ENVIRONMENTAL RESEARCH
LETTERS



OPEN ACCESS

RECEIVED
16 August 2021

REVISED
10 November 2021

ACCEPTED FOR PUBLICATION
24 November 2021

PUBLISHED
14 December 2021

Original Content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.



LETTER

Challenges to aboveground biomass prediction from waveform lidar

Jamis M Bruening[1,*], Rico Fischer[2], Friedrich J Bohn[2], John Armston[1], Amanda H Armstrong[3,4], Nikolai Knapp[2], Hao Tang[1,5], Andreas Huth[2,6,7] and Ralph Dubayah[1]

[1] Department of Geographical Sciences, University of Maryland, College Park, MD 20740, United States of America
[2] Department of Ecological Modeling, Helmholtz Centre for Environmental Research (UFZ), 04318 Leipzig, Germany
[3] Department of Environmental Sciences, University of Virginia, Clark Hall, Charlottesville, VA 22902, United States of America
[4] Universities Space Research Association, Goddard Earth Sciences Technology and Research Studies and Investigations, NASA Goddard Space Flight Center, 8800 Greenbelt Road, Greenbelt, MD 20771, United States of America
[5] Department of Geography, National University of Singapore, Kent Ridge 117570, Singapore
[6] German Centre for Integrative Biodiversity Research (iDiv), 04103 Halle-Leipzig-Jena, Germany
[7] Institute of Environmental Systems Research, University Osnabrück, 49076 Osnabrück, Germany
[*] Author to whom any correspondence should be addressed.

E-mail: jamis@umd.edu

Keywords: GEDI, waveform lidar, forest modeling, aboveground biomass, forest structure

Supplementary material for this article is available online

Abstract

Accurate accounting of aboveground biomass density (AGBD) is crucial for carbon cycle, biodiversity, and climate change science. The Global Ecosystem Dynamics Investigation (GEDI), which maps global AGBD from waveform lidar, is the first of a new generation of Earth observation missions designed to improve carbon accounting. This paper explores the possibility that lidar waveforms may not be unique to AGBD—that forest stands with different AGBD may produce highly similar waveforms—and we hypothesize that non-uniqueness may contribute to the large uncertainties in AGBD predictions. Our analysis integrates simulated GEDI waveforms from 428 *in situ* stem maps with output from an individual-based forest gap model, which we use to generate a database of potential forest stands and simulate GEDI waveforms from those stands. We use this database to predict the AGBD of the 428 *in situ* stem maps via two different methods: a linear regression from waveform metrics, and a waveform-matching approach that accounts for waveform-AGBD non-uniqueness. We find that some *in situ* waveforms are more unique to AGBD than others, which notably impacts AGBD prediction uncertainty (7–411 Mg ha^{-1}, average of 167 Mg ha^{-1}). We also find that forest structure complexity may influence the non-uniqueness effect; stands with low structural complexity are more unique to AGBD than more mature stands with multiple cohorts and canopy layers. These findings suggest that the non-uniqueness phenomena may be introduced by the measuring characteristics of waveform lidar in combination with how forest structure manifests at small scales, and we discuss how this complexity may complicate uncertainty estimation in AGBD prediction. This analysis suggests a limit to the accuracy and precision of AGBD predictions from lidar waveforms seen in empirical studies, and underscores the need for further exploration of the relationships between lidar remote sensing measurements, forest structure, and AGBD.

1. Introduction

One of the most pressing and open questions in climate science is the extent to which forests will act as a net sink or source of carbon in the short- and mid-term future [1, 2]. Accurate baselines of aboveground biomass density (AGBD) within forests are crucial to answering this question, making high resolution mapping of AGBD an immediate need. The Global Ecosystem Dynamics Investigation (GEDI) is the first spaceborne lidar mission specifically intended to map global carbon stocks [3]. GEDI was launched to

© 2021 The Author(s). Published by IOP Publishing Ltd

the International Space Station in December 2018 and provides waveform lidar measurements of forest structure within footprints 25 m in diameter that are used to predict aboveground biomass. Over the course of its prime mission (April 2019–April 2021), GEDI recorded over 10 billion land surface observations within the ISS's orbital extent between 51.5 °N and °S latitude.

The GEDI instrument transmits pulses of light energy towards the Earth's surface and records the intensity of returned energy over time to produce a vertical waveform. Photons reflected by the top of the forest canopy surface are returned to the sensor sooner than others that are reflected lower down in the canopy or by the ground, and more vegetation matter ata given canopy height will yield a larger waveform amplitude at that height. Relative height (RH) metrics are variables derived from the waveform that give the height above the ground at which a certain quantile of returned energy is reached. These metrics are correlated with AGBD [4] and are used as predictors in GEDI's biomass models [3, 5].

Although waveform lidar has proven to be effective for estimating biomass, there remains uncertainty about the accuracies achievable at sub-hectare resolutions, and normalized calibration errors (nRMSE) between 40% and 50% are common [6]. Others have suggested various sources of this error, such as geolocation errors [7, 8], tree crowns that overhang plot boundaries [9], errors and uncertainty in allometric equations [10–12] and differences in environmental and edaphic conditions [13].

Entirely separate from these is the issue of waveform uniqueness with respect to AGBD. Here, we define 'waveform non-uniqueness' as the possibility that a specific waveform shape may be associated with substantially different AGBD values, and 'waveform-AGBD uncertainty' as the likely range of AGBD associated with a specific waveform shape. In other words, the extent of a waveform's non-uniqueness impacts the magnitude of its AGBD uncertainty. In light of this possibility a fundamental question arises: can different configurations of trees and their arrangement spatially and vertically yield highly similar waveforms, yet have very different AGBD?

It is therefore of considerable interest to better understand how uniquely waveforms relate to AGBD. To what extent does non-uniqueness occur and under what conditions? What are the implications for instruments such as GEDI, which rely on the assumption that calibration equations convert waveform metrics to AGBD in an unbiased and accurate fashion? To answer these questions we apply a model-data fusion concept using FORMIND, an individual-based gap model [14] and the integration of simulated GEDI waveforms with the FORMIND simulations. Individual-based gap models, such as FORMIND, are a powerful tool to interpret remote sensing observations ecologically, as they allow a direct link between a patch of forest (real or simulated) and how it may appear to a remote sensing instrument [15]. By using a large number of simulations in conjunction with *in situ* plots, this link can be studied across a wide range of real world conditions. As such, our modeling framework bridges the gap between spaceborne and ground-based estimates of AGBD, and explores the relationship between AGBD and lidar waveforms in a controlled and systematic manner.

In this paper we examine the issue of waveform non-uniqueness with respect to temperate forest AGBD within the Northeast USA. Our objectives are to; (a) quantify the extent to which waveforms can be non-unique with respect to AGBD within 400 m^2 plots; (b) explain possible causes of waveform-AGBD non-uniqueness; and (c) assess the implications of our findings on efforts to predict AGBD from waveform lidar. Through GEDI-FORMIND fusion we employ two distinct methods to estimate AGBD from GEDI waveforms and in doing so characterize the uniqueness of GEDI waveforms with respect to AGBD. We compare the predictions and uncertainties from both methods, and relate patterns in derived AGBD uncertainty to forest stand attributes. Finally, we discuss the relevance of these results on AGBD prediction from lidar waveforms and the possible causes of waveform-AGBD non-uniqueness.

