Earth and Space Science

RESEARCH ARTICLE
10.1029/2022EA002516

Key Points:

- Global Ecosystem Dynamics Investigation (GEDI) aboveground biomass density is from models trained on a comprehensive database of field measurements and simulated GEDI waveforms
- On-orbit prediction requires stratification by plant functional type and world region
- Quality flags and metrics distinguish GEDI measurements that are representative of the conditions under which models were developed

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

J. R. Kellner,
james_r_kellner@brown.edu

Citation:

Kellner, J. R., Armston, J., & Duncanson, L. (2023). Algorithm theoretical basis document for GEDI footprint aboveground biomass density. *Earth and Space Science*, *10*, e2022EA002516. https://doi.org/10.1029/2022EA002516

Received 20 JUL 2022
Accepted 21 OCT 2022

Author Contributions:

Conceptualization: James R. Kellner, John Armston, Laura Duncanson
Data curation: James R. Kellner, John Armston, Laura Duncanson
Formal analysis: James R. Kellner, John Armston, Laura Duncanson
Funding acquisition: James R. Kellner, John Armston, Laura Duncanson
Investigation: James R. Kellner, John Armston, Laura Duncanson

Algorithm Theoretical Basis Document for GEDI Footprint Aboveground Biomass Density

James R. Kellner[1,2], John Armston[3], and Laura Duncanson[3]

[1]Institute at Brown for Environment and Society, Brown University, Providence, RI, USA, [2]Department of Ecology, Evolution and Organismal Biology, Brown University, Providence, RI, USA, [3]Department of Geographical Sciences, University of Maryland College Park, College Park, MD, USA

Abstract The Global Ecosystem Dynamics Investigation (GEDI) lidar is a multibeam laser altimeter on the International Space Station (ISS). GEDI is the first spaceborne instrument designed to measure vegetation height and to quantify aboveground carbon stocks in temperate and tropical forests and woodlands. This document describes the algorithm theoretical basis underpinning the development of the GEDI Level-4A (GEDI04_A) footprint aboveground biomass density (AGBD) data product. The GEDI04_A data product contains estimates of AGBD for individual GEDI footprints and associated prediction intervals. The algorithm uses GEDI02_A relative height metrics and 13 linear models to predict AGBD in 32 combinations of plant functional type and world region within the observation limits of the ISS. GEDI04_A models for the release 1 and release 2 data products were developed using 8,587 quality-filtered simulated GEDI waveforms associated with field estimates of AGBD in 21 countries. Although this is the most geographically comprehensive data available for the development of AGBD models using lidar remote sensing, important regions are underrepresented, including the forests of continental Asia, deciduous broadleaf forests and savannas of the dry tropics, and evergreen broadleaf forests north of Australia. We describe the scientific and statistical assumptions required to develop globally representative estimates of AGBD using GEDI lidar, including generalization beyond training data, and exclusion of GEDI02_A observations that do not meet requirements of the GEDI04_A algorithm. The footprint-level predictions generated by this process provide globally comprehensive estimates of AGBD. These footprint-level predictions are a prerequisite for the GEDI04_B gridded AGBD data product.

Plain Language Summary The amount of carbon stored in aboveground vegetation is uncertain. This uncertainty limits our ability to calculate fluxes of carbon between the land surface and the atmosphere, and prevents rigorous carbon offset crediting in forests. Much of this uncertainty is attributed to inconsistent measurement techniques and the use of Earth-observation methods that were not designed to quantify carbon density. The Global Ecosystem Dynamics Investigation (GEDI) can largely overcome these challenges by producing measurements of vegetation height using a lidar sensor on the International Space Station. This document describes methods developed by the GEDI Science Team to convert spaceborne measurements of vegetation height into estimates of aboveground biomass density. The algorithms depend on the geographic world region and the type of vegetation that is present at a sampled location. For example, evergreen broadleaf forests of the humid tropics in South America and deciduous broadleaf forests of Europe use different algorithms. Statistical models were developed using comprehensive field measurements and simulated GEDI data. This document describes the importance of filtering GEDI data to reduce the impact of measurement artifacts on aboveground biomass predictions. Quality flags and ancillary data contained in the GEDI04_A data product ensure that the best predictions can be used.

