfle, Sebastian Schmidtlein, & Fabian Ewald Fassnacht (2023). 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: An International Journal of Forest Research*, cpad006.

2. Jannika Schäfer, Lukas Winiwarter, Hannah Weiser, Jan Novotný, Bernhard Höfle, Sebastian Schmidtlein, Hans Henniger, Grzegorz Krok, Krzysztof Stereńczak, & Fabian Ewald Fassnacht (2023). Assessing the potential of synthetic and *ex situ* airborne laser scanning and ground plot data to train forest biomass models. *Forestry: An International Journal of Forest Research*, cpad061.

3. Jannika Schäfer, Lukas Winiwarter, Hannah Weiser, Bernhard Höfle, Sebastian Schmidtlein, Jan Novotný, Grzegorz Krok, Krzysztof Stereńczak, Markus Hollaus, & Fabian Ewald Fassnacht (submitted). CNN-based transfer-learning for aboveground biomass prediction from ALS point cloud tomography. Submitted to *European Journal of Remote Sensing*.

The content and structure of Chapter 2, Chapter 3, and Chapter 4 were kept in the form of the original publications or submission, respectively.

Acronyms and Abbreviations

AGB	Aboveground biomass
ALS	Airborne laser scanning
cd	Crown diameter
CNN	Convolutional neural network
$\mathbf{d}_{1.3}$**,** $\mathbf{D}_{1.3}$	Diameter at breast height (measured at 1.3 m above ground)
DAP	Digital aerial photogrammetry
DBH	Diameter at breast height
DGNSS	Differential global navigation satellite system
DN	DendroNET sites
DTM	Digital terrain model
lidar	Light detection and ranging
FAO	Food and Agriculture Organization of the United Nations
GNSS	Global navigation satellite system
IMU	Inertial measurement unit
IPCC	Intergovernmental Panel on Climate Change
ME	Mean error
MF	Milicz Forest
PRF	Petawawa Research Forest
QSM	Quantitative structure model
radar	Radio detection and ranging
RMSE	Root mean squared error
RTK	Real-time kinematic positioning
RTM	Real tree models
SB	Silesian Beskids
STM	Simplified tree models
TLS	Terrestrial laser scanning
UAV	Unoccupied aerial vehicle
ULS	UAV-borne laser scanning

Chapter 1

1 INTRODUCTION

1.1 IMPORTANCE OF FORESTS

Forests provide a wide range of ecosystem services (Bologna & Aquino, 2020; FAO, 2018b; Favero et al., 2020; Turner-Skoff & Cavender, 2019): They produce oxygen, store carbon, and regulate the water cycle. They also prevent soil erosion and filter air and water. Furthermore, they are habitats for humans and animals, and provide food, medicine, fuel, and building materials. The presence of trees has positive effects on the mental and physical health of people (Turner-Skoff & Cavender, 2019). Forested areas affect the climate by decreasing temperature due to evapotranspiration, increasing temperature due to an decrease in surface albedo, and increasing rainfall due to an increase in evaporation (Alkama & Cescatti, 2016; Spracklen et al., 2012). Forests contribute significantly to terrestrial biodiversity and 80% of all known amphibian species, 75% of bird species, and 68% of mammal species find habitat in forests. Tropical forests alone contain 60% of all vascular plants (FAO and UNEP, 2020).

The FAO defines forests as "land spanning more than 0.5 hectares with trees higher than 5 meters and a canopy cover of more than 10%", including areas with trees that have not yet but are expected to reach these thresholds (FAO, 2018a). In 2020, the global forest area was approximately 4.1×10^{12} ha, covering 31% of the global land area. Tropical forests account for 45% of the global forest area, 27% of forests are in boreal regions, 16% in temperate regions, and 11% in subtropical regions. In total, 54% of the global forest area is located in only five countries, namely Russia, Brazil, Canada, the USA, and China (FAO and UNEP, 2020).

While climate change impacts the structure and functioning of forest ecosystems (Achim et al., 2022), forest loss is also in turn one of the main drivers of climate change. Forests can be both a source and a sink of greenhouse gases. Forest-related greenhouse gas emissions are caused by deforestation and forest degradation due to land-use change, droughts, wildfires, and insect infestations (Pan et al., 2011). The dominant greenhouse gas emitted from these actions is CO_2 (Harris et al., 2021). Carbon sequestration is increased by increasing forest area due to afforestation and reforestation, by forest recovery and growth following the abandonment of agricultural use, grazing or harvesting, a change in forest management strategies, and CO_2 fertilization and N deposition (Pan et al., 2011). Overall, global forests are considered a net carbon sink; however, its size is subject to large uncertainties and estimates range from $1.1 \pm 0.8\,\mathrm{Pg\ CO_2eq\,year^{-1}}$ (Pan et al., 2011) to $7.6 \pm 49\,\mathrm{Pg\ CO_2eq\,year^{-1}}$ (Harris et al., 2021).

According to the IPCC, reducing deforestation and forest degradation has the highest economic potential for the short-term mitigation of carbon emissions (IPCC, 2022). Although the global deforestation rate has decreased since 1990, approximately 178 million hectares of forest area have been lost in the following 30 years, mainly due to land-use change to agriculture (FAO, 2020).

Accurate estimation of carbon stocks and fluxes is important to support actions to mitigate climate change (Pan et al., 2011). Annex I parties of the United Nations Framework Convention for Climate Change (industrialised countries and countries with economies in transition) are obliged to annually report on their emissions and removals of direct

greenhouse gases from five sectors, including the sector *land-use change and forestry* (decision 24/CP.19), while non-Annex I parties (mostly developing coutries) have to submit biennial reports on their national greenhouse gas inventories (Annex III of decision 2/CP.17). National forest inventories are a primary source of information for these reports (Tomppo & Schadauer, 2012).
The direct measurement of forest carbon stocks is not feasible. Instead, forest biomass is used as a proxy for forest carbon. Carbon estimates are derived by multiplying the biomass estimates with the C fraction, i.e. the woody tissue carbon concentration. The carbon fraction of aboveground biomass (AGB) is often assumed to be 50% (Doraisami et al., 2022). However, it should be noted that using simplified conversion factors can introduce significant bias to forest carbon estimates, as wood carbon concentration varies among tree species and forest biomes (Martin et al., 2018). In most forests, the main carbon pools are AGB and soil organic matter. Additionally, carbon is also stored in root biomass, woody debris, and the forest floor. Monitoring approaches for forest carbon pools typically focus on AGB, because it is the most dynamic pool and at the same time the easiest to measure (Fahey et al., 2010).

AGB refers to the dry mass of all tree elements above ground, including stems, branches, and leaves, and is commonly reported in megagrams per hectare (White et al., 2013a). The actual measurement of AGB is costly and destructive, as trees must be harvested to be dried and weighed (Andersson et al., 2009). As a result, AGB is usually estimated using species-specific allometric equations based on measurements of diameter at breast height (DBH) and additional parameters such as upper stem diameter or tree height (White et al., 2013a). When information on individual tree dimensions is not available, species-specific biomass expansion factors (BEFs) can be used to derive biomass estimates from volume information at the stand level (Brown, 2002). However, these estimates may be significantly biased because the relationship between volume and biomass depends not only on tree species but also on factors such as age and site characteristics (Ameztegui et al., 2022). Furthermore, estimates derived from allometric equations are also subject to high uncertainty. Ameztegui et al. (2022) showed that using different allometric equations resulted in significant discrepancies in biomass estimates, especially for large trees.

1.2 FOREST INVENTORIES

The primary objective of forest inventories is to gather information about forest area and forest attributes such as age, species distribution, and the growing stock volume, including the amount of timber (Ilvessalo, 1927, as cited in McRoberts et al., 2010). With growing interest in sustainable forest management beyond timber supply, the scope of forest inventories has been broadened to encompass the estimation of carbon stocks and fluxes, as well as information on forest health and biodiversity (Corona et al., 2011; McRoberts et al., 2010). For example, the Austrian national forest inventory collects additional data on the volume of standing and lying deadwood, the abundance and occurrence of shrubs, the stand structure, unusual tree shapes, and rare tree and shrub species (Tomppo et al., 2010).

In Sweden, forest inventories also assess the presence of bent trees, the height of stumps, and damage caused by grazing elk (Fridman et al., 2014).

The first forest inventories were conducted at the end of the Middle Ages, when forest planning became necessary in areas where wood was scarce due to increasing demand for forest resources (Loetsch & Haller, 1973; Gabler & Schadauer 2007, both as cited in McRoberts et al., 2010). Arguments for the need to collect data on forest conditions as part of sustainable forest management, as well as instructions on how to carry out these surveys, can be found in publications by German forest scientists at the end of the 18th and beginning of the 19th century (Cotta, 1804; Hartig, 1795). Systematic forest inventories based on statistical sampling have been conducted since the early 20th century. Norway was the first country to conduct a systematic forest inventory at a national level (1919–1930), followed by Finland (1921–1924) and Sweden (1923–1929). Today, most North American and European countries, as well as other countries such as China and South Korea, regularly conduct systematic forest inventories at the national level. Typical inventory cycles are 5 or 10 years (Tomppo et al., 2010). Due to different inventory traditions, forest conditions, and information needs, inventory methods vary from country to country (Gschwantner et al., 2022).

Due to time and cost constraints, it is impossible to measure all trees in all forested areas (McRoberts & Tomppo, 2007). It is therefore common to use sample-based inventory designs. In the 1980s, the first countries (Austria, Switzerland, Sweden, Germany, Norway, and Spain) established permanent plots, making it easier to monitor changes (Gschwantner et al., 2022). Plot positions are in many countries determined using systematic sampling based on a two-dimensional grid (Tomppo et al., 2010). In Europe, the grid sizes range from 1.0 km $\times$ 0.5 km to 5.0 km $\times$ 5.0 km and 20.0 km $\times$ 20.0 km (Gschwantner et al., 2022).

More than 90% of national forest inventories use circular plots, but data are also collected from square and rectangular plots, along transects, and using angle count sampling (Bitterlich, 1952) resulting in variable radius plots (Tomppo et al., 2010). Circular plots have the advantage of having the smallest ratio of plot boundary to plot area, reducing the number of boundary trees, and their position can be marked with only one marker at the plot centre (Gschwantner et al., 2022). To reduce workload, it is common to not measure all trees in a plot but to sample a subset of trees based on their DBH. This is often done in nested concentric circles, where small trees are only measured in the smallest circle, and trees with a larger DBH are measured in a larger radius. Alternatively, an angle count sampling can be applied. Variables recorded for sampled trees include species, DBH, height, crown base height, and upper stem diameter (Tomppo et al., 2010). The minimum DBH of trees to be recorded varies from 0.0–12.7 cm, depending on the country (McRoberts et al., 2010). While the measurement of DBH is comparatively fast and precise, other measurements such as of tree height and upper diameters are more time-consuming and subject to larger uncertainties. These variables are therefore often measured only for a randomly or systematically selected subsample (10–35%) of the DBH-measured trees. Missing values can then be estimated from species-specific regression models based on the DBH (Gschwantner et al., 2022; White et al., 2013a). The measurements of individual trees are aggregated to derive plot level metrics such as stem number, basal area, mean tree height, Lorey's

height (the average stand height weighted by the basal area of the trees), dominant height, growing stock volume, and AGB (White et al., 2013a).

Field plots are geolocated using global navigation satellite system (GNSS) receivers. For the sampling of trees and the measurement of diameters, Bitterlich relascopes, callipers, and diameter tapes are used, while electronic devices such as (in)clinometers coupled with ultrasonic distance meters are frequently used to measure distances and heights (Tomppo et al., 2010). In the future, terrestrial or mobile sensors for light detection and ranging (lidar) might also be used for these measurements (Bauwens et al., 2016; Ghimire et al., 2017).

The number of variables recorded in forest inventories has increased with rising interest in forest information. In European national forest inventories, it is common to collect data on 100–400 variables. Accordingly, these inventories are becoming more complex, time-consuming, and expensive to conduct. The implementation of remote sensing data in forest inventories can reduce related costs and increase the speed, precision, and timeliness of data collection, while improving the accuracy and spatial resolution of resulting map products (McRoberts & Tomppo, 2007).

1.3 REMOTE SENSING TECHNIQUES FOR FOREST INVENTORIES

The implementation of remote sensing data in forest inventories has evolved from using hand-drawn maps and aerial photographs on paper to satellite imagery with initially coarse and later increasingly fine spatial resolution. Time series of remote sensing data, such as multi-temporal satellite data, enable change analysis over time. Additionally, technologies like stereophotogrammetry, radio detection and ranging (radar), and lidar provide information on forest structure, specifically canopy height (Coops et al., 2021; Lister et al., 2020). The increasing availability of remote sensing data, along with advances in processing technologies and cloud computing systems, has greatly improved the ability to monitor forests (Lister et al., 2020). Therefore, the use of remote sensing data is becoming more and more common in operational programs of many countries.

Remote sensing data used to support forest inventories are collected with both passive and active sensors. Passive sensors detect and measure naturally occurring electromagnetic radiation that has been emitted or reflected from a body. Active sensors emit radiation and measure the amount of radiation that is reflected or backscattered to the sensor. Most commonly, sensors are mounted on aircraft (airborne sensors) or satellites (spaceborne sensors), but in recent years, they have also been increasingly mounted on unoccupied aerial vehicles (UAVs). Platforms can also be terrestrial (i.e. tripods) and mobile (i.e. harvesters or humans). However, due to occlusion effects, sensors mounted on these platforms cannot be used for large-scale data collection in forests.

The only common passive remote sensing data type used in forest inventories is optical imagery. In contrast to active techniques, optical remote sensing relies on solar illumination and is therefore sensitive to cloud cover and daylight. Sensors collecting optical imagery

differ in their spatial and spectral resolution, spatial extent, coverage, temporal resolution, and operational period (Andersson et al., 2009).

Aerial photography acquired by multispectral cameras is the most common remote sensing data source that is operationally used in national forest inventories, in some countries already for more than 50 years (Barrett et al., 2016). Its primary application is for planning forest inventories, i.e. to classify forest and non-forest plots and to design the stratified sampling of field plots. The often very high spatial resolution of these data (below 0.25 m) allows estimation of the cover, development stage, species, and density of trees in forest plots and grouping them accordingly into strata (Barrett et al., 2016; McRoberts et al., 2010). Given sufficient overlap between adjacent images, multiple view-points, and homologous points identifiable in the images and texture, 3D models can be generated from the optical imagery. The digitisation of aerial photography and the development of image matching algorithms have facilitated the generation of aerial photogrammetric point clouds. With many countries regularly acquiring aerial photographs, there is a growing interest in using digital aerial photogrammetry (DAP) products for forest inventory purposes (Stepper et al., 2017). At a countrywide level, DAP has been used, for example, in Switzerland for deriving canopy heights and for the automated mapping of forest cover (Ginzler & Hobi, 2015; Waser et al., 2015). A significant disadvantage of DAP is that information can only be derived from the canopy envelope (White et al., 2013b).

The opening of the Landsat archive in 2008 and the launch of the Sentinel satellites resulted in an increasing use of satellite data to obtain forest information (Gschwantner et al., 2022), progressively replacing aerial photography (McRoberts et al., 2010). In the early 2010s, the most frequently used satellite data in national forest inventories were medium spatial resolution (10–30 m) multispectral data collected from the Landsat and SPOT satellites. Only a few countries used ultrahigh resolution (< 5 m) imagery such as collected from Quickbird and WorldView-2 (Barrett et al., 2016). The temporal resolution of most satellite data is sufficient for forest monitoring; however, the spatial resolution of commonly used satellite sensors is too coarse for some forest inventory requirements (Falkowski et al., 2009). The primary application of satellite data is the prediction of forest attributes over large areas based on field-measured data. Satellite data are also visually interpreted by experts to define areas with similar forest properties, sometimes following an automated segmentation, for example in Canada. One of the first countries to use satellite data in an operational way for forest inventories was Finland, starting in the late 1980s (Barrett et al., 2016).

Active remote sensing techniques such as radar or lidar measure the time between emission and return of a signal to determine the distance to a reflecting/backscattering object. They differ in the wavelength of the emitted electromagnetic radiation (centimeter to meter for radar, micrometer for lidar), causing different scattering mechanisms. Both radar and lidar allow obtaining information on the vertical structure of forests (McRoberts & Tomppo, 2007).

Since its first applications in the late 1990s, ALS has become an increasingly popular remote sensing technology for deriving structural forest information. Until the late 2000s, it had only been applied for inventories at the level of individual stands. The increasing

efficiency of ALS sensors has enabled the collection of data at larger scales and higher point density. Therefore, the integration of ALS data in forest inventories is becoming more frequent (Gschwantner et al., 2022). For example, Denmark started the periodic nationwide mapping of forest attributes based on ALS data in combination with forest inventory data, resulting in maps for the years 2006–2007 and 2014–2015 (Kangas et al., 2018). ALS data are twice to three times more expensive than optical imagery and more difficult to handle due to their large size and the inherent property of being unstructured. Hence, extensive pre-processing is required for further analysis (Lister et al., 2020; White et al., 2013b). While optical imagery provides spectral information, multispectral ALS is still in its infancy and rarely used (Kukkonen et al., 2019). However, in comparison to DAP, ALS has the advantage that it can penetrate forest canopies through small gaps, and thus it also provides information from the sub-canopy layers. The information contained in the three-dimensional point clouds that can be generated by laser scanning is of high value for deriving attributes that are related to the forest structure. It is therefore expected that ALS data will become even more important in future forest inventories (Lister et al., 2020). A comprehensive description of the functioning of ALS and its application in forest inventories is provided in Chapter 1.4.

In contrast to the notable increase in ALS acquisitions in recent decades, there have been only a few spaceborne lidar missions to date, such as ICESat with the Geoscience Laser Altimeter System (GLAS) and ICESat-2 with the Advanced Topographic Laser Altimeter System (ATLAS). The topographic measurements of these sensors can be used to derive vegetation canopy heights in footprints of approximately 50–100 m diameter (Abdalati et al., 2010; Abshire et al., 2005; Schutz et al., 2005). In 2018, the Global Ecosystem Dynamics Investigation (GEDI) was launched, providing data since March 2019. GEDI is the first spaceborne lidar mission specifically designed to map ecosystem structures (Dubayah et al., 2020). The instrument which is attached to the International Space Station (ISS) collects full-waveform lidar data for footprints of 25 m diameter. The openly available data include raw waveforms and information on canopy height, biomass, canopy cover, leaf area index, and topography. However, a drawback of all spaceborne lidar missions to date is that they sample along transects rather than collecting area-wide data, making it difficult to conduct wall-to-wall mapping of forest attributes. Therefore, spaceborne lidar data are usually combined with optical or radar satellite data to extend the derived forest attributes over larger areas (Coops et al., 2021).

Radar techniques such as synthetic aperture radar (SAR) and interferometric SAR (InSAR) allow for the collection of spatially exhaustive data regardless of weather and daylight conditions. The wavelength of the outgoing signal determines the penetration depth into the vegetation cover. The most commonly used SAR in forest remote sensing is L-band SAR (e.g. PALSAR), which is particularly suitable for detecting changes in biomass. Shorter wavelengths such as C-band and X-band (e.g. Sentinel-1 and Tandem X mission) are also used but only provide information from the outer canopy (Coops et al., 2021). The backscatter signals are affected not only by canopy structure but also by other factors such as incident angle and soil moisture, making the interpretation of the signal more complex than, for example, of optical imagery and lidar point clouds (Bae et al., 2019). Many studies have demonstrated the usability of SAR and InSAR data for estimating forest

attributes, often in combination with other remote sensing data types (Coops et al., 2021; McRoberts & Tomppo, 2007). However, according to a survey presented by Barrett et al. (2016), radar data have not been used operationally in national forest inventories in any of the 45 countries included in the study.

Forest inventories are originally aimed at answering questions about the quantity of forest resources. Remote sensing data enable to answer questions about the spatial distribution of these resources (McRoberts & Tomppo, 2007). Due to the sample-based approaches applied in forest inventories, it is not possible to derive wall-to-wall maps of forest attributes based on forest inventory data alone. However, in combination with remote sensing data covering the whole area of interest, it is possible to create full coverage maps (Fridman et al., 2014).

The implementation of remote sensing data can reduce the cost of forest inventories in several ways: (1) the data can be used to select plots to be sampled, i.e. using stratified sampling approaches based on information derived from remote sensing data, also excluding plots in non-forested areas, (2) when plots are difficult to access, interpretation of remote sensing data can replace field measurements, (3) the efficiency of wall-to-wall estimates can be significantly increased. McRoberts and Tomppo (2007) reported that for eight forest areas in eastern Finland, the sampling intensity could be decreased from 500 field plots per 10 000 ha to 35 field plots per 10 000 ha when using Landsat imagery as an additional data source for estimating growing stock volume. For two large forest sites in Minnesota, USA, they estimated that field plot selection based on stratified sampling could save costs of about US $12.4 million compared to simple random sampling.

While remote sensing has proven to be a valuable data source for deriving information on forest attributes, field data collection remains essential for forest inventories. Field measurements are required to develop and train predictive models and to validate the information derived from remote sensing data (Coops et al., 2021). In the future, it may be possible to replace field measurements by collecting spectral information and generating high-density laser scanning point clouds with UAVs. Depending on the data quality and forest type, these data enable the segmentation of individual trees, the identification of tree species, and the determination of DBH and height (Puliti et al., 2020). When linking remote sensing data to field observations, accurate information on plot positions is crucial, especially in diverse and fragmented forests. Time lags between field data collection and remote sensing data acquisition can also contribute to prediction errors when forest characteristics have changed due to harvesting or natural damages such as fire, wind throw, drought, or insect infestations (Barrett et al., 2016). Uncertainties in the field reference data, for example resulting from errors in tree measurements, plot localisation, and modelling of tree attributes, not only affect the quality of prediction models but can also negatively impact the evaluation of remote sensing-based predictions. This means that predictions derived from remote sensing data may be more accurate than suggested by validation based on field data (Persson et al., 2022). Accordingly, the application and development of new methods for using remote sensing data in forestry are highly dependent on the availability of accurate field data.

1.4 AIRBORNE LASER SCANNING FOR FOREST INVENTORIES

Laser scanning is an active remote sensing technology based on electro-optical measurements. Shortly after its development in the early 1960s, land surveyors started using light amplification by stimulated emission of radiation (laser) for distance measurements. The first airborne applications were laser profilers in the 1970s and 1980s, followed by laser scanning in the 1990s, when new technologies enabled direct georeferencing of the measurements (Shan & Toth, 2018).

A laser scanning system consists of a laser scanner, an inertial measurement unit (IMU), an integrated differential GNSS (DGNSS), and a control unit. Laser scanning utilises lidar to collect 3D data of a target surface. It involves emitting a laser beam that is then backscattered by a surface, with the returned signal being received. For continuous wave lasers, the distance between sensor and surface is determined by the phase difference of both signals using amplitude modulation. In ALS, pulse lasers are commonly used, where a short laser pulse of high energy is emitted and the system measures the time-of-flight (Δt) until the backscattered pulse is received. By knowing the speed of light (c), the distance to the surface (R) can be calculated as: $R = \frac{1}{2}c\Delta t$. Combining the distance measurement with information on the orientation of the laser beam allows for determining the vector to the backscattering surface. The orientation of the laser beam is defined by two angles. To create a scanning pattern, a moving mirror is used to change the angle of the laser beam in one direction. A third dimension of scanning can be achieved by moving the platform on which the scanning system is mounted. Different mirror movements result in various ground patterns, e.g. an oscillating mirror creates a zigzag-line, while a rotating polygon mirror creates parallel lines. The IMU and DGNSS provide information on the position and orientation of the laser scanning system, i.e. the origin and direction of the laser beam in a global coordinate frame, which are necessary for computing the coordinates of the point from which the laser beam is returned (Wehr & Lohr, 1999).

According to the mechanism for detecting the returned signal, laser systems are categorised into full waveform, discrete return, or photon counting systems (Shan & Toth, 2018). Full waveform systems digitise the received analogue signal with a high frequency (Hollaus et al., 2014), while discrete return systems detect and record only the peaks in the analogue signal that are above a given threshold (Shan & Toth, 2018). Photon counting systems detect single photons arriving at the sensor, which means that the laser pulses can be of lower energy. Photon counting enables lidar from high altitudes using high pulse repetition frequencies, making the data collection more efficient, especially for larger areas (White et al., 2021). However, due to the low number of available commercial systems, the use of photon counting for forestry applications is still in its infancy.

The footprint size of a laser beam depends on the distance between the sensor and the surface, i.e. the flying altitude, and the beam divergence (White et al., 2013a). The footprint can be circular or elliptical. All surfaces that are within the illuminated footprint contribute to the signal that is returned to the sensor, depending on the surface properties and the respective fraction of the laser beam that interacts with the surfaces (Lefsky et al., 2002;

Shan & Toth, 2018). Multiple reflections can be separated when the spatial distance of the surfaces exceeds a threshold defined by the pulse width. In the case of ALS of forests, a first return of the laser signal is often caused by the tree tops, while a fraction of the laser beam may penetrate the canopy through small gaps, resulting in additional returns from branches, stems, and the ground (Shan & Toth, 2018).

For ALS, the laser system is mounted on a fixed-wing aircraft or a helicopter. Helicopters can fly in more complex terrain, at lower altitudes, and at lower speeds, resulting in higher point densities, but helicopter-borne acquisitions are also more expensive. In the frame of forest inventories, it is common to use discrete return laser systems that operate with short laser pulses (3–10 ns) in the near infrared (800–1550 nm). Pulse repetition frequencies are commonly in the range of 50–200 kHz and the scan angle is typically restricted to be $< 25^{\circ}$ off nadir. At flying altitudes of 500–3 000 m, the resulting pulse density is 0.5–20 pulses/m^2. It is recommended to conduct the acquisitions with a swath overlap of at least 50% to ensure good ground coverage and the option for precise alignment in post-processing, i.e. strip-adjustment (White et al., 2013a).

The usability of ALS for forest inventories has been proven in numerous studies and operational applications. In the most commonly used approach, known as the area-based approach (Næsset, 2002), field inventory measurements of forest attributes are linked to metrics derived from ALS data of the same plots. These metrics are then used to develop predictive models, both parametric and non-parametric, with the forest attributes as the response variable and the ALS metrics as predictors. The resulting models can be applied to the entire area of interest for which ALS data are available (White et al., 2013a). To derive wall-to-wall ALS metrics and corresponding predictions, a grid with a cell size similar to the field plot size is utilised. The predictors used in the models include descriptive statistics of the ALS point coordinates, such as mean height, height percentiles, and cumulated densities (White et al., 2013a), or metrics derived from a canopy height model (Chirici et al., 2016). Metrics can be calculated from first returns, last returns, intermediate returns, or all returns, with many studies applying a height threshold to exclude returns from understorey vegetation (White et al., 2013a).

It has been shown that estimates derived from ALS data can even be more accurate than using traditional methods, such as stand-wise field inventories (Holmgren & Jonsson, 2004; Holopainen et al., 2014). Numerous studies have demonstrated that typical forest inventory attributes, such as top height, Lorey's mean tree height, dominant height, mean stem diameter, stem number, basal area, and growing stock volume can be accurately predicted from ALS data (Holmgren & Jonsson, 2004; McRoberts et al., 2013; Næsset, 2004, 2007; Nord-Larsen & Schumacher, 2012; Sheridan et al., 2015; White et al., 2021). ALS has also been shown to be a valuable data source for estimating stocks and changes in forest AGB (Cao et al., 2016; Chirici et al., 2016; McRoberts et al., 2015; Nord-Larsen & Schumacher, 2012; Sheridan et al., 2015). AGB predictions based on ALS data are more accurate than those based on spaceborne lidar, radar, or passive optical data (Zolkos et al., 2013). Additionally, ALS data can be used to predict forest attributes of high ecological importance, such as understory cover and structure (Korpela et al., 2012; Maltamo et

al., 2005; Wing et al., 2012), forest stand age (Racine et al., 2014), forest edge structure (Bruggisser et al., 2024), and the presence of snags and shrubs (Martinuzzi et al., 2009). Furthermore, ALS data provide information that is difficult to assess in traditional forest inventories, such as estimates of leaf area index (Pope & Treitz, 2013) and soil carbon pools (Strîmbu et al., 2023).

In addition to demonstrating the use cases of ALS data for forest inventories, there has also been extensive research on factors that influence ALS-based predictions of these attributes. Multiple studies have investigated how field plot size and shape, field inventory design, georeferencing errors, and sampling of field plots affect prediction accuracies (Deo et al., 2016; Ene et al., 2017; Frazer et al., 2011; Gobakken & Næsset, 2009; Hawbaker et al., 2009; Maltamo et al., 2007; Packalen et al., 2023; Zhao et al., 2009). When planning ALS campaigns, it is important to know which acquisition settings are best for forest inventory purposes. Therefore, several studies have examined the effects of flying altitude, scanning mode, scan angle, beam divergence, footprint size, pulse repetition frequency, and point density on ALS-based estimates (Chasmer et al., 2006; Gobakken & Næsset, 2008; Goodwin et al., 2006; Hopkinson, 2007; Keränen et al., 2016; Lim et al., 2008; Montaghi, 2013; Morsdorf et al., 2008). Furthermore, as models are needed to derive estimates of forest attributes from ALS data, studies have also compared different prediction methods (e.g. parametric vs. non-parametric models, deep learning approaches vs. traditional machine learning approaches, single tree-based methods vs. area-based methods) and tested various sets of ALS metrics (Bouvier et al., 2015; Chirici et al., 2016; Latifi et al., 2015; Y. Li et al., 2008; Oehmcke et al., 2022; Pearse et al., 2019).

A critical consideration for evaluating methods to derive forest information from ALS data, as well as for the informative value of sensitivity analyses, is the quality of the underlying reference data. This includes their accuracy and representativeness. Additionally, the data requirements for method development and sensitivity analyses can exceed what is typically collected in forest inventories. For instance, information regarding individual tree positions, both within and outside of the field plots, may be necessary to develop new methods or analyse the effects of geolocation errors and the shapes and sizes of field plots. Moreover, conducting sensitivity analyses of ALS acquisition settings may necessitate multiple surveys, which is rarely possible due to high acquisition costs. Simulations offer a cost- and time-efficient alternative to acquiring real data. They also provide the added benefit of allowing for control over all conditions (White et al., 2016). Therefore, synthetic data generated through simulations have significant potential for enhancing the understanding of the relationships between statistical attributes of ALS data and forest characteristics. This understanding is crucial for planning ALS acquisitions as well as for improving existing methods and developing new ones (White et al., 2016).

1.5 SYNTHETIC DATA

Generating synthetic ALS data of forests requires a method to simulate laser scanning and three-dimensional forest scenes that the laser scanning simulation is applied to. To create

the forest scenes, information on the forest composition is needed, including the positions of individual trees and their properties such as species, height, and crown diameter. Tree models can then be created and arranged based on this forest information to build three-dimensional representations of forests. The laser scanning simulation approach, the method for generating forest composition, and the tree models utilised all impact the realism of the synthetic data and their potential applications.

Synthetic data have been used for calibrating and comparing models (Nelson, 1997; Palace et al., 2015) as well as for selecting the most suitable predictors for ALS-based predictions of forest attributes (Fischer et al., 2019; Knapp et al., 2018). They can also be employed for lookup table approaches, e.g. for linking stand diameter distributions to lidar return height distributions (Spriggs et al., 2015) or gross primary productivity to GEDI waveforms (Bauer et al., 2021). Data simulations are highly valuable for sensitivity analyses, where the influence of laser scanning acquisition settings and resulting point density, field plot size, geolocalization errors, edge effects, and forest structure on various ALS-derived estimates, including AGB, canopy closure, and other forest attributes, as well as the retrieval of forest foliage profiles and the accuracy of tree delineation, can be assessed individually and without cross-correlation effects (Disney et al., 2010; Fassnacht et al., 2018; Frazer et al., 2011; Holmgren et al., 2003; Knapp et al., 2021; Qin et al., 2017; Stocker et al., 2023; Wang et al., 2013). Synthetic data also contribute to the comprehension of the relationships between forest structure, laser scanning settings, and the resulting characteristics of ALS data (Bruening et al., 2021; Fischer et al., 2019; Goodwin et al., 2007).

The simulation approaches that have been applied in studies using synthetic ALS data of forests differ in complexity. A few studies used real field inventory data to create forest stands (e.g. Knapp et al., 2021; Nelson et al., 1997; Spriggs et al., 2015). However, most studies used randomly generated stand compositions, with random or regularly distributed tree locations and tree sizes drawn from bounded random distributions. These studies often assumed fixed values or fixed ratios for crown diameters and crown heights (e.g. Goodwin et al., 2007; Holmgren et al., 2003; Kukko & Hyyppä, 2009; Palace et al., 2015; Qin et al., 2017; Stocker et al., 2023; Wang et al., 2013). More complex forest models offer the potential to generate more realistic forest stand compositions. These models must be individual-based and provide spatially explicit output to be suitable for creating three-dimensional forest scenes that are required for laser scanning simulations. Forest growth models simulate the development of forest structure under different environmental conditions and/or forest management strategies. They consider the growth and mortality of individual trees and the competition among them. The forest models that have been used for generating synthetic ALS data of forests include FORMIND (Köhler & Huth, 1998), SILVA (Pretzsch et al., 2002), and Forest Factory (Bohn & Huth, 2017). FORMIND is an individual-based forest gap model. Gap models simulate forest growth on small patches, assuming that the horizontal competition among individual trees is the same for all trees within the area (Shugart et al., 2018). Therefore, they do not explicitly model the positions of individual trees. In studies using FORMIND (Bauer et al., 2021; Fischer et al., 2019; Knapp et al., 2018), tree positions have been randomly assigned within a 20 m $\times$ 20 m patch. SILVA has been used by Fassnacht et al. (2018) to generate synthetic ALS point clouds of forest stands. Unlike gap models, SILVA includes three-dimensional competition

among neighbouring trees (Pretzsch et al., 2002). As a result, the simulation output is spatially explicit at an individual tree level. While FORMIND and SILVA enable long-term simulations of forest stands, Forest Factory is a forest model that does not simulate the development of forest stands over time. Instead, it generates realistic forest states based on the competition and productivity processes implemented in FORMIND. By omitting the time component, Forest Factory is significantly faster than FORMIND, allowing for the generation of a large number of different forest stand compositions in a short time. A first study using Forest Factory to generate synthetic forest lidar data was presented by Bruening et al. (2021).

For generating three-dimensional forest representations, simple geometric objects are often used as tree models. Tree crowns are assumed to be cylindrical (Knapp et al., 2018), spheroidal (Knapp et al., 2021), sinusoidal (Kukko & Hyyppä, 2009), ellipsoidal (Palace et al., 2015; Wang et al., 2013), paraboloid (Frazer et al., 2011), or hemi-ellipsoidal (Goodwin et al., 2007; Holmgren et al., 2003; Nelson et al., 1997). Tree stems are either not explicitly modelled or modelled as cones (Frazer et al., 2011) or cylinders (Wang et al., 2013). A common alternative to these primitive-based tree models are voxel models, which also simplify the data processing. Voxel sizes range from $0.25\,\mathrm{m} \times 0.25\,\mathrm{m} \times 0.25\,\mathrm{m}$ (Goodwin et al., 2007) to $0.5\,\mathrm{m} \times 0.5\,\mathrm{m} \times 1.0\,\mathrm{m}$ (Bauer et al., 2021; Knapp et al., 2018), and $1.0\,\mathrm{m} \times 1.0\,\mathrm{m} \times 1.0\,\mathrm{m}$ (Knapp et al., 2021). Some studies use solid tree models that laser beams cannot penetrate (Holmgren et al., 2003), others assume the same penetration properties for stems and crowns (Wang et al., 2013) or solid stems but penetrable crowns, e.g. by modelling the tree crown as a turbid medium (Knapp et al., 2018) or by using different densities of voxels to simulate a clumped foliage distribution (Goodwin et al., 2007). Much more detailed models with representations of individual needles and leaves have been used by Disney et al. (2010) and Qin et al. (2017). While those tree models were artificially generated, Stocker et al. (2023) used tree models of high detail that were reconstructed from real TLS data using L-Architect (Côté et al., 2009). A different approach of using real data was presented by Fassnacht et al. (2018), who used individual tree point clouds extracted from real ALS data without any further modification but scaling and rotation.

The complexity of the laser scanning simulation approach determines the potential use cases of the synthetic data. Simplistic approaches can generate data that are sufficient for model calibration or the analysis of field plot effects on ALS-based estimations of forest attributes. In contrast, sensitivity analyses of ALS acquisition settings require a methodology that allows for the simulation of different settings. Simplified approaches were, for example, applied by Nelson et al. (1997), who derived airborne laser profiles from the canopy shape of the synthetic stands in a height array of $0.25\,\mathrm{m} \times 0.25\,\mathrm{m}$ (with added noise to create surface texture). Fassnacht et al. (2018) calculated canopy height models directly from the transformed real ALS point clouds of all trees arranged as a forest stand without applying any laser scanning simulation. Frazer et al. (2011) created ALS point clouds based on real laser scanning xy-return coordinates that were used to extract z-values from the synthetic forest stands. The lidar simulation model implemented in FORMIND applies a probabilistic approach: lidar pulses are represented by vertical voxel columns, and the probability of returns is calculated using an exponential decay function based on the Beer-Lambert law of light extinction (Knapp et al., 2018). A probabilistic approach was

also applied by Wang et al. (2013): out of a defined number of lidar footprints per unit area, they randomly sampled which of these footprints resulted in valid returns. A random offset in the xy-direction was added to the footprint coordinates and the height of the returns was determined from the synthetic forest height at that location with an additional random negative offset to simulate penetration to the canopy. A more realistic simulation of laser scanning can be achieved via ray tracing. Holmgren et al. (2003) and Goodwin et al. (2007) applied ray tracing approaches with no beam divergence to analyse the effect of scan angle on the estimation of canopy closure and mean tree height, and on beam interception, respectively. Disney et al. (2010) used Monte Carlo ray tracing of canopy scattering for a sensitivity analysis of multiple scanning settings. Kukko and Hyyppä (2009) presented a simulator for small footprint ALS that incorporates platform movement (including wind-caused drift), includes beam divergence and the modelling of intensity, and allows for the simulation of different scanning patterns, pulse frequencies, and scan angles. A physically based model that has been shown to reproduce small footprint ALS discrete return metrics with high accuracy and precision (Roberts et al., 2020) is the Discrete Anisotropic Radiative Transfer (DART) model (Gastellu-Etchegorry et al., 2015, 2016). DART has for example been applied by Qin et al. (2017) to analyse the effects of ALS acquisition settings on the retrieval of forest foliage profiles. Stocker et al. (2023) used the laser scanning operations simulator HELIOS++ (Winiwarter et al., 2022) to analyse the influence of point density on the performance of semantic segmentations. HELIOS++ is particularly suitable for laser scanning simulations focusing less on the physical accuracy of individual beams but on the laser scanning survey in general. In a comparison of DART and HELIOS++, Winiwarter et al. (2022) found that loading scenes in HELIOS++ takes approximately ten times longer than in DART (55 min vs. 4.4 min), whereas the beam simulations are much faster (200,000 beams/s in HELIOS++ vs. 0.64 beams/s in DART). HELIOS++ allows for the easy configuration of acquisition settings such as movement of the platform, pulse repetition frequency, scan frequency, and scan angle, enabling the reproduction of real laser scanning campaigns and the testing of different settings with low effort.

To summarize, the elements that are required for generating realistic synthetic ALS data of forests, namely tree models reflecting the appearance of real trees, a method for generating forest stand compositions resembling real-world conditions, and a framework for complex laser scanning simulations, have already been developed and applied in multiple studies. However, an approach that combines all of these elements and explores their potential for supporting forest inventories is still missing.

1.6 RESEARCH NEEDS AND THESIS OUTLINE

Ideally, an approach for generating synthetic data should fulfil the following requirements:

1. **Realistic lidar simulations:** The simulation approach should include beam divergence and the generation of multiple returns.
2. **Variation of acquisition settings:** The simulation approach should allow for the definition of platform movement and the scanning parameters.
3. **Realistic forest composition:** The species, locations, and size distribution of trees within a forest stand should resemble real-world conditions.
4. **Variability of forest stands:** The simulation approach should enable the generation of a wide range of differently composed forest stands in short time.
5. **Realistic tree models:** Tree models should represent the inter- and intraspecific variability of real trees. The simulated laser beams should be able to penetrate tree canopies through gaps but not tree stems.

Such an approach enables comprehensive sensitivity analyses, and realistic synthetic data ensure the transferability of methods and findings derived from synthetic data to the real world. For example, a tree segmentation algorithm trained on synthetic data is probably more suitable for application on real data when synthetic forest stands are composed of realistic tree models rather than simple geometric objects. The ability to define and tune acquisition settings allows for the reproduction of real laser scanning campaigns. If the synthetic stands represent the full range of forest composition at a study site, the simulated data can be used to train models that can then be applied on real data.