2. Methods and data

Our methodological approach (figure 1) is a fusion between GEDI waveforms and the FORMIND model. It uses forest simulations to gain new insights into the relationship between GEDI waveforms and AGBD across a network of 428 field sites throughout the northeast US. We calibrated FORMIND to simulate a database of 896 620 potential forest stands that could exist throughout the Northeast US, using an implementation mode called the FF [16]. We then generated lidar point clouds for every simulated FF stand, and simulated GEDI waveforms from each point cloud using a waveform simulator [17]. The result was a database of 896 620 GEDI waveforms from which we developed two different methods of predicting AGBD within *in situ* forest plots. The first is an ordinary least squares (OLS) linear regression model to predict AGBD from lidar waveform RH metrics. The second is a lookup table approach called waveform matching (WFM). For each *in situ* observation, the WFM algorithm identifies the set of the 100 most similar waveforms from the FF database, and infers the AGBD of the *in situ* observation from the distribution of AGBD from those 100 FF stands. We applied both methods (OLS and WFM) to predict the AGBD of 428 *in situ* stem maps from those stands' GEDI waveforms, compared the predictions to one another and to the observed AGBD calculated directly from the inventory data, and contrasted the uncertainties

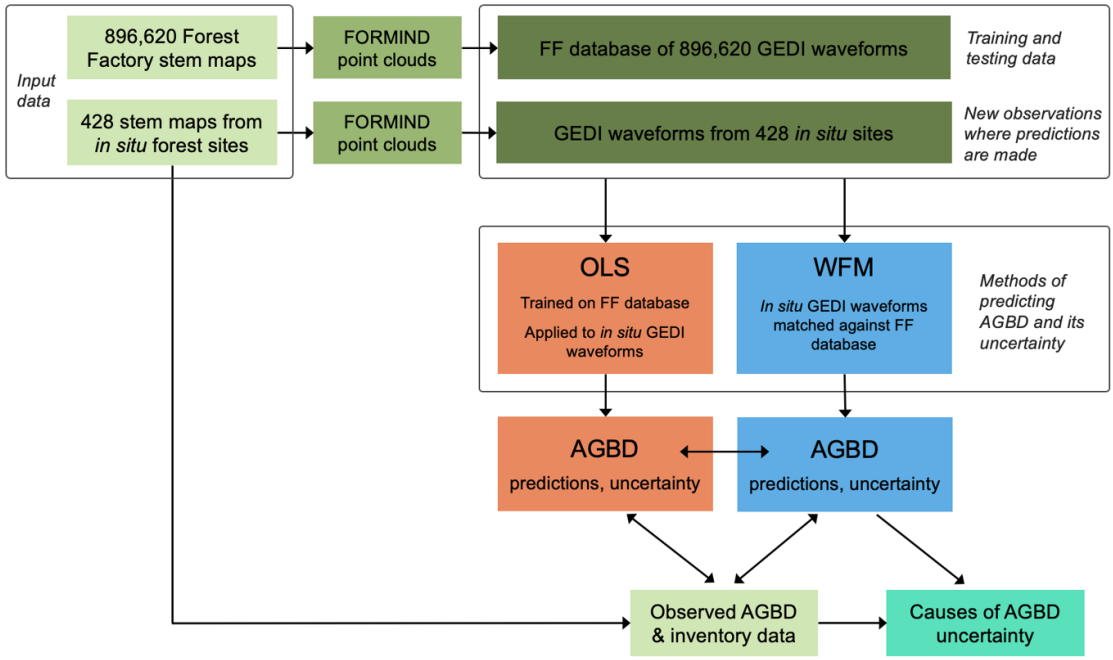


Figure 1. This workflow integrates GEDI waveforms and FORMIND using a waveform simulator to enable fusion between waveforms from *in situ* forest stands and those simulated by FORMIND's Forest Factory (FF). The two rows of green boxes at the top represent the *in situ* and FORMIND-simulated stem maps, point clouds, and GEDI waveforms that serve as the input data for the analysis. The FF database of 896 620 GEDI waveforms is used to predict the AGBD of 428 *in situ* stem maps using two different methods, an OLS regression (orange) and the WFM lookup table approach (blue). We quantify the uncertainty associated with each set of predictions, and compare both sets of predictions and uncertainties to each other and to the observed AGBD calculated from the *in situ* inventory data. Lastly, we identify the potential causes of WFM-derived AGBD prediction uncertainty as a function of stand dynamics and structural complexity within temperate forests.

from these two methods. Finally, we performed a regression tree analysis to explore the relationship between the WFM-derived AGBD prediction uncertainty and forest structure.

The following sections provide an overview of our workflow, and more details are provided in the accompanying supplementary information (available online at stacks.iop.org/ERL/16/125013/mmedia).

2.1. Field sites

The Northeast US (figure 2) was chosen for this study due to the high availability of inventory data (table 1), the large range of potential AGBD values [22], and importance to the global carbon budget through secondary forest regrowth [23, 24]. We use 428 stem mapped forest plots that come from five different research areas, four of which are projects within the GEDI Forest Structure and Biomass Database [3]. Plot size and shape varied across the five areas, which necessitated standardization to a common square plot shape of 20 m × 20 m (such as clipping larger plots to conform to the smaller square shape), however the relative spatial position of all trees within the plot was preserved. Trees greater than 12.7 cm (5 in) in diameter at breast height were measured and their locations recorded within the plot.

2.2. FORMIND and the FF

FORMIND is a forest gap model that simulates growth dynamics at the level of individual trees [14]. It allows simulation of forest dynamics and structure, including gap formation (falling down of large trees) and succession. FORMIND simulates all physiological processes (photosynthesis, respiration, tree growth, mortality, regeneration, competition) at the tree level. Growth of a single tree is calculated on the basis of a carbon balance and depends on the tree size, climate conditions and the shading of surrounding trees. Forests are represented as a collection of square patches (stands) which may vary in successional and structural stage. The size of these stands correspond with the light competition range of trees, which in this case is 400 m^2 [14].

We calibrated a regional version of FORMIND to represent forest composition and structure characteristic of the Northeast US, using all tree-level data from the US Forest Service's Forest Inventory and Analysis (FIA) program's most recent survey of the Northeast USA. We segmented the region's 27 most abundant species into nine different plant function types (PFT) based on each species' maximum size (height and stem diameter), growth rate, and shade tolerance, and then calibrated a set of tree geometry equations for each PFT so that forest stands simulated in FORMIND represent generalized structural characteristics of forests in this region (table S1). We then validated these parameters against the inventory data from the *in situ* field sites.