1. Introduction

The Global Ecosystem Dynamics Investigation (GEDI) is a multibeam waveform lidar on the International Space Station (ISS) (Dubayah et al., 2020). Two objectives of the mission are to quantify the distribution of aboveground carbon in woody vegetation, and to use these estimates to determine the impact of land use and land cover changes on aboveground carbon stocks. Both of these objectives speak to fundamental uncertainties in the spatial distribution of aboveground carbon density (Mitchard et al., 2013; Pan et al., 2011).

Methodology: James R. Kellner, John Armston, Laura Duncanson
Project Administration: James R. Kellner, John Armston, Laura Duncanson
Resources: James R. Kellner, John Armston, Laura Duncanson
Software: James R. Kellner, John Armston, Laura Duncanson
Supervision: James R. Kellner, John Armston, Laura Duncanson
Validation: James R. Kellner, John Armston, Laura Duncanson
Visualization: James R. Kellner, John Armston, Laura Duncanson
Writing – original draft: James R. Kellner
Writing – review & editing: James R. Kellner, John Armston, Laura Duncanson

This document describes the theoretical basis and assumptions that underpin the algorithms developed by the GEDI Science Team to produce estimates of footprint aboveground biomass density (AGBD) from GEDI lidar in release 1 and release 2 of the GEDI04_A data product. It also describes quality assessment and filtering criteria used to minimize differences in measurement characteristics between power and coverage ground tracks that may impact estimates of AGBD. GEDI data are collected by three lasers in eight parallel ground tracks. One laser is split into two beams, called the coverage laser, and two lasers are fired at full power, called power lasers. Beams from coverage and power lasers are optically dithered to produce eight parallel ground tracks (Dubayah et al., 2020).

Footprint AGBD is generated for 32 combinations of plant functional type (PFT) and geographic world region using 13 linear models developed from a comprehensive set of simulated GEDI waveforms associated with field estimates of AGBD from allometric scaling equations in 21 countries. The GEDI04_A data product is publicly available through the Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC). The structure of each data file is described in the GEDI04_A data dictionary available through ORNL DAAC (Dubayah, Armston, Kellner, et al., 2022).

2. Historical Perspective

Developing methods to predict AGBD using lidar requires field estimates of aboveground biomass for individual trees, M_i. These estimates are acquired using one or more allometric models over a fixed area. Summing the M_i over all individuals in a plot or lidar footprint and expressing it per unit ground area produces an estimate of AGBD. Coincident remote sensing data can be used to develop empirical relationships between AGBD and a remotely sensed measurement (Drake et al., 2002; Lefsky et al., 2002).

Many remote sensing technologies have been used to quantify AGBD in forests, including passive optical sensors (Foody et al., 2003), Synthetic Aperture Radar (SAR) systems (Mitchard et al., 2012; Saatchi et al., 2011), discrete return airborne laser scanning (ALS) (Coops et al., 2007; Duncanson et al., 2015; Næsset et al., 2013), airborne waveform lidar systems (Drake et al., 2002; Dubayah et al., 2010; Swatantran et al., 2011), and space-borne waveform lidar (Boudreau et al., 2008; Lefsky et al., 2005; Rosette et al., 2013). Passive optical and SAR backscatter techniques typically saturate with increasing AGBD, with the degree of saturation depending on SAR wavelength (Huete et al., 1997; Luckman et al., 1998). Estimates of AGBD from lidar consistently outperform other technologies (Saatchi et al., 2011; Zolkos et al., 2013).

Most previous efforts have developed site-specific or regional relationships between AGBD and remote sensing measurements (Zolkos et al., 2013). However, GEDI AGBD data products require models and algorithms that perform well throughout the entire observation domain of the ISS. Locally developed or regional relationships between AGBD and height are unlikely to perform well at locations outside the limited geographic extent of training data, unless procedures are developed to evaluate whether models can be generalized to new geographic locations (Friedl et al., 2002; Ploton et al., 2020).

3. Approach to Statistical Model Development

Models to produce GEDI04_A were developed using field estimates of AGBD associated with simulated GEDI waveforms derived from discrete-return airborne lidar (Blair & Hofton, 1999; Hancock et al., 2019). The justification for using simulated GEDI waveforms is that few locations on the land surface are associated with field estimates of AGBD that could be directly linked to on-orbit GEDI data. GEDI geolocation error is also large relative to the size of individual footprints.

An important objective for GEDI04_A models is generalization outside the domain of calibration. Two key components are geographic transferability, meaning that the models can be extrapolated to locations outside the geographic extent of training data, and transferability from simulated to recorded GEDI waveforms. Transferring models from simulated to recorded GEDI waveforms requires that the models are insensitive to errors and uncertainties, including artifacts associated with GEDI waveforms and GEDI02_A processing (Hofton & Blair, 2020).