However, none of the approaches for generating synthetic ALS data of forests presented in Chapter 1.5 fulfils all of these requirements. These approaches include simplified methods for generating ALS point clouds that do not allow for variations in acquisition settings, they use simplified tree models of limited variability, or they are based on randomly generated forest stand compositions that do not necessarily reflect real-world conditions. Therefore, I have developed a new approach for generating synthetic laser scanning data of forests: based on the approach presented by Fassnacht et al. (2018), three-dimensional forest stands are created using a forest simulator to generate realistic stand compositions, and individual tree point clouds extracted from real laser scanning data are used as tree models. Laser scanning simulations of the voxelised 3D scenes are conducted with HELIOS++.

As presented in Chapter 1.4, ALS data have been shown to be a reliable source for estimating AGB. However, a major limitation for widespread use of ALS data in deriving AGB estimates is the need for reference data to train ALS-based regression models. Therefore, this thesis aims to determine whether the presented approach can generate synthetic ALS data that are realistic enough to be used as training data for AGB models and thus help to overcome or at least reduce the need for field data collection. This question is addressed in three studies, two of which have been published in international peer-reviewed scientific journals, while the third is currently under review. An overview of the data and methodologies employed in the studies is depicted in Figure 1.1.

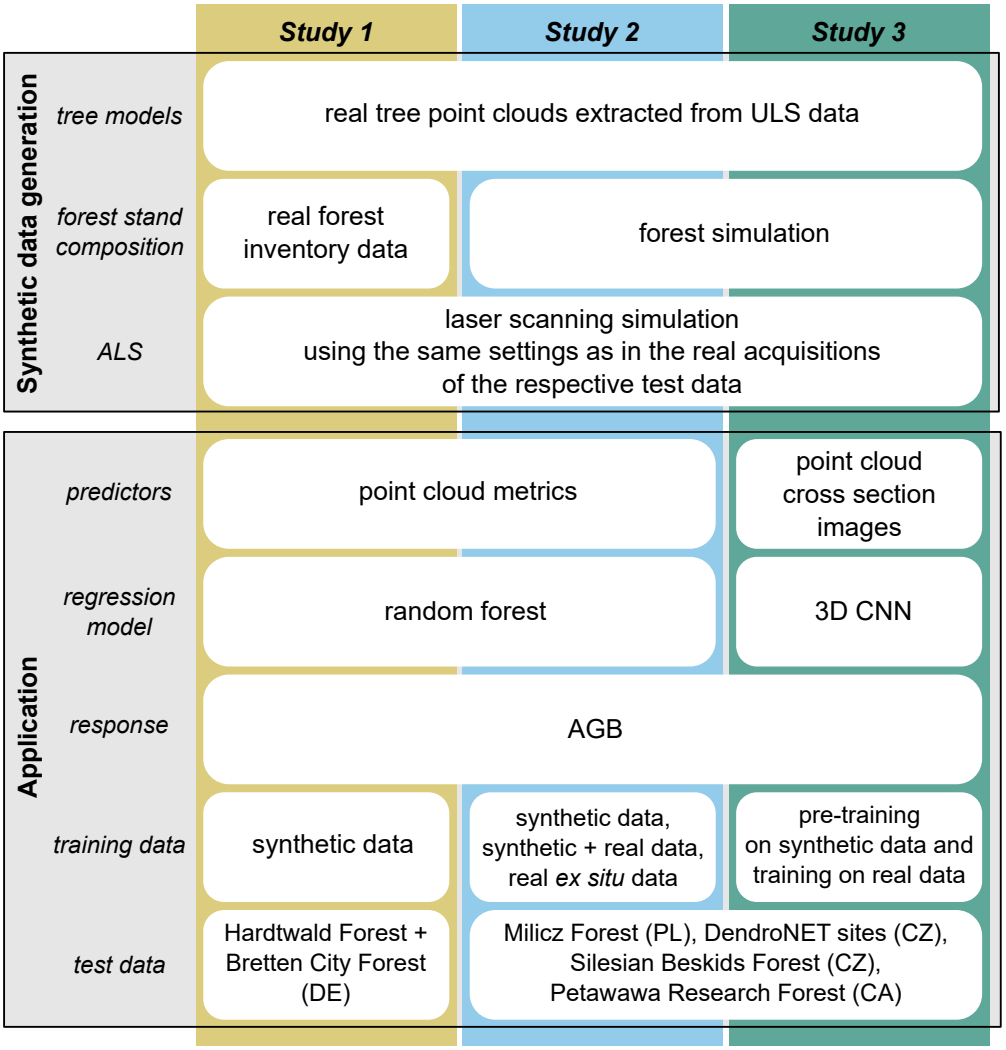

Figure 1.1: Overview of data and methods used in the three studies included in this thesis.

In the first study (Chapter 2), I validate the workflow for simulating ALS data by comparing simulated and real ALS point clouds of mixed forest stands located in the southwest of Germany. The individual tree point clouds that are used as tree models have been extracted from UAV-borne laser scanning data of the same area. I create the synthetic forest stands using forest inventory data from the real forest stands instead of using simulated forest stand compositions, allowing for a direct comparison of simulated and real ALS data. I then apply HELIOS++ on voxelised stand models composed of individual tree point clouds and the resulting simulated point clouds are compared to the real ones based on their height distribution and derived canopy cover. In addition to the real tree point clouds, I use simplified tree models in the form of artificially generated tree point clouds with cylindrical stems and spheroidal crowns to investigate the influence of tree model shape on

the simulated ALS point clouds. AGB models are trained on point cloud metrics derived from both simulated and real ALS data to assess how the differences between real and simulated point clouds affect the accuracy of AGB predictions. The research questions addressed in this study are as follows:

1. Can the combination of HELIOS++ and a forest stand point cloud be used to generate realistic synthetic ALS data?
2. To what extent do the simulated laser scanning data differ when simplified tree models are used instead of real tree point clouds?
3. Can simulated laser scanning data be used to accurately predict forest AGB based on real ALS data?

In the second study (Chapter 3), I investigate the use of fully simulated data for training AGB models. In contrast to the first study, where real forest inventory data are used to generate the synthetic stands, the second study includes the implementation of a forest simulator to generate forest stand compositions. I use Forest Factory 2.0 (Henniger et al., 2023) to simulate a large number of forest stands, which are subsequently used to generate synthetic ALS datasets. Laser scanning simulations are conducted according to the respective acquisition settings of four real laser scanning campaigns located in Poland, the Czech Republic (2×) and Canada. Datasets of field reference data and ALS point clouds from these four acquisitions are used as test data for AGB models trained on the synthetic data. Point clouds of the same trees as in the first study are used as tree models, making the synthetic data completely independent of the real data used for testing. I assess the performance of AGB models trained on these synthetic data compared to models trained on real data derived from the three other sites respectively, and models trained on real data derived from the respective site where they are applied. I also investigate how the training sample size affects model accuracy and whether model accuracy can be increased by extending small datasets of real training data with additional synthetic data. The following research questions are answered in this study:

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

In the third study (Chapter 4), I present a new method for predicting AGB from ALS point clouds. Most existing approaches for ALS-based AGB predictions rely on metrics that describe the vertical distribution of lidar returns as predictor variables. Accordingly, these approaches often omit information provided by the x- and y-coordinates of the returns. For instance, this information could be valuable for determining the number of trees within a plot. In this study, I propose a completely different approach: I use a deep learning algorithm to predict AGB from images depicting cross sections through the ALS point

clouds. Since deep learning methods require a large amount of training data, I further investigate whether synthetic data can be used to address this problem when real data are limited. Accordingly, this final study explores two research questions:

1. To what extent can AGB of forest stands be estimated from stacked cross section images derived from ALS point clouds using a 3D version of the VGG16 convolutional neural network (CNN) with initial weights transferred from the 2D VGG16 pre-trained on the ImageNet dataset?
2. How does additional pre-training on synthetic data influence the prediction performance of the CNN?

Chapter 2

2 GENERATING SYNTHETIC LASER SCANNING DATA OF FORESTS

This chapter has been published as:

Jannika Schäfer, Hannah Weiser, Lukas Winiwarter, Bernhard Höfle, Sebastian Schmidtlein & Fabian Ewald Fassnacht (2023). **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: An International Journal of Forest Research*, cpad006. https://doi.org/10.1093/forestry/cpad006.

ABSTRACT

Airborne laser scanning (ALS) data are routinely used to estimate and map structure-related forest inventory variables. The further development, refinement, and evaluation of methods to derive forest inventory variables from ALS data require extensive datasets of forest stand information on an individual tree-level and corresponding ALS data. A cost-efficient method to obtain such datasets is the combination of virtual forest stands with a laser scanning simulator.

We present an approach to simulate ALS data of forest stands by combining forest inventory information, a tree point cloud database and the laser scanning simulation framework HELIOS++. ALS data of six 1-ha plots were simulated and compared to real ALS data of these plots. The synthetic 3D representations of the forest stands were composed of real laser scanning point clouds of individual trees that were acquired by an uncrewed aerial vehicle (UAV), and, for comparison, simplified tree models with cylindrical stems and spheroidal crowns. The simulated ALS point clouds of the six plots were compared to the real point clouds based on canopy cover, height distribution of returns, and several other point cloud metrics. In addition, the performance of biomass models trained using these synthetic data was evaluated.

The comparison revealed that, in general, both the real tree models and the simplified tree models can be used to generate synthetic data. The results differed for the different study sites and depending on whether all returns or only first returns were considered. The measure of canopy cover was better represented by the data of the simplified tree models, while the height distribution of the returns was – for most of the study sites – better represented by the real tree model data. Training biomass models with metrics derived from the real tree model data led to an overestimation of biomass, while using metrics of the simplified tree model data resulted in an underestimation of biomass. Still, the accuracy of models trained with simulated data was only slightly lower compared to models trained with real ALS data.

Our results suggest that the presented approach can be used to generate ALS data that are sufficiently realistic for many applications. The synthetic data may be used to develop new or refine existing ALS-based forest inventory methods, to systematically investigate the relationship between point cloud metrics and forest inventory variables, and to analyse how this relationship is affected by laser scanning acquisition settings and field reference data characteristics.

2.1 INTRODUCTION

Airborne laser scanning (ALS) data of forests enable a variety of applications. They can be used to estimate, for example, canopy cover (Arumäe & Lang, 2018; Hopkinson & Chasmer, 2009; Smith et al., 2009), leaf area index (Pope & Treitz, 2013), crown base height (Popescu & Zhao, 2008), stand density (Næsset, 2002; Pearse et al., 2019), basal area (Bouvier et al., 2015; Næsset, 2002; Pearse et al., 2019), canopy height (Næsset, 2002; Pearse et al., 2019), stem volume (Bouvier et al., 2015; Pearse et al., 2019), aboveground biomass (Ahmed et al.,

2013; Bouvier et al., 2015; Fassnacht et al., 2014; Næsset et al., 2011; Tsui et al., 2012), and habitat suitability (Graf et al., 2009; Hagar et al., 2014; Johnston & Moskal, 2017). When measured repeatedly, some of these variables can help track forest growth or decline and thus changes in forest carbon stocks (e.g. Ene et al., 2017; McRoberts et al., 2015).

Such forest inventory variables are typically not derived from ALS point clouds directly, but rather with the help of empirical models. In the commonly applied area-based approach, the variable of interest is measured in field reference plots and linked to point cloud metrics calculated from ALS data of the same plots (Næsset, 2002; White et al., 2013a). Using the field- and the ALS data, a parametric or nonparametric regression model can be built that can be used to predict the variable of interest for the entire area for which point cloud metrics are available (Frazer et al., 2011). The collection of field reference data is labour-intensive and costly, so the number of field samples is often limited (Fassnacht et al., 2014; Wulder et al., 2012). Accordingly, the development and evaluation of methods to derive forest inventory variables from ALS data is hampered by the small amount of reference data in most studies (Fassnacht et al., 2018). To better understand the relationship between point cloud metrics and forest characteristics, and to give recommendations for data acquisition and modelling, extensive datasets of ALS data and corresponding field data are needed (Fassnacht et al., 2018; Knapp et al., 2018; White et al., 2016). However, acquisition costs are typically a major obstacle to obtain such datasets.

Simulated laser scanning of virtual forest stands can be a cost-effective alternative method to real acquisitions (White et al., 2016). The key idea is that, based on detailed forest stand information, i.e. species, size and position of every tree in the forest stand, a matching laser scanning point cloud is simulated, resulting in a dataset where both laser scanning data and tree information are available for the entire area. While this might be of limited interest when using real forest stand information (which is rarely available at the required individual tree-level), it offers great potential when combined with synthetic forest inventory data. Realistic, spatially explicit forest stand compositions with varying age structures and species compositions can be simulated using individual tree-based forest growth simulators, such as SILVA (Pretzsch et al., 2002) or FORMIND (Fischer et al., 2016). Simulating laser scanning point clouds of synthetic forest stands can help to overcome the bottlenecks related to limited reference data for forest ALS datasets. In the ideal case, these synthetic datasets can be built for any forest stand composition and laser scanning sensor type and acquisition settings.

Related approaches for simulating airborne laser scanning have been suggested in earlier studies and were for example used to test new methods for tree delineation (Wang et al., 2013) and for the estimation of the effective leaf area index (Zhu et al., 2020), or to identify the best ALS-based predictors for forest biomass estimations (Knapp et al., 2018). Because locations and properties of all trees in the virtual forest are known, the synthetic data also enable to analyse certain effects which are hard to examine with real data. These include for example the influence of plot size and co-registration error between ALS and field reference data on the estimation of forest inventory variables (Frazer et al., 2011) or the effects of different sampling strategies and plot designs. The simulation of the laser scanning process makes it possible to additionally investigate how different sensors and acquisition settings

influence the ALS-based estimation of forest characteristics (Disney et al., 2010; Holmgren et al., 2003).

Synthetic data can also be used to calibrate models which can thereafter be applied to real ALS data to predict forest parameters. For example, ALS point cloud height characteristics can be simulated for a variety of forest stand compositions and stored in a look-up table, along with the related forest stand information. By comparing the height characteristics of the real ALS data with the simulated data in the look-up table, an estimate for the stand information of the real data can be derived (Palace et al., 2015; Spriggs et al., 2017). Alternatively, predictive equations can be derived from the synthetic data and can then be used to infer forest stand composition from the real ALS data. For example, Nelson et al. (1997) developed regression models between simulated canopy height profiles and forest inventory metrics, such as basal area, tree volume and aboveground biomass, and then applied the models to real data.

The success of any workflow using synthetic ALS data naturally depends on the realism of the synthetic datasets. One key component for simulating ALS data of forests on an individual tree-level is the 3D representation of trees implemented in the simulation approach. In earlier studies, tree models of virtual stands were often composed of simple geometric shapes, such as cones, ellipsoids, spheres, paraboloids, and cylinders (e.g. Frazer et al., 2011; Holmgren et al., 2003; Knapp et al., 2018; Milodowski et al., 2021; Nelson, 1997; Palace et al., 2015; Wang et al., 2013). The usage of more complex tree models, e.g. created by the tree generation algorithm of Weber and Penn (1995) as implemented in the Arbaro software or by the OnyxTree software (http://www.onyxtree.com/index.html), is less common, and the studies that followed such an approach often used only a small number of distinct tree models (e.g. Disney et al., 2010; Zhu et al., 2020). One exception is a study by Roberts et al. (2020) who used 700 tree models generated by the OnyxTree software for creating virtual forest scenes. Instead of creating artificial tree models, Fassnacht et al. (2018) proposed the use of real tree representations in their approach to create synthetic forest remote sensing data. The presented GeForse approach uses real laser scanning point clouds of individual trees to create virtual forest stands and simulate canopy height models. In contrast to simple geometric shaped tree models, these tree point clouds reflect the inter- and intraspecific variability in crown structure and habit of real trees. However, a major limitation of the original GeForse approach is the lack of options to actually simulate ALS acquisitions, as in this approach, the generated forest stands' point clouds are merely a rearrangement of the available tree point clouds.

Methods for simulating laser scanning data differ in terms of complexity and computational effort. Simple approaches to simulate ALS data of forests do not model single returns but only the vertical distribution of returns (e.g. Nelson, 1997; Palace et al., 2015; Spriggs et al., 2015). Single returns have been simulated by modelling the statistical probability of returns from the virtual forest stands (e.g. Knapp et al., 2018; Wang et al., 2013) or by intersecting (x, y)-return coordinates of real laser scanning data with canopy heights of the virtual forest (Frazer et al., 2011). The mentioned approaches are suitable to create simplified synthetic ALS data. However, in order to analyse, for example, the effects of different scanner acquisition settings, not only the laser scanning output but also the

scanning process itself must be modelled. Therefore, more complex simulation approaches are required. In this context, Holmgren et al. (2003) presented a laser scanning simulation approach that allows to vary flight and scanning parameters but does not include the modelling of beam divergence or multiple returns, which are especially important in vegetation contexts (Arumäe & Lang, 2018; Grau et al., 2017). More complex simulations of laser scanning have been accomplished via Monte Carlo ray tracing (e.g. Disney et al., 2010), the Discrete Anisotropic Radiative Transfer (DART) model (Gastellu-Etchegorry et al., 2015, 2016; Yin et al., 2016), and the open-source laser scanning simulation framework HELIOS (S. Bechtold & Höfle, 2016) and its successor HELIOS++ (Winiwarter et al., 2022). HELIOS++ supports virtual laser scanning of 3D scenes with various scanners on various platforms, allows a flexible configuration of acquisition settings, and enables reasonably realistic full-waveform laser scanning simulations (Winiwarter et al., 2022).

Up until now, there are few studies that combine realistic tree models with realistic laser scanning simulations to create large, realistic ALS datasets that represent a variety of stand conditions and also mirror the variability in tree habit on the individual tree level (e.g. Roberts et al., 2020). Many studies are limited to simplified tree models, a simplified laser scanning simulation approach, or homogeneous or small-sized forest stands. This paper proposes a new workflow for generating such large amounts of synthetic data that enables complex laser scanning simulations of realistic virtual forest stands with comparatively little computational effort by combining the GeForse approach of Fassnacht et al. (2018) with HELIOS++. 3D representations of forest stands were created based on forest inventory information and matching tree point clouds. The tree point clouds had been extracted from real laser scanning data acquired by an uncrewed aerial vehicle (UAV) and were fed into the workflow from a database of individual trees. Laser scanning of the virtual stand was simulated using HELIOS++.

With this new approach, we attempt to provide highly realistic ALS datasets of forests for which a complete forest inventory dataset is available. In this study, we present this new workflow, and investigate the generated simulated point clouds with respect to their resemblance to real ALS point clouds. The workflow was validated by simulating laser scanning point clouds of six complex temperate forest plots in southwestern Germany of 1 ha size each, for which forest inventory data on individual tree-level and data of real airborne laser scanning had been collected.

We investigate both the quality of the laser scanning simulation itself and the added value of the detailed tree models in comparison to simple geometric objects by comparing canopy cover, the height distribution of returns, and several more point cloud metrics. In addition, we compare the accuracy of biomass models calibrated with the simulated and the real ALS data to evaluate how the simulated data perform in a potential forest inventory application. Three main research questions are addressed in this study:

1. Can the combination of HELIOS++ and a forest stand point cloud be used to generate realistic synthetic ALS data?
2. To what extent do the simulated laser scanning data differ when simplified tree models are used instead of real tree point clouds?

3. Can simulated laser scanning data be used to train models to accurately predict forest aboveground biomass from real ALS data?

2.2 METHODS

2.2.1 Synthetic data generation

The GeForse approach presented by Fassnacht et al. (2018) combines the output of the individual tree-based forest growth simulator SILVA (Pretzsch et al., 2002) with tree point clouds extracted from laser scanning data to create virtual forest stands. Here, we extended the GeForse approach with a laser scanning simulation component by adding the open-source laser scanning simulation framework HELIOS++ (Winiwarter et al., 2022). Figure 2.1 shows the adapted workflow.

The first component for creating the virtual forest scenes is forest stand information on an individual tree-level, including position, species, and size of every tree in the forest stand. These data can, for example, be obtained from forest inventories conducted in the field. An alternative source for these data are individual tree-based forest growth simulators, such as SILVA (Pretzsch et al., 2002). By taking into account site characteristics, competition for resources, and species-specific growth patterns, forest growth simulators can generate realistic forest stand compositions which can then be used to create virtual forest stands.

The second element of the approach is a database of tree models, i.e. individual tree point clouds which were extracted from real laser scanning data. The database contains both the point clouds and additional information such as the species and the point cloud-derived height and crown diameter of the tree models. For each tree in the forest stand, the database is queried for the best matching tree point cloud and the point cloud is inserted at the corresponding location of the tree. All inserted individual tree point clouds together compose the virtual stand.

In a third step, laser scanning of the virtual stand is simulated with HELIOS++. Since points have no volume or surface and therefore cannot interact with simulated laser beams, the merged point cloud is converted into opaque voxels (Weiser et al., 2021). Then, the laser scanning simulation of the voxelised scene is conducted, according to the defined scanner settings and taking into account occlusion effects within the forest stand. The simulation output is a point cloud of the virtual stand, including information on the number of returns of each beam.

HELIOS++ enables simulating laser scanning point clouds for a variety of acquisition settings. Figure 2.2 shows an example of different simulation outputs of simulations with differing pulse repetition frequencies and different trajectories. The cross sections through the simulated point clouds illustrate how both parameters affect the representation of tree stems in the simulated laser scanning data.

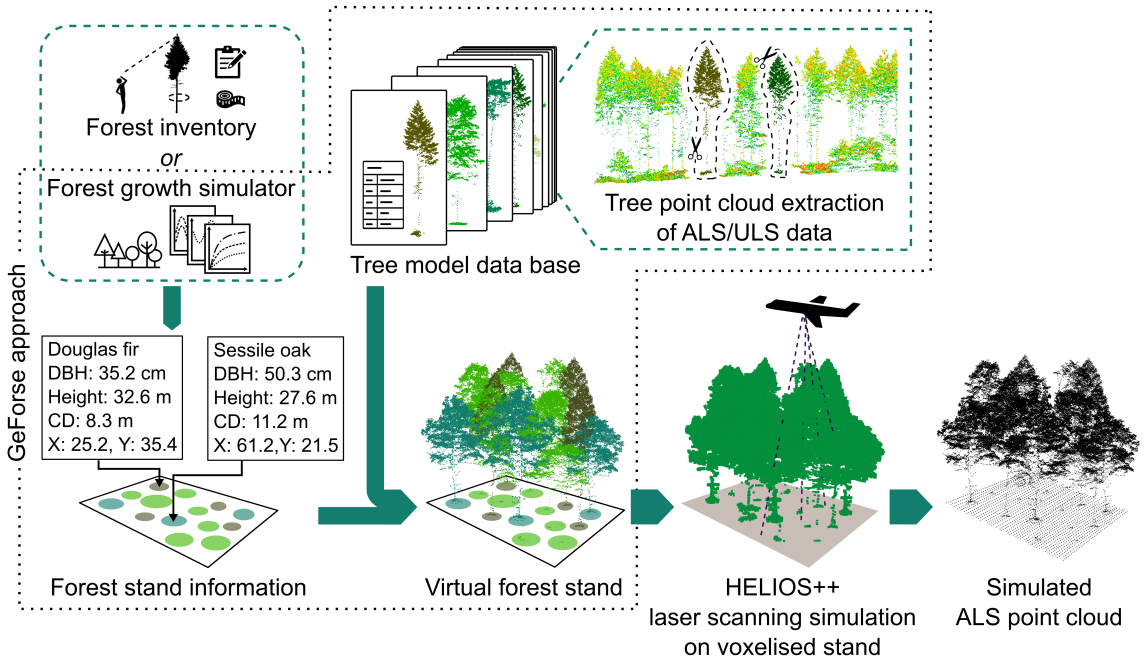

Figure 2.1: Workflow for generating synthetic laser scanning data.

2.2.2 *Case study*

The performance of the presented laser scanning simulation approach was validated by comparing canopy cover, the height distribution, and other point cloud metrics of the simulated and real ALS point clouds of six real forest plots. In order to investigate the added benefit of detailed tree models, virtual stands were created using both real tree models (RTM) extracted from real UAV-borne laser scanning (ULS) point clouds, and simplified tree models (STM) with spheroidal crowns and cylindrical stems. We additionally wanted to evaluate how the synthetic data perform in a potential application by comparing the performance of biomass models trained with simulated and with real ALS data.

2.2.2.1 *Study sites*

We selected six study sites from the dataset published by Weiser et al. (2022b). They are located in a temperate forest region in the federal state of Baden-Württemberg, southwestern Germany. The sites were named according to their location: KA09, KA10, and KA11 are located in the Hardtwald forest in the north of Karlsruhe (49° 02′ 04" N, 8° 25′ 40" E), and BR01, BR03, and BR05 are located in the city forest south of Bretten (49° 00′ 36" N, 8° 41′ 35" E). The IDs correspond to the IDs used in Weiser et al. (2022b). The Karlsruhe sites are characterised by flat terrain. They consist of mixed stands dominated by Scots pine (*Pinus sylvestris* L.), red oak (*Quercus rubra* L.), European hornbeam (*Carpinus betulus* L.), and European beech (*Fagus sylvatica* L.). Black cherry (*Prunus serotina* Ehrh.) and sycamore

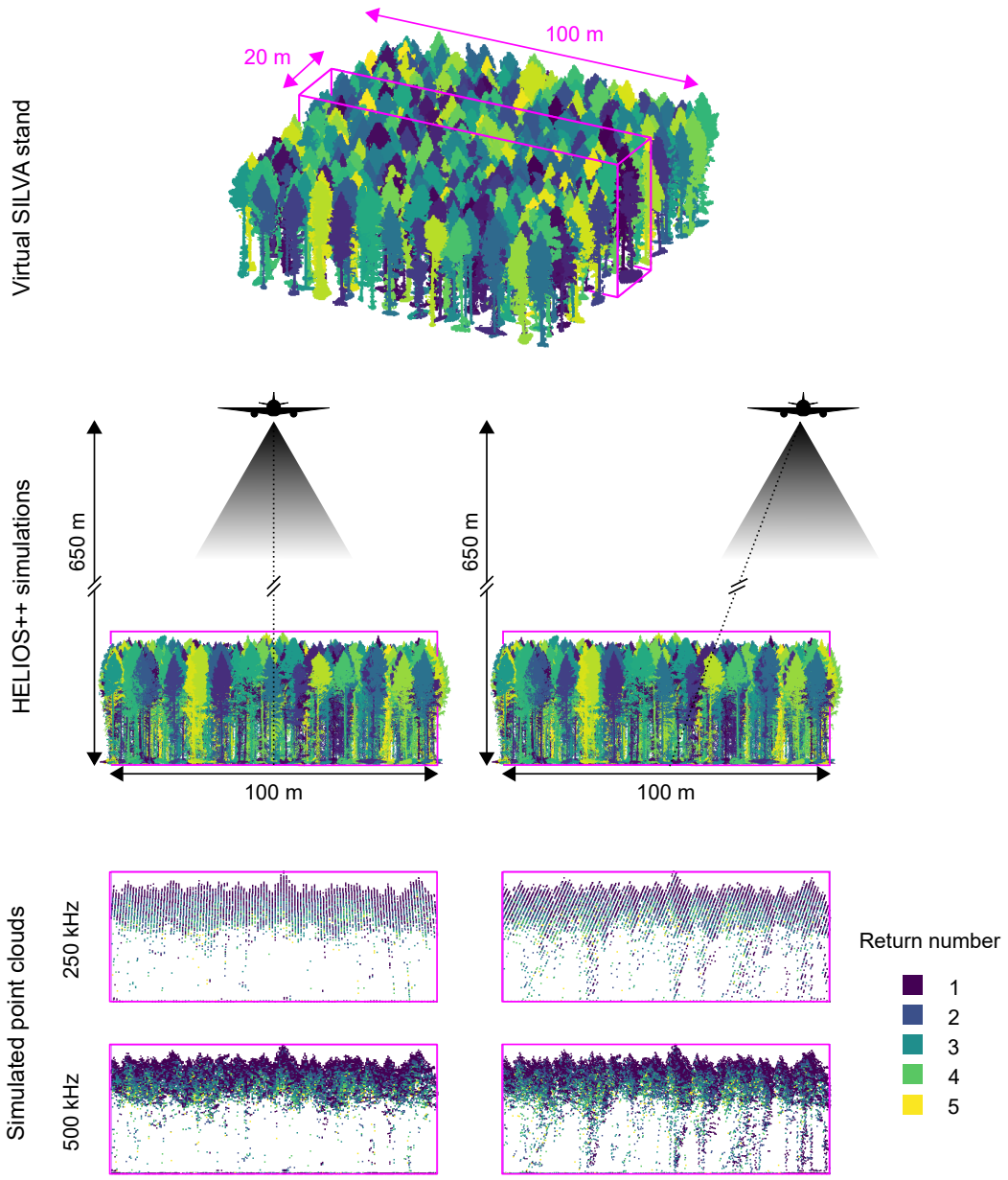

Figure 2.2: Exemplary output of laser scanning simulations for an 80-year-old Douglas fir stand. The stand information was generated with SILVA. The stand was initialised with a random planting pattern of 1 500 trees. Selective thinning with weak intensity was applied in the forest growth simulation. The laser scanning simulations were conducted with two different pulse repetition frequencies: 250 kHz and 500 kHz. The flight pattern was defined as a straight line in north-south direction with an altitude of 650 m above ground. Two different trajectories were applied: one crossing the centre of the forest stand and one located 200 m outside of the forest stand (distance to the centre: 250 m). The simulated point clouds are shown for a section of 100 m $\times$ 20 m. Colours of the simulated points indicate the return number of each point.

maple (*Acer pseudoplatanus* L.) are the most frequent tree species in the understorey layer. The Bretten sites show moderate to steep slopes. Dominant tree species in these stands are European beech, Douglas fir (*Pseudotsuga menziesii* (Mirb.) Franco), sessile oak (*Quercus petraea* (Matt.) Liebl.), Norway spruce (*Picea abies* (L.) H. Karst.), silver fir (*Abies alba* Mill.), and common walnut (*Juglans regia* L.).

All six forest sites are characterised by high within-site heterogeneity in stand structure and species composition. Both mixed and pure, single-layered and multi-layered, even-aged and uneven-aged stands occur. At each site, field data, ULS data, and ALS data were collected.

2.2.2.2 Field data acquisition

Field data were collected from May to October 2019 in six 100 m × 100 m (1 ha) plots. For each tree within these plots with a diameter at breast height ($d_{1.3}$) $\geq$ 5 cm, species, state (dead or alive), location, $d_{1.3}$, and crown diameter (CD) were recorded. The $d_{1.3}$ was measured with a measuring tape at 1.3 m above ground. The CD was measured in two orthogonal directions from the estimated projected crown extent onto the ground. The mean CD was calculated as the mean value of these two measurements. We measured the height (h) of 34% of the trees, mainly trees in the subcanopy layer, where the tree tops were visible from the ground. Height values were additionally extracted from the laser scanning data, if the tree could be identified in the point cloud. Missing height values were estimated by a $d_{1.3}$-based prediction using a Näslund's function, which was found to be suitable for plot-specific height models (Mehtätalo et al., 2015). Models were built based on $d_{1.3}$ and height values of all trees of a species and then used to predict the missing height values for trees of that species. If there were less than three height values for a species, a height model based on all broad-leaved or coniferous trees was used instead of a species-specific model. The plot characteristics are summarised in Table 2.1.

For the measurement of tree locations, plots were divided in 20 m × 20 m subplots. Tree locations were determined relative to one of the subplot corners by measuring the distance and the direction between corners and tree stems. RTK GNSS measurements of some of the plot corners were used to transform relative coordinates to geographic coordinates. For all height and distance measurements, a Haglöf Vertex-IV hypsometer was used. Directions were measured with a magnetic compass.

2.2.2.3 Laser scanning data acquisition and preprocessing

Airborne and UAV-borne laser scanning data were collected in July, August and September 2019 under leaf-on conditions. ALS data were acquired with a RIEGL VQ-780i airborne laser scanner (RIEGL Laser Measurement Systems, 2019). The flight altitude was around 650 m above ground, the flight speed was approximately 51 m/s and the swath overlap was 76%. Laser scanning was performed with a pulse repetition frequency of 1 000 kHz and a scan frequency of 225 lines per second, resulting in a mean point spacing of 28 cm. A

Table 2.1: Stand characteristics of the six 1-ha plots. $d_{1.3}$ is the stem diameter at breast height (1.3 m). σ is the population standard deviation.

Plot	Number of trees	$d_{1.3}$ [cm]	Height [m]	Biomass [t/ha]	Most frequent species
BR01	728	Min: 5.0 Max: 86.7 Mean: 18.6 σ: 14.2	Min: 2.4 Max: 46.4 Mean: 15.6 σ: 8.2	241.4	30% *Fagus sylvatica*, 22% *Picea abies*, 16% *Carpinus betulus*, 16% *Quercus petraea*
BR03	442	Min: 5.0 Max: 99.4 Mean: 29.3 σ: 18.7	Min: 2.3 Max: 42.0 Mean: 22.8 σ: 11.5	299.3	37% *Pseudotsuga menziesii*, 26% *Carpinus betulus*, 10% *Fagus sylvatica*, 6% *Abies alba*
BR05	331	Min: 5.0 Max: 85.7 Mean: 30.5 σ: 18.7	Min: 4.0 Max: 45.8 Mean: 23.5 σ: 10.5	284.8	45% *Fagus sylvatica*, 13% *Pseudotsuga menziesii*, 13% *Picea abies*, 10% *Juglans regia*
KA09	272	Min: 5.0 Max: 79.0 Mean: 28.2 σ: 19.6	Min: 2.6 Max: 43.4 Mean: 18.8 σ: 10.2	170.8	37% *Fagus sylvatica*, 24% *Pinus sylvestris*, 12% *Picea abies*, 6% *Carpinus betulus*
KA10	426	Min: 5.0 Max: 99.4 Mean: 19.9 σ: 17.1	Min: 2.2 Max: 40.6 Mean: 14.2 σ: 8.0	161.9	33% *Prunus serotina*, 25% *Carpinus betulus*, 21% *Quercus rubra*, 7% *Acer pseudoplatanus*
KA11	787	Min: 5.0 Max: 73.5 Mean: 18.6 σ: 10.9	Min: 1.8 Max: 30.7 Mean: 16.3 σ: 6.7	172.0	39% *Fagus sylvatica*, 28% *Quercus rubra*, 15% *Prunus serotina*, 12% *Pinus sylvestris*

digital terrain model (DTM) with a spatial resolution of 0.5 m was calculated from the ALS data for each study site using the TreesVis software system (Weinacker et al., 2004). The DTM was then used to normalise the height of the point cloud data.

ULS data were collected using a RIEGL miniVUX-1UAV sensor mounted on a DJI Matrice 600 Pro Hexacopter. The device was operated with a pulse frequency of 100 kHz. The flights were conducted following a pattern of two overlapping double grids with an offset of 45° in orientation. ALS and ULS acquisition parameters are listed in Table 2.2. A detailed description of the data acquisition and preprocessing can be found in Weiser et al. (2022a).

Table 2.2: Acquisition settings of the real laser scanning campaigns.

	ALS	ULS
Sensor	RIEGL VQ-780i	RIEGL miniVUX-1UAV
Laser beam divergence	0.25 mrad[a]	1.6 mrad × 0.5 mrad[b]
Pulse repetition frequency	1 000 kHz	100 kHz
Scan frequency	225 lines per second	33.4/100 lines per second
Scan angle off nadir	±30°	±90°
Altitude above ground	641–728 m	60–80 m
Flight speed	100 kn (≈ 51 m/s)	3.0–5.0 m/s
Flight line distance	175 m	25–30 m
Mean pulse density	70–78 pulses/m^2	797–1554 pulses/m^2
Flight pattern	Parallel flight strips	Two overlapping double grids

[a] Measured at the $1/e^2$ points.
[b] Measured at 50% peak intensity (≈ 2.72 mrad × 0.85 mrad measured at the $1/e^2$ points).

2.2.2.4 Creation of real tree model forest stands

A 3D model of each plot was created using the field data and matching tree models, i.e. point clouds of real trees. The real tree models (RTM) were derived from the ULS data by manually extracting point clouds belonging to individual trees in the 3D point cloud and mesh processing software CloudCompare (CloudCompare, 2019). Tree metrics including height and crown diameter were estimated from the tree point clouds. Tree species of the tree models were determined based on the field data and additional field surveys. Each tree model was subjectively assigned a quality score from high (q1) to low (q6) by the extracting person, indicating the probability of segmentation and extraction errors. A dataset containing all tree point clouds has been published by Weiser et al. (2022b). A detailed description of the tree models can be found in Weiser et al. (2022a). For the creation of virtual forest stands, only tree models of high to medium quality (q1-q3) were used. In total, 1 245 tree models of 19 different tree species with these quality tags were available.

The habit of dead trees can differ significantly from that of living trees of the same species. As there were only few tree models of dead trees available, dead trees were omitted when creating the synthetic stands. For each living tree within the six study plots, a tree model of the same species, and similar height and crown diameter was selected by filtering all tree models according to the following procedure:

1. Exclude all tree models extracted from the same plot.
2. Include only tree models with matching tree species. If there is no tree model of that species available, include only tree models of the matching species group (broad-leaved or coniferous).
3. Calculate the Euclidean distance d in terms of tree height and crown diameter for all remaining tree models with respect to the query tree (Equation 2.1) and randomly

select a tree model with $d < 4\,\text{m}$. If there is no tree model with $d < 4\,\text{m}$, select the tree model with the minimum d to the query tree.

$$d = \sqrt{(H_{tree\,model} - H_{query\,tree})^2 + (CD_{tree\,model} - CD_{query\,tree})^2}\ [m] \tag{2.1}$$

The selected tree point cloud is scaled along the X-, Y-, and Z-axes so that the CD and H value of the point cloud fit the crown diameter and height of the queried tree. The tree model is randomly rotated around the Z-axis and then placed at the location of the query tree in the forest stand. This way, a 3D geometry model of each plot is created as a composite of multiple tree point clouds.

For 86% of the trees in the six 1-ha plots, tree point clouds of the same species were selected. Tree point clouds were partly missing for tree species which were only extracted from the laser scanning data of one of the plots, i.e. black cherry, common walnut and silver fir, and for some rarely occurring species. For these trees, corresponding tree point clouds from other species were inserted following the process described above.

2.2.2.5 Creation of synthetic simplified tree model forest stands

In order to investigate how the complexity of the tree models affects the simulated ALS point clouds, the laser scanning simulations (see next section) were also applied to additional 3D representations of the six plots that were created using point clouds of simplified tree models (STM) composed of spheroidal crowns and cylindrical stems. This shape is similar to that used by Knapp et al. (2018), who created tree models with cylindrical crowns and cylindrical stems. The STMs have a simplified shape but a realistic structure in the sense that the laser beams are not completely reflected by the outer canopy but may penetrate the crown through gaps, whereas the tree stem is impenetrable to the laser beams. For each tree in a plot, a matching STM was generated. The horizontal semi-axes of the spheroidal crown were defined with a length matching the required crown radius, the length of the vertical semi-axis was fixed as half of the crown height, which was calculated as the difference of tree height and the height of the lowest green branch as measured in the field. The stem cylinder was defined with a diameter matching the $d_{1.3}$ and a height equal to the crown base height. As the simulation is based on voxelised point clouds, we needed to create a point cloud representing the STM shape. To achieve this, points were randomly sampled from a uniform 3D distribution spanning the bounding box of a tree defined by tree height and crown radius. The planar point density was set to $1\,500\,\text{pts}/\text{m}^2$, which is approximately the mean point density of the ULS point clouds of the six plots when ground hits were excluded, as ground is not included in the tree models. The spheroid tree crown was then derived by removing all points which were not within the region bounded by the spheroidal surface. For the tree stem, additional points were created on the cylindrical surface of the stem in a way that the maximum distance of neighbouring points in X-, Y-, and Z-directions was limited to 2 cm, ensuring that a voxelisation with a voxel size of 3 cm would convert the stem point clouds to opaque objects, impenetrable for the laser beams.

2.2.2.6 *HELIOS++ laser scanning simulations*

Airborne laser scanning data of the synthetic forest stands were simulated with the open-source Heidelberg LiDAR operations simulator HELIOS++ (Winiwarter et al., 2022). The HELIOS++ simulations require definitions of the scene to be scanned, the laser scanner, and the platform on which the scanner is mounted, e.g. an airplane. In addition, the position and movement of the platform and the scanner settings must be defined (Winiwarter et al., 2022). The scene is specified as a spatial arrangement of different 3D objects which can be polygon meshes, terrain models, or voxel models. The scene can also include objects in the form of point clouds, which are converted to impenetrable voxels on-the-fly. For each voxel containing at least one point, an opaque cube is created, providing 3D surfaces to be virtually scanned. HELIOS++ simulates the movement of the scanner along the defined trajectories according to the specified flight speed and starting position of the platform. The scan pattern is determined by the laser beam deflector type, the scan angle, the scan speed in lines per second, the pulse repetition frequency, and the aircraft speed. Laser beam divergence is simulated by subrays that are regularly distributed around a central ray. The base intensity of each subray is calculated according to a 2D Gaussian power distribution. A return is generated when a subray hits an object. Additional parameters allow to define the number of subrays, the pulse shape in time, and the detection of local maxima in the waveform, which are used to fit Gaussians. A comprehensive description of the functioning and possible settings of HELIOS++ can be found in Winiwarter et al. (2022) and in the HELIOS++ documentation on GitHub (https://github.com/3dgeo-heidelberg/helios/wiki).