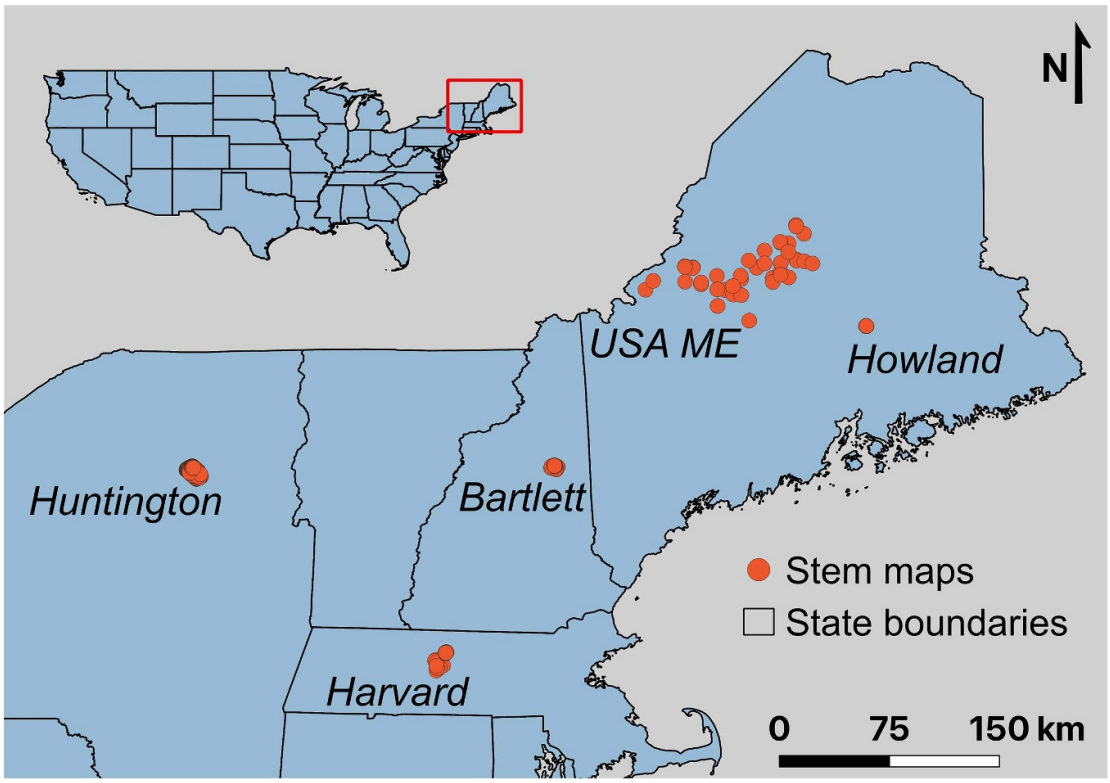


Figure 2. Our study uses five forest research sites and 428 stem map locations distributed throughout the northeast US.

Table 1. Site information for the five forest research sites which contribute stem maps to this analysis. *N* refers to the number of stem maps at each site, 'GEDI' indicates whether the site is part of the GEDI Forest Structure and Biomass Database [3], and 'YEAR' is the plot inventory year.

Abbr.	Site name	Approximate Lat Lon	*N*	Source	GEDI	Year
usme	USA ME	–70.02, 45.58	42	NASA CMS [18, 19]	Yes	2015
harv	Harvard Forest	–72.18, 42.53	35	NEON [20]	Yes	2015–2017
bart	Bartlett Forest	–71.29, 44.05	25	NEON [20]	Yes	2016–2018
howl	Howland Research Forest	–68.73, 45.20	70	Univ. of Maine	Yes	2015
hunt	Huntington Wildlife Forest	–74.22, 43.97	256	SUNY ESF [21]	No	2011

FORMIND has an implementation mode called the Forest Factory (FF), which simulates the structure (not growth) of many unique, 400 m^2 forest stands [16, 25]. The purpose of the FF is to simulate the diversity in forest structure within a region, based on varying the PFT compositions and stem-size distributions across a large number of simulated forest stands. We implemented the FF using the Northeast US calibration of FORMIND to simulate a structural diversity database containing 896 620 unique forest stands that could exist throughout the region. This number of simulations ensured the database covered a wide range of structural and compositional diversity, given the potential occurrence of up to nine different PFTs within a stand, and various potential stem size distributions that have been observed across successional and structural gradients in temperate forests [16, 26]. This database serves as the basis for the OLS and WFM AGBD prediction methods, as it encompasses the myriad of forest stand structural configurations and AGBDs that are possible throughout the Northeast US.

2.3. Lidar simulations

This study is based on GEDI waveform comparisons between real and simulated forests, so it was necessary to standardize these data sources. This section explains our standardization approach to simulate GEDI waveforms for the 428 *in situ* stem maps and all 896 620 stem maps in the FF database.

First we input each *in situ* stem map into FORMIND using its initialization feature, and had the model construct the stand's structure according to the stem map and previously calibrated allometric relationships. We implemented FORMIND's lidar simulator to generate a point cloud representation of

each stem map, according to Knapp *et al* [27], and subsequently simulated a GEDI waveform from each stem map point cloud using the GEDI waveform simulator [17]. The waveform simulations mirror GEDI's calibration process (see Dubayah *et al* [3]), except here the waveform footprint width was set to 20 m to match the size of the 428 *in situ* and 896 620 FF stem maps. The resultant 428 GEDI waveforms represent FORMIND's rendering of forest structure within 428 real forest stands throughout the Northeast US.

We used almost the same process as above to simulate a GEDI waveform from every stand in the FF database. The only difference was an added step to eliminate edge effects between FF stands. The FF outputs 100 stands at once in a 10 × 10 grid, so we separated the stands when generating the lidar point clouds to eliminate the influence of tree crowns that overhang plot boundaries (see Knapp *et al* [9] and supplementary information for details).

The end result is that all 896 620 stem maps in the FF database have an AGBD value and a corresponding GEDI waveform, calculated in the same exact manner as the 428 field sites. If one of the FF stands were to actually exist somewhere in a forest and was inventoried, the AGBD value calculated from the inventory data would be same as the value in the FF database, and ingesting the stem map into FORMIND (as was done for the 428 *in situ* stands) would produce a point cloud and simulated GEDI waveform identical to its counterpart in the FF database. This is crucial to our analysis because it ensures standardization between all data from the *in situ* stands and FF database. This approach removes allometric variability between tree species and across sites, because all trees in the stem maps (the 428 *in situ* and 896 620 FF database) were constructed in our Northeast US regional parameterization of FORMIND.

2.4. Deriving AGBD and its uncertainty from lidar waveforms

We use two methods to predict forest stand AGBD and estimate prediction uncertainty from GEDI waveforms: ordinary least squares regression (OLS), and a lookup table approach we call waveform matching (WFM). Both methods rely on the 896 620 GEDI waveforms from the FF database to predict the AGBD and uncertainty of each of the 428 *in situ* forest stands from their respective GEDI waveforms.

2.4.1. OLS regression modeling

We used an ordinary least squares (OLS) linear regression model to predict AGBD from GEDI waveform RH metrics in 10% increments from RH10 to RH90, with the addition of RH98 (a more stable indicator of top of canopy height than RH100). We developed a set of 18 candidate models based on relevant literature, using a square root transform on the response [5, 28]. We trained each model on half of the FF database (n = 448 310) so that the AGBD models were derived from the same simulated forest stands used in the WFM approach (section 2.4.2). To assess model performance, we randomly split the remaining half of the FF database not used for training into 500 different testing sets (each with n = 897), and applied all 18 candidate models to all 500 testing sets. We selected the final model (table S2, figure S1) based on the lowest average RMSE across the 500 tests, and applied it to the 428 *in situ* stem map GEDI waveforms to generate OLS-based predictions of AGBD and the associated 90% prediction interval for each site.

2.4.2. Waveform matching

WFM is a process that quantifies the similarity in shape between two waveforms [29, 30]. Similarity is defined as the relative overlapping area between two waveforms, expressed as the ratio of area shared by both waveforms to the entire area encompassed by either waveform, as follows

$$r = \sum_{h=0}^{mch} \min(E_x(h), E_y(h)) \left(\sum_{h=0}^{mch} \max(E_x(h), E_y(h) \right)^{-1} \quad (1)$$

where r is the relative overlap, h is height above ground in meters, mch is the maximum canopy height between the waveforms, and E_x and E_y are the returned waveform energies at a given value of h from the stem map and FF-stand waveforms, respectively. We used the same height step of 0.15 m as the waveform simulator [17].

We applied the WFM algorithm individually to the 428 stem map waveforms, and identified the 100 FF-stand waveforms with the largest r for each. When an *in situ* stand did not have 100 FF matches with $r > 0.75$, we removed it from the rest of the analysis to guarantee a high degree of similarity between each *in situ* waveform and its FF matches. The result was a set of 100 best matching FF stands for each *in situ* stem mapped stand, based on waveform shape.

To derive the *in situ* stand's WFM-predicted AGBD value, we use the median AGBD value from the set of 100 FF matches. We represent the WFM AGBD prediction uncertainty as the range in AGBD that encompasses the middle 90% of the 100 FF AGBD values. We define this range as the magnitude of uncertainty in AGBD associated with a stem map's waveform, which represents the extent of non-uniqueness with respect to AGBD for each stem map waveform.