GEDI04_A models were developed using a quality-filtered calibration data set that contains simulated GEDI waveforms and field estimates of AGBD: the Forest Structure and Biomass Database (FSBD). The FSBD is the most geographically comprehensive data available for the development of AGBD models using remote sensing,

Table 1
Numbers of Simulated Global Ecosystem Dynamics Investigation Waveforms Used for Footprint Model Development and Testing for Release 1 and Release 2 of the GEDI04_A Data Product

	DBT	DNT	EBT	ENT	GSW	Total
Africa	490	–	834	0	6	1,330
Australia and Oceania	0	–	213	142	65	420
Europe	333	0	0	417	0	750
North America	873	0	0	1,391	18	2,282
North Asia	2	0	0	36	0	38
South America	0	–	3,441	0	0	3,441
South Asia	0	0	326	0	0	326
Total	1,698	0	4,814	1,986	89	8,587

Note. GEDI04_A models are stratified by combinations of world region and PFT derived from error-corrected and infilled MODIS data product MCD12Q1 V006. These are deciduous broadleaf trees (DBT; class 4), deciduous needleleaf trees (DNT; class 3), evergreen broadleaf trees (EBT, class 2), evergreen needleleaf trees (ENT, class 1), and grasses, shrubs and woodlands (GSW, classes 5, 6, and 11). The DNT stratum does not occur in Africa, Australia and Oceania or South America.

but important regions are under-represented. Underrepresented locations include the forests of continental Asia and the evergreen broadleaf forests throughout the islands of Southeast Asia and north of Australia on the east side of the Wallace line, which defines the floral and faunal boundary between Australia and Asia during the Pleistocene (Mayr, 1944). Underrepresented locations also include the worldwide distributions of deciduous needleleaf forests, savannas and deciduous tropical forests (Table 1).

The approach to model development considered candidates whose performance was evaluated outside the geographic extent of training data (Duncanson et al., 2022). Candidate models were evaluated within sets of 5° grid cells that contain simulated GEDI waveforms with coincident field data. The approach set aside data from one grid cell for testing and trained the model using data within the remaining grid cells. The trained model was used to predict AGBD within the held-out grid cell, and the process was repeated for all grid cells and all models under consideration (Figure 1). This approach to cross validation is explicitly spatial, and evaluates whether candidate models are likely to generalize to locations outside the geographic extent of training data (Le Rest et al., 2014; Pohjankukka et al., 2017; Roberts et al., 2017).

3.1. Stratification of GEDI04_A Models

Building globally representative GEDI04_A models requires stratification (Duncanson et al., 2022). The models are stratified by world region and PFT (Figure 2, Table 1). World regions are the geologically defined continents of Africa and Europe in addition to other continents and locations. The South America world region is the continent of South America, Central America and the Caribbean islands, and geological North America south of southern Mexico. The Australia and Oceania world region is geological Australia and the island regions north of Australia on the east side of the Wallace line. The islands of Micronesia, Melanesia, and Polynesia are associated with the Australia and Oceania world region regardless of political affiliation. The North America world region includes geological North America north of southern Mexico. The continent of Asia is divided into north and south regions that approximately correspond to temperate and tropical forests (Figure 2).

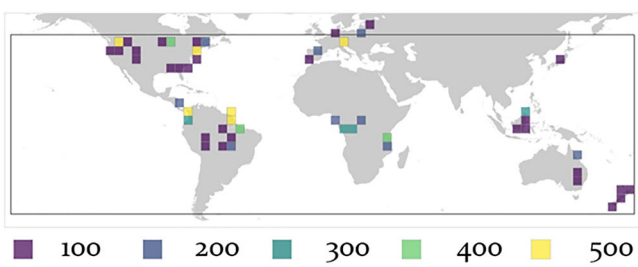

Figure 1. Geographic distribution of the number of simulated Global Ecosystem Dynamics Investigation (GEDI) waveforms used to develop the models for release 1 and release 2 of the GEDI04_A data product. The box inset is the GEDI observation domain of 51.6°N to S latitude.