For the presented approach for synthetic data generation, it is of particular importance whether HELIOS++ can reproduce real laser scanning data, therefore the simulations were designed to replicate the acquisition parameters of the real ALS campaigns, including flight speed, trajectories, and mean flight height above terrain (Table 2.3).

The optimal values for the parameters defining the size of the temporal window for local maxima detection and the voxel size were determined iteratively using a subset of the data. Laser scanning was simulated for a subset of the height-normalised real ALS point cloud covering an area of 22 m $\times$ 24 m in site BR01. The simulations were performed with temporal window sizes of 1.0, 1.1, 1.2, 1.3, 1.4, and 1.5 ns for the maximum detection and voxel sizes of 0.02, 0.03, 0.04, 0.05 m. The simulated point clouds were compared to the real point cloud based on the absolute and relative height distribution of first and all returns (see next section). The analysis revealed that no parameter combination results in a good match for all of the four cases: "all returns", "first returns", "absolute" and "relative height distribution" (see Appendix, Chapter 2.7). Based on the good results for the relative height distribution, a window size of 1.0 ns and a voxel resolution of 0.03 m were selected. For parameters not listed in Table 2.3, the default settings of HELIOS++ version 1.1.0 were used.

The laser scanning simulation was then performed for each forest stand using the identified settings and the 3D forest scenes. The 3D scenes were composed of the point clouds

representing the inserted tree models and a ground layer in form of a horizontal plane matching the horizontal plot extents.

To avoid problems with hardware memory limitations, the point clouds were split into multiple parts. The memory footprint of HELIOS++ simulations depends, among other factors, on the number of primitives in the simulation. In the case of a voxel-based scene derived from a point cloud, this corresponds to the volume of space occupied by the input point cloud, i.e. it does not directly depend on the number of points. Splitting the point clouds into multiple scene parts can reduce the size of the bounding boxes around the scene parts and thus the number of primitives. In our study, the maximum bounding box volume of the scene parts was 519 204 m^3. It took 58 minutes to simulate laser scanning of this scene using HELIOS++ version 1.1.0 on an Intel Xeon CPU E5-2630 v3 @ 2.40 GHz with 256 GB of RAM, resulting in 1 677 666 simulated points.

Table 2.3: Acquisition settings of the laser scanning simulations.

Sensor	RIEGL VQ-780i
Altitude	641–728 m
Flight speed	51 m/s
Pulse repetition frequency	1 000 kHz
Scan frequency	225 Hz
Laser beam deflector	rotating mirror
Scan Angle	±45°
Effective scan angle	±30°
Accuracy	0.02 m
Beam divergence	0.25 mrad
Pulse length	4 ns
Maximum number of returns per pulse	5
Voxel size	0.03 m
Window size for echo detection	1.0 ns

2.2.3 Validation of simulated laser scanning data

The objective of our approach is not to exactly reproduce the existing real laser scanning point cloud of a forest stand, but to generate a point cloud that could potentially be derived from laser scanning of an imaginary forest stand with a similar composition. To judge whether this objective was met, the similarity between simulated and real data was evaluated using an area-based approach, by quantitatively comparing canopy cover and the return height distribution of the 1-ha plots, and point cloud metrics calculated for 20 m × 20 m subplots. In addition, it was tested how biomass models trained with synthetic data perform in comparison to models trained with real data.

A visual comparison between simulated and real point clouds can give a first impression of how well the structural characteristics of a forest stand are represented in the synthetic data. We extracted vertical cross sections, which are particularly suitable for this, since they reveal tree locations and the structure of individual trees.

The canopy cover of each plot was derived from canopy height models with a spatial resolution of $0.2\,\text{m} \times 0.2\,\text{m}$. It was calculated as the proportion of pixels with a value $> 2\,\text{m}$ to the total number of pixels (Ma et al., 2017).

The difference in return height distribution provides information on how well the overall height structure of the stands is reproduced. Point clouds consisting of all returns and point clouds filtered for only first returns were analysed separately. In general, reflections at the top of canopies produce first returns, while reflections within the crowns or reflections from the understorey and ground vegetation are often intermediate and last returns (Fassnacht et al., 2014). Points with a normalised height $\leq 2\,\text{m}$ were not included in the calculations to eliminate potential effects of missing ground vegetation in the synthetic stands.

Vertical profiles of the point distribution were derived by plotting the absolute and relative numbers of returns per 1 m height bin against the height. Following Spriggs et al. (2015) and Knapp et al. (2018), the root mean squared error (RMSE) between simulated and observed relative proportions of returns per height bin, and the relative overlap of the height profiles, i.e. the quotient of intersection area and union area (IoU), were calculated.

For training and evaluation of the biomass models, the six 1-ha plots were split into $20\,\text{m} \times 20\,\text{m}$ subplots. Point cloud metrics were calculated from the returns within a subplot, both from all returns and from first returns only. Following van Lier et al. (2021), we selected metrics describing the canopy height, the vertical structure, the density of the point clouds, and the canopy cover. As canopy height metrics, the maximum height of the returns (Hmax), the mean height (Hmean), the 10th–90th (in steps of 10) height percentiles (H10–H90), and 98th height percentile (H98) were computed. The vertical structure metrics included the standard deviation (SD) and the coefficient of variation (CV) of height, the skewness and kurtosis of the return height profile of each subplot, and the vertical complexity index (VCI) developed by van Ewijk et al. (2011), which was calculated as implemented in the R package "lidR" (Roussel & Auty, 2021; Roussel et al., 2020). Point cloud density metrics were calculated as the proportion of returns above the 10th–90th height percentiles to the total number of returns (D1–D9) (Næsset, 2002). As a measure of canopy cover, the ratio of the number of returns above 2 m height to the total number of returns (Pgr2) was used (Smith et al., 2009). In addition, the total number of returns (NP) and the ratio of returns above the mean height to the total number of returns (Pgrmean) were calculated. Similar to the return height profiles, all metrics except Pgr2 were calculated from returns above 2 m height.

Biomass reference data of the subplots were the summed aboveground biomass of all trees within a subplot divided by the area of the subplot. The biomass of a tree was calculated based on DBH and tree height according to the species-specific equations developed for the German national forest inventory, available in the R package "rBDAT" (Vonderach et al., 2021).

The random forests algorithm (Breiman, 2001) was used to predict biomass from the point cloud metrics. The models were trained with 500 trees to grow (ntree) and the optimal number of variables considered at each split (mtry) was identified for each model using the function "tuneRF" of the R package "randomForest" (Liaw & Wiener, 2002). Each model was built from point cloud metrics and corresponding biomass data of 135 randomly selected subplots and it was tested on the remaining 15 subplots. The models were trained with a) the point cloud metrics derived from the original data, b) the point cloud metrics derived from the simulated RTM data, and c) the point cloud metrics derived from the simulated STM data. For testing, only the original data were used. The same split between training and testing was used throughout the experiment. The process of model building and testing was repeated 500 times for each training data type.

The performance of the biomass models was assessed by calculating the RMSE and the mean error (ME) between predicted and reference biomass values for each test sample. Data processing, analysis and visualisation were carried out in R 4.0.4 (R Core Team, 2021) within the RStudio interface (RStudio Team, 2016) using the packages "data.table" (Dowle & Srinivasan, 2021) and "ggplot2" (Wickham, 2016).

2.3 RESULTS

2.3.1 Cross sections

Cross sections through the simulated laser scanning point clouds and the original ALS point clouds illustrate that the habit of individual trees and the structural variety was better represented in the RTM point clouds than in the STM point clouds. The highlighted parts in Figure 2.3 reveal mismatches between cross sections of RTM and original ALS point clouds that indicate some challenges in the creation of the virtual stands. These mismatches include differences in the crown structure (Figure 2.3 d), differences in the canopy closure (Figure 2.3 e), and differences in the subcanopy layers (Figures 2.3 f, g, h).

2.3.2 Canopy cover

The canopy cover was 80% for plot KA09 and between 94% and 97% for the other plots. The computed canopy cover of the simulated RTM data was significantly lower. It was 66% for KA09 and ranged from 76% to 89% for the other plots. The canopy cover of the simulated STM data was in between the values derived from the STM data and the values derived from the real ALS data (Table 2.4).

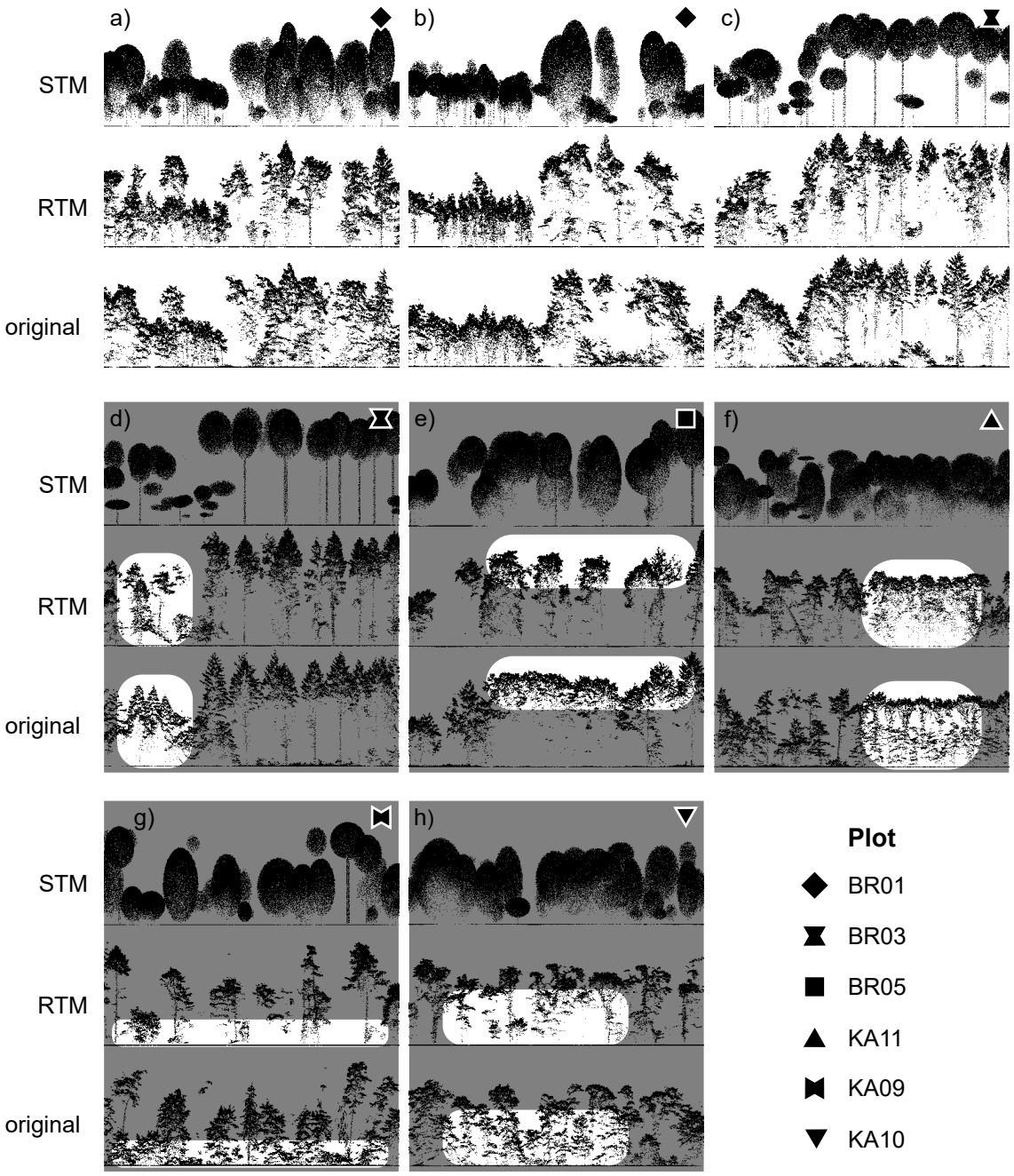

Figure 2.3: Cross sections through the simulated STM (top), the simulated RTM (centre), and the original (bottom) point clouds (thickness = 4 m). The cross sections a-c show a good match between original and RTM point clouds. The highlighted parts in cross sections d-h reveal some problems in the composition of the virtual stands compared to the real forest data: false tree model species (d), artificial canopy gaps (e), different tree habit of same species (f), missing undergrowth (g), and differences in crown base height (h).

Table 2.4: Canopy cover of the 1-ha plots calculated from canopy height models of the original ALS data and the simulated RTM and STM data.

	original	RTM	STM
BR01	97%	83%	93%
BR03	97%	86%	95%
BR05	94%	84%	92%
KA09	80%	66%	78%
KA10	95%	76%	91%
KA11	96%	89%	95%

2.3.3 Profiles of the return height distribution

The similarity of the return height distribution between simulated and real laser scanning data differed for the six plots and depended on whether all or only first returns were considered. In most cases, the laser scanning simulations of the RTM stands generated fewer returns than the real airborne laser scanning (Table 2.5). The simulated point clouds of the STM stands had quite similar numbers of first returns as in the real data, but much higher numbers of all returns.

A comparison of the relative frequencies of returns per 1 m height bins between simulated and original data revealed that it depends on the study site, whether the RTM or the STM data better represented the vertical return distribution of all returns. The relative overlap of the return height profiles using all returns of the simulated and the original data was higher for the RTM data than for the STM data for plots BR01, BR03, BR05 and KA11, and lower for plots KA09 and KA10. When investigating first returns only, the relative overlap of the RTM height profiles with the original data was higher than with the STM data in most cases, the only exception was the height profile of plot KA11 with a relative overlap of 71% for the RTM point cloud and a relative overlap of 80% for the STM point cloud (Table 2.5).

Figure 2.4 shows the return height profiles of all plots. For most of the plots, the shape of the profiles of the simulated and the original data matched well, even if some of them were slightly shifted. There were large divergences between the return height profiles of all returns for plots KA09, KA10, and KA11. These divergences were less pronounced for the first return height profiles of these plots. In most cases, the height distributions of the RTM and the STM point clouds were quite similar, but there were also significant differences, e.g. for all returns of plots KA09 and KA10, as well as for the first returns of BR03. The mean relative overlap of the relative height distribution profiles was 79% for RTM stands (for both all returns and first returns) and 79% and 75% for STM stands (all returns and first returns, respectively). When aggregating the return distribution over all plots, the relative overlap increased to 87% and 90% (all returns and first returns, respectively) for the RTM

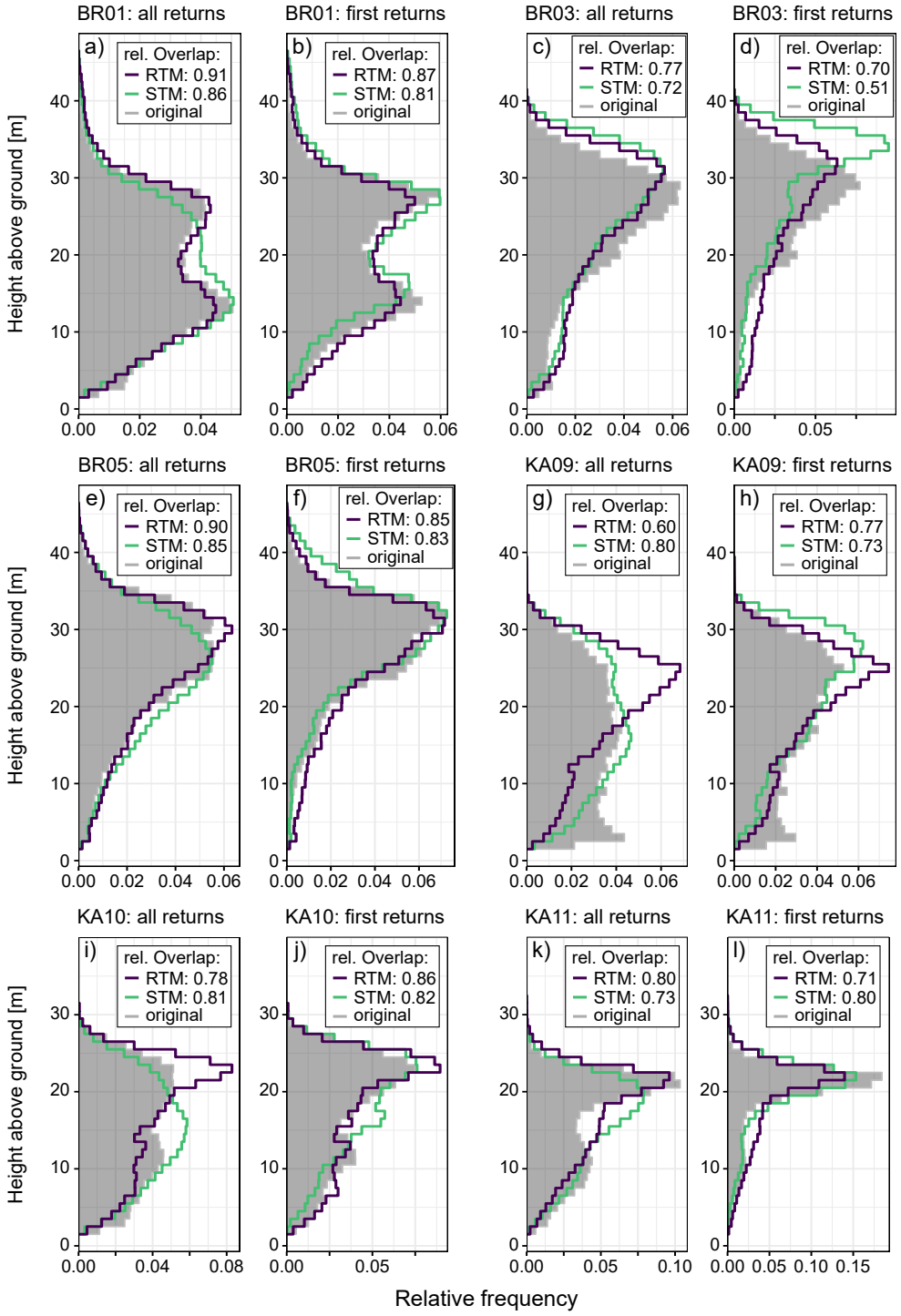

Figure 2.4: Relative frequencies of returns per 1 m height bins for all returns and first returns of original and simulated point clouds for all plots. The height distribution of the original point clouds is depicted in grey. The relative overlap between simulated and original height profiles is given in the text boxes.

Table 2.5: Comparison of the total number of returns and the relative height distribution of returns (> 2 m) of original and simulated data. *mean* is the mean value calculated from the plot specific results. *all* is the value for the returns summed across all plots. Percentages are given relative to the number of returns in the original dataset.

Returns	Plot	Total number of returns ($\times 10^6$)			Overlap		RMSE	
		Original	RTM	STM	RTM	STM	RTM	STM
all	BR01	1.52	1.19 (78%)	2.28 (189%)	0.91	0.86	0.0024	0.0048
	BR03	1.39	1.36 (98%)	2.54 (182%)	0.77	0.72	0.0078	0.0104
	BR05	1.32	1.37 (104%)	2.84 (215%)	0.90	0.85	0.0035	0.0054
	KA09	1.19	0.86 (72%)	2.23 (188%)	0.60	0.80	0.0176	0.0085
	KA10	1.37	1.02 (74%)	3.02 (221%)	0.78	0.81	0.0118	0.0087
	KA11	1.54	1.36 (88%)	3.26 (211%)	0.80	0.73	0.0087	0.0152
	mean	1.39	1.19	2.79	0.79	0.79	0.0086	0.0088
	all	8.33	7.16 (86%)	16.76 (201%)	0.87	0.85	0.0041	0.0048
first	BR01	0.78	0.61 (78%)	0.68 (88%)	0.87	0.81	0.0041	0.0061
	BR03	0.75	0.64 (86%)	0.69 (93%)	0.70	0.51	0.0105	0.0248
	BR05	0.71	0.63 (89%)	0.68 (96%)	0.85	0.83	0.0047	0.0057
	KA09	0.58	0.48 (83%)	0.58 (100%)	0.77	0.73	0.0096	0.0115
	KA10	0.72	0.58 (81%)	0.71 (99%)	0.86	0.82	0.0071	0.0088
	KA11	0.74	0.70 (95%)	0.75 (101%)	0.71	0.80	0.0181	0.0112
	mean	0.71	0.61	0.68	0.79	0.75	0.0090	0.0114
	all	4.27	3.64 (85%)	4.09 (96%)	0.90	0.83	0.0031	0.0051

stands and to 85% and 83% for the STM stands. The comparison of the RMSE values led to the same results as the comparison of the relative overlap.

2.3.4 Point cloud metrics

Point cloud metrics were calculated from all returns and from first returns of the original and simulated data for 20 m × 20 m subplots. Figure 2.5 shows the deviation of metrics calculated from the simulated point clouds compared to those calculated from the original point clouds of the same subplots. For most metrics calculated from all returns, the difference from the original data was similar for both simulated data types. The median absolute difference of the height percentiles of original and simulated data was not higher than 1.13 m. In general, the median deviation from the STM data to the original data was smaller than the median deviation from the RTM data to the original data. For the height percentiles and the density metrics, the median deviation of the RTM data was always negative, corresponding to higher values derived from the RTM data than derived from the original data. The total number of returns (NP) and the point cloud-derived measure of canopy cover (Pgr2) were the only two metrics for which large differences between RTM and STM data could be observed. NP was always higher for the STM data than for the

original data. In contrast, in the RTM point clouds, there were – in general – fewer returns than in the original point clouds. Pgr2 calculated from the original data was lower than when calculated from the STM data and higher than when calculated from the RTM data.

The comparison of metrics calculated from first returns showed a different pattern than when all returns were considered. The lower height percentiles (H10–H50) and the intermediate density metrics (D3–D7) were smaller when calculated from the RTM data than from the original data. In contrast, most metrics of the STM data showed slightly higher values than the same metrics of the original data, leading to negative differences. The median difference in the number of first returns was close to zero for the STM data, whereas there were fewer first returns in the RTM data than in the original data.

2.3.5 Biomass models

The six study sites showed high heterogeneity in terms of tree species composition, number of trees, and age structure, and consequently also in biomass values (see Table 2.1). The total aboveground biomass of all trees was lowest for plot KA10 (161.9 t/ha) and highest for plot BR03 (299.3 t/ha). The mean biomass values per individual tree ranged from 218.6 kg to 860.5 kg (plot KA11 and plot BR05, respectively). The aggregated biomass of the 20 m $\times$ 20 m subplots ranged from 37.6 t/ha (in plot BR05) to 483.9 t/ha (also in plot BR05), the mean biomass was 217.1 t/ha and the standard deviation was 101.3 t/ha. Biomass models were trained with metrics derived from the original point clouds as well as with metrics derived from the simulated RTM and STM point clouds. The models were tested by predicting biomass from point cloud metrics derived from the original point clouds. Figure 2.6 shows the reference biomass plotted against the mean predicted biomass of each subplot for each type of training data. All models overestimated the biomass of subplots with very low biomass and underestimated the biomass of subplots with high biomass. The RMSE and the mean error between reference and predicted biomass for each test dataset are illustrated in Figure 2.7. They reveal that the type of training data had only a small effect on the model accuracy. The median RMSE of the predictions from models trained with all return metrics of the original data was 67.1 t/ha. It was slightly higher for the RTM models at 72.4 t/ha and highest for the STM models at 76.2 t/ha. The differences in model accuracy were even smaller if only first returns were used in the metrics calculation: the median RMSE of the models trained with original data was 66.5 t/ha, it was 70.6 t/ha for the RTM models, and 71.7 t/ha for the STM models. On average, the prediction accuracy of models trained with original data was the highest, while it was lowest for STM models. However, the lowest RMSE of all predictions based on all return metrics was achieved by a model trained with STM data. As an indicator of bias, the mean error (ME) of reference and predicted biomass was calculated. The median ME of the models trained with the original data was close to zero (i.e. no evident bias). In contrast, the RTM models tended to overestimate biomass (median ME -9.1 t/ha for all returns, -18.9 t/ha for first returns),

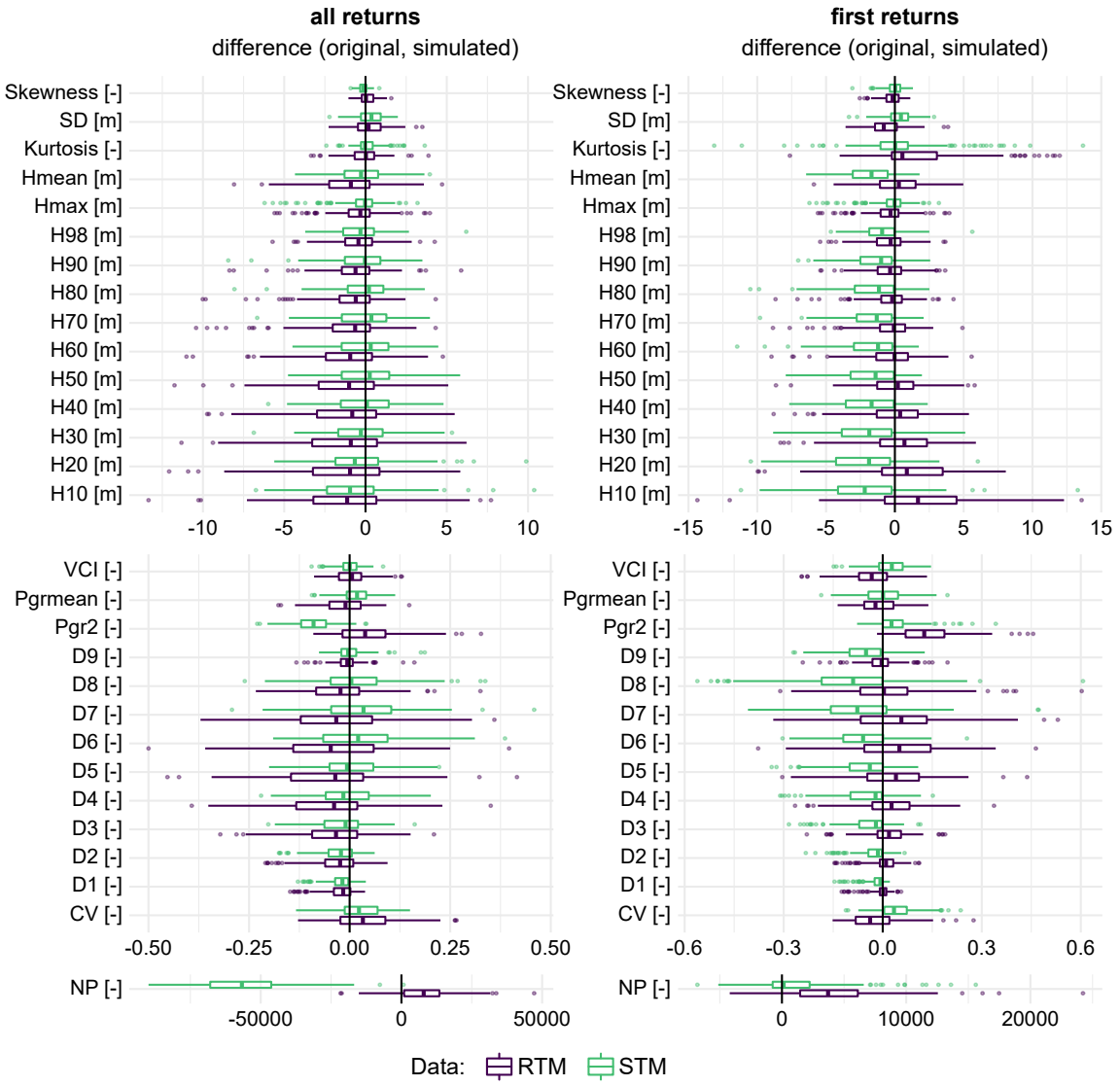

Figure 2.5: Difference of metrics calculated from the original point clouds and the same metrics of the same subplots calculated from the simulated point clouds for all returns (left) and first returns (right). For illustrative purposes, the metrics were grouped by value range.

while the STM models underestimated the biomass (median ME 29.5 t/ha for all returns, 17.2 t/ha for first returns).

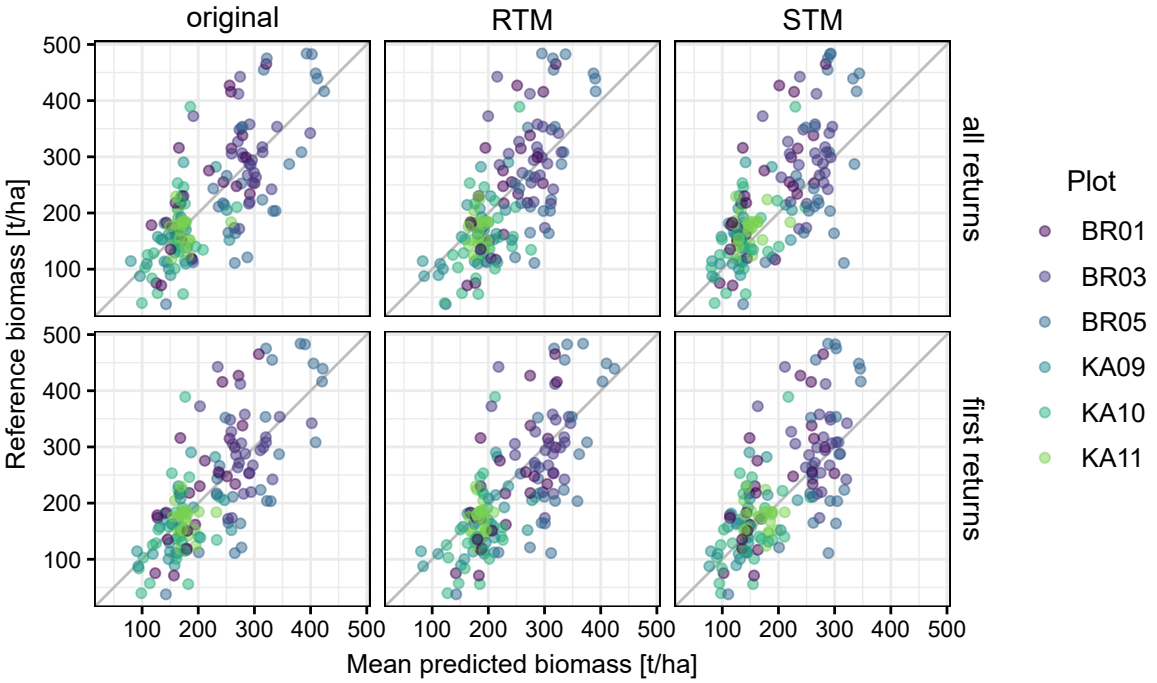

Figure 2.6: Reference plotted against mean predicted biomass of 20 m × 20 m subplots for models trained with point cloud metrics from the original data (left), the simulated RTM data (centre) and the simulated STM data (right), including all returns (top) and only first returns (bottom). The solid lines indicate the one-to-one line.

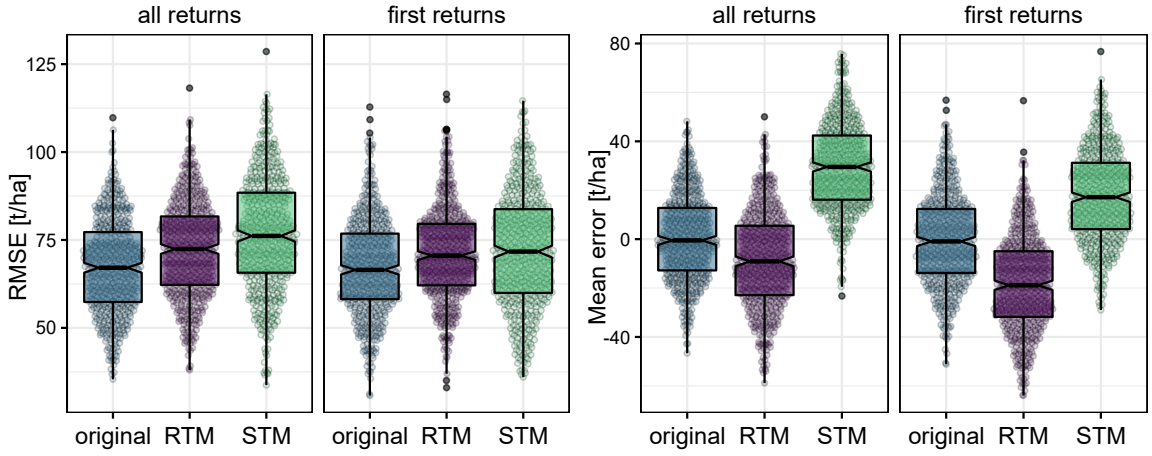

Figure 2.7: Root mean squared error (left) and mean error (right) of reference and predicted biomass of the randomly sampled test data for each training data type.

2.4 DISCUSSION

In this study, we present a new approach for generating synthetic laser scanning datasets that combines forest inventory data, detailed individual tree models and the laser scanning simulation framework HELIOS++. The approach was evaluated with regard to its ability to generate realistic laser scanning point clouds and its usability for calibrating biomass models. Furthermore, the added value of the detailed tree models compared to simple geometric objects was investigated.

2.4.1 Laser scanning simulations

The application of HELIOS++ allowed us to simulate exactly the same flight trajectories and ALS acquisition settings (scanning system, scan angle, scan frequency, etc.) used to acquire the real ALS dataset to which our simulations were compared. At the same time, HELIOS++ can be used to simulate alternative acquisition settings and hence allows to specifically assess the influence of acquisition settings on point cloud characteristics and subsequent modelling exercises using sensitivity analyses. Furthermore, the HELIOS++ approach provides a compromise between computational demand and an accurate simulation of the geometric and physical scanning process. For example, laser beams are represented by multiple subrays in HELIOS++ to simulate the illuminated area of the conical real laser beams (Winiwarter et al., 2022). In this way, the simulation of a single beam can result in multiple returns, comparable to a full-waveform laser scanning system. This is particularly relevant in forest canopies which form semi-permeable volumes where multiple returns are very common (Fassnacht et al., 2014). In theory, our workflow could also be coupled with other approaches for virtual laser scanning, e.g. the photon tracing approach implemented in the Discrete Anisotropic Radiative Transfer (DART) model (Gastellu-Etchegorry et al., 2015).

Apart from laser scanner type, flight trajectories and other acquisition settings that were adopted from the real ALS acquisitions, HELIOS++ requires the definition of the voxel size for modelling the trees and the window size in the local maxima detection. Both parameters particularly influence the number of returns and the height distribution of returns: the smaller the voxel size and the window size, the more likely are multiple returns. A comparison of real ALS data and simulated data with different combinations of voxel size values and window size values revealed that the optimal values differ depending on what is analysed: the number of returns or the height distribution of returns, all returns or first returns only. The optimal parameters of the voxel size and the window size further depend on the point density of the virtual scene, the pulse density of the laser scanning simulation, and the footprint size of the simulated laser beam.

Airborne laser scanning data often contain not only information on the location of the returns but also on the intensity of the backscattered light (Wehr & Lohr, 1999). The return intensity can contribute valuable information for classifying tree species (Ørka et al., 2009), discriminating vegetation strata (Morsdorf et al., 2010), or estimating aboveground biomass

of forests (García et al., 2010). The simulation of intensities requires information on the reflectivity of the concerned surfaces and a much more complex ray-scene interaction model than what is required for simulating the location of returns. As reliable information on reflectivity was not available for our tree model dataset, we opted to exclude any intensity-based aspects from our analysis, instead focusing on geometric features of the point clouds. Simulating realistic return intensities of forest scenes should be part of future research.

2.4.2 Synthetic forest stands

Synthetic forest stands should ideally reflect the structural characteristics of real forest stands, which are determined not only by the average traits of the forest stands, but also by the variety of individual habits and crown structure of each tree. Therefore, we used individual tree point clouds extracted from real ULS data as tree models in our approach. In this way, the 3D representations of the forest stand scenes adequately represent the inter- and intraspecific variability in tree growth forms.

The visual comparison of cross sections through simulated and real point clouds showed a high similarity between both data types in some cases. However, there are also clear differences visible, which indicate problems in the composition of virtual stands (see highlighted parts in Figure 2.3). The silver fir stand in plot BR03 (Figure 2.3 d) is not well represented in the virtual RTM stand, because there were no silver fir tree models available. The selected Scots pine tree models have a different habit than the original silver fir trees, which is clearly visible in the cross section. In the highlighted parts in Figure 2.3 f and Figure 2.3 h, tree models of the same species as in the original data were selected. Still, the habit of the individual trees in the virtual and the real stands differ significantly. The tree growth form does not only depend on the species, but also on stand structure, the social class of a tree (e.g. dominant or suppressed tree (Assman, 1961)) and other growth conditions. In this study, tree models were selected based on species, tree height and crown diameter. If crown diameter measurements are not available for a forest stand, the synthetic stand can either be created based solely on tree position, tree species, and height, or crown diameter values can be estimated from stem diameters (W. A. Bechtold, 2003). Additional filtering criteria such as $d_{1.3}$, crown base height and social class could improve the tree model selection, if these data are available for both tree models and the forest stand or if they could be deduced from available stand data. Further research might explore the matching between tree models and real trees depending on the applied filtering criteria and available data.

Two main characteristics of natural forests are currently not taken into account in the creation of virtual stands: canopy closure and undergrowth. In natural forests in central Europe, trees usually grow in a way that enables them to maximise the amount of absorbed incoming solar radiation (Getzin & Wiegand, 2007), which normally leads to a closed canopy. By randomly rotating the tree models at their designated position, without taking into account the growth form of neighbouring trees, artificial gaps are introduced in the canopy of the virtual stands (see Figure 2.3 e). This issue is also reflected in the lower

canopy cover of the synthetic stands compared to the canopy cover values derived from the real data. It could be solved by taking into account the crown shape of neighbouring trees when placing a new tree model. In addition to canopy closure, undergrowth elements are not yet implemented in the virtual stands (Figure 2.3 g). The incorporation of undergrowth requires both information on where undergrowth occurs and 3D models or point clouds representing this vegetation. The latter can be easily extracted from laser scanning point clouds of forest stands, but as undergrowth is typically not recorded in forest inventories and it is not included in the synthetic stands generated by forest growth simulators, there is no information on where undergrowth elements should be placed. Since undergrowth characteristics, such as height, cover and biomass, have been shown to be correlated to overstory light transmission (Bartemucci et al., 2006; Vales, 1985; Zavitkovski, 1976), possible undergrowth occurrences could be predicted from the point cloud of the synthetic stand. However, this would require more empirical studies to derive models or rules enabling a sound allocation of undergrowth elements in the forest scene. Here, we addressed the problem of lack of undergrowth in the synthetic data in a pragmatic way. To make the simulated point clouds of the synthetic stands (without undergrowth) comparable to the original ALS point clouds of the real forest stands (with undergrowth), returns from the undergrowth in the original data were excluded by dropping points below 2 m normalised height. Those were assumed to be representing ground vegetation (McRoberts et al., 2013; Næsset, 2002). However, it is clear that in most forest stands, especially those with high structural diversity, a fixed height threshold will not be sufficient to properly distinguish between undergrowth returns and returns from trees.

Apart from tree models and undergrowth, topography is a relevant element of forest scenes which can influence point cloud characteristics. In this study, we did not examine how topography affects the simulation process. However, in theory, height information in the form of digital terrain models can be included in the synthetic stands and hence the laser scanning simulation (Winiwarter et al., 2022). The tree models would then be inserted on a 3D surface representing the topography instead of on a flat surface.

The neighbourhood of a forest stand might also affect the laser scanning point cloud of that stand. Shadowing effects of neighbouring forest stands will reduce the number of incoming laser beams and decrease the number of returns in the subcanopy layers. For simplicity, virtual forest stands in this study were simulated as isolated, without neighbouring stands. Dummy scene parts composed of transmissive voxels could be inserted around the virtual stands to emulate the effects of neighbouring forest stands.

2.4.3 Tree models

In addition to the real tree models extracted from ULS data, simplified tree models were used to create the virtual scenes in order to investigate whether the level of detail and the accompanying added computational complexity provided by the real tree models is required to generate realistic synthetic laser scanning data. As expected, the added benefit of the individual tree point clouds as detailed tree models compared to the simplified spheroid tree models became clear in the visual comparison. In contrast to the simplified

tree models, the real tree models are all different in their crown structure and even single branches are visible. While this may be of limited value for analyses following area-based approaches, it is intuitively clear that such datasets offer advantages for example in the context of creating benchmark datasets for individual tree detection algorithms or other applications where fine-scale structures matter (e.g. individual tree delineation, tree species classification, predicting light availability on the ground, mapping habitat structures, etc.).