2.5. Explaining AGBD uncertainty

WFM yields predicted values of AGBD for each stem map, as well as the uncertainty around each estimate. We performed a regression tree analysis, using the *rpart* package in R [31], to identify the extent to which forest structure variables explain patterns in the WFM-derived AGBD uncertainty across the

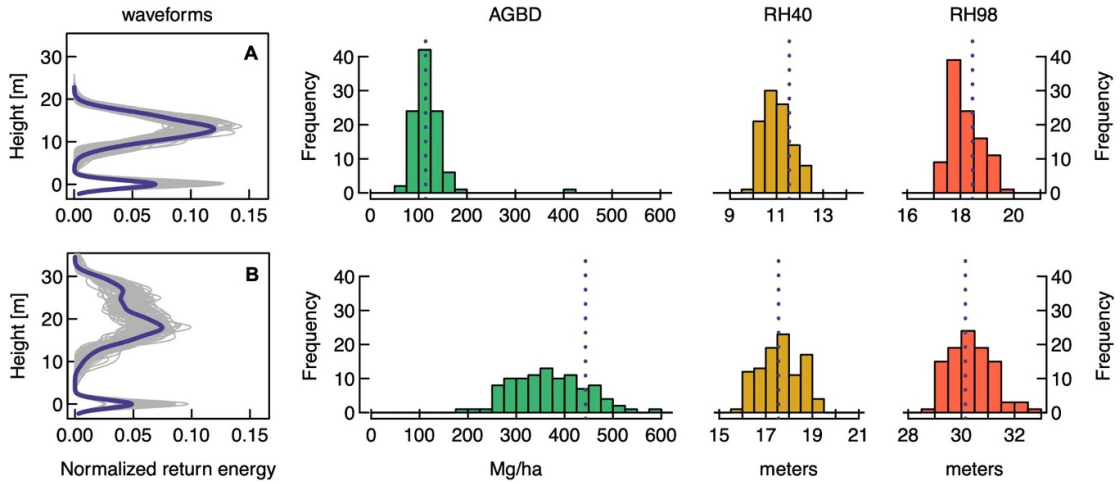


Figure 3. The distributions of AGBD (green) from the 100 matching waveforms (grey lines) for two example stem map waveforms (blue lines) are highly variable. Waveform A has less variability in AGBD across its 100 matches than waveform B. Waveform B has a larger uncertainty in WFM-derived AGBD than waveform A, despite similar variability in RH40 and RH98 across the 100 matches for these two examples. According to the WFM method, waveform A is more unique to AGBD than waveform B. The vertical dotted lines represent the *in situ* stem mapped stand's value for each variable.

in situ stem maps. The explanatory variables used were waveform entropy and skewness, the standard deviation in tree heights, standard deviation of tree diameters, and stand basal area. To ensure a simple and interpretable model, we set the maximum tree depth to three, and only allowed a split from nodes with at least 15% of the total sample. This means the final model could have had a maximum of seven possible splits, and eight possible classes, and that a split was not allowed if a node had less than 15% of the total observations.

3. Results

3.1. Stem map waveform matches

Waveform matching (WFM) revealed 100 best matches with $r > 0.75$ for 380 of the 428 stem maps (figure S2). The distribution of biomass across these 380 *in situ* stands (figure S4A) is very similar to the distribution of biomass across all 38 000 FF stands that were identified as a match to the 380 *in situ* stands (figure S4B). Across these 380 sites, the WFM-derived uncertainty in AGBD ranged from 7 to 411 Mg ha^{-1} (mean of 167 Mg ha^{-1}), and from 0.56 to 5.57 (mean of 1.29) as a ratio relative to the WFM-predicted AGBD. There was considerable variability in the WFM-derived distributions of AGBD associated with different waveform shapes (e.g. figure 3), as some waveforms had a larger range of possible AGBD values than others (figure S3).

3.2. AGBD predictions and uncertainty

The OLS and WFM methods performed similarly with respect to prediction accuracy (R^2 of 0.64 and 0.66 respectively) and overall error (nRMSE of 32.1% and 31.6% respectively) (figures 4(A) and (B)). The agreement between each set of predictions was higher than between either set and the observed values, and there was substantial agreement between the residuals (figures 4(C) and (D)). However, the uncertainty associated with WFM- and OLS-predicted AGBD were different (vertical lines in figures 4(A) and (B)). Across these sites, AGBD uncertainty was lower from WFM than from OLS, despite a larger degree of variability and some extremely large values in the WFM prediction uncertainty (figure 5). When the AGBD prediction uncertainty from each method was compared directly, and the magnitude of uncertainty from WFM was lower than from OLS at 313 of the 380 sites with 100 FF matches.

3.3. WFM AGBD uncertainty as a function of forest stand attributes

The regression tree analysis ($R^2 = 0.4$) partitioned the stem maps into five groups of increasing AGBD uncertainty (derived from WFM) (figure 6). Total stand basal area had the most explanatory power with respect to predicting the AGBD uncertainty, as stands with low basal area (<10 m^2 ha^{-1}) tended to have a smaller uncertainty than stands with a larger basal area. Among stands with basal area >10 m^2 ha^{-1}, those with more variation in tree height (height standard deviation >4 m) tended to have larger AGBD uncertainty than stands with less variation in tree height. Lastly, within each resulting group (both above and below height standard deviation of four meters), stands with more variation in tree diameter tended to have larger AGBD uncertainty than stands with less variation in tree diameter. Forest attributes differed substantially across the five groups (figure 7).

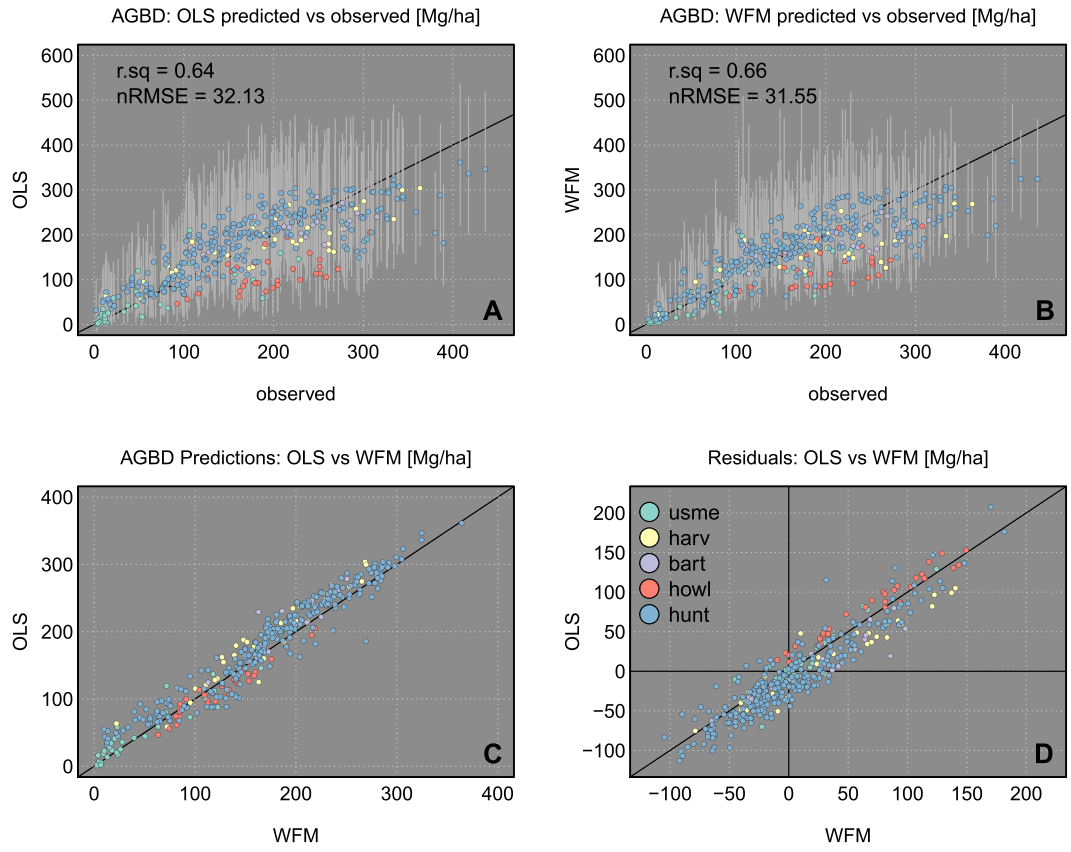


Figure 4. Ordinary least squares (OLS) (A) and waveform matching (WFM) (B) perform similarly in AGBD prediction compared to the observed values. There is strong agreement between the predicted values from both methods (C) and the residuals (D), despite differences in reported uncertainty (grey lines in A and B).