GEDI04_A models are stratified by combinations of PFT derived from an infilled and error-corrected version of MODIS data product MCD12Q1 V006 (Friedl et al., 2002, 2010). These are deciduous broadleaf trees (DBT; class 4), deciduous needleleaf trees (DNT; class 3), evergreen broadleaf trees (EBT, class 2), evergreen needleleaf trees (ENT, class 1), and grasses,

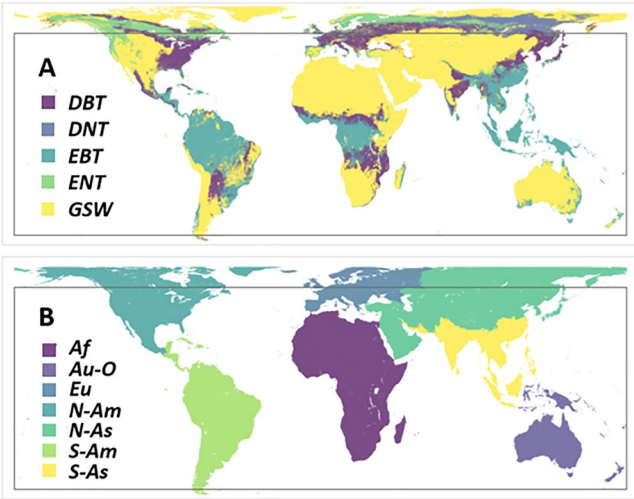

Figure 2. Global stratification by five combinations of error-corrected and infilled MODIS MCD12Q1 V006 PFT (a) and world region (b) to produce GEDI04_A models. The box inset is the Global Ecosystem Dynamics Investigation observation domain of 51.6°N to S latitude. DBT (deciduous broadleaf trees), DNT (deciduous needleleaf trees), EBT (evergreen broadleaf trees), ENT (evergreen needleleaf trees), GSW (grasses, shrubs and woodlands). Af (Africa), Au-O (Australia and Oceania), Eu (Europe), N-Am (North America north of southern Mexico), N-As (North Asia), S-Am (South America, Central America, southern Mexico, and the Caribbean), S-As (South Asia).

shrubs, and woodlands (GSW, classes 5, 6, and 11; Figure 2). In MCD12Q1 V006, class 5 is shrub, class 6 is grass, and class 11 is barren.

3.2. Training Data Quality-Control Filters

The data being used to develop GEDI04_A models accumulate over time as new data are assimilated and improvements are made to existing records. The unfiltered database used for releases 1 and 2 of GEDI04_A contains 31,414 simulated GEDI waveforms. After excluding incomplete projects and others that cannot be used, the unfiltered database contains 12,140 simulated GEDI waveforms. After applying quality-control filters, the database used to develop releases 1 and 2 of the GEDI04_A data product contains 8,587 simulated waveforms from 21 countries (Table 1). The analysis below indicates the number of simulated waveforms that were flagged by each quality-control filter. Because some waveforms were flagged by multiple filters, the total number of flagged waveforms does not sum to the 3,553 waveforms that were removed from the unfiltered data set.

3.2.1. Incongruent AGBD and Height

Footprints were excluded when there was an incongruence between field-estimated AGBD and simulated RH98. In particular, when AGBD was <1 Mg/ha and RH98 was >5 m, the footprint was excluded (113 footprints, or 0.93% of the unfiltered database). When AGBD was >150 Mg/ha and RH98 was <5 m, the footprint was excluded (7 footprints, or 0.06%).

3.2.2. Incongruent AGBD and Canopy-Cover Fraction (CCF)

When Canopy-cover Fraction was 0 and RH98 >5 m, the footprint was excluded (158 footprints, or 1.30%). The GEDI along-beam laser intensity half-width results in estimates of RH100 close to 4.5 m on surfaces of uniform reflectance observed orthogonally to the beam path. One implication of this filter is that waveforms with 0 AGBD on sloped terrain were excluded from training data.

3.2.3. Incongruent Field-Measured or Modeled Height and Lidar Height

Some field data include measurements of individual tree height. When field measurements of height were not available, tree height was predicted using regional height-diameter allometric scaling equations. This is necessary because some allometric models used to predict M_i require tree height. When the difference between measured or predicted height and RH98 was >10 m, the footprint was excluded (997 footprints, or 8.21%).