The quantitative analysis showed small differences between the simulated RTM and STM data. Depending on the purpose of a study, the simplified tree models could be an easy-to-create alternative to the real tree models. They could further be improved with comparatively low effort, e.g. by adjusting the point density of the tree crown, or by changing the spheroid crown shape to other geometric shapes, such as cones, depending on the tree species.

The real tree models were extracted manually from high density ULS point clouds. Tree models extracted from terrestrial laser scanning (TLS) or high density ALS point clouds could also be used, if the tree structure is well represented in the data. Because of the sensitivity of the optimal HELIOS++ parameters to the point density of the input data, we deduce that only tree models of similar point density should be used in a virtual scene. In addition, the point density of the tree models should not be lower than the resulting point density of the laser scanning simulations.

It is important to mention that the quality of the synthetic ALS data produced with the presented workflow depends on the availability of a sufficient number of high quality tree models for all relevant species and a good coverage of within-species variability. The manual extraction of tree models is a rather time-demanding yet feasible approach, as the tree models – once extracted – can serve for many simulations and scenarios. With the further development of tree segmentation algorithms (e.g. Ayrey et al., 2017; Burt et al., 2019; Krisanski et al., 2021; Luo et al., 2022; Vega et al., 2014; Windrim & Bryson, 2020), it may also be possible to extract tree models from less complex forest stands automatically. A limitation of this approach is that high resolution laser scanning datasets are not yet available for many parts of the world. Furthermore, the extraction of individual trees directly from the data may become more challenging when focusing on forests with very complex structures and a high diversity in tree species, such as in the tropics. In such situations, it might be more suitable to summarise tree models of several species into physiognomic types. Making tree point cloud datasets publicly available (e.g. Seidel et al., 2021; Weiser et al., 2022b, https://3dforecotech.eu), is a promising way to solve the problem of limited tree model availability in the future.

2.4.4 Quantitative comparison of the point clouds and biomass models

The simulation process was evaluated with regard to its ability to generate realistic ALS data of forests. This could only be done by comparing the simulated point clouds to real ALS point clouds, even if the objective of the presented approach is not to replicate existing point clouds but to generate simulated point clouds of (synthetic) forest stands for which

no laser scanning data are available. The comparison was thus not conducted point-wise but on a plot and subplot level. We thereby omit small scale differences between trees and selected tree models, because it is not expected (and also not envisaged) that the inserted tree point clouds perfectly match the real trees of the original ALS data. In addition, a plot-wise comparison can reduce the influence of spatial mismatch between the datasets that occurs due to errors in the field measurement of tree positions. A drawback of the plot-wise comparison of the point distribution is that it disregards individual tree structures, which can lead to compensation effects of false present and false missing returns.

The canopy cover of the synthetic forest stands was always lower than the canopy cover of the real forest plots. As already discussed, this can be explained by the missing implementation of canopy closure in the creation of synthetic stands. Unlike the simplified tree models, the real tree models may have unevenly shaped crowns or lean in one direction, the canopy cover of the RTM stands is therefore lower than the canopy cover of the STM stands. Consequently, as long as canopy closure is not considered in the creation of synthetic stands, the simplified tree models are more suitable for simulating realistic values for canopy cover. In the real forest sites, neighbouring trees with their stem positions outside of the 1-ha plots also contribute to the canopy cover. The fact that these edge effects are missing in the synthetic data further contributes to the lower canopy cover of the synthetic stands.
The height distribution profiles (Figure 2.4) of the returns demonstrate that it is possible to simulate ALS point clouds with a realistic return height distribution. Furthermore, they also reveal that the accuracy differs between study sites. Differences in the return height profiles of the RTM and STM data show that the tree habit affects the return height distribution. In some plots (e.g. BR01), the use of species specific real tree models resulted in a better matching return height distribution than the simplified ones, whereas in others (e.g. KA09), the simplified tree models led to better results. Where the simplified tree models performed better, this was probably because the trees there had a different growth form, e.g. a different crown base height, than the available real tree models of these species, a fact that underlines the importance of accounting for within-species variability. A larger number and diversity of available tree models and the addition of further filter criteria in the tree model selection procedure could solve these problems. Some of the visible differences in the shape of the return height profiles of the simulated and the original data can again be explained by the missing implementation of undergrowth and canopy closure in the synthetic stands. For example, the high canopy cover at plot KA11 led to a small number of first returns in the subcanopy layers. Because the canopy cover was lower in the simulated RTM stand than in reality (see Table 2.4), there were fewer first returns from the canopy layer and more first returns from the layers below (see Figure 2.4 l). High deviations in the height distribution of all returns, visible in the profiles of the RTM stands of plot KA09 and plot KA10 (see Figure 2.4 g & Figure 2.4 i), indicate that the subcanopy layers were not well represented in the simulated data.

The comparison of point cloud metrics derived from the original and the simulated data supports the previous findings that the height distribution of the returns was not equally well represented in the simulated point clouds of all (sub-) plots. Still, the median deviation of the metrics was small. Accordingly, the differences in the point cloud metrics of the

original and simulated data had only minor effects on the RMSE of the biomass predictions. Both the RTM and the STM data were shown to be suitable for calibrating biomass models. The significant differences in the mean error of reference and predicted biomass of the RTM and STM models indicate that the metrics that clearly differed between the RTM and STM datasets (e.g. Pgr2 and NP) had a strong influence on the predictions. Adjusting the parameters of the HELIOS++ simulations so that the total number of returns and the percentage of returns above 2 m of the simulated data better match the original data could potentially increase the model accuracy.

2.4.5 Outlook

The idea of creating synthetic laser scanning datasets is to generate extensive amounts of data which can be used for method development and sensitivity analyses. The presented approach requires stand information on an individual tree-level, which is rarely available from common forest inventories. Roberts et al. (2020) presented an approach to reconstruct 400 m^2 forest plots based on forest inventory data by estimating tree positions and properties through Monte Carlo simulations. An alternative method for generating such data is the use of forest growth simulators. They enable to simulate realistic forest stand data for different species compositions, management practices and environmental conditions. Examples for such growth simulators are FORMIND (Fischer et al., 2016) and SILVA (Pretzsch et al., 2002). FORMIND was developed for simulating dynamics in tropical rain forests, but there is also an adapted version for European temperate zones including eight tree species (Bohn et al., 2014). SILVA was developed based on long-term forest growth experiments in Germany and Switzerland. It is parameterised for Norway spruce, silver fir, Scots pine, European beech, and sessile oak (Pretzsch et al., 2002). Even if these forest growth simulators have a limited number of species and do not incorporate all processes occurring in natural forests, they allow to easily generate realistic data of forest stands with low complexity. A big advantage of these synthetic data is that tree positions and properties are known, thus effects of measurement errors do not need to be considered in further analyses.

Coupling forest stand information (from real forest inventories or forest growth simulators) with real tree models and the laser scanning simulation approach HELIOS++ enables generating laser scanning point clouds for varying laser scanning acquisition settings and forest stand compositions with full information on every individual tree. This approach can therefore be used for sensitivity analyses, e.g. to investigate how laser scanning pulse density, geolocation errors, border effects, or methods for field reference data collection affect the laser scanning-based estimations of forest attributes. Coupling the laser scanning simulation approach with a forest growth simulator allows for analyses of the effects of temporal discrepancy between field data collection and laser scanning acquisition, by simulating the forest stand composition and corresponding laser scanning data of two consecutive time steps. Methods for the identification and removal of statistical outliers could be developed by artificially introducing outliers to the synthetic data. In addition, the created data are suitable for the development and testing of methods for the detection and

segmentation of individual trees (e.g. Hamraz et al., 2017; Maltamo et al., 2004; Morsdorf et al., 2004) and tree species classification (e.g. Brandtberg, 2007).

As it has been shown in this study, simulated laser scanning data can be used for the calibration of regression models which can then be applied on real data. The synthetic data also have high potential as training data for deep learning applications, e.g. convolutional neural networks (CNNs), which are increasingly used to segment individual trees and to derive forest characteristics from airborne and UAV-borne laser scanning data (e.g. Ayrey & Hayes, 2018; Balazs et al., 2022; Briechle et al., 2021; Luo et al., 2022; Windrim & Bryson, 2020). The high demand for labeled training data, which is rarely available for ALS (Kölle et al., 2021), can be satisfied with virtual laser scanning datasets.

2.5 CONCLUSIONS

This study shows that combining forest stand information, real tree point clouds and the laser scanning simulation framework HELIOS++ can generate synthetic ALS point clouds of complex forest scenes that are sufficiently realistic for many applications. The laser scanning simulation approach could reproduce the relative height distribution of first and multiple returns, but the workflow did not work equally well for all forest stand compositions, mainly because undergrowth, canopy closure and tree interaction are not yet implemented in the creation of virtual stands. Training biomass models with point cloud metrics derived from simulated data instead of real ALS data only slightly decreased the model accuracy. The synthetic data therefore seem suitable for use in commonly-used area-based approaches.

Our analyses could not confirm the hypothesis that real tree point clouds add significant value over simplified penetrable spheroid tree models for area-based approaches, when using opaque voxels to represent the trees. Both tree model types can be used for creating synthetic point cloud data suitable to train empirical models. Which type of tree model is more suitable differs depending on the use case and on the forest stand composition. Canopy cover was better represented by the simplified tree model stands, whereas using the real tree models led to slightly better biomass models, in terms of their RMSE and ME. The real tree model approach would likely benefit from a larger number of available tree point clouds and additional criteria for the selection of tree models, enhancing the fit between real trees and tree models with regard to the within-species variability. For applications for which fine scale structure and diversity of tree growth forms are less important, the simplified tree models are an adequate and easy to generate alternative to the real tree models.

Future studies should explore further applications of the synthetic data, e.g. sensitivity analyses, or method development for individual tree detection and segmentation.

2.6 STATEMENTS

Data availability statement

The data underlying this article (i.e. the synthetic forest stands, the normalised real ALS point clouds and the files to reproduce the HELIOS++ simulations) are available in the KITopen repository, at https://doi.org/10.5445/IR/1000147797. The individual tree point clouds extracted from the real ULS data are available in PANGAEA, at https://doi.org/10.1594/PANGAEA.942856. Exemplary code for the creation of the synthetic forest stands using real tree models or simplified tree models is available on GitHub, at https://github.com/JannikaSchaefer/Syssifoss.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the frame of the project SYSSIFOSS - 411263134 / 2019-2022.

Acknowledgements

We thank Andressa Soarez Braz, Angelo Mayer, Christian Seitz, Daria Baete, Denis Debroize, Johannes Brand, Lisa-Maricia Schwarz, Lioba Martin, Marian Schimka, Nina Krašovec, and Thorben Schrempp for their assistance with data acquisition and processing.

2.7 APPENDIX

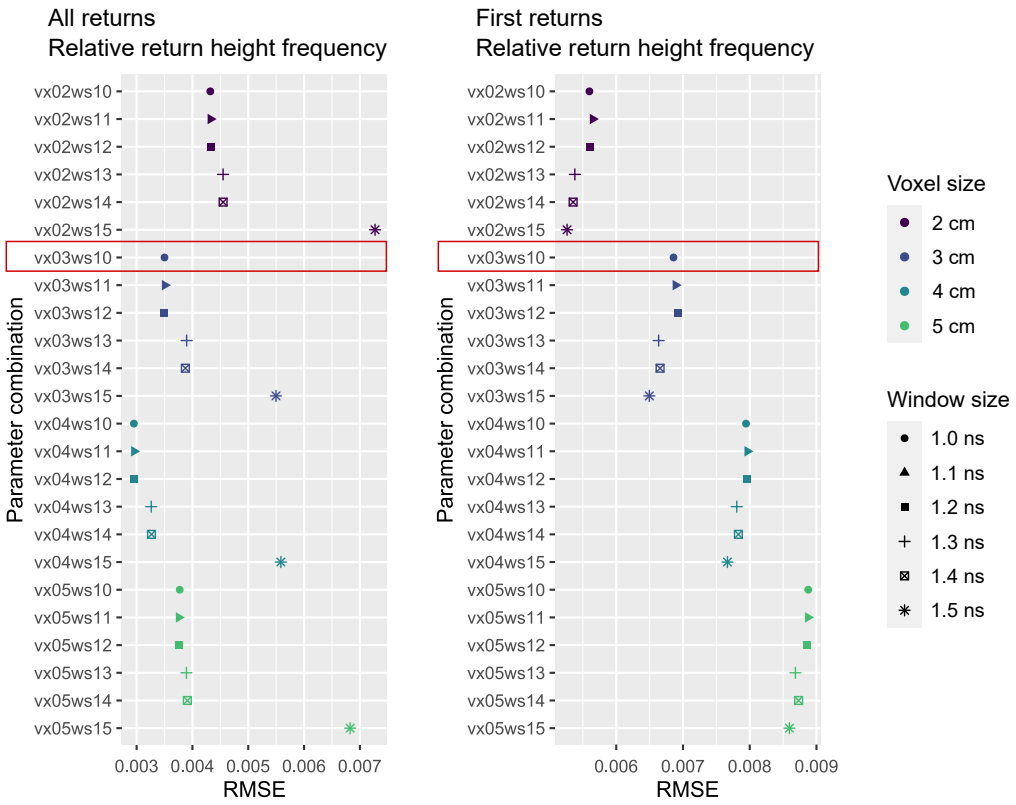

Figure 2.8: Root mean square error between simulated and observed return frequencies per 1 m height bins for different combinations of voxel size and window size applied in the HELIOS++ simulations. The red box highlights the selected parameter combination.

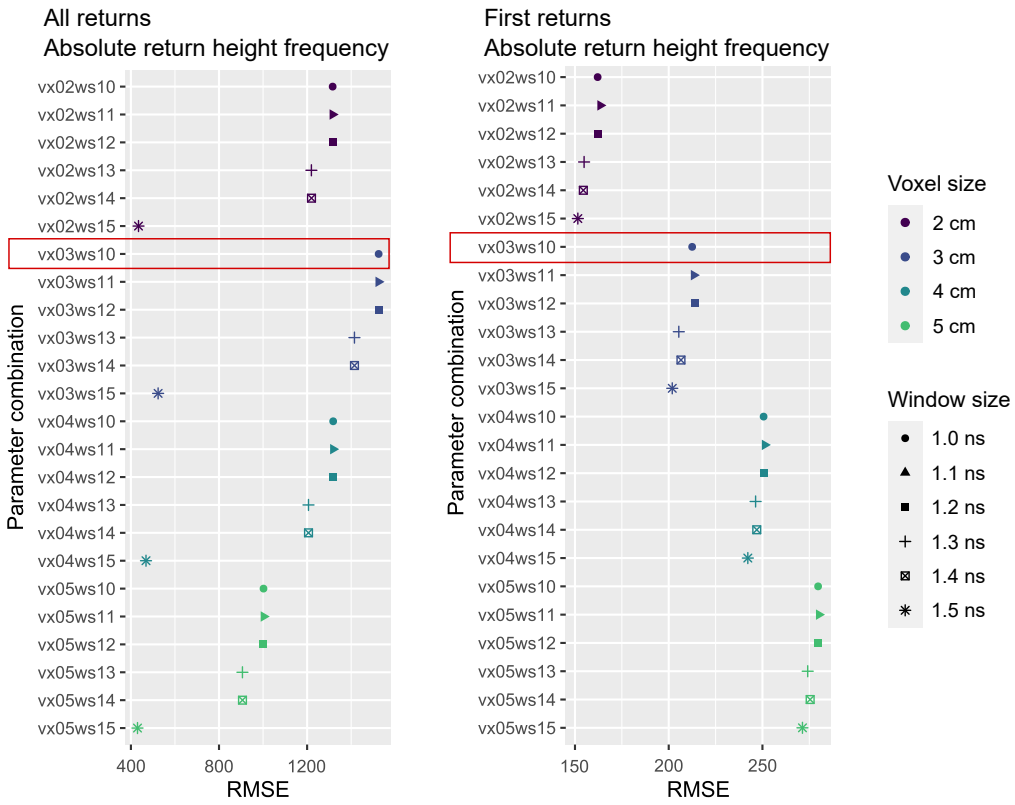

Figure 2.8 (cont.)**:** Root mean square error between simulated and observed return frequencies per 1 m height bins for different combinations of voxel size and window size applied in the HELIOS++ simulations. The red box highlights the selected parameter combination.

Chapter 3

3 ASSESSING THE POTENTIAL OF SYNTHETIC AND EX SITU DATA TO TRAIN AGB MODELS

This chapter has been published as:

Jannika Schäfer, Lukas Winiwarter, Hannah Weiser, Jan Novotný, Bernhard Höfle, Sebastian Schmidtlein, Hans Henniger, Grzegorz Krok, Krzysztof Stereńczak, Fabian Ewald Fassnacht (2023). **Assessing the potential of synthetic and *ex situ* airborne laser scanning and ground plot data to train forest biomass models.** *Forestry: An International Journal of Forest Research*, cpad061. https://doi.org/10.1093/forestry/cpad061.

ABSTRACT

Airborne laser scanning (ALS) data are increasingly used to predict forest biomass over large areas. Biomass information cannot be derived directly from ALS 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 a) simulated data only, b) a combination of simulated and real data, c) real data collected from different study sites, and d) 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.

3.1 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., 2013a). 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 (Frazer et al., 2011; Gobakken & Næsset, 2008; Lisańczuk et al., 2020; Packalen et al., 2023). However, field measurements are time consuming and costly, especially when field plots are remote or difficult to access (Hawbaker et al., 2009; Rana et al., 2016), and dense canopy and complex topography can reduce Global Navigation Satellite System (GNSS) positioning accuracy (Dalponte et al., 2011; Næsset & Gjevestad, 2008). For cost reasons it is hence of interest to keep the number of field plots small (Dalponte et al., 2011; Gobakken & Næsset, 2009; 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; Fekety et al., 2015; Maltamo et al., 2011).

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 (de Lera Garrido et al., 2020; Domingo et al., 2019). Temporal transferability of ALS-based biomass models has been demonstrated in several studies (de Lera Garrido et al., 2020; Domingo et al., 2019; Fekety et al., 2015; Zhao et al., 2018). However, the temporally transferred models often performed worse than models trained with field and ALS data acquired at the same time (de Lera Garrido et al., 2020; Domingo et al., 2019).

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 (Dalponte et al., 2011; Hawbaker et al., 2009; 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. Frazer et al., 2011; Knapp et al., 2018; Nelson, 1997; Palace et al., 2015), 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., 2023a). Laser scanning can be simulated using simplified statistical models (e.g. Knapp et al., 2018; Nelson, 1997; Palace et al., 2015; Spriggs et al., 2015; Wang et al., 2013), or using computationally more intensive approaches (e.g. Disney et al., 2010; Holmgren et al., 2003; Roberts et al., 2020; Schäfer et al., 2023a; Zhu et al., 2020) which 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 (Disney et al., 2010; Holmgren et al., 2003; Roberts et al., 2020), or to analyse the influence of field plot size and co-registration error on ALS-based biomass predictions (Fassnacht et al., 2018; Frazer et al., 2011). 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. (2023a) 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 a) simulated data only, and b) 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)?

3.2 MATERIALS AND METHODS

3.2.1 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 3.1 gives an overview of the laser scanning acquisition settings and resulting mean pulse densities and mean planar point densities of all study sites.

3.2.2 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 2 500 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

Table 3.1: Laser scanning acquisition settings and resulting mean pulse density and mean planar point density. Numbers in square brackets indicate values of the simulations differing from reported values of the real acquisitions.

	Milicz Forest	Silesian Beskids	DendroNET sites	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]

[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.

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 $\pm$ 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. (2023a) 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 $\times$ 20 m Forest Factory stands were arranged to larger scenes of 100 m $\times$ 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 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., 2023a), 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 3.1), 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 values 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.

3.2.3 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 DBH-based sampling), and whether individual tree positions were included in the data. Therefore, data were extracted in ways specific to the study sites (Figure 3.1).

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 $\times$ 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 $\times$ 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 $\times$ 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

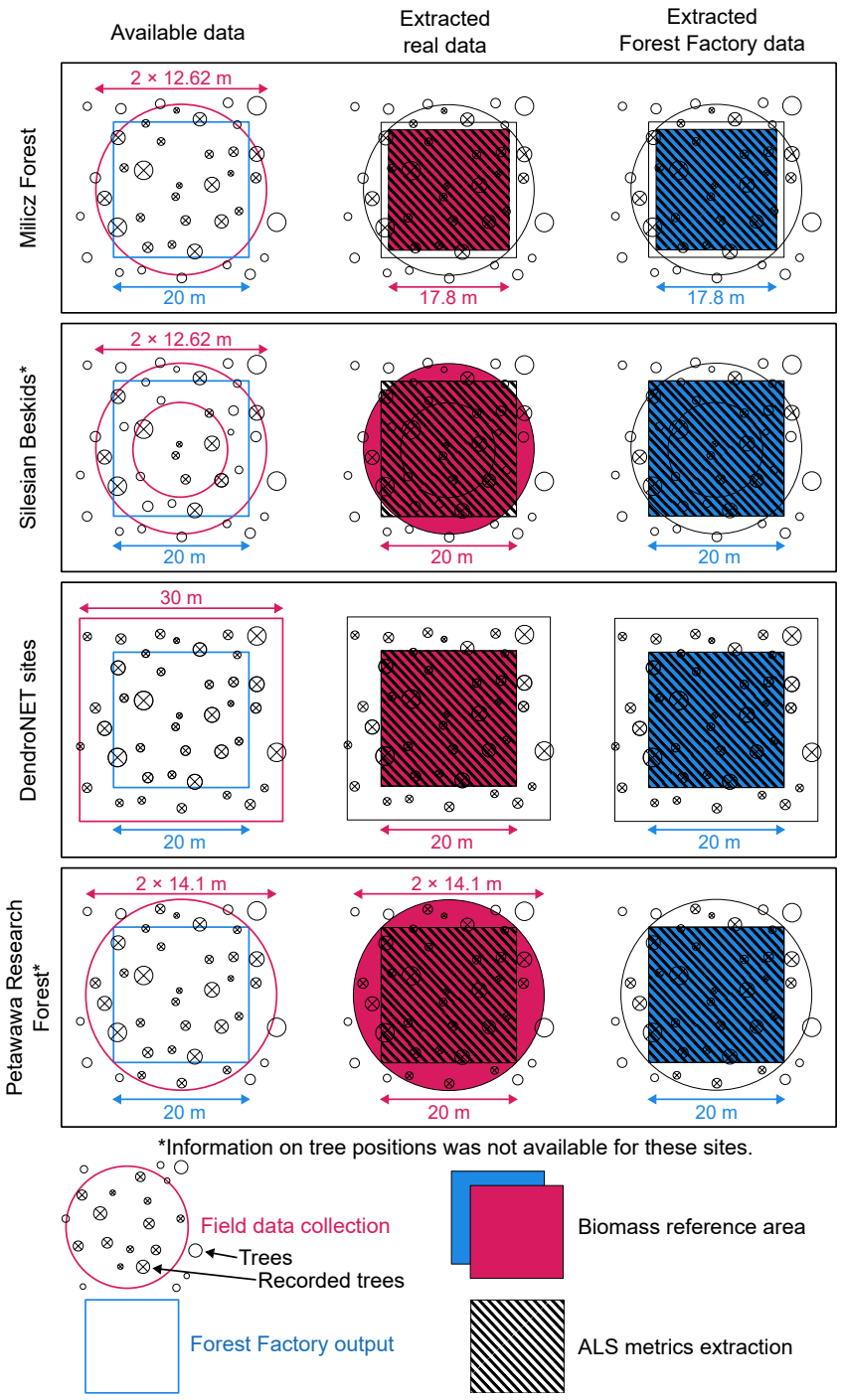

Figure 3.1: Schematic overview of field plot sizes and shapes of the four study sites in comparison to the Forest Factory plots (left). Areas from which biomass reference data were collected and areas from which ALS metrics were extracted are highlighted by colour and stripe pattern, respectively, for the real-world data (centre) and the Forest Factory data (right).

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 & Auty, 2021; Roussel et al., 2020). The point clouds were cropped to the plot extents and the height values were normalised using the *normalize_height()* function included in the "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 1^{st}–5^{th} 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 10th–90th (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.

3.2.4 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 t/ha to 1 323.81 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 2 500 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 3.2 shows two examples of the return height distributions of real forest plots and the corresponding Forest Factory plots.

3.2.5 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 root mean squared error (RMSE), the squared Pearson correlation coefficient (r^2), and – as a measure of bias – the mean error (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 a) all simulated data, b) a randomly selected subset of the simulated data, and c) 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 (b) 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

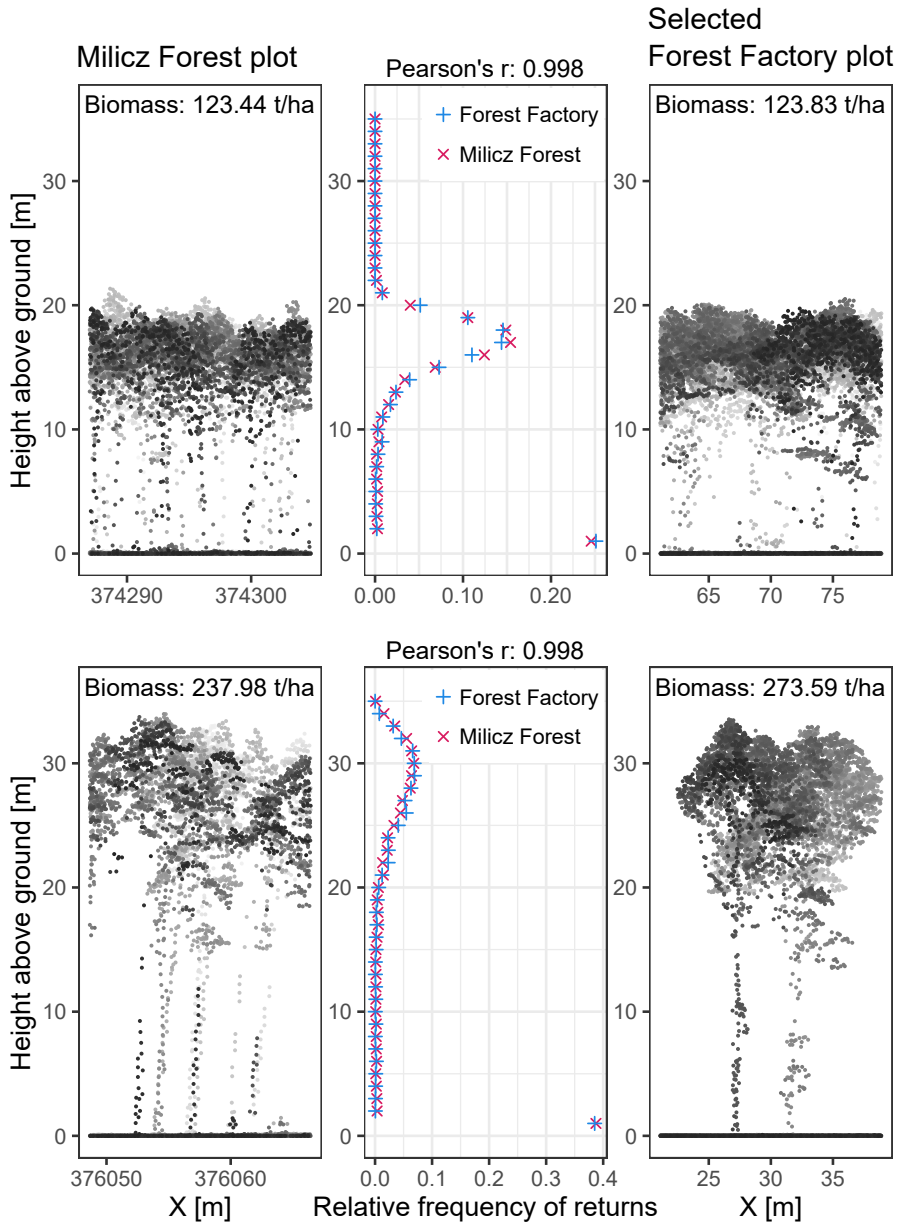

Figure 3.2: Two examples of the applied waveform matching approach for selecting the best matching Forest Factory plot for each real plot. The left images show vertical sections of real ALS point clouds of plots located in the Milicz Forest, the right images show the simulated ALS point clouds of the 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 t/ha and 123.83 t/ha), but have a higher difference for the lower example (237.98 t/ha and 273.59 t/ha). Points are coloured according to their position in Y-direction.

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 et al., 2021), "ggplot2" (Wickham, 2016), "viridis" (Garnier et al., 2021), "ggpubr" (Kassambara, 2020), and their dependencies.

3.3 RESULTS

3.3.1 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 (0.26–1 323.81 t/ha for 20 m × 20 m plots, 0.00–1 488.57 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 14 897 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 (2 500 trees/ha). The DendroNet sites had the smallest mean stand density (846 trees/ha) and also the smallest range (89–1 600 trees/ha). The stand density of the Milicz Forest plots ranged from 32 trees/ha to 4 261 trees/ha, with a mean of 951 trees/ha. Information on the stand density of the Silesian Beskids plots was not available. Figure 3.3 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 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 the Hmean show that the simulated data covered a wider range of structural diversity as expressed by these two metrics (Figure 3.4). Compared to the real data, the simulated data show lower biomass values in relation to Hmean (Figure 3.4, left column). This trend is less pronounced but still visible in the best matching subset of the simulated data (Figure 3.4, 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 (Figure 3.4, right column).

3.3.2 Biomass models

3.3.2.1 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 3.5 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 mean error, 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 (Figure 3.5, 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 (Figure 3.5, 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 Figure 3.5, 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),

Figure 3.3: 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.

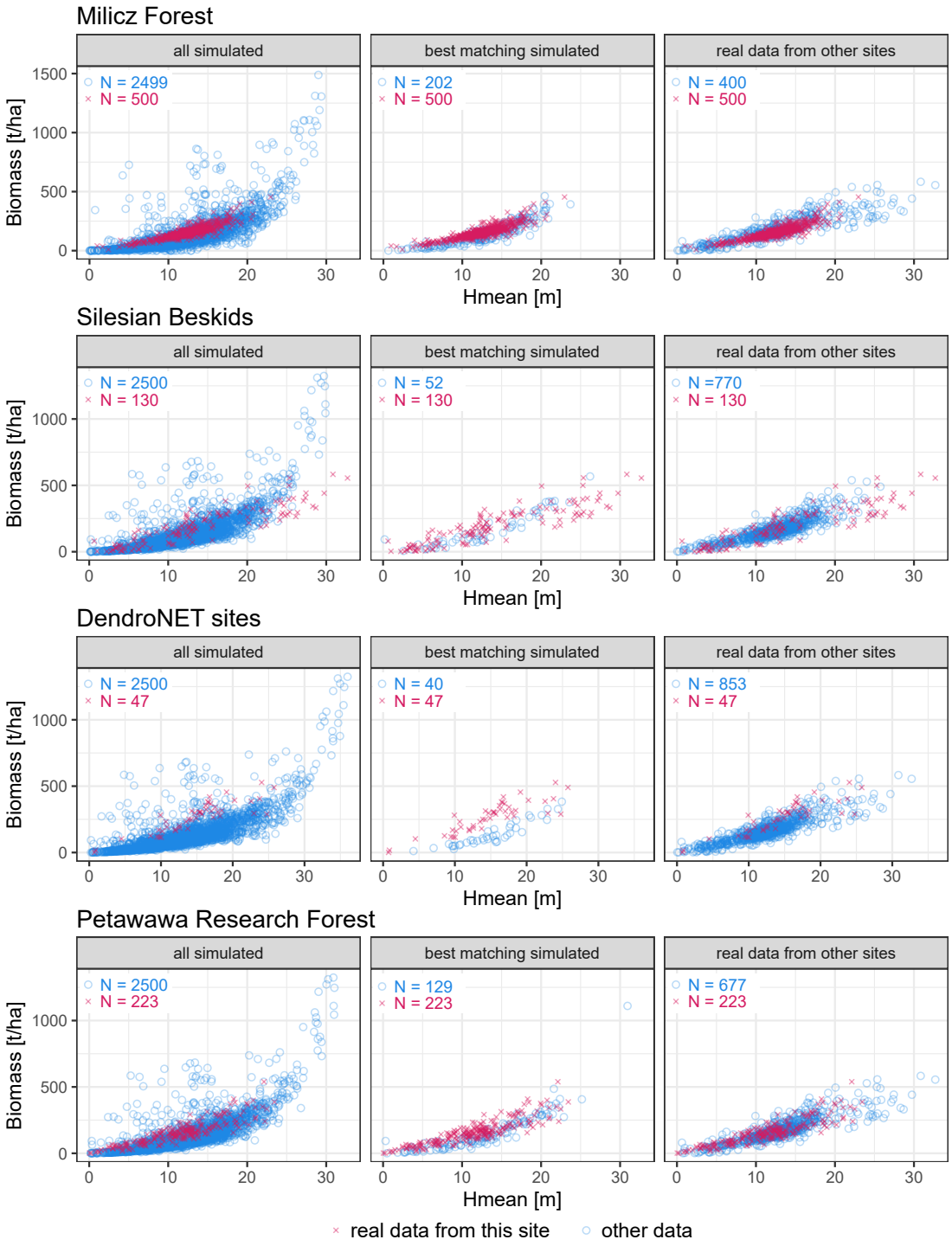

Figure 3.4: Biomass and mean height of returns (Hmean) of real and simulated forest plots. The real data collected from a site is compared to all simulated data (left), the best matching subset of the simulated data (centre), and the real data collected from the other sites (right).

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 (Figure 3.5, 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 3.6 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 Figure 3.5. 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 3.2 and the dashed vertical lines in Figure 3.6 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.

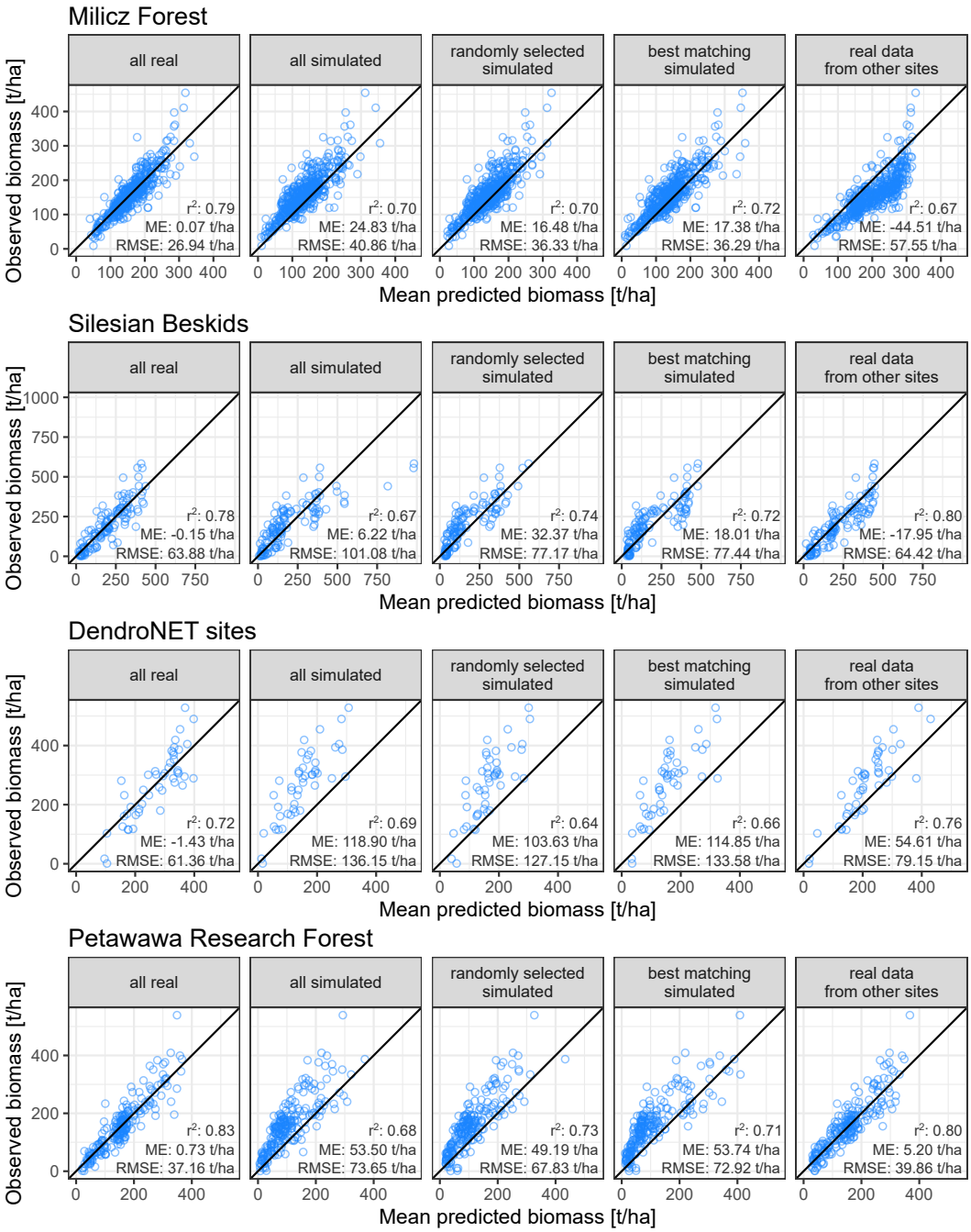

Figure 3.5: Mean predicted biomass and observed biomass of all field plots by study site and training data type. Model building and predictions were repeated 500 times for each training data type. In each of the 500 iterations, the real data were randomly split into 30% test data and 70% training data, 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 mean error (ME) and the root mean squared error (RMSE) are given.

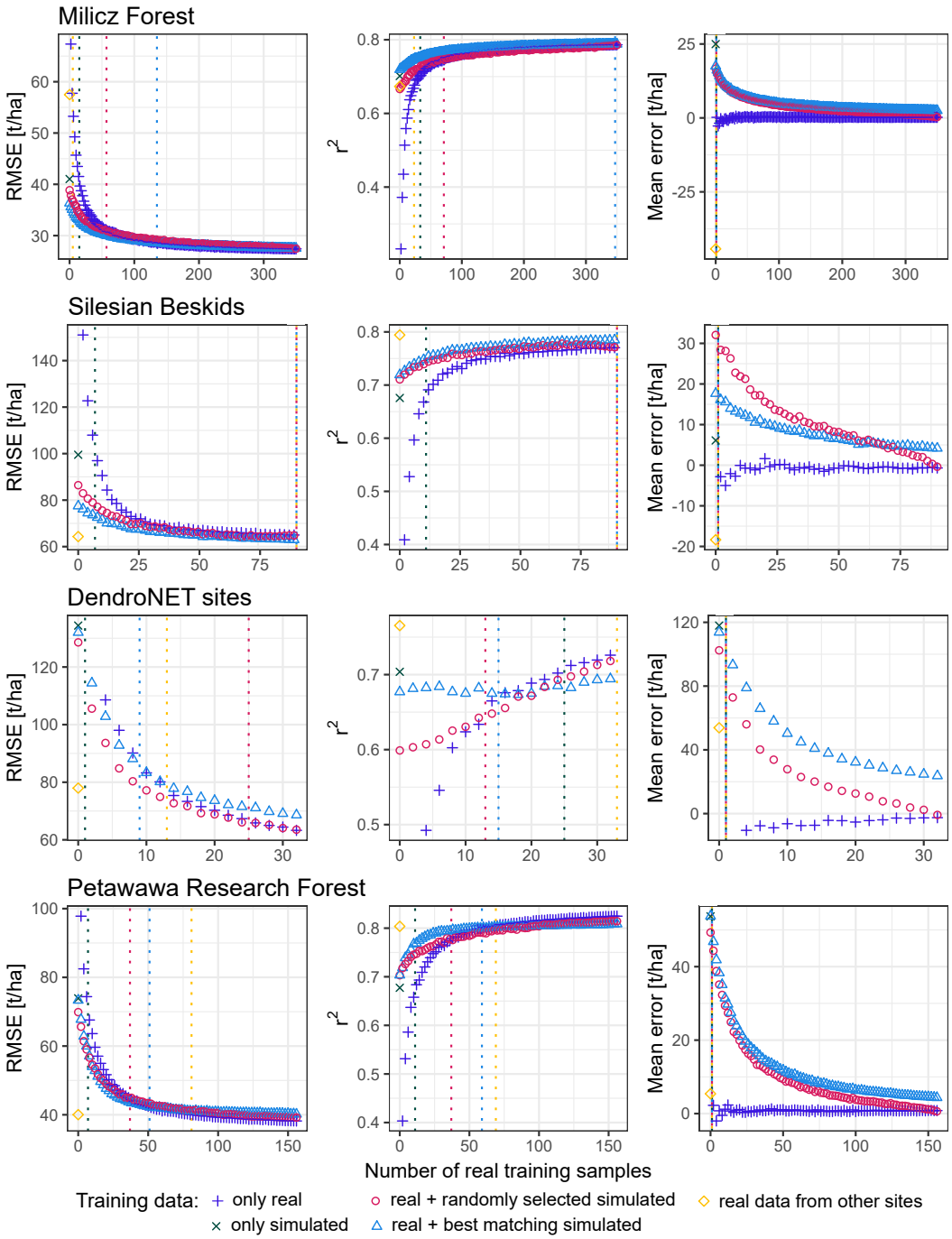

Figure 3.6: Mean root mean squared error (RMSE), mean squared Pearson correlation coefficient (r^2), and mean error (ME) of the biomass predictions for different training data types. Model building and predictions were repeated 500 times for each training data type and each number of real training samples. 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 3.2: Number of training samples up to which models trained with real *in situ* data perform 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

3.3.2.2 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 (Figure 3.6). 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 3.2.