4. Discussion

The outcome of the WFM analysis indicates that lidar waveforms are not unique to AGBD, but instead are associated with a distribution of possible AGBDs (e.g. figure 3), and the median of this distribution tends to be a reliable predictor of the observed AGBD within the waveform footprint (figure 4(B)) when compared to the OLS predictions. Additionally, the WFM-derived distributions of AGBD associated with different waveform shapes are themselves variable (e.g. figure 3). Thus some waveforms are associated with a wider range of possible AGBD than others (figure 5). Our results also support that a specific value of AGBD can be associated with multiple different waveform shapes. Two stands with different tree configurations may have the same AGBD, yet these stands would produce considerably different waveforms (figure 8). This is evidenced by the overlapping AGBD uncertainties for different waveforms (figures 4(A) and (B)). AGBD is not unique to a specific waveform shape, perhaps because multiple pathways of forest stand development can lead to a given AGBD. In the following sections we discuss how waveform-AGBD non-uniqueness may present limitations in AGBD prediction from lidar, and we provide a possible explanation for the observed pattern in WFM-derived AGBD uncertainty across the stem map field sites used in this analysis.

4.1. Limits to AGBD prediction and uncertainty estimation from lidar

A consistent and perplexing issue associated with modeling AGBD is the heteroscedastic predictions and high nRMSEs in AGBD calibration equations for small plot sizes [6]. While the community has searched for solutions to this problem through the inclusion of more and more complex metrics [27, 32] and fusion with other remote sensing data [33, 34], the problem persists. Our results suggest the heteroscedasticity may not be solvable; rather it may be introduced by the measurement properties of waveform lidar at small scales.

Waveforms represent an aggregate measure of vertical structure throughout the footprint, reducing the information from multiple trees into a single observation. There is no horizontal differentiation within a waveform, as it represents the amount of combined plant matter at a given canopy elevation across the entire footprint area. Further, the laser pulse follows a Gaussian distribution within the footprint, so vegetation matter at the center of the

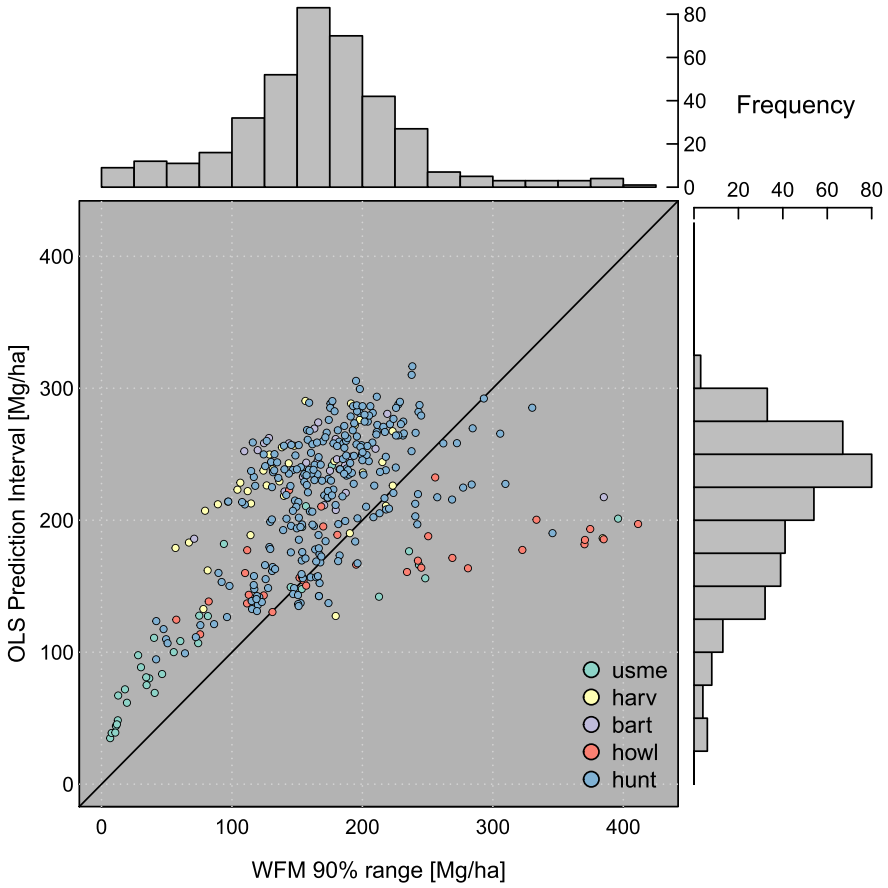


Figure 5. The WFM prediction uncertainty is typically lower than the OLS prediction uncertainty. The WFM prediction uncertainty is the range of AGBD that encompasses 90% of the 100 FF matches AGBD values for a given *in situ* waveform, and the OLS prediction uncertainty is the width of the 90% prediction interval for each prediction. The solid line shows the 1:1 relationship, and the bin width for both histograms is 25 Mg ha^{-1}.

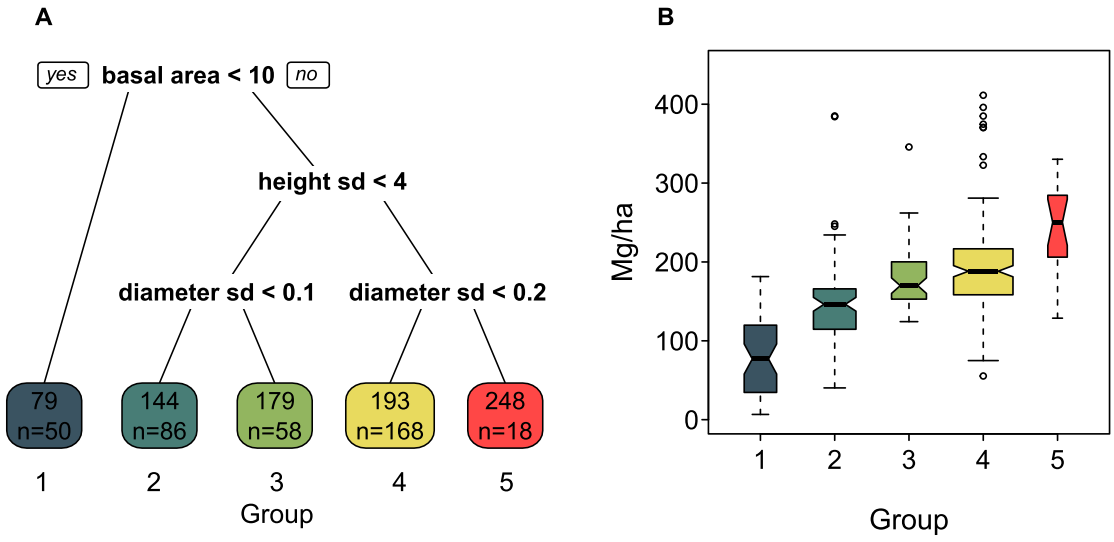


Figure 6. The AGBD prediction uncertainty derived from WFM is related to a stand's basal area (m^2 ha^{-1}) and heterogeneity in tree height, represented here as the standard deviation in tree height (m) and diameter (m). The regression tree (A) partitions the stem maps into groups based on the influence of various forest stand attributes on WFM-derived AGBD uncertainty (B). The top value in each terminal node of the tree is the group's mean AGBD uncertainty, and the bottom number is the number of stands in the group. The boxplot widths are weighted according to the square root of the number of observations in each group.