3.2.4. Extrapolation of Allometric Scaling Equations Beyond Measured Range

Some of the allometric scaling models used to predict M_i have a reported tree size domain over which predictions are valid. These tree size domains are defined by the data used to develop the equations (Chave et al., 2014; Forrester et al., 2017; Jenkins et al., 2003; Paul et al., 2016; Roxburgh et al., 2019; Ung et al., 2008). If a footprint contained at least one tree with a diameter, height, or wood specific gravity outside the range defined by the original data, the footprint was excluded (640 footprints, or 5.27%). This filter may exclude training data from old-growth conditions in particular. Of the 640 footprints excluded by this filter, 322 were from the Zofin forest site, a temperate old-growth forest in the southwest Czech Republic, and six footprints were excluded from Barro Colorado Island, Panama, an old-growth tropical moist forest. The remaining 312 footprints were distributed among 15 secondary forest projects, or projects of unknown forest age.

3.2.5. Overlap Between Simulated Footprints and Field Data

Some simulated GEDI footprints are not completely contained within the boundaries of field-inventory plots. This occurs when the diameter of the plot is smaller than the dimeter of a simulated GEDI footprint, or when the center location of the simulated footprint happens to be <12.5 m from the edge of the plot. When this occurs, information about AGBD within the footprint is incomplete. Previous work has demonstrated that inclusion of these observations in statistical models causes relationships to be biased toward zero (Rejou-Mechain

et al., 2014). If >10% of the area of a simulated footprint was outside the boundaries of a field inventory plot, it was excluded (129 footprints, or 1.06%).

3.2.6. Large Field Plots

The data are organized into spatial units by project and then by plot. A project is single contribution from a given research group. For example, La Selva, Costa Rica and Robson Creek, Australia are individual projects. Some projects contain multiple plots. Because the number and size of plots is variable, a small number of large plots contributes disproportionately to the total number of observations. Because these observations would overwhelm model fitting and evaluation at the expense of plots with fewer samples (and broader geographic coverage), we placed an upper limit of 200 footprints on the contribution of each plot (not project) to the filtered data set used to develop the release 1 and release 2 GEDI04_A data products. When the number of footprints in a plot that passed other filters was <200, all footprints were included. When the number of footprints in a plot was >200 after applying other filters, a stratified random sample of 200 footprints was generated, where the per-footprint probability of inclusion was inversely proportional to the number of footprints in each of 20 equally spaced AGBD bins between the minimum and maximum footprint AGBD in the plot, and probabilities were scaled so that each bin had an equal probability of contributing footprints to the sample.

3.2.7. Candidate Model Selection and Specification

Models were developed for every combination of PFT and world region in Table 1 with >50 footprints, and for every PFT and world region independently. For example, the model for EBT in the South America world region was developed using 3,441 footprints, and the global ENT model was developed using 1,986 footprints. The approach to model selection considered candidates with square-root or natural-logarithm transformations on the response and either the same transformation or no transformation on the predictors, for a total of four transformation scenarios. Candidate predictors were simulated relative height (RH) metrics in increments of 10% and RH98 in addition to all possible products of RH metrics.

The GEDI04_A data product refers to combinations of PFT and world region as strata. The approach to model selection distinguishes between prediction strata and fit strata when developing GEDI04_A models. Prediction strata are the combination of PFT and world region within which a given model is tested and applied to on-orbit GEDI data (e.g., DBT × Europe). The fit stratum refers to the PFT, world region, or combination of PFT and world region from which data were used to develop a candidate model. For example, a model could be evaluated in the DBT × Europe prediction stratum in three ways: first, using only DBT × Europe fit data ($n = 333$ simulated footprints; Table 1); second, using worldwide DBT fit data, where model parameters are estimated using 1,698 simulated footprints in the global DBT stratum, and the model is tested against the subset of 333 simulated footprints in DBT × Europe; and third by using fit data from all 750 simulated footprints in the Europe world region and testing the model against the subset of 333 simulated footprints in DBT × Europe. Note that the model is tested against the same 333 footprints in each scenario. This approach varies the data used to estimate model parameters, not to evaluate performance.