3.3.2.3 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 (Figure 3.5, 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: $\geq$ 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 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 a ME of -44.51 t/ha.

3.4 DISCUSSION

3.4.1 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 (Figure 3.6). The underprediction was highest for the DendroNET sites (mean error 103.63–118.90 t/ha) and the Petawawa Research Forest (mean error 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 (Figures 3.3 and 3.4). 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 (Figure 3.5, 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 (Figure 3.4, 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 a) the simulation of the forest stand composition or b) 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 3.1). 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 (Ameztegui et al., 2022; Zianis et al., 2005). 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 which, 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 (Figure 3.4). 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 Figure 3.2) 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.

3.4.2 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.

3.4.3 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 and 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 mean error 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 and 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 the 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.

3.5 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.

3.6 STATEMENTS

CRediT author 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).

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.

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.

Chapter 4

4 AGB PREDICTION FROM ALS POINT CLOUD TOMOGRAPHY

This chapter is based on a journal article submitted to *European Journal of Remote Sensing*:

Jannika Schäfer, Lukas Winiwarter, Hannah Weiser, Bernhard Höfle, Sebastian Schmidtlein, Jan Novotný, Grzegorz Krok, Krzysztof Stereńczak, Markus Hollaus, Fabian Ewald Fassnacht. **CNN-based transfer-learning for aboveground biomass prediction from ALS point cloud tomography.**

ABSTRACT

Airborne laser scanning (ALS) can provide information on the three-dimensional structure of forests over large areas. ALS data are therefore increasingly being used for the prediction of forest aboveground biomass (AGB). In the area-based approach, it is common to build regression models from metrics describing the ALS point cloud distribution and corresponding AGB reference data. With this study, we present an alternative method: We predicted AGB from sequences of images depicting vertical cross sections through the ALS point clouds using convolutional neural networks (CNN). We used a 3D version of the VGG16 CNN with initial weights derived from pre-training on the ImageNet dataset. We tested the approach on four datasets collected in Canada, Poland, and the Czech Republic. The models were trained on differently sized samples ranging from 10 up to 375 ground plots (including validation data), depending on the study site. We compared the predictive performance of the CNNs with random forest models (RF) that were trained on traditional point cloud metrics and found that in most cases, the CNNs performed slightly worse at small training sizes and slightly better at the maximum number of training samples. At the maximum number of training samples, the difference in root mean squared error between observed and predicted AGB of CNNs and RFs ranged from -2 t/ha to 5 t/ha, and the difference in squared Pearson correlation coefficient ranged from -0.05 to 0.06. The absolute mean error was 2–6 t/ha larger for the CNNs compared to the RFs. We investigated whether additional pre-training on synthetic data derived from virtual laser scanning of simulated forest stands could improve the prediction performance of the CNNs, which was the case when very few real training samples (10–40, depending on the study site) were available. For larger training sample sizes, no benefit was observed. While the 3D CNNs trained on cross section images derived from real data showed promising results, RF models based on point cloud metrics remain a competitive alternative with lower computational requirements and faster model training times.

4.1 INTRODUCTION

Forests play an important role in the global carbon cycle, being the main terrestrial carbon sink (Pan et al., 2011). Deforestation and forest degradation contribute to anthropogenic carbon emissions, while carbon is sequestered through forest growth and the expansion of forest areas (Dixon et al., 1994). To effectively monitor forests, and to investigate the effects of anthropogenic and non-anthropogenic influences on forest status, large-scale data on forest structure is required. Airborne laser scanning (ALS) allows to non-destructively obtain information on the three-dimensional structure of forests over large areas. It is therefore increasingly used for estimating stocks and changes of forest aboveground biomass (AGB) (Strîmbu et al., 2023).

In the most commonly applied area-based approach (Næsset, 2002), ground measurements of AGB are linked to metrics describing the distribution of the spatially co-located laser scanning returns. Regression models can subsequently be employed for wall-to-wall

AGB prediction from the ALS data. Both linear models and non-parametric machine learning methods, such as nearest neighbour interpolation, support vector machines, and random forest (RF), are frequently applied for relating the point cloud metrics to the AGB observations (Fassnacht et al., 2014). While many studies have shown that AGB can be predicted from point cloud metrics (e. g. Bouvier et al., 2015; Sheridan et al., 2015; Zhao et al., 2009), it remains an open question whether AGB estimates could be further improved by using information from individual returns, i. e. the xyz-coordinates of the returns inherent in discrete return ALS data, rather than aggregated point cloud metrics. Depending on the point density, information about individual trees (e.g. their location, height, and volume) may be present in the raw point cloud data that is lost when using aggregated metrics at the plot level. Deep learning algorithms offer the potential to test this hypothesis without the need to specifically detect the individual trees as done in earlier studies (e.g. Dalponte et al., 2018; Jucker et al., 2017), as they do not require handcrafted and pre-extracted features. They can process raw data such as images or point clouds, thereby learning a latent representation of the data during the model optimization (LeCun et al., 2015).

Previous studies demonstrated that deep learning on ALS data can be used to classify tree species, coniferous and deciduous trees, as well as dead trees and snags (Briechle et al., 2021; Hamraz et al., 2019; Hell et al., 2022), and to estimate forest attributes such as growing stock volume and AGB (Ayrey & Hayes, 2018; Ayrey et al., 2019; Balazs et al., 2022; Oehmcke et al., 2022). The applied methods include 2D convolutional neural networks (CNNs) that are applied on 2D projections of the point clouds (Balazs et al., 2022; Briechle et al., 2021; Hamraz et al., 2019), 3D CNNs for which the point clouds are binned into a voxel space (Ayrey & Hayes, 2018; Ayrey et al., 2019; Balazs et al., 2022; Oehmcke et al., 2022), and deep learning algorithms that take the raw point clouds as input, such as PointNet, KPConv, 3DmFV-Net, or PointCNN (Hell et al., 2022; Oehmcke et al., 2022).

The limiting factor for the development of deep learning applications for inferring forest attributes from ALS data is the high demand for labelled training data (Hamedianfar et al., 2022). In the aforementioned studies that predicted forest attributes on a plot level, 1 044–17 432 plots were used for model training, an additional 225–1 000 plots for model validation and 225–3 000 plots for model testing. Such large sample sizes are rarely available in forestry applications. In a review on remote sensing-based forest AGB estimations, Fassnacht et al. (2014) reported that only 9 of 90 reviewed studies had a sample size between 200 and 500 plots, and 66 of 90 studies had a sample size smaller than 100 plots.

Common techniques for dealing with limited data availability are data augmentation and transfer learning (Hamedianfar et al., 2022). Data augmentation can increase the number of training data, for example by flipping, rotating, and cropping images (Shorten & Khoshgoftaar, 2019), or rotating, scaling, jittering, and point-wise displacement of point clouds (R. Li et al., 2020). As the neural network learns how to represent the data in the optimization process, these examples show transformations of the data that are invariate with respect to the output. However, many of these methods were developed for classification or object detection but not for regression tasks and may not be directly applicable for some datasets, for example if the scales in the image are related to the

response variable (Hwang & Whang, 2022). In transfer learning, models pre-trained on other data are further trained on a small sample of the target data, reducing the amount of labelled target data required (Hamedianfar et al., 2022). A lot of training efforts are required to recognize basic shapes like edges and corners, and these efforts can be transferred across domains. Transfer is generally more successful the closer the target and the source domain are, but it has been shown to work even across quite contrasting domain pairs (Niu et al., 2021). There are many predefined CNN architectures available with weights derived from training on large image datasets, such as those from the ImageNet database (Deng et al., 2009; Kattenborn et al., 2021). In contrast, pre-trained models do not yet exist for point-based deep learning methods for vegetation analysis (Winiwarter et al., 2022).

Pre-trained CNNs have been successfully applied for forestry applications, e.g. Briechle et al. (2021) used ResNet-18 models with pre-trained ImageNet weights for the classification of tree species and standing dead trees. However, there is a domain gap between the ImageNet images and those images that can be derived from laser scanning or other remote sensing techniques of forest. Accordingly, Fuller et al. (2022) found that pre-training on satellite images instead of ImageNet images improved the performance of land-cover classification from satellite images when using a vision transformer architecture. Another possibility for model pre-training is the use of simulated data (Winiwarter et al., 2022). Data simulation is a cost- and time-efficient way to generate large amounts of labelled training data. Luo et al. (2022) used synthetic forest scenes composed of randomly placed individual tree point clouds to train a deep learning model for tree detection in forest laser scanning point clouds acquired by an unoccupied aerial vehicle (UAV) and Sun et al. (2022) trained a deep learning model on synthetic images generated by Generative Adversarial Networks to segment individual tree crowns from ALS canopy height models.

In this study, we investigated whether deep learning can be applied to predict AGB from point clouds when the size of the training dataset is rather small. To compensate for the limited amount of training data, we used a transfer learning approach. We employed a CNN architecture developed by Solovyev et al. (2022) that is fed with sequences of 2D frames. Solovyev et al. (2022) created 3D versions of popular 2D CNNs that have been pre-trained on ImageNET images. These 3D CNNs have been successfully used to detect stalled brain capillaries in stacks of mouse brain images. Here, we apply this method to predict AGB from ALS tomography, i.e. sequences of images derived from vertical cross sections through ALS point clouds. Since the cross section images look quite different from the ImageNet images, we tested whether implementing an additional pre-training step with cross section images derived from simulated laser scanning point clouds can help in the domain transfer and increase prediction performance. These synthetic data were generated by combining virtual forest stands and a laser scanning simulator. The performance of the 3D CNNs was evaluated based on datasets obtained from four study sites, using 10–375 ground plots for training and validation. As a benchmark of model performance, AGB was also predicted from point cloud metrics using a random forest model. The research questions we addressed in this study were:

1. To what extent can AGB of forest stands be estimated from stacked cross section images derived from ALS point clouds using a 3D version of the VGG16 CNN pre-trained on the ImageNet dataset?
2. How does additional pre-training on synthetic data influence prediction performance of the CNN?

4.2 MATERIAL AND METHODS

4.2.1 Study sites

We used ALS point clouds and corresponding forest inventory data from four sites that were collected 1) in the Petawawa Research Forest (PRF) in Ontario, Canada, 2) in the Milicz forest district (MF) in the south-west of Poland, 3) in the Silesian Beskids (SB) in the east of the Czech Republic, and 4) in the DendroNET sites (DN) that are spread across the Czech Republic. Table 4.1 provides an overview on ground data including main tree species and ALS data acquisitions at the four sites. For MF, individual tree information (species, diameter at breast height ($D_{1.3}$), and tree height) was available and AGB was estimated using the same models as for the synthetic data (see next section). For the other sites, we used the AGB reference values that were provided by the original data owners.

PRF is a remote sensing supersite that covers approximately 10 000 ha of mixed-wood forests. The open-access data of PRF have been described in detail by White et al. (2019). The stand density in the 223 circular ground plots ranged from 32–13 024 trees/ha, with a mean value of 2 500 trees/ha (Figure 4.1). AGB ranged from 1–529 t/ha, with a mean value of 158 t/ha.

In MF, 70% of the 500 circular ground plots were located in pure stands of *Pinus sylvestris* L. (Stereńczak et al., 2018). The stand density ranged from 20–3 957 trees/ha and AGB ranged from 15-368 t/ha, with mean values of 952 trees/ha and 160 t/ha, respectively.

For SB, ground data of 130 plots were available. Tree data were collected in nested circular plots. Trees with a $D_{1.3} > 7$ cm were sampled within a radius of 3 m, whereas trees with a $D_{1.3} > 12$ cm were sampled within a radius of 12.62 m. Information on stand density was not available. The average AGB was 198 t/ha, ranging from 2–583 t/ha. More information on the SB data can be found in Brovkina et al. (2022).

The 47 DN plots were mostly located in pure forest stands. Ground data were collected in square plots of 30 m $\times$ 30 m. ALS data were available for 20 m $\times$ 20 m plots, therefore plot AGB values have been estimated from trees within these smaller plots based on individual tree positions. The stand density (measured in 30 m $\times$ 30 m) ranged from 89–1 600 trees/ha, with a mean value of 846 trees/ha. AGB ranged from 1–528 t/ha, with a mean value of 268 t/ha.

Table 4.1: Ground data collection and laser scanning acquisition settings, and resulting mean pulse density and mean planar point density of the four study sites. Numbers in square brackets indicate settings/values of the laser scanning simulations differing from the reported settings/values of the real acquisitions.

	Petawawa Research Forest (PRF)	Milicz Forest (MF)	Silesian Beskids (SB)	DendroNET (DN)
Forest inventory time	2014	Summer 2015	July 2019	October 2021
Ground plot design	circular plots, radius: 14.1 m	circular plots, radius: 12.62 m	nested circular plots, max. radius: 12.62 m	square plots, side length: 30 m
Number of ground plots	223	500	130	47
Main tree species	*Pinus strobus* L., *Populus tremuloides* Michx., *Quercus rubra* L., *Pinus resinosa* Ait., *Betula papyrifera*, *Picea glauca* Moench	*Pinus sylvestris* L., *Fagus sylvatica* L., *Quercus* spp. L.	*Picea abies* (L.) H. Karst, *Fagus sylvatica* L.	*Picea abies* (L.) H. Karst, *Pinus sylvestris* L., *Fagus sylvatica* L.
ALS acquisition time	2012	Summer 2015	July 2019	October 2021
Sensor	*RIEGL* LMS-Q680i	*RIEGL* LMS-Q680i	*RIEGL* LMS-Q780	*RIEGL* LMS-Q780
Altitude above ground	750 m	480–620 m [480 m]	819 m	515 m
Flight speed	? [54 m/s]	54 m/s	56 m/s	56 m/s
Flight line distance	250 m [≈ 242 m]	? [≈ 296 m]	440 m	-
Flight pattern	parallel[a]	parallel[a]	parallel[a]	perpendicular[a]
Mean pulse density	5.4 pulses/m^2 [4.3 pulses/m^2]	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]
Mean planar point density	11.7 points/m^2 [8.3 points/m^2]	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]

[a] The laser scanning simulations were performed with flight strips that were not perfectly parallel/perpendicular to reflect deviations from the flight pattern in the real data.

4.2.2 Synthetic data

Four datasets of synthetic forest inventory information and corresponding simulated ALS point clouds were generated, one for each study site. Forest stand compositions were simulated using Forest Factory 2.0 (Henniger et al., 2023), a software that makes use of the individual-based forest gap model FORMIND (Köhler & Huth, 1998). The original Forest Factory 2.0 version is calibrated to generate square forest plots of 20 m × 20 m. Because the ground plots of the real data exceeded this size, we used a modified version of Forest Factory 2.0 that enables to generate forest plots of 30 m × 30 m. The virtual forest stands were composed of *Pinus sylvestris* L., *Picea abies* (L.) H. Karst, *Fagus sylvatica* L., and *Quercus* spp. L.. Tree biomass was estimated using species-specific allometric models of the German National Forest Inventory that are implemented in the R package "rBDAT" (Vonderach et al., 2021). By default, Forest Factory 2.0 generates many more small AGB plots than large AGB plots (Schäfer et al., 2023b). To avoid effects of unbalanced training data, we sought to simulate data with AGB values equally distributed over the range of the real datasets. We therefore simulated a large number of forest stands (4 565 200). Of these virtual stands, all stands with an AGB of 0–600 t/ha were grouped into 12 bins, each bin with an AGB range of 50 t/ha. We then randomly sampled 200 stands per bin, resulting in a selection of 2 400 forest stands in total. Forest Factory 2.0 randomly assigns tree locations within a forest stand, which means that trees can be located unrealistically close together. We therefore implemented a workflow to generate new tree locations: For each forest stand, we created a grid of possible tree locations with an Euclidean distance of 1 cm between the locations and a minimum distance of 20 cm to the plot borders. We randomly selected one of these possible locations and assigned it to the first tree in the plot. All locations that were within the crown radius of that tree were then excluded from the remaining possible locations. This allowed for partial overlaps of trees, as the stem of the second tree could be placed directly at the edge of the first tree's crown. This procedure was continued for all trees within the plot.

Laser scanning of the virtual stands was simulated with the open-source laser scanning simulation framework HELIOS++ version 1.1.2 (Winiwarter et al., 2022). 3D representations of the virtual forest stands were created using tree point clouds that were extracted from real laser scanning data of temperate forests in the south-west of Germany. These data were acquired by a RIEGL miniVUX-1UAV mounted on a UAV during leaf-on conditions. The dataset of tree point clouds and corresponding tree metrics has been published by Weiser et al. (2022b) and described in detail by Weiser et al. (2022a). For each tree in the virtual forest stands, a tree point cloud of matching tree species and a tree height within ±4 m of the virtual tree's height was randomly selected and placed at the location of the virtual tree. If no tree point cloud was available in the specified height range, the one with the smallest difference in height was selected. The tree point clouds were randomly rotated around the Z-axis and uniformly scaled in all three dimensions such that the height of the point cloud matched the height of the virtual tree. All points outside of the 30 m × 30 m stands were removed. Laser scanning was simulated for scenes of 90 m × 90 m composed of nine virtual forest stands. The forest point clouds were voxelized with a voxel size of 3 cm. Filled voxels (with at least one point) were set to be opaque, and empty voxels were

transparent for the simulation (following Weiser et al., 2021). The ground was represented by a horizontal plane. Airborne laser scanning of the virtual forest scenes was simulated according to the acquisition settings of the real campaigns (Table 4.1), i.e. four different ALS simulations were conducted over the same scenes.

The virtual forest stands and simulated ALS point clouds were cropped to match the real ground plots (circular plots with radii of 12.62 m or 14.1 m, and square plots with side lengths of 20 m, depending on the dataset used, see Table 4.1). When there was no tree located within the cropped plot area, the synthetic forest stand was removed from the dataset. This resulted in 2 379 synthetic plots for PRF, 2 362 for SB, and 2 340 for DN. Due to an error in the sampling of the virtual forest stands, the dataset for MF consisted of 2 426 synthetic plots (of which 2 362 resulted from the sampling with regards to a uniform AGB distribution). Since the distribution was barely influenced by the additional 64 plots, the sampling was not repeated. Plot AGB was calculated as the sum of AGB of all trees within a plot divided by the plot area. The characteristics of the synthetic forest stands differed slightly depending on plot size and shape. The stand density ranged from 16–9414 trees/ha and the AGB ranged from 1–1 028 t/ha, with mean values of 377 trees/ha and 298 t/ha, respectively (Figure 4.1).

4.2.3 Cross section images

In order to feed point cloud data into a CNN, a rasterisation is required. Therefore, vertical cross section images were extracted from real and simulated ALS point clouds. To that end, the point clouds were height normalized using the "normalize_height" function of the R package "lidR" (Roussel & Auty, 2021; Roussel et al., 2020), and cut into 1 m thick vertical non-overlapping slices, both in x-direction and in y-direction. The slices were 45 m high (corresponding to the tallest trees) and the width corresponded to the respective plot extent (20–28.2 m). A binary image of 128 (width) $\times$ 256 (height) pixels was generated from each slice. Pixel values were set to black (0) if the volume represented by the pixel contained at least one ALS return, and to white (1) otherwise. Preliminary experiments on the MF dataset revealed that using RGB-images with a height-related colour map (viridis) did not improve the results. Figure 4.2 shows exemplary cross section images for one MF ground plot.

4.2.4 Experimental set-up

All experiments were conducted individually for each of the four study sites. The real data were split into 25% test and 75% training data using a stratified sampling approach, so that both test and training datasets covered the full range of AGB values of the respective study site. This resulted in 56, 125, 33, and 12 test plots for PRF, MF, SB, and DN, respectively. The remaining plots were used for model training and validation.

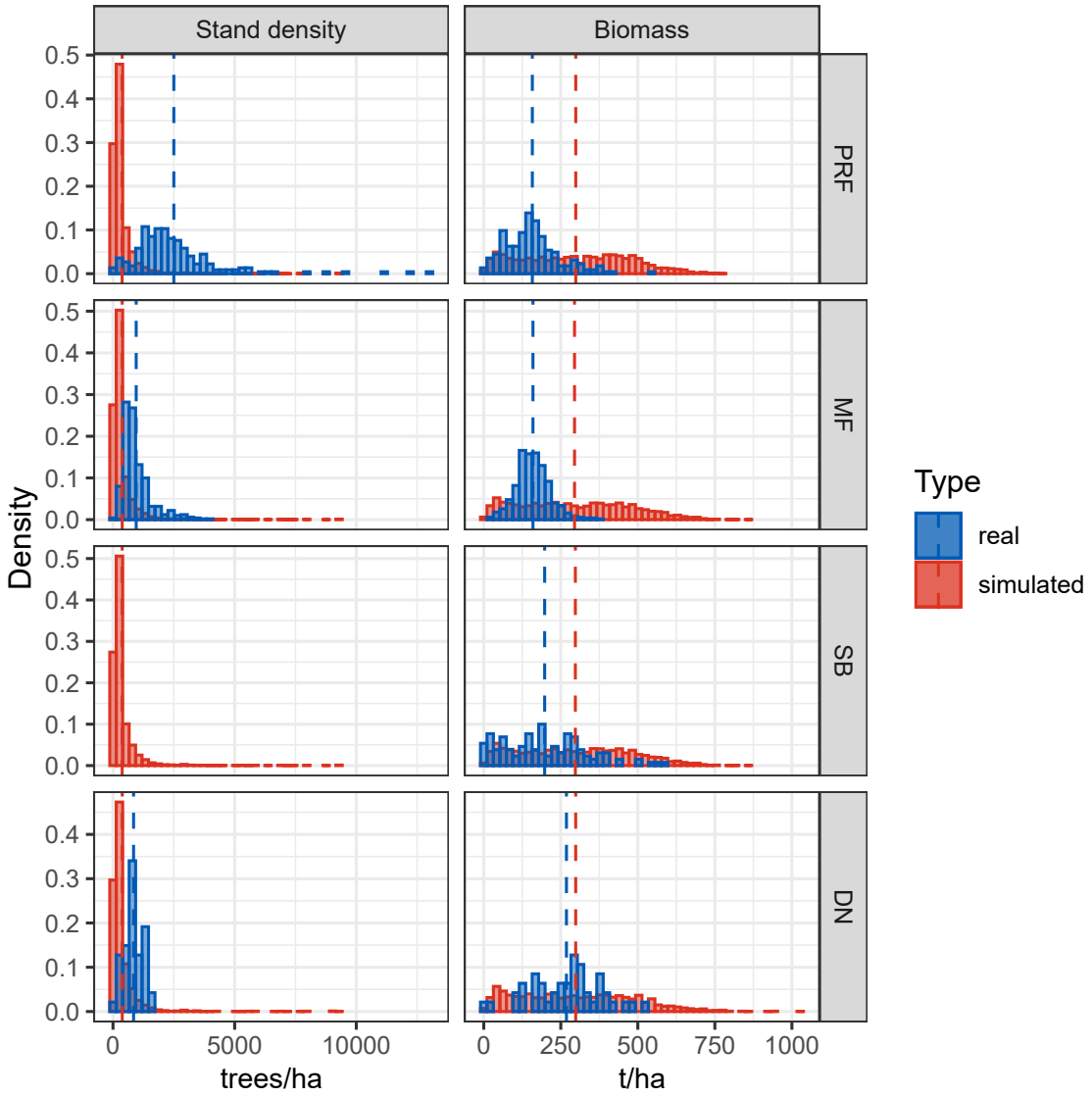

Figure 4.1: Stand density and aboveground biomass (AGB) of the real and simulated forest plots for the four study areas of Petawawa Research Forest (PRF), Milicz Forest (MF), Silesian Beskids (SB), and DendroNET sites (DN). Mean values are indicated by the dashed vertical lines. Information on stand density was not available for the Silesian Beskids (SB).

We evaluated the performance of CNNs trained only on cross sections derived from real data, compared to CNNs pre-trained on cross sections derived from synthetic data. As a benchmark, we used random forest models (RFs) (Breiman, 2001) with point cloud metrics as AGB predictors. Both CNNs and RFs were trained on the entire training datasets of each study site and, additionally, on randomly selected subsamples of different sizes. The size ranged from 10–100 (in increments of 10) training plots for PRF and MF, 10–90 plots for

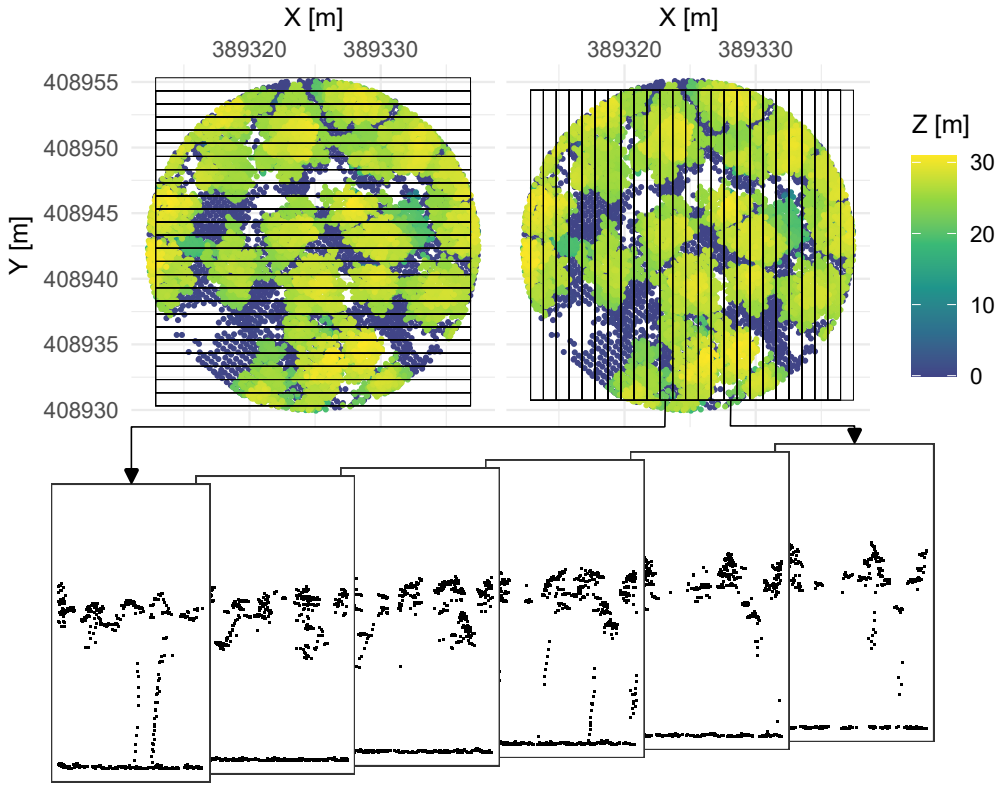

Figure 4.2: Generation of cross section images from an ALS point cloud of a ground plot located in the Milicz Forest. For better visualisation, the black pixels in the cross section images have been enlarged compared to the actual images and only a selection of images is shown. The vertical and horizontal black lines indicate the borders of the transects extracted from each plot point cloud.

SB, and 10–30 plots for DN (due to the different numbers of available plots for each study site). CNNs and RFs were furthermore trained on synthetic data only. For each sample size, training data sampling and subsequent model training were repeated 10 times. For the CNNs, the training datasets were further randomly split into 80% actual training data and 20% validation data, used for optimizing the model hyperparameters. Model performance was assessed using the root mean squared error (RMSE), the squared Pearson correlation coefficient (r^2) and, as a measure of systematic error, the mean error (ME) of observed and predicted AGB values for the test datasets.

The CNNs were conducted on a system with an NVIDIA RTX A4000 GPU, 16 GB VRAM, 256 GB RAM, and an Intel® Xeon(R) Silver 4210R CPU @ 2.40GHz × 40. The RFs were conducted on a system with 256 GB RAM, an Intel® Xeon(R) CPU E5-2630 v3 @ 2.40GHz × 16, and no dedicated GPU. We did not systematically assess the run time of the models. CNN training took between 4 minutes and 6 hours, depending on the training sample size and the number of epochs before early stopping. Because of the large training data size,

pre-training on the synthetic data took more than a day. In contrast, the RF training took only a few seconds for each run.

4.2.5 Neural network architecture

The backbone of our method is a VGG16 2D CNN (Simonyan & Zisserman, 2015) pre-trained on the ImageNet dataset (Deng et al., 2009). In order to apply it on 3D data, we used a re-designed architecture by Solovyev et al. (2022) in which all 2D convolutions are replaced by 3D convolutions. Thereby, the convolutions are not only applied on a single image but on a stack of images. The initial weights of the 3D kernel were transferred from the weights of the pre-trained VGG16. We fed the cross sections, ordered by depth, in both x- and y-directions separately through the CNN and concatenated the two feature vectors resulting from the two directions. We then passed the concatenated vector through a dense layer to obtain a scalar AGB value. The network architecture is shown in Figure 4.3. For optimisation of the weights, we used the Adam algorithm (Kingma & Ba, 2017), employing an exponentially decaying learning rate (initial learning rate = 10^{-7}, decay steps = 100 000, decay rate = 0.96, staircase = TRUE). Because of the limited GPU memory, the batch size was set to 1. Training was carried out for up to 600 epochs using the mean squared error as loss metric. After each epoch, the validation RMSE was evaluated and early stopping was applied if it did not decline over 20 consecutive epochs. For the pre-training on the synthetic data, the same neural network architecture was first applied on cross section images derived from the simulated point clouds. The weights of the best model according to the validation on the synthetic dataset were then used as initial weights in the further training on the real data cross section images.

4.2.6 Random forest models

For the benchmark models, point cloud metrics were derived from all returns, first returns, and all returns with a normalized height $> 2\,\mathrm{m}$. We used a subset of the pre-defined standard metrics from the "cloud_metrics" function implemented in the R package "lidR" (Roussel et al., 2020), precisely: the mean and the maximum of return heights, the standard deviation, the entropy, the kurtosis, and the skewness of the height distribution, the percentage of returns with a height $> 2\,\mathrm{m}$, the percentage of returns above the mean height, the percentage of the 1st–5th returns, the percentage of ground returns, the 5th to 95th height percentiles in increments of 5%, the cumulative percentage of returns in each of nine equally spaced height layers, and the total number of returns.

The point cloud metrics served as predictors in a random forest regression. The function "tuneRF" of the "randomForest" package (Liaw & Wiener, 2002) in R was employed to search for the optimal number of predictors to sample at each split. The number of trees was set to 500.

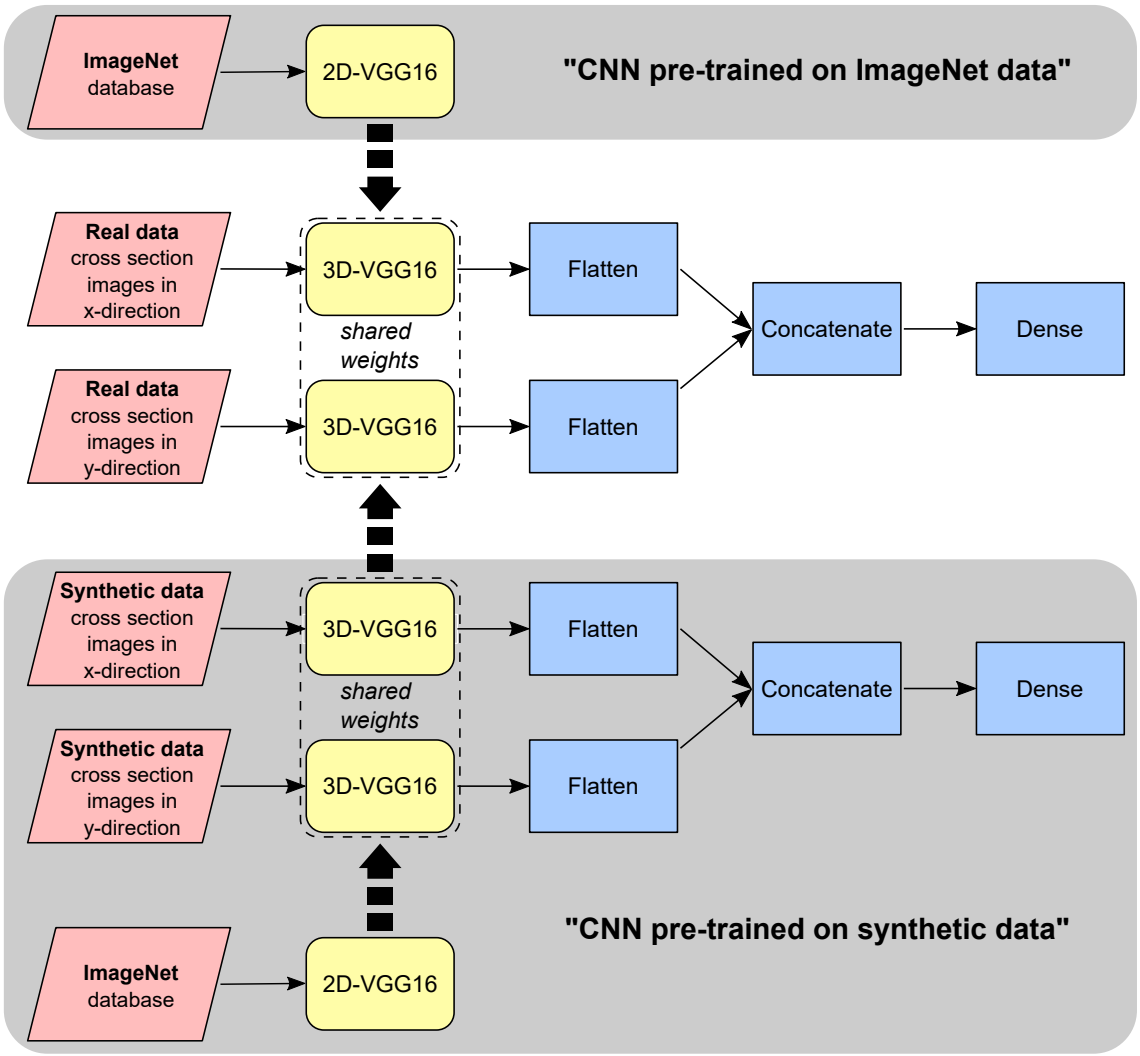

Figure 4.3: Neural network architecture. For a single data point, a block of images in x- and y-directions are fed through the 3D-VGG16 separately, but with shared weights between the networks. The output is flattened and concatenated, before being fed through a dense layer, further reducing the dimensionality to 1, i.e. the scalar AGB value as the regression target. Initial weights were always derived from pre-training on the ImageNet database. In case of the "CNNs pre-trained on synthetic data", an additional pre-training step on synthetic forest data was performed.

4.3 RESULTS

4.3.1 Model performance of CNNs compared to random forest models

The difference between CNN and RF performance varied between the study sites. Performance metrics were aggregated by taking the median over 10 repetitions, for each of which

the training/test split was randomized. Figure 4.4 shows the performance metrics for the different models and training sample sizes (tables of the accuracy metrics can be found in the Appendix, Chapter 4.7). When using the maximum number of training samples, differences in prediction performance and systematic error between CNNs and RFs were small. The RMSE was 20–71 t/ha for RFs and 24–70 t/ha for CNNs, and r^2 was 0.77–0.87 for RFs and 0.78–0.88 for CNNs, depending on the study site. The ME of the CNNs ranged from -23 t/ha for DN (indicating an overprediction of AGB) to +18 t/ha for SB (indicating an underprediction of AGB), and from -21 t/ha to +12 t/ha when using RFs. The absolute ME of the RFs was 2–6 t/ha smaller than of the CNNs. For PRF, SB, and DN, the CNNs performed similarly or slightly better than the RFs with regards to RMSE and r^2, whereas they performed slightly worse for MF.

The prediction performance of both CNNs and RFs depended on the number of samples that were used for model training. When training sample sizes were small, the CNNs had – in most cases – a slightly smaller prediction performance than RFs, but the predictive performance of CNNs and RFs tended to converge as the training sample size increased. For PRF and MF, the RFs performed better than the CNNs when trained on 10–100 samples. The difference in model performance was more pronounced in r^2 than in RMSE. For both model types, the systematic error was rather small, especially for MF. Only at a training sample size of 10, the CNNs showed a much larger mean error than the RFs did.

For SB, the RFs outperformed the CNNs when trained on 10–20 samples. At larger sample sizes, the difference in prediction performance between both model types was small. At some sample sizes, the CNNs resulted in a higher prediction performance, at other sample sizes, the RF predictions were better. The systematic error was in most cases slightly smaller for the RFs than for the CNNs.

The results for the DN dataset differed the most from the other study sites. Because of the small number of ground plots, models could only be trained on 10–35 samples (including validation data). The RMSE of the CNNs and the RFs were similar for sample sizes > 10, while it was slightly higher for the CNNs at a sample size of 10. The systematic error of the CNNs was much smaller than of the RFs for sample sizes of 10–20 (ME of -3 to -2 t/ha for the CNNs, -23 to -21 t/ha for the RFs), and similar for sample sizes of 30–35. With regard to the r^2, CNNs did not reach the prediction performance of RFs when trained on 10–20 samples, but outperformed RFs when trained on 30 and 35 samples.

4.3.2 Pre-training on synthetic data

Using CNNs that were pre-trained on synthetic data only improved the prediction performance when the number of real training samples was very small (Figure 4.4, blue and red cross marks). For PRF and SB, the performance of the CNNs in terms of RMSE and r^2 only improved by pre-training when no more than 10 and 20 samples were used for model training, respectively. A positive effect of pre-training on the systematic error could be observed when using 10–30 training samples for PRF, and 10–40 samples for MF. For SB, RMSE and r^2 of the predictions could be improved by pre-training on synthetic data

for 10–20 and 10–30 training samples, respectively. Pre-training could only reduce the systematic error at a sample size of 10.

The results for MF, PRF, and SB datasets showed that pre-training often not only failed to improve the models, but rather substantially worsened them. At larger sample sizes, pre-training strongly increased the RMSE for PRF, decreased r^2 for PRF and MF, and increased the underprediction of AGB for SB. The most positive effect of pre-training on synthetic data was observed for the DN dataset. Here, pre-training decreased the RMSE by 4–29 t/ha and increased r^2 by 0.13–0.28 for training sample sizes of 10–30 and 10–20, respectively. For larger training sample sizes, the pre-trained CNNs performed slightly worse than the non-pre-trained models, but the differences were small (0.5 t/ha for RMSE, 0.02-0.03 for r^2). In contrast, pre-training strongly increased the overprediction of the CNNs for 10–30 training samples.

When models were trained only on synthetic data, RFs performed better than the CNNs for PRF and SB (Figure 4.4, values for 0 real training samples). For MF, CNNs and RFs performed similarly in terms of RMSE, while r^2 was better for the RFs and the ME was better for the CNNs. For DN, the CNNs outperformed the RFs when no real data were used for model training.

4.3.3 Model stability

Figure 4.5 shows scatter plots of predicted and observed AGB of the four study sites resulting from the best and worst models for each model type (CNN pre-trained on ImageNet data, CNN pre-trained on synthetic forest data, and RF) and the minimum (10) and maximum (dependent on the study site) number of training samples. Differences in the prediction performance between the best and the worst of the randomized repetitions of models indicate how the models were influenced by the selection of training data and random processes within the models. The samples that were used for model training were randomly sampled from the training datasets. In case of the CNNs (both ImageNet pre-trained and synthetic forest pre-trained), 20% of these data were further removed to serve as validation data. Hence, when 10 training samples were available, eight were used as actual training samples and two as validation samples. When comparing models trained and validated on 10 samples, it is striking that in case of the CNNs that were not pre-trained on synthetic data, the worst models resulted in a very narrow range of predicted AGB values, albeit not the mean value of the training data AGB (Figure 4.5, top left panel of each study site: red markers appearing in an approximate vertical line). For example, for PRF, the worst CNN trained on 10 samples predicted AGB values of 43–66 t/ha while the reference was in the range of 1–399 t/ha. This effect did not occur for the largest sample size nor for CNNs that were pre-trained on synthetic data. We could not find a direct relation between prediction range and training data samples, as the AGB range of the training data of these models was much wider than the predicted AGB. However, using only two samples for validation could also have negatively affected model performance if validation accuracy lead to stopping model training too early.