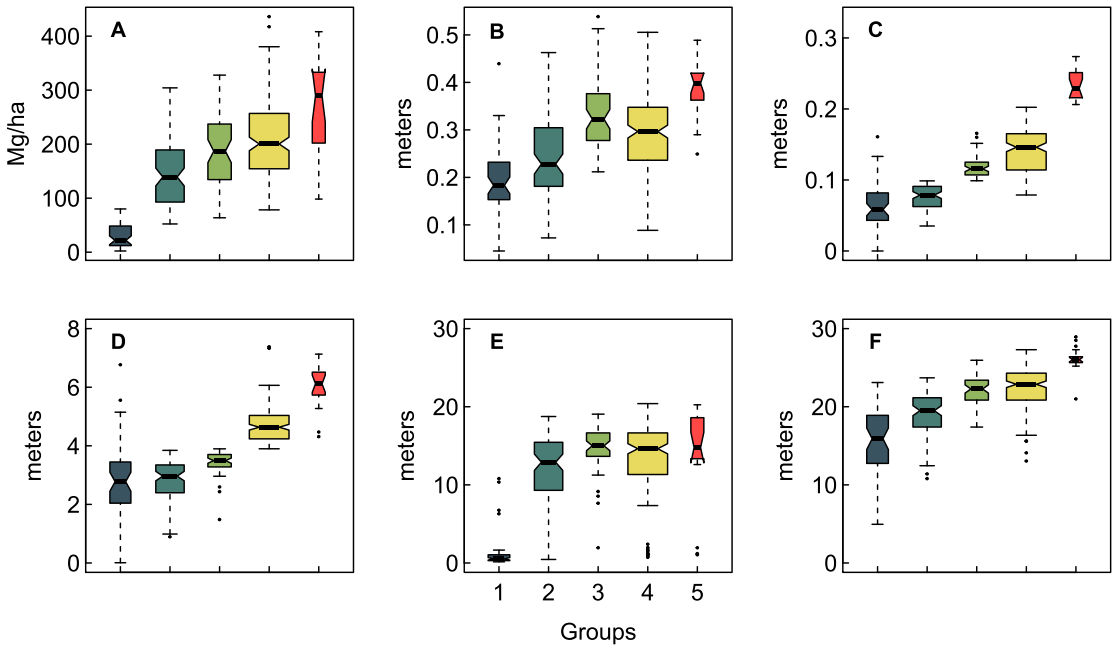


Figure 7. Forest stand attributes differ across the groups of increasing AGBD uncertainty. Panels are: (A) AGBD, (B) mean tree diameter, standard deviation of tree diameters (C) and heights (D), (E) RH50, (F) RH95. The boxes are colored according to group, to match figure 6.

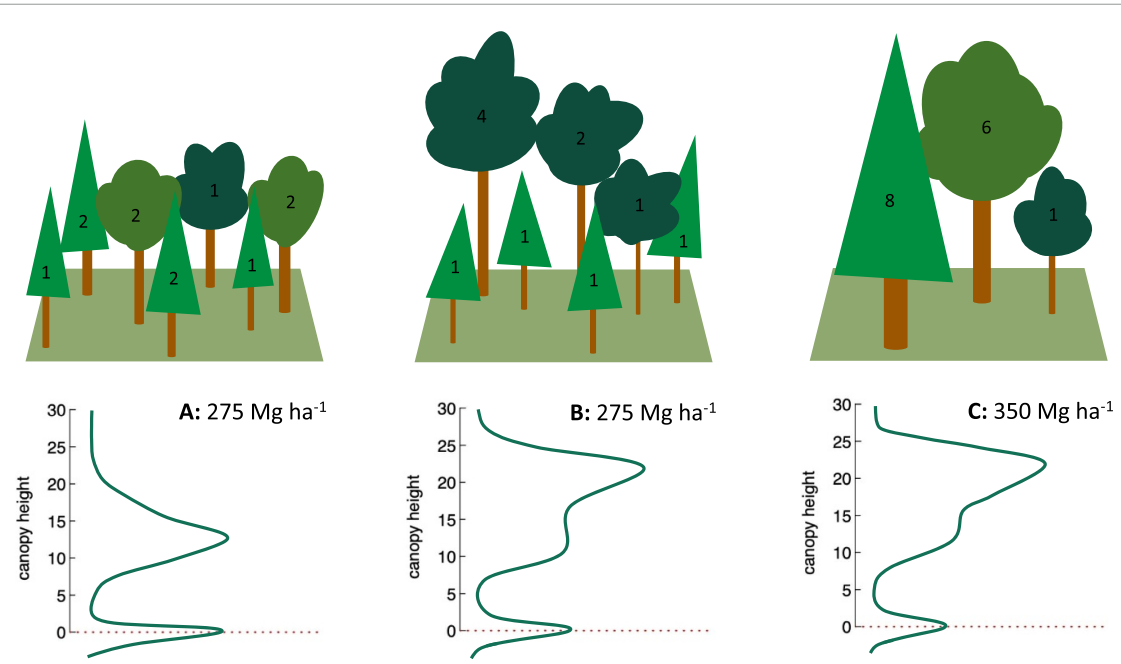


Figure 8. Forest stands (400 m^2) with unique tree configurations can have similar AGBD yet different waveforms (A) and (B), or different AGBD yet similar waveforms (B) and (C). The numbers in each tree represent its AGB Mg, and the reported densities here have been scaled to 1 ha. These stands are only conceptual, used to illustrate the non-uniqueness phenomenon, and not actual stands within our *in situ* or simulated databases.

footprint influences the waveform more than matter near the edges. We argue this aggregation across space may introduce the heteroscedasticity between waveform metrics and AGBD, and is more likely for small plot sizes. It is generally known that variability in AGBD increases with decreasing plot size, and that smaller area plots may also have much larger AGBDs than larger area plots. The scale-dependent variability in forest structure and biomass may explain why, at small scales, a specific waveform shape can be produced by stands with substantially different AGBDs. A decrease in AGBD variability in larger scales would decrease this effect, resulting in smaller ranges of AGBD associated with a given waveform shape. This would help explain the large decrease in nRMSEs observed in biomass models as calibration plot size

increases [6]. Indeed, Knapp *et al* [27] demonstrate that as plot sizes increase, the ability of lidar structural metrics to predict AGBD increases concurrently with a decrease in the variability in both AGBD and structural metrics. In effect, the uniqueness between LiDAR metrics and AGBD may increase with plot size.

The non-uniqueness effect may complicate efforts to estimate the uncertainty of AGBD predictions from statistical models. Uncertainty in regression models is often represented as a prediction interval, defined as the range in which the predicted value from a new observation will fall given what has been observed in the model training sample, for any desired probability. In the case of OLS, the size of the prediction interval for a new observation depends on the training sample mean and variance, and how far from the mean that new observation is. While statistically valid, prediction intervals do not recognize that some lidar waveforms may be more unique to AGBD than others. As such, the OLS prediction interval measures something inherently different than the WFM-derived AGBD uncertainty (the 90% range of possible AGBD associated with the 100 matches to a given stem map's waveform). The motivation for WFM is to quantify the variability in AGBD associated with a given waveform shape from a specific forest stand, absent all other influences, especially other forest stands' waveforms and AGBDs. In doing so this method accounts for the possibility that some waveforms may be associated with much larger ranges of AGBD than others, something not accounted for in OLS prediction intervals. This would help explain the differences between the OLS prediction intervals and WMF-derived uncertainty (figure 5).