There are 24 out of 32 prediction strata represented by <50 footprints in the filtered training data set (Table 1). In these cases, the selected model is from the best PFT prediction stratum in 17 cases and an alternative model stratified by PFT and world region in seven cases (Tables 2 and 3). These seven cases represent two EBT strata, two DBT strata, and three DNT strata. In the EBT prediction stratum within Europe and North America the selected models are the corresponding DBT by world region candidates. This selection assumes that models trained using data from DBT in the northern-latitude temperate zone will perform better in these prediction strata than models developed using EBT data. In the data set used to develop the release 1 and release 2 GEDI04_A models, EBT samples are exclusively tropical or within the Australia and Oceania world region, and thus not representative of EBT in North America or Europe. In the DBT stratum within the South America world region and the Australia and Oceania world region, the selected model is the EBT model for the associated world region. In the Australia and Oceania world region, the DBT classification is probably an error in MCD12Q1, because Australia lacks forests and woodlands dominated by upper-canopy deciduous trees. In the South America world region, DBT forests are likely to be tropical moist or dry forests that are more similar to EBT of South America than to DBT of other world regions. Finally, the current version of the FSBD does not contain training data in DNT anywhere. The selected model is a corresponding ENT by world region candidate for two of these strata (Europe, and North

Table 2
Associations Between 14 Models and 32 Prediction Strata in the GEDI Domain

Model name	Prediction strata
DBT	DBT × North Asia, EBT North Asia
EBT	DBT × South Asia, EBT × South Asia
ENT	DNT × North Asia, DNT × South Asia, ENT × Africa, ENT × North Asia, ENT × South America, ENT × South Asia
GSW	GSW × Africa, GSW × Europe, GSW × North America, GSW × North Asia, GSW × South America, GSW × South Asia
DBT × Africa	DBT × Africa
DBT × Europe	DBT × Europe, EBT × Europe
DBT × North America	DBT × North America, EBT × North America
EBT × Africa	EBT × Africa
EBT × Australia	DBT × Australia, EBT × Australia
EBT × South America	DBT × South America, EBT × South America
ENT × Australia	ENT × Australia
ENT × Europe	DNT × Europe, ENT × Europe
ENT × North America	DNT × North America, ENT × North America
GSW × Australia and Oceania	GSW × Australia and Oceania

Note. DBT, deciduous broadleaf trees; EBT, evergreen broadleaf trees; ENT, evergreen needleleaf trees; GEDI, Global Ecosystem Dynamics Investigation; GSW, grasses, shrubs, and woodlands. These associations refer to the release 1 and release 2 GEDI04_A data product.

Table 3
Selected Models and Back-Transformation Corrections Used to Generate AGBD Predictions in 32 Prediction Strata in the GEDI Domain

Model name	Model
DBT	$\text{AGBD} = 1.017 \times (-110.059 + 5.134 \times \sqrt{RH60 + 100} + 6.172 \times \sqrt{\text{RH98} + 100})^2$
EBT	$\text{AGBD} = 1.113 \times (-104.965 + 6.802 \times \sqrt{RH50 + 100} + 3.955 \times \sqrt{\text{RH98} + 100})^2$
ENT	$\text{AGBD} = 1.018 \times (-118.411 + 7.777 \times \sqrt{RH60 + 100} + 4.378 \times \sqrt{RH98 + 100})^2$
GSW	$\text{AGBD} = 1.118 \times (-124.832 + 12.426 \times \sqrt{RH98 + 100})^2$
DBT × Africa	$\text{AGBD} = 1.092 \times (-118.408 + 1.957 \times \sqrt{RH50 + 100} + 9.962 \times \sqrt{RH98 + 100})^2$
DBT × Europe	$\text{AGBD} = 0.963 \times (-96.531 + 7.175 \times \sqrt{RH70 + 100} + 2.921 \times \sqrt{RH98 + 100})^2$
DBT × North America	$\text{AGBD} = 1.052 \times (-120.777 + 5.508 \times \sqrt{RH50 + 100} + 6.808 \times \sqrt{RH98 + 100})^2$
EBT × Africa	$\text{AGBD} = 1.113 \times (-104.965 + 6.802 \times \sqrt{RH50 + 100} + 3.955 \times \sqrt{RH98 + 100})^2$
EBT × Australia	$\text{AGBD} = 1.018 \times (-155.414 + 7.817 \times \sqrt{RH70 + 100} + 7.710 \times \sqrt{RH98 + 100})^2$
EBT × South America	$\text{AGBD} = 1.106 \times (-134.770 + 6.654 \times \sqrt{RH50 + 100} + 6.687 \times \sqrt{RH98 + 100})^2$
ENT × Australia	$\text{AGBD} = 0.898 \times (-101.984 + 6.397 \times \sqrt{RH40 + 100} + 4.259 \times \sqrt{RH98 + 100})^2$
ENT × Europe	$\text{AGBD} = 0.963 \times (-96.531 + 7.175 \times \sqrt{RH70 + 100} + 2.921 \times \sqrt{RH98 + 100})^2$
ENT × North America	$\text{AGBD} = 1.013 \times (-114.355 + 8.401 \times \sqrt{RH70 + 100} + 3.346 \times \sqrt{RH98 + 100})^2$
GSW × Australia and Oceania	$\text{AGBD} = 1.128 \times (-151.383 + 4.491 \times \sqrt{RH50 + 100} - 2.347 \times \sqrt{RH80 + 100} + 12.941 \times \sqrt{RH98 + 100})^2$