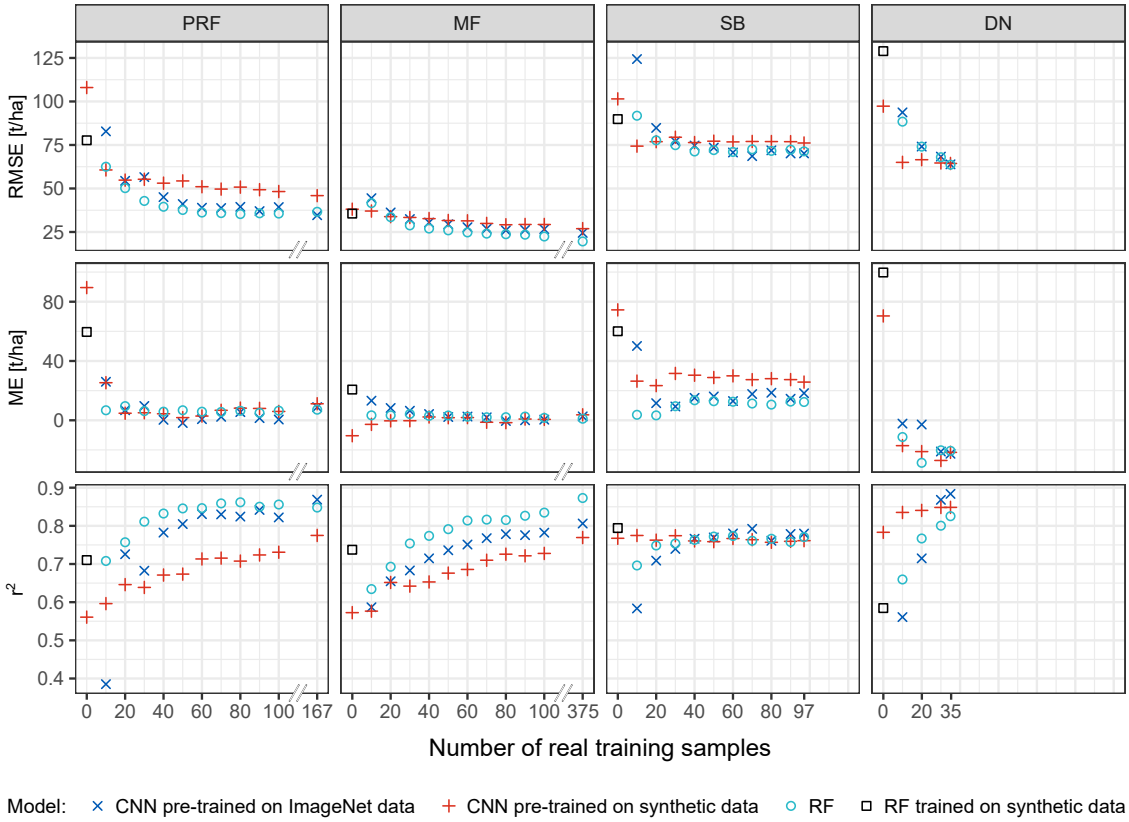

Figure 4.4: Median root mean squared error (RMSE), mean error (ME) and squared Pearson correlation coefficient (r^2) of the AGB predictions for different model types and sample sizes. Training sample count included 20% validation data for the CNNs. Model training and testing was repeated 10 times for each training sample size, except for the CNNs that were only trained on synthetic data. For these models, model training and testing was not repeated due to the long computing times.

In most cases, the difference in performance between best and worst prediction was highest for the ImageNet pre-trained CNNs and smallest for the synthetic forest pre-trained CNNs. Accordingly, pre-training on synthetic data had a stabilizing effect on the CNNs. The positive effect of the pre-training diminishes with increasing training sample size.

4.4 DISCUSSION

In our study, CNNs using images of ALS point cloud cross sections performed similar to RFs using traditional point cloud metrics in the prediction of AGB on plot level, but there were differences in model performance between study sites and depending on how many samples were used for model training. In an attempt to identify patterns of when CNNs

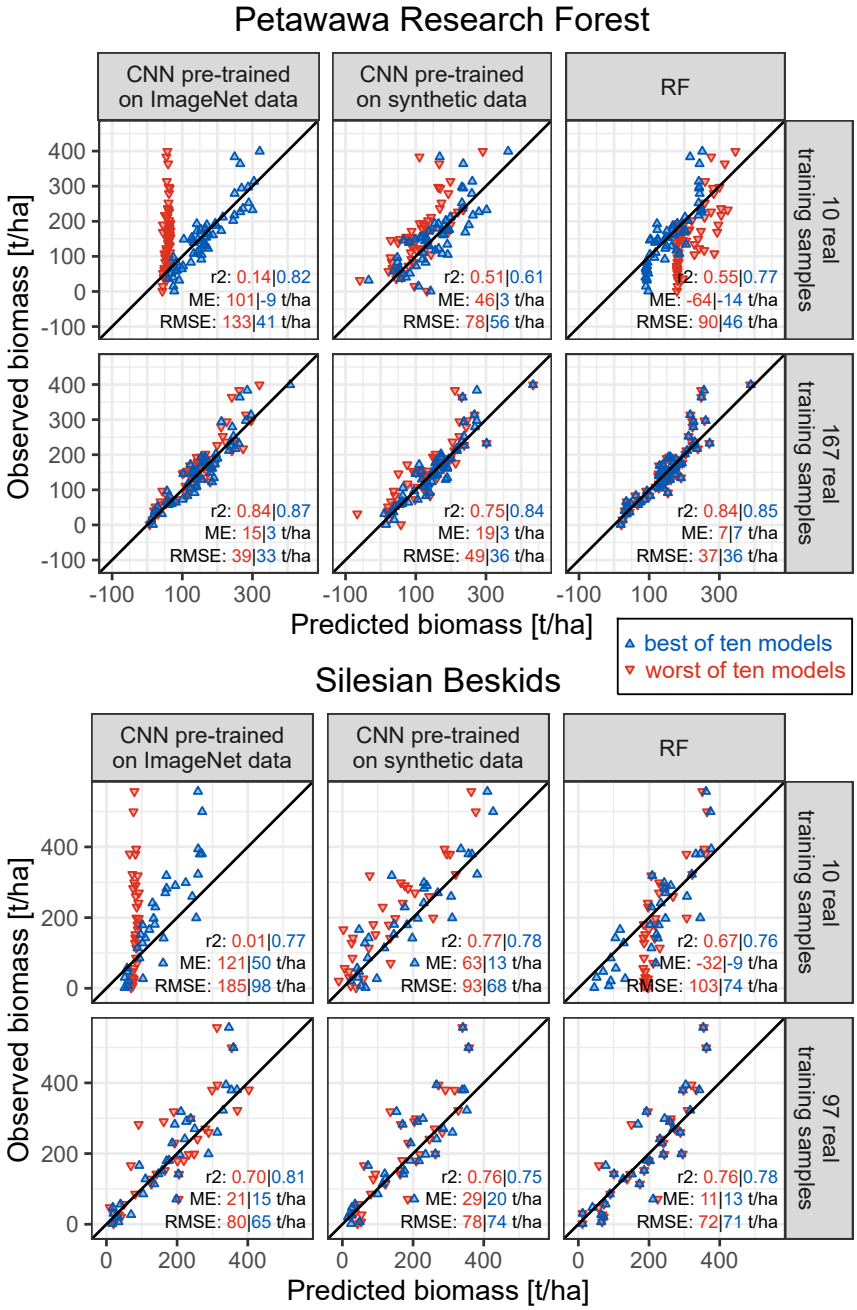

Figure 4.5: Predicted and observed AGB of the ground plots in the test datasets of each study site. Results are shown for models trained on 10 real training samples and on the maximum number of available training samples. Model training and prediction were repeated 10 times for each model type and training sample size, resulting in 10 predictions per ground observation. The training datasets consisting of 10 samples were randomly selected from the total training data in each run. Only the best and the worst predictions (in terms of RMSE) of the 10 model iterations are depicted.

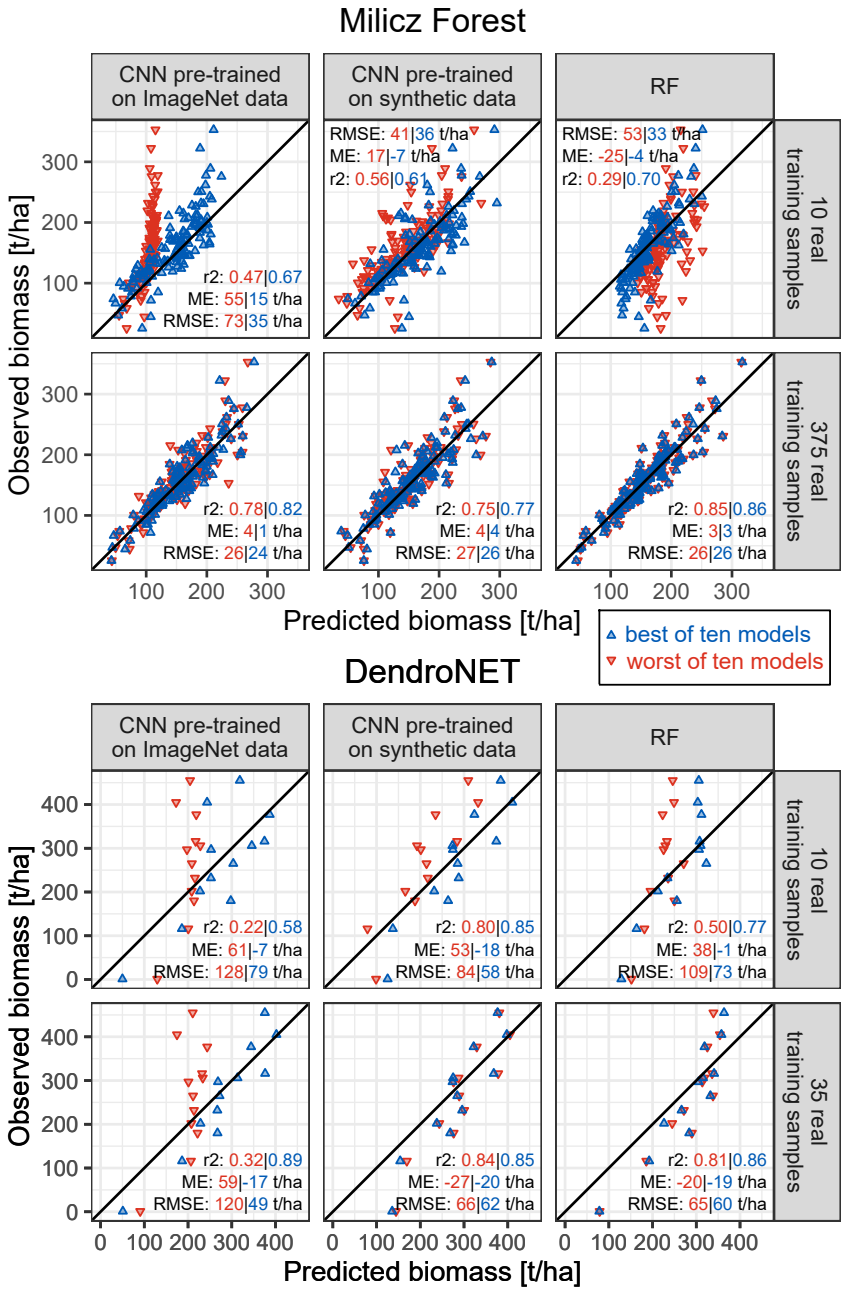

Figure 4.5 (cont.)**:** Predicted and observed AGB of the ground plots in the test datasets of each study site. Results are shown for models trained on 10 real training samples and on the maximum number of available training samples. Model training and prediction were repeated 10 times for each model type and training sample size, resulting in 10 predictions per ground observation. The training datasets consisting of 10 samples were randomly selected from the total training data in each run. Only the best and the worst predictions (in terms of RMSE) of the 10 model iterations are depicted.

performed substantially better or worse than RFs, we visually examined point clouds of some of the test plots. However, there were no obvious connections of performance difference and forest structure type, e.g. clearings, stand density, and subcanopy layers.

Regarding the use of pre-trained CNNs, our results showed that the use of simulated data for additional pre-training (on top of the ImageNet weights) did not improve model performance with the exception of cases of extremely limited training data availability. In all other cases, performance was actually decreased by the addition of the pre-training step with simulated data, suggesting that the CNN finds patterns in the synthetic data that do not exist in a similar way in the real data. This effect may be caused by either the synthetic forest stands or by the simulation of laser scanning. The simulated forest stands that were selected for the synthetic training datasets differ in their composition from the real forest stands at our study sites: They have a different species composition and they have on average larger AGB values but a smaller number of trees per hectare than the real stands (cf. Figure 4.1). In addition, they lack understorey elements, and the crowns of neighbouring trees may overlap unrealistically due to our simple approach of assigning tree positions. It should be tested whether synthetic datasets generated using alternative forest growth simulators, such as SILVA (Pretzsch et al., 2002), which incorporate competition between neighbouring trees at the individual tree level, output actual realistic tree positions, and allow the simulation of different forest management strategies, would improve the performance of the pre-trained CNNs. The differences between simulated and real pulse density and planar point density (cf. Table 4.1) indicate that our simulations could not exactly replicate the real laser scanning of the four study sites with our simulations, which can in part be explained by the missing understorey. In a previous study, we also observed that the height distribution of the simulated laser scanning returns can differ substantially from the real one, depending on the stand characteristics (Schäfer et al., 2023a). Other studies using HELIOS++ for virtual laser scanning have found that the quality of the generated point cloud is also subject to the representation of the 3D scene and can be improved either by precise fine-tuning of the voxelisation model (Weiser et al., 2021) or by the use of procedurally generated, highly-detailed mesh models of trees (Esmorís et al., 2024). In the latter study, a successful transfer of a deep-learning model trained on purely synthetic data to a real dataset was shown for the case of leaf-wood separation. We conclude that more effort is needed to fine-tune the scene model (e.g. use a different representation, such as high-detail mesh models of trees) and the parameters for the HELIOS++ simulations to make the resulting point clouds more realistic. To investigate whether the poor results for the pre-training on synthetic data are more affected by the forest stand simulations or the laser scanning simulations, two potential pathways exist: 1) the use of virtual laser scanning based on real forest inventory data, thereby excluding effects from the forest stand synthetization, and 2) the use of real laser scanning point clouds of trees stitched together based on the compositions given from the synthetic forest stands, excluding the laser scanning simulation. For the latter case, the flight- and sensor-parameters of the available tree point cloud database would have to match the ones of the real training and test data, which was not the case in our study.

To precisely evaluate model performance, accurate reference values (i.e. AGB) are necessary. AGB values in this study were calculated using allometric models, which have been shown

to contain significant uncertainty (Vorster et al., 2020). Additionally, tree crowns reaching out of or into the ground plots contribute to errors in the reference AGB (Knapp et al., 2021), as the plot AGB is estimated as the total AGB of all trees with a stem position within the plot. While these sources of uncertainty can only be removed by extensive and potentially destructive fieldwork with real data, the use of 3D mesh models for synthetic data allows the accurate quantification of wood volume and thus the derivation of AGB estimates that are not affected by allometric errors. Synthetic data also include information on tree parts exceeding or reaching into the plot, making it easy to precisely quantify the amount of AGB that is within the plot area. Accounting for these boundary effects in real data is much more difficult as it requires detailed tree information and is subject to uncertainties due to assumptions about tree shape and crown projection area that need to be made (Kleinn et al., 2020).

In a previous study on ALS-based AGB predictions that solely investigated RFs and their response to simulated training data using the same real-world ALS and ground data, we also observed that results differed substantially between the four study sites (Schäfer et al., 2023a). We note that these differences partially result from different AGB distributions, which are not represented equally well by the simulated data. Most substantially, in the current study, the mean AGB value of the simulated data of DN was much closer to the one of the real data than for the other study sites (cf. Figure 4.1). This was also the dataset for which pre-training of the CNNs on synthetic data was most successful.

While the metrics used in the RF models usually describe the vertical distribution of the returns, and the horizontal distribution is less frequently taken into account (Bouvier et al., 2015), the CNNs are able to consider both vertical and horizontal distributions in the convolutions. Limited data augmentation was carried out to achieve larger training sets but could be exploited more in the future, e.g. by mirroring images or by extracting cross sections in other directions. As this would not solve any issues related to the domain of AGB values present in the training data, the effect of such efforts may, however, be limited.

Although the predictive performance of CNNs and RFs was similar, there are several reasons to use RFs rather than CNNs, both in terms of data pre-processing and the modeling itself. Generating images of point cloud cross sections is much more complicated, takes more time and needs more disk space than extracting point cloud metrics. In addition, CNNs require a high performance GPU to satisfy the computational demands and still take much longer to train than RFs. A drawback of using CNNs is also the "black box" characteristic of the approach that makes it more difficult to interpret the results, e.g. to explain why the pre-trained CNNs sometimes predicted negative AGB values for one plot in PRF (cf. Figure 4.5).

Other studies comparing deep learning methods to traditional machine learning models for AGB predictions from ALS data often found that deep learning results in higher prediction performance. Ayrey and Hayes (2018) adapted several 2D CNNs (LeNet, AlexNet, GoogLenet, Inception-V3, and ResNet-50) to run on 3D voxel representations of the ALS point clouds and compared the model performance to RF and linear mixed models trained on point cloud metrics. In their study, all deep learning models except for AlexNet resulted in smaller RMSE but larger or similar systematic error compared to RF and linear mixed

models. Oehmcke et al. (2022) applied PointNet, the kernel point convolution (KPConv) approach, and the Minkowski CNN in comparison to linear regression and power regression models for ALS-based AGB predictions and found that their adaptations of the Minkowski CNN and KPConv clearly surpassed the linear regression and power regression, while PointNet performed worse. When comparing RF to an Octree CNN-HRNet and a Dynamic Graph CNN, Seely et al. (2023) showed that AGB predictions with RF had a slightly smaller R^2 and a slightly larger RMSE than the deep learning predictions. In contrast to our study, Ayrey and Hayes (2018) used 15 373 samples for model training and 1 000 samples for validation, Oehmcke et al. (2022) used 4 271 and 919 samples, and Seely et al. (2023) used 1 635 and 350, respectively. Compared to our sample sizes of up to 35–375 plots (including 20% validation samples), these datasets are much better suited for deep learning approaches. We expect that the performance of the AGB prediction on cross section images with CNNs could be improved when using more training samples. While we hypothesized that synthetic data could be used to extend training datasets when limited data are available, our results did not support that claim.

Finally, we see potential for future research in multiple directions:

- Improvement of a) the synthetic forest stand composition and the positions of the individual trees within a plot, b) the 3D models of individual trees, and c) the laser scanning simulation parameters.
- Systematic investigations on how the CNN performance is affected by forest structure as well as by ground plot and point cloud characteristics (e.g. stand density, tree species, plot size and shape, point density, and penetration into subcanopy layers).
- Experiments with hyperparameter tuning and different deep learning network architectures for which pre-trained ImageNet weights are available (e.g. ResNet, EfficientNet, DenseNet), as well as investigation of the effect of these pre-training efforts by running models on randomly initialized weights using the same model architecture for comparison.
- Additional data augmentation by rotation, mirroring, and random jittering of points, as shown in previous studies using CNNs for point cloud tasks (Briechle et al., 2021; H. Li et al., 2020; Oehmcke et al., 2022).

4.5 CONCLUSION

This study demonstrated that CNNs can predict AGB from cross section images and achieve similar accuracies as RFs trained on traditional point cloud metrics. When the maximum number of available training samples was used, the CNN performance slightly surpassed the performance of the RFs for three of the four study sites, indicating that the CNN performance could be further improved by increasing the training sample size. We investigated whether the need of deep learning models for large amounts of training data could be satisfied by data simulations but found that pre-training on synthetic data did only improve model performance when very little training data were available. Notably, pre-training on synthetic data even decreased model performance at larger training sample

sizes. Since the use of simulated data has been shown to provide a benefit in other applications, even in the domain of forestry remote sensing, there is reason to believe that a gap between real and simulated data needs to be closed before such transfer can be successful for our use-case. For the time being, RF remains a competitive alternative to data-hungry deep learning models.

4.6 STATEMENTS

CRediT author statement

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

Acknowledgements

We are grateful to Joanne White and the Canadian Forest Service for providing data of the Petawawa Research Forest, and to technicians from the Institute of Forest Ecosystem Research and the Global Change Research Institute for data collection at the DendroNetwork and Silesian Beskids sites. We thank Roman Solovyev for sharing the pre-trained 3D CNNs and Hans Henniger for providing the modified version of Forest Factory 2.0. The first author is extremely grateful for her research stay in the Photogrammetry research unit at TU Wien, funded by the Graduate School for Climate and Environment of KIT.

Data availability statement

The individual tree point clouds used for generating the synthetic scenes are published at PANGAEA, at https://doi.pangaea.de/10.1594/PANGAEA.942856. R code for the creation synthetic forest stands is available on GitHub, at https://github.com/JannikaSchaefer/Syssifoss.

4.7 APPENDIX

Table 4.2: Model accuracies for DendroNET sites (DN). N is the number of real training samples used for model training, RMSE is the median root mean squared error, ME is the median mean error, and r^2 is the median squared Pearson correlation coefficient. A positive mean error indicates an underprediction of biomass.

ModelType	N	RMSE	ME	r^2
CNN pre-trained on ImageNet data	10	93.72	-2.24	0.56
CNN pre-trained on ImageNet data	20	74.08	-2.86	0.71
CNN pre-trained on ImageNet data	30	68.31	-21.08	0.87
CNN pre-trained on ImageNet data	35	63.87	-22.58	0.88
CNN pre-trained on synthetic data	0	97.3	70.4	0.78
CNN pre-trained on synthetic data	10	65.01	-17.12	0.84
CNN pre-trained on synthetic data	20	66.54	-21.13	0.84
CNN pre-trained on synthetic data	30	64.65	-27.11	0.85
CNN pre-trained on synthetic data	35	64.37	-21.65	0.85
RF trained on real data	10	88.4	-11.32	0.66
RF trained on real data	20	74.05	-28.69	0.77
RF trained on real data	30	67.96	-20.27	0.8
RF trained on real data	35	63.64	-20.64	0.83
RF trained on syntetic data	0	128.92	99.71	0.58

Table 4.3: Model accuracies for Petawawa Research Forest (PRF). N is the number of real training samples used for model training, RMSE is the median root mean squared error, ME is the median mean error, and r^2 is the median squared Pearson correlation coefficient. A positive mean error indicates an underprediction of biomass.

ModelType	N	RMSE	ME	r^2
CNN pre-trained on ImageNet data	10	82.82	25.97	0.39
CNN pre-trained on ImageNet data	20	54.41	6.76	0.73
CNN pre-trained on ImageNet data	30	56.53	9.64	0.68
CNN pre-trained on ImageNet data	40	45.04	0.30	0.78
CNN pre-trained on ImageNet data	50	41.06	-1.84	0.80
CNN pre-trained on ImageNet data	60	38.94	0.75	0.83
CNN pre-trained on ImageNet data	70	38.76	2.34	0.83
CNN pre-trained on ImageNet data	80	39.46	5.52	0.82
CNN pre-trained on ImageNet data	90	37.00	1.46	0.84
CNN pre-trained on ImageNet data	100	39.30	0.52	0.82
CNN pre-trained on ImageNet data	167	34.62	8.90	0.87
CNN pre-trained on synthetic data	0	108.00	89.53	0.56
CNN pre-trained on synthetic data	10	60.65	25.35	0.60
CNN pre-trained on synthetic data	20	54.80	4.75	0.65
CNN pre-trained on synthetic data	30	55.30	5.06	0.64
CNN pre-trained on synthetic data	40	53.05	4.38	0.67
CNN pre-trained on synthetic data	50	54.34	1.80	0.67
CNN pre-trained on synthetic data	60	51.02	2.85	0.71
CNN pre-trained on synthetic data	70	49.65	6.74	0.72
CNN pre-trained on synthetic data	80	50.78	8.18	0.71
CNN pre-trained on synthetic data	90	49.30	8.00	0.72
CNN pre-trained on synthetic data	100	48.25	5.95	0.73
CNN pre-trained on synthetic data	167	45.90	11.15	0.77
RF trained on real data	10	62.50	6.74	0.71
RF trained on real data	20	50.20	9.54	0.76
RF trained on real data	30	42.84	6.19	0.81
RF trained on real data	40	39.50	5.80	0.83
RF trained on real data	50	37.65	6.76	0.85
RF trained on real data	60	36.17	5.79	0.85
RF trained on real data	70	35.86	5.60	0.86
RF trained on real data	80	35.28	6.39	0.86
RF trained on real data	90	35.74	5.01	0.85
RF trained on real data	100	35.59	6.47	0.86
RF trained on real data	167	36.61	7.13	0.85
RF trained on syntetic data	0	77.70	59.60	0.71

Table 4.4: Model accuracies for Milicz Forest (MF). N is the number of real training samples used for model training, RMSE is the median root mean squared error, ME is the median mean error, and r^2 is the median squared Pearson correlation coefficient. A positive mean error indicates an underprediction of biomass.

ModelType	N	RMSE	ME	r^2
CNN pre-trained on ImageNet data	10	44.42	13.22	0.59
CNN pre-trained on ImageNet data	20	36.21	8.28	0.65
CNN pre-trained on ImageNet data	30	32.42	6.43	0.68
CNN pre-trained on ImageNet data	40	30.44	4.21	0.71
CNN pre-trained on ImageNet data	50	29.44	2.20	0.74
CNN pre-trained on ImageNet data	60	27.74	2.52	0.75
CNN pre-trained on ImageNet data	70	26.87	1.99	0.77
CNN pre-trained on ImageNet data	80	25.94	-0.48	0.78
CNN pre-trained on ImageNet data	90	26.19	-0.02	0.78
CNN pre-trained on ImageNet data	100	26.61	0.42	0.78
CNN pre-trained on ImageNet data	375	24.48	2.63	0.81
CNN pre-trained on synthetic data	0	38.10	-10.41	0.57
CNN pre-trained on synthetic data	10	37.04	-2.76	0.58
CNN pre-trained on synthetic data	20	33.82	-0.37	0.65
CNN pre-trained on synthetic data	30	33.35	-0.29	0.64
CNN pre-trained on synthetic data	40	32.67	2.23	0.65
CNN pre-trained on synthetic data	50	31.59	1.74	0.68
CNN pre-trained on synthetic data	60	31.40	1.83	0.69
CNN pre-trained on synthetic data	70	29.91	-1.32	0.71
CNN pre-trained on synthetic data	80	29.15	-1.60	0.73
CNN pre-trained on synthetic data	90	29.31	0.82	0.72
CNN pre-trained on synthetic data	100	29.31	0.61	0.73
CNN pre-trained on synthetic data	375	26.91	3.67	0.77
RF trained on real data	10	41.59	3.33	0.63
RF trained on real data	20	33.54	3.35	0.69
RF trained on real data	30	28.79	4.21	0.75
RF trained on real data	40	26.96	2.95	0.77
RF trained on real data	50	25.94	3.22	0.79
RF trained on real data	60	24.75	2.43	0.81
RF trained on real data	70	24.05	2.03	0.82
RF trained on real data	80	23.71	2.07	0.82
RF trained on real data	90	23.45	2.50	0.83
RF trained on real data	100	22.40	1.61	0.83
RF trained on real data	375	19.57	1.05	0.87
RF trained on syntetic data	0	35.57	20.72	0.74

Table 4.5: Model accuracies for Silesian Beskids (SB). N is the number of real training samples used for model training, RMSE is the median root mean squared error, ME is the median mean error, and r^2 is the median squared Pearson correlation coefficient. A positive mean error indicates an underprediction of biomass.

ModelType	N	RMSE	ME	r^2
CNN pre-trained on ImageNet data	10	124.36	50.13	0.58
CNN pre-trained on ImageNet data	20	84.80	11.55	0.71
CNN pre-trained on ImageNet data	30	77.05	9.30	0.74
CNN pre-trained on ImageNet data	40	74.68	15.13	0.77
CNN pre-trained on ImageNet data	50	73.50	16.06	0.77
CNN pre-trained on ImageNet data	60	70.69	12.97	0.78
CNN pre-trained on ImageNet data	70	68.58	17.70	0.79
CNN pre-trained on ImageNet data	80	71.93	18.57	0.76
CNN pre-trained on ImageNet data	90	70.12	14.52	0.78
CNN pre-trained on ImageNet data	97	70.16	18.29	0.78
CNN pre-trained on synthetic data	0	101.48	74.49	0.77
CNN pre-trained on synthetic data	10	74.34	26.40	0.77
CNN pre-trained on synthetic data	20	76.89	23.39	0.76
CNN pre-trained on synthetic data	30	79.44	31.60	0.77
CNN pre-trained on synthetic data	40	76.38	30.41	0.76
CNN pre-trained on synthetic data	50	77.17	28.84	0.76
CNN pre-trained on synthetic data	60	76.76	29.97	0.77
CNN pre-trained on synthetic data	70	77.01	27.41	0.76
CNN pre-trained on synthetic data	80	76.95	28.08	0.76
CNN pre-trained on synthetic data	90	76.89	27.44	0.76
CNN pre-trained on synthetic data	97	76.13	25.76	0.76
RF trained on real data	10	91.81	3.70	0.70
RF trained on real data	20	77.67	3.30	0.75
RF trained on real data	30	74.89	9.49	0.75
RF trained on real data	40	71.25	13.49	0.76
RF trained on real data	50	72.06	12.87	0.77
RF trained on real data	60	71.00	12.52	0.77
RF trained on real data	70	72.52	11.30	0.76
RF trained on real data	80	71.60	10.57	0.77
RF trained on real data	90	72.31	12.58	0.76
RF trained on real data	97	71.34	12.34	0.77
RF trained on syntetic data	0	89.86	60.07	0.79

Chapter 5

5 SYNTHESIS

The backbone of this thesis is a novel approach to produce realistic synthetic datasets of ALS point clouds and underlying forest stand information. The three studies incorporated in this thesis have demonstrated that the data generated by this approach can be utilised to train models for predicting AGB from real ALS point clouds. However, they have also revealed that the prediction accuracy of models trained on synthetic data is typically lower than of models trained on real data, and the prediction accuracy depends on the study site where the model is applied. Consequently, there are general but also site-dependent disparities between synthetic and real data that impact the quality of the models. When assessing an approach for generating realistic synthetic data, three key questions should be considered:

1. What defines data as realistic?
2. How can the realism of data be assessed?
3. What degree of realism is necessary?

To answer the first question, one can state that synthetic ALS data of forests are considered realistic if they could have hypothetically been derived from a real acquisition of a real forest. Answering the second question is much more complex and is related to the third one. A measure of realism could be that synthetic data cannot at all be distinguished from real data. Synthetic and real data could be presented to a human expert, a traditional machine learning model or a deep learning model, and they could be asked to classify the data as "synthetic" or "real". Considering the vast amount of information contained within ALS datasets, I anticipate that both humans and computer models would struggle to complete this task without narrowing it down to specific features upon which the classification should be based. These features could be related to the ALS point clouds, e.g. return height metrics, as well as to the forest stand characteristics, e.g. the number of trees, tree locations, and the distribution of tree sizes. While this might be a simple solution to determine whether generated data are realistic, the information content of such a validation method is questionable. Even if an expert or algorithm fails to distinguish between synthetic and real data, one could argue that the expert or algorithm is merely not good enough and discrimination would be possible with further training. A related approach would be to test whether the feature space of the synthetic data falls within the feature space of real data. However, a significant drawback of this method is that the validation results rely heavily on the available real data for comparison. For example, if the stand density in the real dataset ranges from 100 to 1 500 trees/ha, a synthetic stand with a stand density of 2 000 trees/ha would be classified as unrealistic, even if there are many forest stands with that stand density in the real world—they are just not included in the sample provided by the real data. Consequently, a dataset covering all possible real-world conditions would be necessary for a comprehensive validation following this approach. However, if such data were available, it would be more practical to generate forest stands using the inherent information on forest composition rather than using a forest simulator.

To evaluate the laser scanning simulations themselves, they could be applied to real forest stands for which real laser scanning data are available. The quality of the simulations could be assessed based on the difference between simulated and real point clouds or metrics derived from the point clouds. This approach is particularly useful for fine-tuning the

simulation parameters; for example, it has been applied by Weiser et al. (2022a) to determine the optimum voxel resolution for HELIOS++ ALS simulations of tree groups represented by TLS point clouds. Nevertheless, when using the difference between synthetic and real point clouds or metrics as a measure of realism, it is necessary to define a threshold that indicates when the difference or deviation is too large, suggesting that the simulated data differ too much from the real data to be considered realistic. This leads to the third question: to what degree do synthetic data have to be realistic?

A fixed threshold to distinguish between what is realistic and what is not realistic will always be somewhat arbitrary. For example, if the mean height of simulated returns differs from the mean height of real returns by just 2 cm, we would likely consider the simulated data to be realistic. However, if the difference in mean point height is 2 m, the answer to whether the simulated data are realistic becomes more complex. This difference could be due to a systematic height offset, meaning that the height values of the simulated points are consistently 2 m higher than the real points. In this case, the absolute height values would not be considered realistic, but the height distribution of the simulated returns would be identical to that of the real data, and would therefore be considered realistic. Accordingly, whether synthetic data can be considered realistic or not depends on their potential use case. Synthetic data are always generated with a specific purpose. Therefore, instead of asking whether synthetic data are realistic or not, it is more appropriate to ask whether they are realistic enough for a potential application.

In this thesis, the presented approach for creating synthetic ALS data was evaluated based on its ability to generate data that are realistic enough for the training of AGB regression models. To be more specific, the accuracy of models trained on synthetic data was assessed and compared to models trained on real data. However, no final judgement was made regarding whether the achieved level of accuracy is sufficient to consider the data realistic enough, i.e. suitable for this purpose.

The studies included in this thesis have shown that the generated synthetic data can be used to train models for predicting AGB from real ALS point clouds. They have revealed that the accuracy of the models depends on the dataset, i.e. the study site to which the models are applied. In general, models trained on synthetic data did not achieve as high accuracies as models trained on real-world data. This leads to the conclusion that the data are not yet realistic enough to adequately replace real-world data for training AGB models. Therefore, there is still room for improvement in the approach for generating synthetic data. In accordance with the results of the presented studies, all three components of the approach, i.e. the laser scanning simulations, the tree models, and the forest simulations, affect the realism of the generated data. These components will be discussed separately in the following sections.

5.1 LASER SCANNING SIMULATIONS

Realistic laser scanning simulations are necessary for replicating actual laser scanning acquisitions and for conducting sensitivity analyses of acquisition settings. To assess the accuracy of laser scanning simulations independent of tree models and forest simulations, the laser scanning simulations must be performed on real forest scenes. In the first study (Chapter 2), a real forest laser scanning point cloud was used as input for the HELIOS++ simulations to determine the optimal parameters for the temporal window size for local maxima detection and the voxel resolution. The simulated and real point clouds were compared based on their height distributions. The analysis revealed that the optimal values varied depending on whether all returns or only first returns were considered, as well as whether the absolute or relative height distribution was taken into account. Although this examination was limited to a small forest patch covering an area of 22 m $\times$ 24 m and focused solely on the point height distribution, it demonstrated that HELIOS++ was not able to perfectly reproduce the point cloud characteristics of real ALS data.

A systematic investigation of the influence of HELIOS++ parameters and forest scene characteristics (i.e. the stand structure and the density of the tree point clouds) on the resulting simulated point clouds could offer insights into how the simulations could be improved. Such an analysis should be focused on a voxel level, specifically comparing the number of returns in small-scale voxels, rather than relying on point cloud metrics or the point height distribution. I anticipate that in order to enhance the realism of the simulations, it will be necessary not only to fine-tune the simulation parameters but also to improve the algorithms implemented in HELIOS++. Additionally, to expand the potential use cases of the synthetic data, HELIOS++ should be further developed to simulate realistic intensity values, which can provide valuable information for tree species identification, for example.

One factor that was not explored in the presented studies but could contribute to the simulated data being less realistic is the absence of shadowing effects at the edges of the scene. The size of forest scenes that can be used as input for the HELIOS++ simulations is limited by computer memory. At forest edges, the laser beams are not shadowed by neighbouring trees, resulting in more returns from the subcanopy layers. It would be worthwhile to investigate whether a wall composed of both opaque and penetrable voxels could be implemented around the forest scenes to emulate shadowing effects.

When the synthetic data are generated for a specific purpose, such as training AGB models, it would be sufficient to optimize the laser scanning simulation accordingly. A feature selection process could be utilized to identify the most relevant features for the models. This would allow for a focus on specific metrics when searching for the optimal HELIOS++ parameters, while non-relevant metrics could be excluded when comparing synthetic and real data.

5.2 TREE MODELS

The comparison of real tree point clouds and simplified tree point clouds in the first study (Chapter 2) has shown that the type of tree model affects the characteristics of the simulated point clouds. Real laser scanning point clouds of individual trees are inherently realistic and capture the inter- and intraspecific variability of real trees. However, using real tree point clouds as tree models also presents certain drawbacks. Firstly, the successful application of this approach depends on the availability of tree point clouds for all species of interest and for diversity of tree sizes, tree shapes, and social status (dominant or non-dominant). The point clouds should have sufficient point density and be segmented with high quality to accurately represent the tree habit. Information on tree species should also be provided with the point clouds. However, some other metrics, such as tree height and crown diameter, can be derived directly from the point clouds when they are not available from field measurements. With advancements in automatic segmentation methods for individual tree point clouds and the increasing use of UAVs for laser scanning data collection (Puliti et al., 2020), as well as a general trend to open data policies driven by the open science movement, tree model availability is expected to improve in the future.

The forest scenes composed of individual tree point clouds need to be voxelised for the laser scanning simulations because the points themselves do not have a surface the laser beams can interact with. The voxel size plays a crucial rule in the resulting simulated laser scanning point clouds. Larger voxel sizes lead to unrealistic, "blocky" tree shapes and an underestimation of returns from the subcanopy layers. This is due to the reduced probability of returns penetrating the canopy as the voxel size increases. On the other hand, smaller voxel sizes create unrealistic gaps, especially at the the tree stems, and can result in an excessive number of laser beams reaching the ground without generating any returns from the trees. When using laser scanning point clouds as input for laser scanning, a double sampling is applied. This means that in the second scanning, returns can only be generated from surfaces that already generated returns in the first scanning. One potential solution to this issue would be to utilise tree point clouds with very high point density and a small voxel size. However, the memory footprint of laser scanning simulations typically increases as the voxel size decreases. Therefore, this approach is constrained by both memory limitations and the availability of high-density tree point clouds. High-resolution point clouds could be derived by fusing TLS and ULS data, ensuring that both the canopy and non-canopy parts of the tree are represented well in the point cloud (Fekry et al., 2022). If high-density tree point clouds are available, a voxel scaling approach as suggested by Weiser et al. (2021) could be employed to reduce memory requirements. Alternatively, quantitative structure models (QSMs) could be generated from the high-density tree point clouds (Fekry et al., 2022). Using QSMs as tree models has the advantage of allowing accurate estimation of AGB directly from the QSMs independent of allometric equations (Disney et al., 2018), making the AGB estimates of the synthetic stands more reliable.

In the first study (Chapter 2), tree models were selected based on species, height, and crown diameter of the trees in the real forest stands. In the other two studies (Chapter 3 and Chapter 4), Forest Factory 2.0 was used to generate forest stand information. Therefore, the

selection of tree point clouds was based only on species and height, as the crown diameter of Forest Factory trees is fully determined by species and height. However, in reality, the crown diameter is not solely determined by species and height but is also influenced by other factors, such as the stand structure (Bragg, 2001; Holdaway, 1986; Krajicek et al., 1961). Isolated or dominant trees tend to have a higher ratio of crown diameter to tree height compared to non-dominant trees or trees in stands of high density. Additional information such as crown base height or social status of a tree could also improve the realism of the synthetic stands. The crown base height, in particular, affects the number of returns from the subcanopy layers. The first study (Chapter 2) has shown that the simulated return height distribution of some forest plots was more similar to the real one when using simplified tree point clouds instead of real tree point clouds. This can be explained by the fact that the simplified tree point clouds were created with a crown base height that matched the crown base height measured in the field. Therefore, I recommend using both crown diameter and crown base height as additional filter criteria when selecting the tree point clouds. However, this requires a forest simulation approach that computes realistic values for these metrics, considering the stand structure, which is not the case for Forest Factory. Additionally, reliable values for these metrics must be available for the tree point clouds. In the dataset used in this thesis, crown base height was defined as the height of the lowest branch with a minimum length of 1 m (Weiser et al., 2022a). This definition was chosen for practical reasons as it was easy to apply in field measurements and automated extraction from the tree point clouds. However, in reality, there are many trees with branches longer 1 m that are not part of the crown. Therefore, a more appropriate definition of crown base height should be established before implementing this variable in the selection process.