In both WFM and OLS, a given waveform shape will always result in the same predicted value of AGBD—there is no stochasticity in either prediction method. However, the knowledge that multiple forest stands with different structures and waveform shapes can have the same AGBD (e.g. figure 8) highlights another important difference between OLS and WFM. In OLS, a predicted value of AGBD is unique to the set of predictor values used in the regression equation to make that prediction. In other words, the number of waveform shapes that could result in a specific predicted AGBD is directly related to the number of terms in the model. In a model that only uses RH98 to predict AGBD, a specific value of RH98 can only result in one predicted AGBD value, and that prediction is associated with a single prediction interval. However this is not the case in WFM, in which the prediction uncertainty is a function of the variability in AGBD associated with the matched set of waveforms.

4.2. Drivers of waveform-AGBD uncertainty

A conceptual model may characterize forest structure according to the interaction between three attributes: (a) the number and spatial position of trees within a stand (b) the sizes of the trees (e.g. maximum height, crown shape, stem diameter), and (c) the variation in tree sizes within the stand. A forest stand's structural stage can then be classified according to these attributes [22, 26, 35], and it has been shown that discrete airborne lidar observations can effectively discriminate between stands in various structural classes [36, 37]. Our results indicate that the extent to which lidar waveforms are unique to AGBD depends on the structural stage of the forest stand.

The regression tree (figure 6) segmented the *in situ* stands into in five groups of increasing WFM-derived AGBD uncertainty, and the magnitude of that uncertainty is related to a stands basal area and variation in tree sizes. Stands with larger basal area and more variation in tree height and diameter tend to produce waveforms that are less unique to biomass than stands with smaller basal area and more homogeneous tree sizes. Together, the five groups of forest stands represent a progression in structural complexity driven by dynamic forest processes [35].

Investigations of temperate forest stand dynamics have resulted in multiple classifications of forest stand structural development (e.g. [35, 38, 39]), however the general trends in structural development over time are consistent. Following a stand replacing disturbance or at the start of old-field succession, the stand-initiation phase is characteristic of open canopies with a single stratum of saplings, rapid growth, small basal area, and low biomass density [35]. The group with the lowest WFM-derived AGBD uncertainty embodies these characteristics (figures 6 and 7).

Upon stand initiation, most available growing space is quickly occupied and trees start to directly compete for resources during the stem exclusion phase [35]. Boles become larger and a dense canopy shades the forest floor, precluding new seedlings and perpetuating a single stratum of homogeneous tree sizes. As individuals in the overstory start to die, growing space becomes available and access to resources lower down in the canopy enables a transition to the understory reinitiation stage. Overtime, the canopy may stratify into various layers, and successive mortality events sustain a multi-layered canopy composed of different cohorts, stems of various sizes, and the potential for high biomass density in the multi-strata stage [40]. Groups 2–5 embody a general transition from young initiated stands through the various stages of structural development to more mature, mosaiced stands with established understories and multiple canopy strata.

The characteristics of these five groups illustrate how patterns of structural development within a forest may impact the uniqueness of waveforms with respect to AGBD. It appears that the ability of lidar waveforms to uniquely represent biomass within the footprint decreases as the structural complexity inside

the footprint increases. A waveform that represents a forest stand with a single, clearly defined canopy layer and moderate ground return tends to map to a small range of AGBD because the forest stands that could produce such a waveform must be composed of trees of similar sizes that form a single canopy layer, and those stands are similar in structure and AGBD (figure 3(A)). Conversely, a waveform that indicates multiple canopy strata and more canopy cover could theoretically come from a set of stands with larger compositional and structural variety, resulting in a relatively wider range of AGBD (figure 3(B)). Waveforms capture a finite amount of structural information about a stand, which is only a fraction of the stand's total structural information (e.g. maximum height, crown shape, and stem diameter of every tree in the footprint). We argue that this fraction likely varies based on the structural complexity of the observed stand, and in the context of AGBD prediction, the higher the fraction of structural information captured, the more unique a waveform may be to AGBD.

4.3. Forest simulations

The heavy reliance on simulated data in this analysis is intentional, as this experiment would not be possible using field data alone, and each stage of the workflow has been extensively tested and validated [14, 16, 17, 27, 30]. The use of FORMIND and the FF allows us to systematically explore the relationship between forest structure and AGBD across a wide range of structural conditions that exist in real forests, and to make novel inferences about AGBD prediction uncertainty. FORMIND captures general trends in forest structure across time and space, and it is not intended to predict the exact AGBD value of a single stand. Instead, we use it to assess the likely variability in AGBD predictions based on a given forest structure, rather than reporting our predicted AGBD values as truth. Additionally, FORMIND does not represent stochastic variability in tree allometry, or other differences due to environmental gradients or other factors known to influence forest structure (past disturbances, land use transitions, etc). As such the point cloud representation of the *in situ* forest stands do not exactly reproduce the structural complexity of these stands in the real world, as shrubs, downed trees and debris, and small trees are not represented in the stem maps or FORMIND-generated point clouds, nor are the potential impacts of topography represented in the point clouds or waveforms. However, all of these aspects are absent from both the *in situ* and FF stem maps and waveforms, ensuring a like-to-like comparison.

We have attempted to reproduce results common to empirical studies (increasing variation in predictions as AGBD increases and large calibration nRMSEs) in a modeling framework, and in doing so expose waveform-AGBD non-uniqueness as a contributing factor. Our experimental design allowed for a controlled setting to explore this possibility, while eliminating various proposed causes of heteroscedastic predictions and large calibration nRMSEs, as follows. An unrepresentative training sample is not likely a factor, as 896 620 FF stands were used to train and test the OLS model, and expanding the training sample would not act to reduce the non-uniqueness effect. There is no geolocation error between the FORMIND-generated point clouds and the *in situ* or FF-derived stem maps, nor do any of FF point clouds have tree crowns that overhang the plot boundaries, both of which have been suggested to add to calibration equation error and prediction scatter [7–9]. All the FF and *in situ* stands are constructed according to the same set of PFT-specific allometric equations, so there is no allometric variability within a PFT or between stands, and the same equations used to estimate the field biomass values of the stem maps were used to obtain the biomass of the simulated FF stands. We do not claim these factors have no influence on AGBD predictions from LiDAR waveforms in general, however by controlling for their influence in this analysis, we conclude that the non-uniqueness effect may account for a substantial amount of AGBD prediction uncertainty and error, and should not be overlooked in future studies.

5. Conclusions

In this paper we used GEDI-FORMIND fusion to explore the possibility that lidar waveforms are not unique to AGBD. Our results support that lidar waveforms may instead be associated with a range of potential AGBD values, and that this range varies among waveform shapes. We have demonstrated that within the study extent, the range of AGBD associated with specific a waveform may be a function of the stand's structural characteristics. Forest stands in early structural development tend to be relatively homogeneous and similar to one another, resulting in low biomass variation for a given structural signature (waveform shape). Over time, stand dynamics drive changes in forest composition and structure. This process of structural development over time may result in substantial differences in biomass between structurally mature stands, yet due to the measurement properties of waveform lidar, some stands may still produce similar waveform shapes. The result is that some waveform shapes are likely associated with a small range of possible AGBD, while others may be associated with a greater range of possible AGBD.

The phenomena of waveform non-uniqueness with respect to AGBD presents challenges in the context of traditional approaches to modeling AGBD from lidar waveforms. However, these challenges are not new and are well documented [6]. Acknowledging this phenomena may help explain the heteroscadasticity and large calibration errors

present in empirical studies, although further investigation into waveform-AGBD non-uniqueness across scales is necessary. The non-uniqueness effect is perhaps intuitive to some extent, however we have attempted to isolate its influence from other factors known to impact AGBD predictions. In doing so, this work highlights limitations to AGBD prediction from waveform lidar at fine scales.