In summary, the quality and availability of tree models and tree information have a significant impact on the synthetic laser scanning data. A larger number of tree point clouds to select from, along with supplementary information on crown diameter, crown base height, and social status that can be used to find the best matching tree point cloud, and a high point density that allows for accurate reconstruction of the tree shape from the point cloud, can contribute to making the synthetic data more realistic.

5.3 FOREST STANDS

Evaluating the synthetic forest stands is a more complex task than evaluating the quality of the laser scanning simulations. Comparing the stand characteristics of synthetic and real data can provide a general idea of the realism of the synthetic stands. However, this approach offers limited insight into the realism of individual stands, particularly in terms of tree positions and attributes. The second and third study (Chapter 3 and Chapter 4) revealed notable disparities in stand density and biomass distribution between forest plots simulated with Forest Factory 2.0 and the real datasets used for comparison. Training AGB models using synthetic data based on Forest Factory plots resulted in an underprediction of AGB for the test datasets of Milicz Forest (Poland), Silesian Beskids (Czech Republic), DendroNET sites (Czech Republic), and Petawawa Research Forest

(Canada). Conversely, when AGB models were trained on synthetic data based on real forest inventory data, an overprediction of AGB was observed when testing the models with data collected in southwestern Germany (first study, Chapter 2). It is important to interpret the findings of this comparison cautiously, as the models were not tested on the same datasets. Nevertheless, it appears that using Forest Factory stands for generating training data introduces a significant bias to the AGB models, which surpasses biases resulting from laser scanning simulations and tree models. Regarding the stand characteristics, it can be hypothesised that Forest Factory simulates an unrealistically high number of forest plots with high biomass but low stand density, at least when compared to the test datasets from Poland, the Czech Republic, and Canada. This is a significant drawback when aiming to substitute real data with synthetic data for training AGB models.

The decision to use Forest Factory instead of SILVA for creating the synthetic forest stands was based on several factors. SILVA simulations are more time-consuming, necessitate configuration through a graphical user interface, and require the definition of an initial stand composition and management strategy. Consequently, generating a large number of different forest stands becomes more labor-intensive and demands more knowledge of forest management as opposed to when employing Forest Factory. However, if information regarding the typical initial stand composition and management strategy is available for a study site, SILVA provides the advantage of simulating forest stands accordingly. Future studies should investigate whether SILVA produces synthetic data that are better suited for training models to be applied to real data than Forest Factory.

Unrealistic features of the forest stands could also be observed at the individual plot level. The synthetic stands lack understory and have a lower canopy cover compared to real forest stands (first study, Chapter 2). Furthermore, in the Forest Factory stands, trees can be located unrealistically close to each other. This issue is not present in SILVA stands, as SILVA simulates explicit tree positions, taking into account three-dimensional competition among neighbouring trees. However, due to the fact that real trees do not necessarily have an axially symmetrical shape, the random rotation of tree point clouds around the z-axis can cause overlapping placement of tree point clouds. Deep learning techniques could potentially address this issue by improving the realism of the tree point cloud placement. An algorithm could be trained using real forest point clouds to learn how to position individual tree point clouds in a way that minimises overlap while maximising canopy cover.

All three studies revealed that the performance of AGB models trained on synthetic data strongly depends on the study site to which the models are applied. However, the outcomes were not consistent across the different experiments (training models solely on synthetic data, training models on mixed datasets comprised of synthetic and real data, and pre-training models on synthetic data). For instance, in case of the DendroNET sites, pre-training CNNs on synthetic data resulted in high prediction accuracies (third study, Chapter 4), whereas training random forest models on synthetic data resulted in low prediction accuracies (second study, Chapter 3). More datasets are required to systematically analyse the relationship between study site characteristics and the suitability of synthetic data for

model training. This would provide valuable insights on how to improve the synthetic data generation in order to make the data applicable to a wider range of study sites.

5.4 CONCLUSIONS AND OUTLOOK

The answer to the question of how realistic synthetic data must be depends on the specific application and turns into a question of *fitness for use*. In the case of area-based approaches, the realism of individual tree shapes and positions is less critical than the realism of forest stand compositions. For sensitivity analyses, it is probably less important to use synthetic data with realistic forest stand compositions compared to when the data are used to train models for application on real data. Realistic laser scanning simulations are necessary for sensitivity analyses aiming to identify optimal laser scanning acquisition settings. When synthetic data are utilised for individual tree-based approaches, such as the development of tree segmentation algorithms, ensuring a high level of realism in tree models and tree placement is likely to be the most crucial factor for successful application on real-world data.

My research has shown that synthetic data enable AGB predictions from ALS data without the need for field reference data collection. However, models trained on synthetic data do not reach the accuracy of models trained on real data. It appears that the quality of the synthetic data as they are is not sufficient to substitute real data without a decrease in model accuracy. The three studies have shown that when employing an area-based approach, training AGB models on real data results in higher prediction accuracies than training on synthetic data, often even if the real data were collected from different sites. Therefore, instead of attempting to further improve the simulations, more effort could be dedicated to collecting and sharing real data. A database could be established where real ALS data and corresponding forest stand information can be uploaded and made publicly accessible. Exemplary cases of data sharing are the openly available dataset of the Petawawa Research Forest (White et al., 2019), which includes ALS and forest inventory data from multiple years, and the pytreedb (https://pytreedb.geog.uni-heidelberg.de/), a database of single-tree laser scanning and forest inventory data.

Given that forest inventory data often do not include information on tree positions, synthetic data can still be highly valuable for approaches based on individual trees. For instance, the data could be used to train a model that, in the first step, detects individual trees in the ALS point clouds, identifies the tree species and estimates the tree size. In the second step, the AGB of each tree could be predicted and then summed up to derive plot-level predictions. Additionally, the presented approach for generating synthetic data is well-suited for sensitivity analyses. In contrast to real-world acquisitions, it allows for extensive testing of different laser scanning settings and field sampling designs. The insights derived from such analyses are particularly useful for planning the operational use of ALS in national forest inventories. For instance, they can help identify the most efficient acquisition settings or establish acceptable precision tolerances for location measurements.

Further research is required to explore the potential of the presented approach in relation to these areas of application.

References

Abdalati, W., Zwally, H. J., Bindschadler, R., Csatho, B., Farrell, S. L., Fricker, H. A., Harding, D., Kwok, R., Lefsky, M., Markus, T., Marshak, A., Neumann, T., Palm, S., Schutz, B., Smith, B., Spinhirne, J., & Webb, C. (2010). The ICESat-2 Laser Altimetry Mission. *Proceedings of the IEEE*, *98*(5), 735–751. https://doi.org/10.1109/JPROC.2009.2034765

Abshire, J. B., Sun, X., Riris, H., Sirota, J. M., McGarry, J. F., Palm, S., Yi, D., & Liiva, P. (2005). Geoscience Laser Altimeter System (GLAS) on the ICESat Mission: On-orbit Measurement Performance. *Geophysical Research Letters*, *32*(21). https://doi.org/10.1029/2005GL024028

Achim, A., Moreau, G., Coops, N. C., Axelson, J. N., Barrette, J., Bédard, S., Byrne, K. E., Caspersen, J., Dick, A. R., D'Orangeville, L., Drolet, G., Eskelson, B. N. I., Filipescu, C. N., Flamand-Hubert, M., Goodbody, T. R. H., Griess, V. C., Hagerman, S. M., Keys, K., Lafleur, B., … White, J. C. (2022). The Changing Culture of Silviculture. *Forestry: An International Journal of Forest Research*, *95*(2), 143–152. https://doi.org/10.1093/forestry/cpab047

Ahmed, R., Siqueira, P., & Hensley, S. (2013). A Study of Forest Biomass Estimates from Lidar in the Northern Temperate Forests of New England. *Remote Sensing of Environment*, *130*, 121–135. https://doi.org/10.1016/j.rse.2012.11.015

Alkama, R., & Cescatti, A. (2016). Biophysical Climate Impacts of Recent Changes in Global Forest Cover. *Science*, *351*(6273), 600–604. https://doi.org/10.1126/science.aac8083

Ameztegui, A., Rodrigues, M., & Granda, V. (2022). Uncertainty of Biomass Stocks in Spanish Forests: A Comprehensive Comparison of Allometric Equations. *European Journal of Forest Research*, *141*(3), 395–407. https://doi.org/10.1007/s10342-022-01444-w

Andersson, K., Evans, T. P., & Richards, K. R. (2009). National Forest Carbon Inventories: Policy Needs and Assessment Capacity. *Climatic Change*, *93*(1), 69–101. https://doi.org/10.1007/s10584-008-9526-6

Arumäe, T., & Lang, M. (2018). Estimation of Canopy Cover in Dense Mixed-Species Forests Using Airborne Lidar Data. *European Journal of Remote Sensing*, *51*(1), 132–141. https://doi.org/10.1080/22797254.2017.1411169

Assman, E. (1961). Waldertragskunde. BLV, München. English Translation by SH Gardiner (1970): The Principles of Forest Yield Study.

Ayrey, E., Fraver, S., Kershaw, J. A., Kenefic, L. S., Hayes, D., Weiskittel, A. R., & Roth, B. E. (2017). Layer Stacking: A Novel Algorithm for Individual Forest Tree Segmentation from LiDAR Point Clouds. *Canadian Journal of Remote Sensing*, *43*(1), 16–27. https://doi.org/10.1080/07038992.2017.1252907

Ayrey, E., & Hayes, D. J. (2018). The Use of Three-Dimensional Convolutional Neural Networks to Interpret LiDAR for Forest Inventory. *Remote Sensing*, *10*(4), 649. https://doi.org/10.3390/rs10040649

Ayrey, E., Hayes, D. J., Kilbride, J. B., Fraver, S., Kershaw, J. A., Cook, B. D., & Weiskittel, A. R. (2019). Synthesizing Disparate LiDAR and Satellite Datasets through Deep Learning to Generate Wall-to-Wall Regional Forest Inventories. https://doi.org/10.1101/580514

Bae, S., Levick, S. R., Heidrich, L., Magdon, P., Leutner, B. F., Wöllauer, S., Serebryanyk, A., Nauss, T., Krzystek, P., Gossner, M. M., Schall, P., Heibl, C., Bässler, C., Doerfler, I., Schulze, E.-D., Krah, F.-S., Culmsee, H., Jung, K., Heurich, M., … Müller, J. (2019). Radar Vision in the Mapping of Forest Biodiversity from Space. *Nature Communications*, *10*(1), 4757. https://doi.org/10.1038/s41467-019-12737-x

Balazs, A., Liski, E., Tuominen, S., & Kangas, A. (2022). Comparison of Neural Networks and K-Nearest Neighbors Methods in Forest Stand Variable Estimation Using Airborne Laser Data. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, *4*, 100012. https://doi.org/10.1016/j.ophoto.2022.100012

Barrett, F., McRoberts, R. E., Tomppo, E., Cienciala, E., & Waser, L. T. (2016). A Questionnaire-Based Review of the Operational Use of Remotely Sensed Data by National Forest Inventories. *Remote Sensing of Environment*, *174*, 279–289. https://doi.org/10.1016/j.rse.2015.08.029

Bartemucci, P., Messier, C., & Canham, C. D. (2006). Overstory Influences on Light Attenuation Patterns and Understory Plant Community Diversity and Composition in Southern Boreal Forests of Quebec. *Canadian Journal of Forest Research*, *36*(9), 2065–2079. https://doi.org/10.1139/x06-088

Bauer, L., Knapp, N., & Fischer, R. (2021). Mapping Amazon Forest Productivity by Fusing GEDI Lidar Waveforms with an Individual-Based Forest Model. *Remote Sensing*, *13*(22), 4540. https://doi.org/10.3390/rs13224540

Bauwens, S., Bartholomeus, H., Calders, K., & Lejeune, P. (2016). Forest Inventory with Terrestrial LiDAR: A Comparison of Static and Hand-Held Mobile Laser Scanning. *Forests*, *7*(6), 127. https://doi.org/10.3390/f7060127

Bechtold, S., & Höfle, B. (2016). HELIOS: A multi-purpose lidar simulation framework for research, planning and training of laser scanning operations with airborne, ground-based mobile and stationary platforms. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *III-3*, 161–168. https://doi.org/10.5194/isprs-annals-III-3-161-2016

Bechtold, W. A. (2003). Crown-Diameter Prediction Models for 87 Species of Stand-Grown Trees in the Eastern United States. *Southern Journal of Applied Forestry*, *27*(4), 269–278. https://doi.org/10.1093/sjaf/27.4.269

Bitterlich, W. (1952). Die Winkelzählprobe: Ein optisches Meßverfahren zur raschen Aufnahme besonders gearteter Probeflächen für die Bestimmung der Kreisflächen pro Hektar an stehenden Waldbeständen. *Forstwissenschaftliches Centralblatt*, *71*(7), 215–225.

Bivand, R., Keitt, T., & Rowlingson, B. (2022). *rgdal: Bindings for the 'Geospatial' Data Abstraction Library* [R package version 1.5-29]. https://CRAN.R-project.org/package=rgdal

Blair, J., & Hofton, M. (1999). Modeling Laser Altimeter Return Waveforms over Complex Vegetation Using High-Resolution Elevation Data. *Geophysical Research Letters*, *26*(16), 2509–2512. https://doi.org/10.1029/1999GL010484

Bohn, F. J., Frank, K., & Huth, A. (2014). Of Climate and Its Resulting Tree Growth: Simulating the Productivity of Temperate Forests. *Ecological Modelling*, *278*, 9–17. https://doi.org/10.1016/j.ecolmodel.2014.01.021

Bohn, F. J., & Huth, A. (2017). The Importance of Forest Structure to Biodiversity–Productivity Relationships. *Royal Society Open Science*, *4*(1), 160521. https://doi.org/10.1098/rsos.160521

Bologna, M., & Aquino, G. (2020). Deforestation and World Population Sustainability: A Quantitative Analysis. *Scientific Reports*, *10*(1), 7631. https://doi.org/10.1038/s41598-020-63657-6

Bouvier, M., Durrieu, S., Fournier, R. A., & Renaud, J.-P. (2015). Generalizing Predictive Models of Forest Inventory Attributes Using an Area-Based Approach with Airborne LiDAR Data. *Remote Sensing of Environment*, *156*, 322–334. https://doi.org/10.1016/j.rse.2014.10.004

Bragg, D. C. (2001). A Local Basal Area Adjustment for Crown Width Prediction. *Northern Journal of Applied Forestry*, *18*(1), 22–28. https://doi.org/10.1093/njaf/18.1.22

Brandtberg, T. (2007). Classifying Individual Tree Species under Leaf-off and Leaf-on Conditions Using Airborne Lidar. *ISPRS Journal of Photogrammetry and Remote Sensing*, *61*(5), 325–340. https://doi.org/10.1016/j.isprsjprs.2006.10.006

Breidenbach, J., Kublin, E., McGaughey, R., Andersen, H.-E., & Reutebuch, S. (2008). Mixed-Effects Models for Estimating Stand Volume by Means of Small Footprint Airborne Laser Scanner Data. *Photogrammetric Journal of Finland*, *21*(1), 4–15.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Briechle, S., Krzystek, P., & Vosselman, G. (2021). Silvi-Net – A Dual-CNN Approach for Combined Classification of Tree Species and Standing Dead Trees from

Remote Sensing Data. *International Journal of Applied Earth Observation and Geoinformation*, *98*, 102292. https://doi.org/10.1016/j.jag.2020.102292

Brovkina, O., Navrátilová, B., Novotný, J., Albert, J., Slezák, L., & Cienciala, E. (2022). Influences of Vegetation, Model, and Data Parameters on Forest Aboveground Biomass Assessment Using an Area-Based Approach. *Ecological Informatics*, *70*, 101754. https://doi.org/10.1016/j.ecoinf.2022.101754

Brown, S. (2002). Measuring Carbon in Forests: Current Status and Future Challenges. *Environmental Pollution*, *116*(3), 363–372. https://doi.org/10.1016/S0269-7491(01)00212-3

Bruening, J. M., Fischer, R., Bohn, F. J., Armston, J., Armstrong, A. H., Knapp, N., Tang, H., Huth, A., & Dubayah, R. (2021). Challenges to Aboveground Biomass Prediction from Waveform Lidar. *Environmental Research Letters*, *16*(12), 125013. https://doi.org/10.1088/1748-9326/ac3cec

Bruggisser, M., Wang, Z., Ginzler, C., Webster, C., & Waser, L. T. (2024). Characterization of Forest Edge Structure from Airborne Laser Scanning Data. *Ecological Indicators*, *159*, 111624. https://doi.org/10.1016/j.ecolind.2024.111624

Burt, A., Disney, M., & Calders, K. (2019). Extracting Individual Trees from Lidar Point Clouds Using Treeseg. *Methods in Ecology and Evolution*, *10*(3), 438–445. https://doi.org/10.1111/2041-210X.13121

Cao, L., Coops, N. C., Innes, J. L., Sheppard, S. R. J., Fu, L., Ruan, H., & She, G. (2016). Estimation of Forest Biomass Dynamics in Subtropical Forests Using Multi-Temporal Airborne LiDAR Data. *Remote Sensing of Environment*, *178*, 158–171. https://doi.org/10.1016/j.rse.2016.03.012

Chasmer, L., Hopkinson, C., Smith, B., & Treitz, P. (2006). Examining the Influence of Changing Laser Pulse Repetition Frequencies on Conifer Forest Canopy Returns. *Photogrammetric Engineering & Remote Sensing*, *72*(12), 1359–1367. https://doi.org/10.14358/PERS.72.12.1359

Chirici, G., McRoberts, R. E., Fattorini, L., Mura, M., & Marchetti, M. (2016). Comparing Echo-Based and Canopy Height Model-Based Metrics for Enhancing Estimation of Forest Aboveground Biomass in a Model-Assisted Framework. *Remote Sensing of Environment*, *174*, 1–9. https://doi.org/10.1016/j.rse.2015.11.010

CloudCompare. (2019). CloudCompare, version 2.10.2. https://www.cloudcompare.org/

Coops, N. C., Tompalski, P., Goodbody, T. R. H., Queinnec, M., Luther, J. E., Bolton, D. K., White, J. C., Wulder, M. A., van Lier, O. R., & Hermosilla, T. (2021). Modelling Lidar-Derived Estimates of Forest Attributes over Space and Time: A Review of Approaches and Future Trends. *Remote Sensing of Environment*, *260*, 112477. https://doi.org/10.1016/j.rse.2021.112477

Corona, P., Chirici, G., McRoberts, R. E., Winter, S., & Barbati, A. (2011). Contribution of Large-Scale Forest Inventories to Biodiversity Assessment and

Monitoring. *Forest Ecology and Management*, *262*(11), 2061–2069. https://doi.org/10.1016/j.foreco.2011.08.044

Côté, J.-F., Widlowski, J.-L., Fournier, R. A., & Verstraete, M. M. (2009). The Structural and Radiative Consistency of Three-Dimensional Tree Reconstructions from Terrestrial Lidar. *Remote Sensing of Environment*, *113*(5), 1067–1081. https://doi.org/10.1016/j.rse.2009.01.017

Cotta, H. (1804). *Systematische Anleitung Zur Taxation Der Waldungen*. Sander.

Dalponte, M., Frizzera, L., Ørka, H. O., Gobakken, T., Næsset, E., & Gianelle, D. (2018). Predicting Stem Diameters and Aboveground Biomass of Individual Trees Using Remote Sensing Data. *Ecological Indicators*, *85*, 367–376. https://doi.org/10.1016/j.ecolind.2017.10.066

Dalponte, M., Martinez, C., Rodeghiero, M., & Gianelle, D. (2011). The Role of Ground Reference Data Collection in the Prediction of Stem Volume with LiDAR Data in Mountain Areas. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(6), 787–797. https://doi.org/10.1016/j.isprsjprs.2011.09.003

de Lera Garrido, A., Gobakken, T., Ørka, H. O., Næsset, E., & Bollandsås, O. M. (2020). Reuse of Field Data in ALS-assisted Forest Inventory. *Silva Fennica*, *54*(5).

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Deo, R. K., Froese, R. E., Falkowski, M. J., & Hudak, A. T. (2016). Optimizing Variable Radius Plot Size and LiDAR Resolution to Model Standing Volume in Conifer Forests. *Canadian Journal of Remote Sensing*, *42*(5), 428–442. https://doi.org/10.1080/07038992.2016.1220826

Disney, M. I., Boni Vicari, M., Burt, A., Calders, K., Lewis, S. L., Raumonen, P., & Wilkes, P. (2018). Weighing Trees with Lasers: Advances, Challenges and Opportunities. *Interface Focus*, *8*(2), 20170048. https://doi.org/10.1098/rsfs.2017.0048

Disney, M. I., Kalogirou, V., Lewis, P., Prieto-Blanco, A., Hancock, S., & Pfeifer, M. (2010). Simulating the Impact of Discrete-Return Lidar System and Survey Characteristics over Young Conifer and Broadleaf Forests. *Remote Sensing of Environment*, *114*(7), 1546–1560. https://doi.org/10.1016/j.rse.2010.02.009

Dixon, R. K., Solomon, A. M., Brown, S., Houghton, R. A., Trexier, M. C., & Wisniewski, J. (1994). Carbon Pools and Flux of Global Forest Ecosystems. *Science*, *263*(5144), 185–190. https://doi.org/10.1126/science.263.5144.185

Domingo, D., Alonso, R., Lamelas, M. T., Montealegre, A. L., Rodríguez, F., & de la Riva, J. (2019). Temporal Transferability of Pine Forest Attributes Modeling Using Low-Density Airborne Laser Scanning Data. *Remote Sensing*, *11*(3), 261. https://doi.org/10.3390/rs11030261

Doraisami, M., Kish, R., Paroshy, N. J., Domke, G. M., Thomas, S. C., & Martin, A. R. (2022). A Global Database of Woody Tissue Carbon Concentrations. *Scientific Data*, *9*(1), 284. https://doi.org/10.1038/s41597-022-01396-1

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

Dubayah, R., Blair, J. B., Goetz, S., Fatoyinbo, L., Hansen, M., Healey, S., Hofton, M., Hurtt, G., Kellner, J., Luthcke, S., Armston, J., Tang, H., Duncanson, L., Hancock, S., Jantz, P., Marselis, S., Patterson, P. L., Qi, W., & Silva, C. (2020). The Global Ecosystem Dynamics Investigation: High-resolution Laser Ranging of the Earth's Forests and Topography. *Science of Remote Sensing*, *1*, 100002. https://doi.org/10.1016/j.srs.2020.100002

Ene, L. T., Næsset, E., Gobakken, T., Bollandsås, O. M., Mauya, E. W., & Zahabu, E. (2017). Large-Scale Estimation of Change in Aboveground Biomass in Miombo Woodlands Using Airborne Laser Scanning and National Forest Inventory Data. *Remote Sensing of Environment*, *188*, 106–117. https://doi.org/10.1016/j.rse.2016.10.046

Esmorís, A. M., Weiser, H., Winiwarter, L., Cabaleiro, J. C., & Höfle, B. (2024). *Deep learning with simulated laser scanning data for 3D point cloud classification* [Preprint available on Earth ArXiv. doi.org/10.31223/X53Q3Q].

Fahey, T. J., Woodbury, P. B., Battles, J. J., Goodale, C. L., Hamburg, S. P., Ollinger, S. V., & Woodall, C. W. (2010). Forest Carbon Storage: Ecology, Management, and Policy. *Frontiers in Ecology and the Environment*, *8*(5), 245–252. https://doi.org/10.1890/080169

Falkowski, M. J., Wulder, M. A., White, J. C., & Gillis, M. D. (2009). Supporting Large-Area, Sample-Based Forest Inventories with Very High Spatial Resolution Satellite Imagery. *Progress in Physical Geography: Earth and Environment*, *33*(3), 403–423. https://doi.org/10.1177/0309133309342643

FAO. (2018a). *Global Forest Resources Assessment 2020: Terms and Definitions*. Retrieved March 20, 2024, from https://www.fao.org/3/I8661EN/i8661en.pdf

FAO. (2018b). *The State of the World's Forests 2018: Forest Pathways to Sustainable Development*. Retrieved November 14, 2023, from https://www.fao.org/3/i9535en/i9535en.pdf

FAO. (2020). *Global Forest Resources Assessment 2020: Main Report*. https://doi.org/10.4060/ca9825en

FAO and UNEP. (2020). *The State of the World's Forests 2020: Forests, Biodiversity and People*. https://doi.org/10.4060/ca8642en

Fassnacht, F. E., Hartig, F., Latifi, H., Berger, C., Hernández, J., Corvalán, P., & Koch, B. (2014). Importance of Sample Size, Data Type and Prediction Method for Remote Sensing-Based Estimations of Aboveground Forest Biomass. *Remote Sensing of Environment*, *154*, 102–114. https://doi.org/10.1016/j.rse.2014.07.028

Fassnacht, F. E., Latifi, H., & Hartig, F. (2018). Using Synthetic Data to Evaluate the Benefits of Large Field Plots for Forest Biomass Estimation with LiDAR. *Remote Sensing of Environment*, *213*, 115–128. https://doi.org/10.1016/j.rse.2018.05.007

Favero, A., Daigneault, A., & Sohngen, B. (2020). Forests: Carbon Sequestration, Biomass Energy, or Both? *Science Advances*, *6*(13), eaay6792. https://doi.org/10.1126/sciadv.aay6792

Fekety, P. A., Falkowski, M. J., & Hudak, A. T. (2015). Temporal Transferability of LiDAR-based Imputation of Forest Inventory Attributes. *Canadian Journal of Forest Research*, *45*(4), 422–435. https://doi.org/10.1139/cjfr-2014-0405

Fekry, R., Yao, W., Cao, L., & Shen, X. (2022). Ground-Based/UAV-LiDAR Data Fusion for Quantitative Structure Modeling and Tree Parameter Retrieval in Subtropical Planted Forest. *Forest Ecosystems*, *9*, 100065. https://doi.org/10.1016/j.fecs.2022.100065

Fischer, R., Bohn, F., Dantas de Paula, M., Dislich, C., Groeneveld, J., Gutiérrez, A. G., Kazmierczak, M., Knapp, N., Lehmann, S., Paulick, S., Pütz, S., Rödig, E., Taubert, F., Köhler, P., & Huth, A. (2016). Lessons Learned from Applying a Forest Gap Model to Understand Ecosystem and Carbon Dynamics of Complex Tropical Forests. *Ecological Modelling*, *326*, 124–133. https://doi.org/10.1016/j.ecolmodel.2015.11.018

Fischer, R., Knapp, N., Bohn, F., Shugart, H. H., & Huth, A. (2019). The Relevance of Forest Structure for Biomass and Productivity in Temperate Forests: New Perspectives for Remote Sensing. *Surveys in Geophysics*, *40*(4), 709–734. https://doi.org/10.1007/s10712-019-09519-x

Frazer, G. W., Magnussen, S., Wulder, M. A., & Niemann, K. O. (2011). Simulated Impact of Sample Plot Size and Co-Registration Error on the Accuracy and Uncertainty of LiDAR-derived Estimates of Forest Stand Biomass. *Remote Sensing of Environment*, *115*(2), 636–649. https://doi.org/10.1016/j.rse.2010.10.008

Fridman, J., Holm, S., Nilsson, M., Nilsson, P., Ringvall, A. H., & Ståhl, G. (2014). Adapting National Forest Inventories to Changing Requirements – the Case of the Swedish National Forest Inventory at the Turn of the 20th Century. *Silva Fennica*, *48*(3).

Fuller, A., Millard, K., & Green, J. R. (2022). SatViT: Pretraining Transformers for Earth Observation. *IEEE Geoscience and Remote Sensing Letters*, *19*, 1–5. https://doi.org/10.1109/LGRS.2022.3201489

García, M., Riaño, D., Chuvieco, E., & Danson, F. M. (2010). Estimating Biomass Carbon Stocks for a Mediterranean Forest in Central Spain Using LiDAR Height and Intensity Data. *Remote Sensing of Environment*, *114*(4), 816–830. https://doi.org/10.1016/j.rse.2009.11.021

Garnier, S., Ross, N., Rudis, R., Camargo, A. P., Sciaini, M., & Scherer, C. (2021). *viridis - Colorblind-Friendly Color Maps for R* [R package version 0.6.2]. https://doi.org/10.5281/zenodo.4679424

Gastellu-Etchegorry, J.-P., Yin, T., Lauret, N., Cajgfinger, T., Gregoire, T., Grau, E., Feret, J.-B., Lopes, M., Guilleux, J., Dedieu, G., Malenovský, Z., Cook, B. D., Morton, D., Rubio, J., Durrieu, S., Cazanave, G., Martin, E., & Ristorcelli, T. (2015). Discrete Anisotropic Radiative Transfer (DART 5) for Modeling Airborne and Satellite Spectroradiometer and LIDAR Acquisitions of Natural and Urban Landscapes. *Remote Sensing*, *7*(2), 1667–1701. https://doi.org/10.3390/rs70201667

Gastellu-Etchegorry, J.-P., Yin, T., Lauret, N., Grau, E., Rubio, J., Cook, B. D., Morton, D. C., & Sun, G. (2016). Simulation of Satellite, Airborne and Terrestrial LiDAR with DART (I): Waveform Simulation with Quasi-Monte Carlo Ray Tracing. *Remote Sensing of Environment*, *184*, 418–435. https://doi.org/10.1016/j.rse.2016.07.010

Getzin, S., & Wiegand, K. (2007). Asymmetric Tree Growth at the Stand Level: Random Crown Patterns and the Response to Slope. *Forest Ecology and Management*, *242*(2), 165–174. https://doi.org/10.1016/j.foreco.2007.01.009

Ghimire, S., Xystrakis, F., & Koutsias, N. (2017). Using Terrestrial Laser Scanning to Measure Forest Inventory Parameters in a Mediterranean Coniferous Stand of Western Greece. *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, *85*(4), 213–225. https://doi.org/10.1007/s41064-017-0024-1

Ginzler, C., & Hobi, M. L. (2015). Countrywide Stereo-Image Matching for Updating Digital Surface Models in the Framework of the Swiss National Forest Inventory. *Remote Sensing*, *7*(4), 4343–4370. https://doi.org/10.3390/rs70404343

Gobakken, T., & Næsset, E. (2008). Assessing Effects of Laser Point Density, Ground Sampling Intensity, and Field Sample Plot Size on Biophysical Stand Properties Derived from Airborne Laser Scanner Data. *Canadian Journal of Forest Research*, *38*(5), 1095–1109. https://doi.org/10.1139/X07-219

Gobakken, T., & Næsset, E. (2009). Assessing Effects of Positioning Errors and Sample Plot Size on Biophysical Stand Properties Derived from Airborne Laser Scanner Data. *Canadian Journal of Forest Research*, *39*(5), 1036–1052. https://doi.org/10.1139/X09-025

Goodbody, T. R. H., Coops, N. C., Queinnec, M., White, J. C., Tompalski, P., Hudak, A. T., Auty, D., Valbuena, R., LeBoeuf, A., Sinclair, I., McCartney, G., Prieur, J.-F., & Woods, M. E. (2023). sgsR: A Structurally Guided Sampling Toolbox for LiDAR-based Forest Inventories. *Forestry: An International Journal of Forest Research*, cpac055. https://doi.org/10.1093/forestry/cpac055

Goodwin, N. R., Coops, N. C., & Culvenor, D. S. (2007). Development of a Simulation Model to Predict LiDAR Interception in Forested Environments. *Remote*

Sensing of Environment, *111*(4), 481–492. https://doi.org/10.1016/j.rse.2007.04.001

Goodwin, N. R., Coops, N. C., & Culvenor, D. S. (2006). Assessment of Forest Structure with Airborne LiDAR and the Effects of Platform Altitude. *Remote Sensing of Environment*, *103*(2), 140–152. https://doi.org/10.1016/j.rse.2006.03.003

Graf, R. F., Mathys, L., & Bollmann, K. (2009). Habitat Assessment for Forest Dwelling Species Using LiDAR Remote Sensing: Capercaillie in the Alps. *Forest Ecology and Management*, *257*(1), 160–167. https://doi.org/10.1016/j.foreco.2008.08.021

Grau, E., Durrieu, S., Fournier, R., Gastellu-Etchegorry, J.-P., & Yin, T. (2017). Estimation of 3D Vegetation Density with Terrestrial Laser Scanning Data Using Voxels. A Sensitivity Analysis of Influencing Parameters. *Remote Sensing of Environment*, *191*, 373–388. https://doi.org/10.1016/j.rse.2017.01.032

Gschwantner, T., Alberdi, I., Bauwens, S., Bender, S., Borota, D., Bosela, M., Bouriaud, O., Breidenbach, J., Donis, J., Fischer, C., Gasparini, P., Heffernan, L., Hervé, J.-C., Kolozs, L., Korhonen, K. T., Koutsias, N., Kovácsevics, P., Kučera, M., Kulbokas, G., … Tomter, S. M. (2022). Growing Stock Monitoring by European National Forest Inventories: Historical Origins, Current Methods and Harmonisation. *Forest Ecology and Management*, *505*, 119868. https://doi.org/10.1016/j.foreco.2021.119868

Hagar, J. C., Eskelson, B. N. I., Haggerty, P. K., Nelson, S. K., & Vesely, D. G. (2014). Modeling Marbled Murrelet (Brachyramphus Marmoratus) Habitat Using LiDAR-derived Canopy Data. *Wildlife Society Bulletin*, *38*(2), 237–249. https://doi.org/10.1002/wsb.407

Hamedianfar, A., Mohamedou, C., Kangas, A., & Vauhkonen, J. (2022). Deep Learning for Forest Inventory and Planning: A Critical Review on the Remote Sensing Approaches so Far and Prospects for Further Applications. *Forestry: An International Journal of Forest Research*, cpac002. https://doi.org/10.1093/forestry/cpac002

Hamraz, H., Contreras, M. A., & Zhang, J. (2017). Vertical Stratification of Forest Canopy for Segmentation of Understory Trees within Small-Footprint Airborne LiDAR Point Clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, *130*, 385–392. https://doi.org/10.1016/j.isprsjprs.2017.07.001

Hamraz, H., Jacobs, N. B., Contreras, M. A., & Clark, C. H. (2019). Deep Learning for Conifer/Deciduous Classification of Airborne LiDAR 3D Point Clouds Representing Individual Trees. *ISPRS Journal of Photogrammetry and Remote Sensing*, *158*, 219–230. https://doi.org/10.1016/j.isprsjprs.2019.10.011

Hancock, S., Armston, J., Hofton, M., Sun, X., Tang, H., Duncanson, L. I., Kellner, J. R., & Dubayah, R. (2019). The GEDI Simulator: A Large-Footprint Waveform Lidar Simulator for Calibration and Validation of Spaceborne

Missions. *Earth and Space Science*, *6*(2), 294–310. https://doi.org/10.1029/2018EA000506

Harris, N. L., Gibbs, D. A., Baccini, A., Birdsey, R. A., de Bruin, S., Farina, M., Fatoyinbo, L., Hansen, M. C., Herold, M., Houghton, R. A., Potapov, P. V., Suarez, D. R., Roman-Cuesta, R. M., Saatchi, S. S., Slay, C. M., Turubanova, S. A., & Tyukavina, A. (2021). Global Maps of Twenty-First Century Forest Carbon Fluxes. *Nature Climate Change*, *11*(3), 234–240. https://doi.org/10.1038/s41558-020-00976-6

Hartig, G. L. (1795). *Anweisung zur Taxation der Forste, oder zur Bestimmung des Holzertrags der Wälder*. Heyer.

Hawbaker, T. J., Keuler, N. S., Lesak, A. A., Gobakken, T., Contrucci, K., & Radeloff, V. C. (2009). Improved Estimates of Forest Vegetation Structure and Biomass with a LiDAR-optimized Sampling Design. *Journal of Geophysical Research: Biogeosciences*, *114*(G2), n/a–n/a. https://doi.org/10.1029/2008JG000870

Hayashi, R., Kershaw, J. A., & Weiskittel, A. (2015). Evaluation of alternative methods for using LiDAR to predict aboveground biomass in mixed species and structurally complex forests in northeastern North America. *Mathematical and Computational Forestry & Natural-Resource Sciences (MCFNS)*, *7*(2), 49–65.

Hell, M., Brandmeier, M., Briechle, S., & Krzystek, P. (2022). Classification of Tree Species and Standing Dead Trees with Lidar Point Clouds Using Two Deep Neural Networks: PointCNN and 3DmFV-Net. *PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, *90*(2), 103–121. https://doi.org/10.1007/s41064-022-00200-4

Henniger, H., Huth, A., Frank, K., & Bohn, F. J. (2023). Creating Virtual Forests around the Globe and Analysing Their State Space. *Ecological Modelling*, *483*, 110404. https://doi.org/10.1016/j.ecolmodel.2023.110404

Holdaway, M. R. (1986). Modeling Tree Crown Ratio. *The Forestry Chronicle*, *62*(5), 451–455. https://doi.org/10.5558/tfc62451-5

Hollaus, M., Mücke, W., Roncat, A., Pfeifer, N., & Briese, C. (2014). Full-Waveform Airborne Laser Scanning Systems and Their Possibilities in Forest Applications (Chapter 3). In M. Maltamo, E. Næsset, & J. Vauhkonen (Eds.), *Forestry Applications of Airborne Laser Scanning: Concepts and Case Studies* (pp. 43–62, Vol. 27). Springer Netherlands. https://doi.org/10.1007/978-94-017-8663-8

Holmgren, J., & Jonsson, T. (2004). Large Scale Airborne Laser Scanning of Forest Resources in Sweden. *International Archives of photogrammetry, remote sensing and spatial information sciences*, *36*(8), W2.

Holmgren, J., Nilsson, M., & Olsson, H. (2003). Simulating the Effects of Lidar Scanning Angle for Estimation of Mean Tree Height and Canopy Closure. *29*(5), 11.

Holopainen, M., Vastaranta, M., & Hyyppä, J. (2014). Outlook for the Next Generation's Precision Forestry in Finland. *Forests*, *5*(7), 1682–1694. https://doi.org/10.3390/f5071682

Hopkinson, C. (2007). The Influence of Flying Altitude, Beam Divergence, and Pulse Repetition Frequency on Laser Pulse Return Intensity and Canopy Frequency Distribution. *Canadian Journal of Remote Sensing*, *33*(4), 312–324. https://doi.org/10.5589/m07-029

Hopkinson, C., & Chasmer, L. (2009). Testing LiDAR Models of Fractional Cover across Multiple Forest Ecozones. *Remote Sensing of Environment*, *113*(1), 275–288. https://doi.org/10.1016/j.rse.2008.09.012

Hwang, S.-H., & Whang, S. E. (2022). RegMix: Data Mixing Augmentation for Regression [Preprint available on arXiv. https://doi.org/10.48550/arXiv.2106.03374].

IPCC. (2022). *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (P. Shukla, J. Skea, R. Slade, A. A. Khourdajie, R. van Diemen, D. McCollum, M. Pathak, S. Some, P. Vyas, R. Fradera, M. Belkacemi, A. Hasija, G. Lisboa, S. Luz, & J. Malley, Eds.). Cambridge University Press. https://doi.org/10.1017/9781009157926

Johnston, A. N., & Moskal, L. M. (2017). High-Resolution Habitat Modeling with Airborne LiDAR for Red Tree Voles. *The Journal of Wildlife Management*, *81*(1), 58–72. https://doi.org/10.1002/jwmg.21173

Jucker, T., Caspersen, J., Chave, J., Antin, C., Barbier, N., Bongers, F., Dalponte, M., van Ewijk, K., Forrester, D. I., Haeni, M., Higgins, S. I., Holdaway, R. J., Iida, Y., Lorimer, C., Marshall, P. L., Momo, S., Moncrieff, G. R., Ploton, P., Poorter, L., … Coomes, D. A. (2017). Allometric Equations for Integrating Remote Sensing Imagery into Forest Monitoring Programmes. *Global Change Biology*, *23*(1), 177–190. https://doi.org/10.1111/gcb.13388

Kangas, A., Astrup, R., Breidenbach, J., Fridman, J., Gobakken, T., Korhonen, K. T., Maltamo, M., Nilsson, M., Nord-Larsen, T., Næsset, E., & Olsson, H. (2018). Remote Sensing and Forest Inventories in Nordic Countries – Roadmap for the Future. *Scandinavian Journal of Forest Research*, *33*(4), 397–412. https://doi.org/10.1080/02827581.2017.1416666

Kassambara, A. (2020). *ggpubr: 'ggplot2' Based Publication Ready Plots* [R package version 0.4.0]. https://CRAN.R-project.org/package=ggpubr

Kattenborn, T., Leitloff, J., Schiefer, F., & Hinz, S. (2021). Review on Convolutional Neural Networks (CNN) in Vegetation Remote Sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, *173*, 24–49. https://doi.org/10.1016/j.isprsjprs.2020.12.010

Keränen, J., Maltamo, M., & Packalen, P. (2016). Effect of Flying Altitude, Scanning Angle and Scanning Mode on the Accuracy of ALS Based Forest Inventory. *International Journal of Applied Earth Observation and Geoinformation*, *52*, 349–360. https://doi.org/10.1016/j.jag.2016.07.005

Kingma, D. P., & Ba, J. (2017). Adam: A Method for Stochastic Optimization [Preprint available on arXiv. https://doi.org/10.48550/arXiv.1412.6980].