Data availability statement

The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

This work was primarily supported by funding from the Global Ecosystem Dynamics Investigation and NASA FINESST Grant (80NSSC21K1626). The authors would like to thank Donal O'Leary and Steve Hancock for thoughtful discussions that improved the analysis, and Shawn Fraver for sharing the inventory data from Howland, ME. The authors are also appreciative of the comments from two anonymous reviewers that greatly contributed to the quality of this manuscript.

Authorship roles

J B: Conceptualization-equal, Methodology-lead, Data Curation, Investigation, Analysis, Validation-lead, Writing original draft and review & editing. R F: Conceptualization-equal, Methodology-supporting, Software-lead, Validation-supporting. F B: Methodology-supporting, Software-supporting, Validation-supporting. J A: Methodology-supporting, Supervision-supporting. A A: Methodology-supporting. N K: Software-supporting. H T: Methodology-supporting, Writing original draft-supporting. A H: Conceptualization-supporting, Methodology-supporting. R D: Conceptualization-equal, Supervision-lead, Writing original draft and review & editing.

Conflict of interest

The authors declare that there were no conflicts of interest.

ORCID iDs

Jamis M Bruening https://orcid.org/0000-0002-9750-7806
Rico Fischer https://orcid.org/0000-0002-0482-0095
Friedrich J Bohn https://orcid.org/0000-0002-7328-1187
John Armston https://orcid.org/0000-0003-1232-3424
Amanda H Armstrong https://orcid.org/0000-0002-9123-8924
Nikolai Knapp https://orcid.org/0000-0001-5065-9979
Hao Tang https://orcid.org/0000-0001-7935-5848
Ralph Dubayah https://orcid.org/0000-0003-1440-6346

References

[1] McDowell N G *et al* 2020 *Science* **368** eaaz9463
[2] Maréchaux I *et al* 2021 *Ecol. Evol.* **11** 3746–70
[3] Dubayah R *et al* 2020 *Sci. Remote Sens.* **1** 100002
[4] Drake J B, Dubayah R O, Knox R G, Clark D B and Blair J B 2002 *Remote Sens. Environ.* **81** 378–92
[5] Dubayah R, Armston J, Kellner J, Duncanson L, Hofton M, Blair J and Luthcke S 2021 Gedi l4a footprint level aboveground biomass density, version 1 (https://doi.org/10.3334/ORNLDAAC/1907)
[6] Zolkos S, Goetz S and Dubayah R 2013 *Remote Sens. Environ.* **128** 289–98
[7] Frazer G, Magnussen S, Wulder M and Niemann K 2011 *Remote Sens. Environ.* **115** 636–49
[8] Rejou-Mechain M *et al* 2014 *Biogeosciences* **11** 6827–40
[9] Knapp N, Huth A and Fischer R 2021 *Remote Sens.* **13** 1592
[10] Zhao F, Guo Q and Kelly M 2012 *Agric. Forest Meteorol.* **165** 64–72
[11] Ahmed R, Siqueira P, Hensley S and Bergen K 2013 *Remote Sens.* **5** 3007–36
[12] Vorster A G, Evangelista P H, Stovall A E and Ex S 2020 *Carbon Balance Manage.* **15** 1–20
[13] Shao G, Shao G, Gallion J, Saunders M R, Frankenberger J R and Fei S 2018 *Remote Sens. Environ.* **204** 872–82
[14] Fischer R *et al* 2016 *Ecol. Model.* **326** 124–33
[15] Shugart H H *et al* 2015 *Front. Ecol. Environ.* **13** 503–11
[16] Bohn F J and Huth A 2017 *R. Soc. Open Sci.* **4** 160521
[17] Hancock S, Armston J, Hofton M, Sun X, Tang H, Duncanson L I, Kellner J R and Dubayah R 2019 *Earth Space Sci.* **6** 294–310
[18] Cohen W, Yang Z, Healey S and Andersen H 2020 Disturbance history and forest biomass from landsat for six US sites, 1985-2014 (available at: https://doi.org/10.3334/ORNLDAAC/1679)
[19] Legner K, Andersen H E, Cooke A and Cohen W 2020 *Gen. Tech. Rep.* PNW-GTR-984. vol 66 (Portland, OR: US Department of Agriculture, Forest Service, Pacific Northwest Research Station) p 984
[20] National Ecological Observatory Network (NEON) 2021 Woody plant vegetation structure (dp1.10098.001) (available at: https://data.neonscience.org/data-products/DP1.10098.001/RELEASE-2021)
[21] Breitmeyer B W, Gooden M K, Appleby M J, Ash R and Rahn J 2019 Continuous forest inventory (CFI), 1970-2017, long-term forest property monitoring by state university of New York college of environmental science and forestry, New York, USA (available at: https://portal.edirepository.org/nis/mapbrowse?packageid=edi.410.1)
[22] Keeton W S, Whitman A A, McGee G C and Goodale C L 2011 *Forest Sci.* **57** 489–505
[23] Hurtt G, Pacala S W, Moorcroft P R, Caspersen J, Shevliakova E, Houghton R and Moore B 2002 *Proc. Natl Acad. Sci.* **99** 1389–94
[24] Pugh T A M, Lindeskog M, Smith B, Poulter B, Arneth A, Haverd V and Calle L 2019 *PNAS* **116** 4382–7
[25] Bohn F J, Frank K and Huth A 2014 *Ecol. Model.* **278** 9–17
[26] Lorimer C G and Halpin C R 2014 *Forest Ecol. Manage.* **334** 344–57
[27] Knapp N, Fischer R and Huth A 2018 *Remote Sens. Environ.* **205** 199–209

[28] Duncanson L et al 2020 *Remote Sens. Environ.* **242** 111779
[29] Inman H F and Bradley J E L 1989 *Commun. Stat.-Theory Methods* **18** 3851–74
[30] Rödig E, Knapp N, Fischer R, Bohn F J, Dubayah R, Tang H and Huth A 2019 *Nat. Commun.* **10** 1–7
[31] Therneau T M, Atkinson B and Ripley M B 2010 *R Foundation for Statistical Computing: Oxford, UK*
[32] Ni-Meister W, Lee S, Strahler A H, Woodcock C E, Schaaf C, Yao T, Ranson K J, Sun G and Blair J B 2010 *J. Geophys. Res.: Biogeosci.* **115**
[33] Lu D, Chen Q, Wang G, Moran E, Batistella M, Zhang M, Vaglio Laurin G and Saah D 2012 *Int. J. Forestry Res.* **2012** 436537
[34] Qi W, Saarela S, Armston J, Ståhl G and Dubayah R 2019 *Remote Sens. Environ.* **232** 111283
[35] Oliver C D et al 1996 *Forest Stand Dynamics: Updated Edition* (New York: Wiley)
[36] Falkowski M J, Evans J S, Martinuzzi S, Gessler P E and Hudak A T 2009 *Remote Sens. Environ.* **113** 946–56
[37] Kane V R, Bakker J D, McGaughey R J, Lutz J A, Gersonde R F and Franklin J F 2010 *Can. J. Forest Res.* **40** 774–87
[38] Spies T 1997 *Creating a Forestry for the 21st Century* (Washington, DC: Island Press) pp 11–30
[39] Palik B J, D'Amato A W, Franklin J F and Johnson K N 2020 *Ecological Silviculture: Foundations and Applications* (Long Grove, IL: Waveland Press)
[40] Barton A M and Keeton W S 2018 *Ecology and Recovery of Eastern Old-Growth Forests* (Washington, DC: Island Press)