Kleinn, C., Magnussen, S., Nölke, N., Magdon, P., Álvarez-González, J. G., Fehrmann, L., & Pérez-Cruzado, C. (2020). Improving Precision of Field Inventory Estimation of Aboveground Biomass through an Alternative View on Plot Biomass. *Forest Ecosystems*, *7*(1), 57. https://doi.org/10.1186/s40663-020-00268-7

Knapp, N., Fischer, R., & Huth, A. (2018). Linking Lidar and Forest Modeling to Assess Biomass Estimation across Scales and Disturbance States. *Remote Sensing of Environment*, *205*, 199–209. https://doi.org/10.1016/j.rse.2017.11.018

Knapp, N., Huth, A., & Fischer, R. (2021). Tree Crowns Cause Border Effects in Area-Based Biomass Estimations from Remote Sensing. *Remote Sensing*, *13*(8), 1592. https://doi.org/10.3390/rs13081592

Köhler, P., & Huth, A. (1998). The Effects of Tree Species Grouping in Tropical Rainforest Modelling: Simulations with the Individual-Based Model Formind. *Ecological Modelling*, *109*(3), 301–321. https://doi.org/10.1016/S0304-3800(98)00066-0

Kölle, M., Walter, V., Schmohl, S., & Soergel, U. (2021). Remembering Both the Machine and the Crowd When Sampling Points: Active Learning for Semantic Segmentation of ALS Point Clouds. In A. Del Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, H. J. Escalante, & R. Vezzani (Eds.), *Pattern Recognition. ICPR International Workshops and Challenges* (pp. 505–520). Springer International Publishing. https://doi.org/10.1007/978-3-030-68787-8_37

Korpela, I., Hovi, A., & Morsdorf, F. (2012). Understory Trees in Airborne LiDAR Data — Selective Mapping Due to Transmission Losses and Echo-Triggering Mechanisms. *Remote Sensing of Environment*, *119*, 92–104. https://doi.org/10.1016/j.rse.2011.12.011

Kotivuori, E., Korhonen, L., & Packalen, P. (2016). Nationwide Airborne Laser Scanning Based Models for Volume, Biomass and Dominant Height in Finland. *Silva Fennica*, *50*(4).

Krajicek, J. E., Brinkman, K. A., & Gingrich, S. F. (1961). Crown competition—a measure of density. *Forest science*, *7*(1), 35–42.

Krisanski, S., Taskhiri, M. S., Gonzalez Aracil, S., Herries, D., Muneri, A., Gurung, M. B., Montgomery, J., & Turner, P. (2021). Forest Structural Complexity Tool—An Open Source, Fully-Automated Tool for Measuring Forest Point Clouds. *Remote Sensing*, *13*(22). https://doi.org/10.3390/rs13224677

Kukko, A., & Hyyppä, J. (2009). Small-Footprint Laser Scanning Simulator for System Validation, Error Assessment, and Algorithm Development. *Photogrammetric Engineering & Remote Sensing*, *75*(10), 1177–1189. https://doi.org/10.14358/PERS.75.10.1177

Kukkonen, M., Maltamo, M., Korhonen, L., & Packalen, P. (2019). Comparison of multispectral airborne laser scanning and stereo matching of aerial images as

a single sensor solution to forest inventories by tree species. *Remote Sensing of Environment*, *231*, 111208. https://doi.org/https://doi.org/10.1016/j.rse.2019.05.027

Lang, N., Kalischek, N., Armston, J., Schindler, K., Dubayah, R., & Wegner, J. D. (2022). Global Canopy Height Regression and Uncertainty Estimation from GEDI LIDAR Waveforms with Deep Ensembles. *Remote Sensing of Environment*, *268*, 112760. https://doi.org/10.1016/j.rse.2021.112760

Latifi, H., Fassnacht, F. E., Müller, J., Tharani, A., Dech, S., & Heurich, M. (2015). Forest Inventories by LiDAR Data: A Comparison of Single Tree Segmentation and Metric-Based Methods for Inventories of a Heterogeneous Temperate Forest. *International Journal of Applied Earth Observation and Geoinformation*, *42*, 162–174. https://doi.org/10.1016/j.jag.2015.06.008

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep Learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lefsky, M. A., Cohen, W. B., Harding, D. J., Parker, G. G., Acker, S. A., & Gower, S. T. (2002). Lidar Remote Sensing of Above-Ground Biomass in Three Biomes. *Global Ecology and Biogeography*, *11*(5), 393–399. https://doi.org/10.1046/j.1466-822x.2002.00303.x

Li, H., Hu, B., Li, Q., & Jing, L. (2020). CNN-Based Tree Species Classification Using Airborne Lidar Data and High-Resolution Satellite Image. *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, 2679–2682. https://doi.org/10.1109/IGARSS39084.2020.9324011

Li, R., Li, X., Heng, P.-A., & Fu, C.-W. (2020). PointAugment: An Auto-Augmentation Framework for Point Cloud Classification. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6377–6386. https://doi.org/10.1109/CVPR42600.2020.00641

Li, Y., Andersen, H.-E., & McGaughey, R. (2008). A Comparison of Statistical Methods for Estimating Forest Biomass from Light Detection and Ranging Data. *Western Journal of Applied Forestry*, *23*(4), 223–231. https://doi.org/10.1093/wjaf/23.4.223

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22. https://cran.r-project.org/web/packages/randomForest

Lim, K., Hopkinson, C., & Treitz, P. (2008). Examining the Effects of Sampling Point Densities on Laser Canopy Height and Density Metrics. *The Forestry Chronicle*, *84*. https://doi.org/10.5558/tfc84876-6

Lisańczuk, M., Mitelsztedt, K., Parkitna, K., Krok, G., Stereńczak, K., Wysocka-Fijorek, E., & Miścicki, S. (2020). Influence of Sampling Intensity on Performance of Two-Phase Forest Inventory Using Airborne Laser Scanning. *Forest Ecosystems*, *7*(1), 65. https://doi.org/10.1186/s40663-020-00277-6

Lister, A. J., Andersen, H., Frescino, T., Gatziolis, D., Healey, S., Heath, L. S., Liknes, G. C., McRoberts, R., Moisen, G. G., Nelson, M., Riemann, R., Schleeweis, K., Schroeder, T. A., Westfall, J., & Wilson, B. T. (2020). Use of Remote Sensing

Data to Improve the Efficiency of National Forest Inventories: A Case Study from the United States National Forest Inventory. *Forests*, *11*(12), 1364. https://doi.org/10.3390/f11121364

Luo, Z., Zhang, Z., Li, W., Chen, Y., Wang, C., Nurunnabi, A. A. M., & Li, J. (2022). Detection of Individual Trees in UAV LiDAR Point Clouds Using a Deep Learning Framework Based on Multichannel Representation. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–15. https://doi.org/10.1109/TGRS.2021.3130725

Ma, Q., Su, Y., & Guo, Q. (2017). Comparison of Canopy Cover Estimations From Airborne LiDAR, Aerial Imagery, and Satellite Imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *10*(9), 4225–4236. https://doi.org/10.1109/JSTARS.2017.2711482

Maltamo, M., Bollandsas, O. M., Naesset, E., Gobakken, T., & Packalen, P. (2011). Different Plot Selection Strategies for Field Training Data in ALS-assisted Forest Inventory. *Forestry*, *84*(1), 23–31. https://doi.org/10.1093/forestry/cpq039

Maltamo, M., Eerikäinen, K., Pitkänen, J., Hyyppä, J., & Vehmas, M. (2004). Estimation of Timber Volume and Stem Density Based on Scanning Laser Altimetry and Expected Tree Size Distribution Functions. *Remote Sensing of Environment*, *90*(3), 319–330. https://doi.org/10.1016/j.rse.2004.01.006

Maltamo, M., Korhonen, K., Packalén, P., Mehtätalo, L., & Suvanto, A. (2007). Testing the usability of truncated angle count sample plots as ground truth in airborne laser scanning-based forest inventories. *Forestry*, *80*(1), 73–81. https://doi.org/10.1093/forestry/cpl045

Maltamo, M., Packalén, P., Yu, X., Eerikäinen, K., Hyyppä, J., & Pitkänen, J. (2005). Identifying and Quantifying Structural Characteristics of Heterogeneous Boreal Forests Using Laser Scanner Data. *Forest Ecology and Management*, *216*(1), 41–50. https://doi.org/10.1016/j.foreco.2005.05.034

Martin, A. R., Doraisami, M., & Thomas, S. C. (2018). Global Patterns in Wood Carbon Concentration across the World's Trees and Forests. *Nature Geoscience*, *11*(12), 915–920. https://doi.org/10.1038/s41561-018-0246-x

Martinuzzi, S., Vierling, L. A., Gould, W. A., Falkowski, M. J., Evans, J. S., Hudak, A. T., & Vierling, K. T. (2009). Mapping Snags and Understory Shrubs for a LiDAR-based Assessment of Wildlife Habitat Suitability. *Remote Sensing of Environment*, *113*(12), 2533–2546. https://doi.org/10.1016/j.rse.2009.07.002

McRoberts, R. E., Næsset, E., & Gobakken, T. (2013). Inference for Lidar-Assisted Estimation of Forest Growing Stock Volume. *Remote Sensing of Environment*, *128*, 268–275. https://doi.org/10.1016/j.rse.2012.10.007

McRoberts, R. E., Næsset, E., Gobakken, T., & Bollandsås, O. M. (2015). Indirect and Direct Estimation of Forest Biomass Change Using Forest Inventory and Airborne Laser Scanning Data. *Remote Sensing of Environment*, *164*, 36–42. https://doi.org/10.1016/j.rse.2015.02.018

McRoberts, R. E., & Tomppo, E. O. (2007). Remote Sensing Support for National Forest Inventories. *Remote Sensing of Environment*, *110*(4), 412–419. https://doi.org/10.1016/j.rse.2006.09.034

McRoberts, R. E., Tomppo, E. O., & Næsset, E. (2010). Advances and Emerging Issues in National Forest Inventories. *Scandinavian Journal of Forest Research*, *25*(4), 368–381. https://doi.org/10.1080/02827581.2010.496739

Mehtätalo, L., de-Miguel, S., & Gregoire, T. G. (2015). Modeling Height-Diameter Curves for Prediction. *Canadian Journal of Forest Research*. https://doi.org/10.1139/cjfr-2015-0054

Meyer, H., & Pebesma, E. (2021). Predicting into Unknown Space? Estimating the Area of Applicability of Spatial Prediction Models. *Methods in Ecology and Evolution*, *12*(9), 1620–1633. https://doi.org/10.1111/2041-210X.13650

Milodowski, D. T., Coomes, D. A., Swinfield, T., Jucker, T., Riutta, T., Malhi, Y., Svátek, M., Kvasnica, J., Burslem, D. F. R. P., Ewers, R. M., Teh, Y. A., & Williams, M. (2021). The Impact of Logging on Vertical Canopy Structure across a Gradient of Tropical Forest Degradation Intensity in Borneo. *Journal of Applied Ecology*, *n/a*(n/a). https://doi.org/10.1111/1365-2664.13895

Montaghi, A. (2013). Effect of Scanning Angle on Vegetation Metrics Derived from a Nationwide Airborne Laser Scanning Acquisition. *Canadian Journal of Remote Sensing*, *39*(sup1), S152–S173. https://doi.org/10.5589/m13-052

Morsdorf, F., Frey, O., Meier, E., Itten, K. I., & Allgöwer, B. (2008). Assessment of the Influence of Flying Altitude and Scan Angle on Biophysical Vegetation Products Derived from Airborne Laser Scanning. *International Journal of Remote Sensing*, *29*(5), 1387–1406. https://doi.org/10.1080/01431160701736349

Morsdorf, F., Mårell, A., Koetz, B., Cassagne, N., Pimont, F., Rigolot, E., & Allgöwer, B. (2010). Discrimination of Vegetation Strata in a Multi-Layered Mediterranean Forest Ecosystem Using Height and Intensity Information Derived from Airborne Laser Scanning. *Remote Sensing of Environment*, *114*(7), 1403–1415. https://doi.org/10.1016/j.rse.2010.01.023

Morsdorf, F., Meier, E., Kötz, B., Itten, K. I., Dobbertin, M., & Allgöwer, B. (2004). LIDAR-based Geometric Reconstruction of Boreal Type Forest Stands at Single Tree Level for Forest and Wildland Fire Management. *Remote Sensing of Environment*, *92*(3), 353–362. https://doi.org/10.1016/j.rse.2004.05.013

Moudrý, V., Cord, A. F., Gábor, L., Laurin, G. V., Barták, V., Gdulová, K., Malavasi, M., Rocchini, D., Stereńczak, K., Prošek, J., Klápště, P., & Wild, J. (2023). Vegetation Structure Derived from Airborne Laser Scanning to Assess Species Distribution and Habitat Suitability: The Way Forward. *Diversity and Distributions*, *29*(1), 39–50. https://doi.org/10.1111/ddi.13644

Næsset, E. (2002). Predicting Forest Stand Characteristics with Airborne Scanning Laser Using a Practical Two-Stage Procedure and Field Data. *Remote Sensing of Environment*, *80*(1), 88–99. https://doi.org/10.1016/S0034-4257(01)00290-5

Næsset, E. (2004). Practical Large-Scale Forest Stand Inventory Using a Small-Footprint Airborne Scanning Laser. *Scandinavian Journal of Forest Research*, *19*(2), 164–179. https://doi.org/10.1080/02827580310019257

Næsset, E. (2007). Airborne Laser Scanning as a Method in Operational Forest Inventory: Status of Accuracy Assessments Accomplished in Scandinavia. *Scandinavian Journal of Forest Research*, *22*(5), 433–442. https://doi.org/10.1080/02827580701672147

Næsset, E., & Gjevestad, J. G. (2008). Performance of GPS Precise Point Positioning Under Conifer Forest Canopies. *Photogrammetric Engineering & Remote Sensing*, *74*(5), 661–668. https://doi.org/10.14358/PERS.74.5.661

Næsset, E., & Gobakken, T. (2008). Estimation of Above- and below-Ground Biomass across Regions of the Boreal Forest Zone Using Airborne Laser. *Remote Sensing of Environment*, *112*(6), 3079–3090. https://doi.org/10.1016/j.rse.2008.03.004

Næsset, E., Gobakken, T., Solberg, S., Gregoire, T. G., Nelson, R., Ståhl, G., & Weydahl, D. (2011). Model-Assisted Regional Forest Biomass Estimation Using LiDAR and InSAR as Auxiliary Data: A Case Study from a Boreal Forest Area. *Remote Sensing of Environment*, *115*(12), 3599–3614. https://doi.org/10.1016/j.rse.2011.08.021

Nelson, R. (1997). Modeling Forest Canopy Heights: The Effects of Canopy Shape. *Remote Sensing of Environment*, *60*(3), 327–334. https://doi.org/10.1016/S0034-4257(96)00214-3

Nelson, R., Oderwald, R., & Gregoire, T. G. (1997). Separating the Ground and Airborne Laser Sampling Phases to Estimate Tropical Forest Basal Area, Volume, and Biomass. *Remote Sensing of Environment*, *60*(3), 311–326. https://doi.org/10.1016/S0034-4257(96)00213-1

Niu, S., Liu, M., Liu, Y., Wang, J., & Song, H. (2021). Distant Domain Transfer Learning for Medical Imaging. *IEEE Journal of Biomedical and Health Informatics*, *25*(10), 3784–3793. https://doi.org/10.1109/JBHI.2021.3051470

Nord-Larsen, T., & Schumacher, J. (2012). Estimation of Forest Resources from a Country Wide Laser Scanning Survey and National Forest Inventory Data. *Remote Sensing of Environment*, *119*, 148–157. https://doi.org/10.1016/j.rse.2011.12.022

Oehmcke, S., Li, L., Revenga, J. C., Nord-Larsen, T., Trepekli, K., Gieseke, F., & Igel, C. (2022). Deep Learning Based 3D Point Cloud Regression for Estimating Forest Biomass. *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, 1–4. https://doi.org/10.1145/3557915.3561471

Ørka, H. O., Næsset, E., & Bollandsås, O. M. (2009). Classifying species of individual trees by intensity and structure features derived from airborne laser scanner data. *Remote Sensing of Environment*, *113*(6), 1163–1174. https://doi.org/https://doi.org/10.1016/j.rse.2009.02.002

Packalen, P., Strunk, J., Maltamo, M., & Myllymäki, M. (2023). Circular or Square Plots in ALS-based Forest Inventories—Does It Matter? *Forestry: An International Journal of Forest Research*, *96*(1), 49–61. https://doi.org/10.1093/forestry/cpac032

Palace, M. W., Sullivan, F. B., Ducey, M. J., Treuhaft, R. N., Herrick, C., Shimbo, J. Z., & Mota-E-Silva, J. (2015). Estimating Forest Structure in a Tropical Forest Using Field Measurements, a Synthetic Model and Discrete Return Lidar Data. *Remote Sensing of Environment*, *161*, 1–11. https://doi.org/10.1016/j.rse.2015.01.020

Pan, Y., Birdsey, R. A., Fang, J., Houghton, R., Kauppi, P. E., Kurz, W. A., Phillips, O. L., Shvidenko, A., Lewis, S. L., Canadell, J. G., Ciais, P., Jackson, R. B., Pacala, S. W., McGuire, A. D., Piao, S., Rautiainen, A., Sitch, S., & Hayes, D. (2011). A Large and Persistent Carbon Sink in the World's Forests. *Science*, *333*(6045), 988–993. https://doi.org/10.1126/science.1201609

Pearse, G. D., Watt, M. S., Dash, J. P., Stone, C., & Caccamo, G. (2019). Comparison of Models Describing Forest Inventory Attributes Using Standard and Voxel-Based Lidar Predictors across a Range of Pulse Densities. *International Journal of Applied Earth Observation and Geoinformation*, *78*, 341–351. https://doi.org/10.1016/j.jag.2018.10.008

Persson, H. J., Ekström, M., & Ståhl, G. (2022). Quantify and Account for Field Reference Errors in Forest Remote Sensing Studies. *Remote Sensing of Environment*, *283*, 113302. https://doi.org/10.1016/j.rse.2022.113302

Pope, G., & Treitz, P. (2013). Leaf Area Index (LAI) Estimation in Boreal Mixedwood Forest of Ontario, Canada Using Light Detection and Ranging (LiDAR) and WorldView-2 Imagery. *Remote Sensing*, *5*(10), 5040–5063. https://doi.org/10.3390/rs5105040

Popescu, S. C., & Zhao, K. (2008). A Voxel-Based Lidar Method for Estimating Crown Base Height for Deciduous and Pine Trees. *Remote Sensing of Environment*, *112*(3), 767–781. https://doi.org/10.1016/j.rse.2007.06.011

Pretzsch, H., Biber, P., & Ďurský, J. (2002). The Single Tree-Based Stand Simulator SILVA: Construction, Application and Evaluation. *Forest Ecology and Management*, *162*(1), 3–21. https://doi.org/10.1016/S0378-1127(02)00047-6

Puliti, S., Breidenbach, J., & Astrup, R. (2020). Estimation of Forest Growing Stock Volume with UAV Laser Scanning Data: Can It Be Done without Field Data? *Remote Sensing*, *12*(8), 1245. https://doi.org/10.3390/rs12081245

Qin, H., Wang, C., Xi, X., Tian, J., & Zhou, G. (2017). Simulating the Effects of the Airborne Lidar Scanning Angle, Flying Altitude, and Pulse Density for Forest Foliage Profile Retrieval. *Applied Sciences*, *7*(7), 712. https://doi.org/10.3390/app7070712

R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Racine, E. B., Coops, N. C., St-Onge, B., & Bégin, J. (2014). Estimating Forest Stand Age from LiDAR-Derived Predictors and Nearest Neighbor Imputation. *Forest Science*, *60*(1), 128–136. https://doi.org/10.5849/forsci.12-088

Rana, P., Gautam, B., & Tokola, T. (2016). Optimizing the Number of Training Areas for Modeling Above-Ground Biomass with ALS and Multispectral Remote Sensing in Subtropical Nepal. *International Journal of Applied Earth Observation and Geoinformation*, *49*, 52–62. https://doi.org/10.1016/j.jag.2016.01.006

RIEGL Laser Measurement Systems. (2019). RIEGL VQ-780i, Data Sheet [Accessed: 2021-10-22]. http://www.riegl.com/uploads/tx_pxpriegldownloads/RIEGL_VQ-780i_Datasheet_2019-09-02.pdf

Roberts, O., Bunting, P., Hardy, A., & McInerney, D. (2020). Sensitivity Analysis of the DART Model for Forest Mensuration with Airborne Laser Scanning. *Remote Sensing*, *12*(2), 247. https://doi.org/10.3390/rs12020247

Roussel, J.-R., & Auty, D. (2021). *Airborne LiDAR Data Manipulation and Visualization for Forestry Applications* [R package version 3.2.3]. https://cran.r-project.org/package=lidR

Roussel, J.-R., Auty, D., Coops, N. C., Tompalski, P., Goodbody, T. R., Meador, A. S., Bourdon, J.-F., de Boissieu, F., & Achim, A. (2020). lidR: An R package for analysis of Airborne Laser Scanning (ALS) data. *Remote Sensing of Environment*, *251*, 112061. https://doi.org/10.1016/j.rse.2020.112061

RStudio Team. (2016). *RStudio: Integrated Development Environment for R*. RStudio, Inc. Boston, MA. http://www.rstudio.com/

Schäfer, J., Weiser, H., Winiwarter, L., Höfle, B., Schmidtlein, S., & Fassnacht, F. E. (2023a). 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: An International Journal of Forest Research*, cpad006. https://doi.org/10.1093/forestry/cpad006

Schäfer, J., Winiwarter, L., Weiser, H., Novotný, J., Höfle, B., Schmidtlein, S., Henniger, H., Krok, G., Stereńczak, K., & Fassnacht, F. E. (2023b). Assessing the Potential of Synthetic and Ex Situ Airborne Laser Scanning and Ground Plot Data to Train Forest Biomass Models. *Forestry: An International Journal of Forest Research*, cpad061. https://doi.org/10.1093/forestry/cpad061

Schutz, B. E., Zwally, H. J., Shuman, C. A., Hancock, D., & DiMarzio, J. P. (2005). Overview of the ICESat Mission. *Geophysical Research Letters*, *32*(21). https://doi.org/10.1029/2005GL024009

Seely, H., Coops, N. C., White, J. C., Montwé, D., Winiwarter, L., & Ragab, A. (2023). Modelling Tree Biomass Using Direct and Additive Methods with Point Cloud Deep Learning in a Temperate Mixed Forest. *Science of Remote Sensing*, *8*, 100110. https://doi.org/10.1016/j.srs.2023.100110

Seidel, D., Annighöfer, P., Thielman, A., Seifert, Q. E., Thauer, J.-H., Glatthorn, J., Ehbrecht, M., Kneib, T., & Ammer, C. (2021). Predicting Tree Species From

3D Laser Scanning Point Clouds Using Deep Learning. *Frontiers in Plant Science*, *12*, 141. https://doi.org/10.3389/fpls.2021.635440

Shan, J., & Toth, C. K. (2018). *Topographic Laser Ranging and Scanning: Principles and Processing, Second Edition*. CRC Press.

Sheridan, R. D., Popescu, S. C., Gatziolis, D., Morgan, C. L. S., & Ku, N.-W. (2015). Modeling Forest Aboveground Biomass and Volume Using Airborne LiDAR Metrics and Forest Inventory and Analysis Data in the Pacific Northwest. *Remote Sensing*, *7*(1), 229–255. https://doi.org/10.3390/rs70100229

Shorten, C., & Khoshgoftaar, T. M. (2019). A Survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, *6*(1), 60. https://doi.org/10.1186/s40537-019-0197-0

Shugart, H. H., Wang, B., Fischer, R., Ma, J., Fang, J., Yan, X., Huth, A., & Armstrong, A. H. (2018). Gap Models and Their Individual-Based Relatives in the Assessment of the Consequences of Global Change. *Environmental Research Letters*, *13*(3), 033001. https://doi.org/10.1088/1748-9326/aaaacc

Signorell, A., et al. (2021). *DescTools: Tools for Descriptive Statistics* [R package version 0.99.44]. https://cran.r-project.org/package=DescTools

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition [Preprint available on arXiv. https://doi.org/10.48550/arXiv.1409.1556].

Smith, A. M. S., Falkowski, M. J., Hudak, A. T., Evans, J. S., Robinson, A. P., & Steele, C. M. (2009). A Cross-Comparison of Field, Spectral, and Lidar Estimates of Forest Canopy Cover. *Canadian Journal of Remote Sensing*, *35*(5), 447–459. https://doi.org/10.5589/m09-038

Solovyev, R., Kalinin, A. A., & Gabruseva, T. (2022). 3D Convolutional Neural Networks for Stalled Brain Capillary Detection. *Computers in Biology and Medicine*, *141*, 105089. https://doi.org/10.1016/j.compbiomed.2021.105089

Spracklen, D. V., Arnold, S. R., & Taylor, C. M. (2012). Observations of Increased Tropical Rainfall Preceded by Air Passage over Forests. *Nature*, *489*(7415), 282–285. https://doi.org/10.1038/nature11390

Spriggs, R. A., Coomes, D. A., Jones, T. A., Caspersen, J. P., & Vanderwel, M. C. (2017). An Alternative Approach to Using LiDAR Remote Sensing Data to Predict Stem Diameter Distributions across a Temperate Forest Landscape. *Remote Sensing*, *9*(9), 944. https://doi.org/10.3390/rs9090944

Spriggs, R. A., Vanderwel, M. C., Jones, T. A., Caspersen, J. P., & Coomes, D. A. (2015). 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. *Canadian Journal of Forest Research*, *45*(10), 1338–1350. https://doi.org/10.1139/cjfr-2015-0018

Stepper, C., Straub, C., Immitzer, M., & Pretzsch, H. (2017). Using Canopy Heights from Digital Aerial Photogrammetry to Enable Spatial Transfer of Forest Attribute Models: A Case Study in Central Europe. *Scandinavian Journal of*

Forest Research, *32*(8), 748–761. https://doi.org/10.1080/02827581.2016.1261935

Stereńczak, K., Lisańczuk, M., Parkitna, K., Mitelsztedt, K., Mroczek, P., & Miścicki, S. (2018). The Influence of Number and Size of Sample Plots on Modelling Growing Stock Volume Based on Airborne Laser Scanning. *Drewno. Prace Naukowe. Doniesienia. Komunikaty*, *61*(201).

Stocker, O., Kouhi, R. M., Guilbert, E., Ferraz, A., & Badard, T. (2023). Investigating the Impact of Point Cloud Density on Semantic Segmentation Performance Using Virtual Lidar in Boreal Forest. *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, 978–981. https://doi.org/10.1109/IGARSS52108.2023.10282100

Strîmbu, V. F., Næsset, E., Ørka, H. O., Liski, J., Petersson, H., & Gobakken, T. (2023). Estimating Biomass and Soil Carbon Change at the Level of Forest Stands Using Repeated Forest Surveys Assisted by Airborne Laser Scanner Data. *Carbon Balance and Management*, *18*(1), 10. https://doi.org/10.1186/s13021-023-00222-4

Sun, C., Huang, C., Zhang, H., Chen, B., An, F., Wang, L., & Yun, T. (2022). Individual Tree Crown Segmentation and Crown Width Extraction From a Heightmap Derived From Aerial Laser Scanning Data Using a Deep Learning Framework. *Frontiers in Plant Science*, *13*.

Suvanto, A., & Maltamo, M. (2010). Using Mixed Estimation for Combining Airborne Laser Scanning Data in Two Different Forest Areas. *Silva Fennica*, *44*(1). https://doi.org/10.14214/sf.164

Tompalski, P., White, J. C., Coops, N. C., & Wulder, M. A. (2019). Demonstrating the Transferability of Forest Inventory Attribute Models Derived Using Airborne Laser Scanning Data. *Remote Sensing of Environment*, *227*, 110–124. https://doi.org/10.1016/j.rse.2019.04.006

Tomppo, E., Gschwantner, T., Lawrence, M., & McRoberts, R. E. (Eds.). (2010). *National Forest Inventories: Pathways for Common Reporting*. Springer Netherlands. https://doi.org/10.1007/978-90-481-3233-1

Tomppo, E., & Schadauer, K. (2012). Harmonization of National Forest Inventories in Europe: Advances under COST Action E43. *Forest Science*, *58*(3), 191–200. https://doi.org/10.5849/forsci.10-091

Tsui, O. W., Coops, N. C., Wulder, M. A., Marshall, P. L., & McCardle, A. (2012). Using Multi-Frequency Radar and Discrete-Return LiDAR Measurements to Estimate above-Ground Biomass and Biomass Components in a Coastal Temperate Forest. *ISPRS Journal of Photogrammetry and Remote Sensing*, *69*, 121–133. https://doi.org/10.1016/j.isprsjprs.2012.02.009

Turner-Skoff, J. B., & Cavender, N. (2019). The Benefits of Trees for Livable and Sustainable Communities. *PLANTS, PEOPLE, PLANET*, *1*(4), 323–335. https://doi.org/10.1002/ppp3.39

Vales, D. J. (1985). *Functional Relationships between Salal Understory and Forest Overstory* [Doctoral dissertation, University of British Columbia]. https://doi.org/10.14288/1.0075147

van Ewijk, K., Tompalski, P., Treitz, P., Coops, N. C., Woods, M., & Pitt, D. (2020). Transferability of ALS-Derived Forest Resource Inventory Attributes Between an Eastern and Western Canadian Boreal Forest Mixedwood Site. *Canadian Journal of Remote Sensing*, *46*(2), 214–236. https://doi.org/10.1080/07038992.2020.1769470

van Ewijk, K., Treitz, P. M., & Scott, N. A. (2011). Characterizing Forest Succession in Central Ontario Using Lidar-derived Indices. *Photogrammetric Engineering & Remote Sensing*, *77*(3), 261–269. https://doi.org/10.14358/PERS.77.3.261

van Lier, O. R., Luther, J. E., White, J. C., Fournier, R. A., & Côté, J.-F. (2021). Effect of Scan Angle on ALS Metrics and Area-Based Predictions of Forest Attributes for Balsam Fir Dominated Stands. *Forestry: An International Journal of Forest Research*, (cpab029). https://doi.org/10.1093/forestry/cpab029

Vega, C., Hamrouni, A., El Mokhtari, S., Morel, J., Bock, J., Renaud, J.-P., Bouvier, M., & Durrieu, S. (2014). PTrees: A point-based approach to forest tree extraction from lidar data. *International Journal of Applied Earth Observation and Geoinformation*, *33*, 98–108. https://doi.org/https://doi.org/10.1016/j.jag.2014.05.001

Vonderach, C., Kublin, E., Bösch, B., & Kändler, G. (2021). *rBDAT: Implementation of BDAT Tree Taper Fortran Functions* [R package version 0.9.8]. https://CRAN.R-project.org/package=rBDAT

Vorster, A. G., Evangelista, P. H., Stovall, A. E. L., & Ex, S. (2020). Variability and Uncertainty in Forest Biomass Estimates from the Tree to Landscape Scale: The Role of Allometric Equations. *Carbon Balance and Management*, *15*(1), 8. https://doi.org/10.1186/s13021-020-00143-6

Wang, L., Birt, A. G., Lafon, C. W., Cairns, D. M., Coulson, R. N., Tchakerian, M. D., Xi, W., Popescu, S. C., & Guldin, J. M. (2013). Computer-Based Synthetic Data to Assess the Tree Delineation Algorithm from Airborne LiDAR Survey. *GeoInformatica*, *17*(1), 35–61. https://doi.org/10.1007/s10707-011-0148-1

Waser, L. T., Fischer, C., Wang, Z., & Ginzler, C. (2015). Wall-to-Wall Forest Mapping Based on Digital Surface Models from Image-Based Point Clouds and a NFI Forest Definition. *Forests*, *6*(12), 4510–4528. https://doi.org/10.3390/f6124386

Weber, J., & Penn, J. (1995). Creation and Rendering of Realistic Trees. *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '95*, 119–128. https://doi.org/10.1145/218380.218427

Wehr, A., & Lohr, U. (1999). Airborne Laser Scanning—an Introduction and Overview. *ISPRS Journal of Photogrammetry and Remote Sensing*, *54*(2), 68–82. https://doi.org/10.1016/S0924-2716(99)00011-8

Weinacker, H., Koch, B., & Weinacker, R. (2004). TREESVIS: A software system for simultaneous ED-real-time visualisation of DTM, DSM, laser raw data, multispectral data, simple tree and building models. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, *36*, 90–95.

Weiser, H., Schäfer, J., Winiwarter, L., Krašovec, N., Fassnacht, F. E., & Höfle, B. (2022a). Individual Tree Point Clouds and Tree Measurements from Multi-Platform Laser Scanning in German Forests. *Earth System Science Data*, *14*(7), 2989–3012. https://doi.org/10.5194/essd-14-2989-2022

Weiser, H., Schäfer, J., Winiwarter, L., Krašovec, N., Seitz, C., Schimka, M., Anders, K., Baete, D., Braz, A. S., Brand, J., Debroize, D., Kuss, P., Martin, L. L., Mayer, A., Schrempp, T., Schwarz, L.-M., Ulrich, V., Fassnacht, F. E., & Höfle, B. (2022b). Terrestrial, UAV-borne, and airborne laser scanning point clouds of central European forest plots, Germany, with extracted individual trees and manual forest inventory measurements. https://doi.org/10.1594/PANGAEA.942856

Weiser, H., Winiwarter, L., Anders, K., Fassnacht, F. E., & Höfle, B. (2021). Opaque Voxel-Based Tree Models for Virtual Laser Scanning in Forestry Applications. *Remote Sensing of Environment*, *265*, 112641. https://doi.org/10.1016/j.rse.2021.112641

Wetzel, S., Swift, D. E., Burgess, D., & Robinson, C. (2011). Research in Canada's National Research Forests—Past, Present and Future. *Forest Ecology and Management*, *261*(5), 893–899. https://doi.org/10.1016/j.foreco.2010.03.020

White, J. C., Chen, H., Woods, M. E., Low, B., & Nasonova, S. (2019). The Petawawa Research Forest: Establishment of a Remote Sensing Supersite. *The Forestry Chronicle*, *95*(03), 149–156. https://doi.org/10.5558/tfc2019-024

White, J. C., Coops, N. C., Wulder, M. A., Vastaranta, M., Hilker, T., & Tompalski, P. (2016). Remote Sensing Technologies for Enhancing Forest Inventories: A Review. *Canadian Journal of Remote Sensing*, *42*(5), 619–641. https://doi.org/10.1080/07038992.2016.1207484

White, J. C., Penner, M., & Woods, M. (2021). Assessing Single Photon LiDAR for Operational Implementation of an Enhanced Forest Inventory in Diverse Mixedwood Forests. *The Forestry Chronicle*, *97*(01), 78–96. https://doi.org/10.5558/tfc2021-009

White, J. C., Wulder, M. A., Varhola, A., Vastaranta, M., Coops, N. C., Cook, B. D., Pitt, D., & Woods, M. (2013a). A Best Practices Guide for Generating Forest Inventory Attributes from Airborne Laser Scanning Data Using an Area-Based Approach. *The Forestry Chronicle*, *89*(06), 722–723. https://doi.org/10.5558/tfc2013-132

White, J. C., Wulder, M. A., Vastaranta, M., Coops, N. C., Pitt, D., & Woods, M. (2013b). The Utility of Image-Based Point Clouds for Forest Inventory: A Comparison with Airborne Laser Scanning. *Forests*, *4*(3), 518–536. https://doi.org/10.3390/f4030518

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Windrim, L., & Bryson, M. (2020). Detection, Segmentation, and Model Fitting of Individual Tree Stems from Airborne Laser Scanning of Forests Using Deep Learning. *Remote Sensing*, *12*(9), 1469. https://doi.org/10.3390/rs12091469

Wing, B. M., Ritchie, M. W., Boston, K., Cohen, W. B., Gitelman, A., & Olsen, M. J. (2012). Prediction of Understory Vegetation Cover with Airborne Lidar in an Interior Ponderosa Pine Forest. *Remote Sensing of Environment*, *124*, 730–741. https://doi.org/10.1016/j.rse.2012.06.024

Winiwarter, L., Esmorís Pena, A. M., Weiser, H., Anders, K., Martínez Sánchez, J., Searle, M., & Höfle, B. (2022). Virtual Laser Scanning with HELIOS++: A Novel Take on Ray Tracing-Based Simulation of Topographic Full-Waveform 3D Laser Scanning. *Remote Sensing of Environment*, *269*, 112772. https://doi.org/10.1016/j.rse.2021.112772

Wu, J., Zou, D., Braverman, V., Gu, Q., & Kakade, S. M. (2022, August). The Power and Limitation of Pretraining-Finetuning for Linear Regression under Covariate Shift [Preprint available on arXiv. https://doi.org/10.48550/arXiv.2208.01857].

Wulder, M. A., White, J. C., Nelson, R. F., Næsset, E., Ørka, H. O., Coops, N. C., Hilker, T., Bater, C. W., & Gobakken, T. (2012). Lidar Sampling for Large-Area Forest Characterization: A Review. *Remote Sensing of Environment*, *121*, 196–209. https://doi.org/10.1016/j.rse.2012.02.001

Yang, T.-R., Kershaw, J. A., Jr., Weiskittel, A. R., Lam, T. Y., & McGarrigle, E. (2019). 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*, *92*(3), 311–323. https://doi.org/10.1093/forestry/cpz014

Yin, T., Lauret, N., & Gastellu-Etchegorry, J.-P. (2016). Simulation of Satellite, Airborne and Terrestrial LiDAR with DART (II): ALS and TLS Multi-Pulse Acquisitions, Photon Counting, and Solar Noise. *Remote Sensing of Environment*, *184*, 454–468. https://doi.org/10.1016/j.rse.2016.07.009

Zavitkovski, J. (1976). Ground Vegetation Biomass, Production, and Efficiency of Energy Utilization in Some Northern Wisconsin Forest Ecosystems. *Ecology*, *57*(4), 694–706. https://doi.org/10.2307/1936183

Zhao, K., Popescu, S., & Nelson, R. (2009). Lidar Remote Sensing of Forest Biomass: A Scale-Invariant Estimation Approach Using Airborne Lasers. *Remote Sensing of Environment*, *113*(1), 182–196. https://doi.org/10.1016/j.rse.2008.09.009

Zhao, K., Suarez, J. C., Garcia, M., Hu, T., Wang, C., & Londo, A. (2018). Utility of Multitemporal Lidar for Forest and Carbon Monitoring: Tree Growth, Biomass Dynamics, and Carbon Flux. *Remote Sensing of Environment*, *204*, 883–897. https://doi.org/10.1016/j.rse.2017.09.007

Zhu, X., Liu, J., Skidmore, A. K., Premier, J., & Heurich, M. (2020). A Voxel Matching Method for Effective Leaf Area Index Estimation in Temperate Deciduous Forests from Leaf-on and Leaf-off Airborne LiDAR Data. *Remote Sensing of Environment*, *240*, 111696. https://doi.org/10.1016/j.rse.2020.111696

Zianis, D., Muukkonen, P., Mäkipää, R., & Mencuccini, M. (2005). Biomass and Stem Volume Equations for Tree Species in Europe. *Silva Fennica Monographs*, *2005*(4), 1–63. https://doi.org/10.14214/sf.sfm4

Zolkos, S. G., Goetz, S. J., & Dubayah, R. (2013). A Meta-Analysis of Terrestrial Aboveground Biomass Estimation Using Lidar Remote Sensing. *Remote Sensing of Environment*, *128*, 289–298. https://doi.org/10.1016/j.rse.2012.10.017

Eidesstattliche Versicherung

Eidesstattliche Versicherung gemäß § 13 Absatz 2 Satz 1 Ziffer 4 der Promotionsordnung des Karlsruher Instituts für Technologie (KIT) für die KIT-Fakultät für Bauingenieur-, Geo- und Umweltwissenschaften

1. Bei der eingereichten Dissertation zu dem Thema **"Using Synthetic Datasets to predict Forest Aboveground Biomass from Airborne Laser Scanning Data"** handelt es sich um meine eigenständig erbrachte Leistung.
2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.
3. Die Arbeit oder Teile davon habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.
4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.
5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt.

Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Karlsruhe, den 09. April 2024