Note. AGBD, aboveground biomass density; DBT, deciduous broadleaf trees; EBT, evergreen broadleaf trees; ENT, evergreen needleleaf trees; GEDI, Global Ecosystem Dynamics Investigation; GSW, grasses, shrubs, and woodlands. These refer to the release 1 and release 2 GEDI04_A data product. The DBT × Europe and ENT × Europe models are identical because the model-selection procedure identified models trained using the Europe fit stratum in both cases.

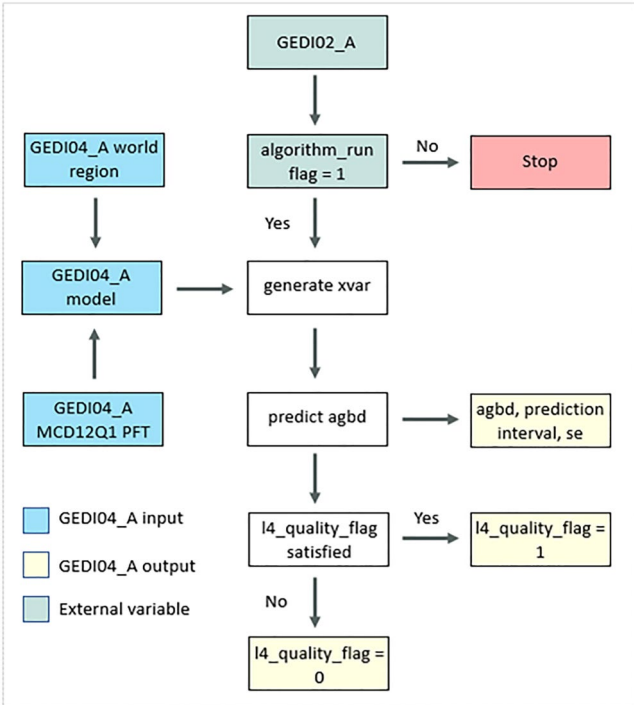

Figure 3. GEDI04_A algorithm flow. The GEDI04_A algorithm assimilates external data from GEDI02_A and other sources. A prediction is generated for every Global Ecosystem Dynamics Investigation shot where algorithm_run_flag = 1. The algorithm looks up the GEDI04_A model using a world region grid and error-corrected and infilled MODIS MCD12Q1 PFT. xvar is the transformed and scaled predictor data (GEDI02_A RH metrics). AGBD and associated uncertainty are outputs of the GEDI04_A algorithm for every GEDI02_A algorithm selection setting.

America). In the remaining DNT strata there is no corresponding ENT by world region candidate, and the selected model is the global ENT model (Tables 2 and 3).

The approach to model specification computed all possible 1, 2, 3, and 4 variable predictor matrices. Models were considered that contained the product of RH metric pairs (e.g., RH50 × RH98), even when the individual RH metrics were not included in the model as independent variables. Predictor matrices that were multicollinear were eliminated from further consideration. Multicollinearity was quantified by computing the Pearson correlation matrix and the variance inflation factor (VIF) for the candidate predictor matrix. If the maximum absolute value of the correlation coefficient was >0.9, or VIF was >10, the predictor matrix was excluded. For all candidates that passed both multicollinearity tests, a weighted linear model was fitted by regressing the transformation of AGBD on the predictors. Weights were inversely proportional to the number of simulated footprints in each 5° grid cell used under geographic cross validation and scaled to sum to 1, so that training data in every grid cell contributed equally to the model, regardless of the number of observations within the grid cell.

3.2.8. Benchmarking the Candidate Models

The performance of all candidate models was evaluated by ranking every model in order of smallest mean residual error rounded down to the nearest Mg/ha, the smallest percentage root mean squared error (RMSE) rounded down to the nearest 5%, the maximum RH metric in the model, the number of coefficients in the model, and the number of RH metrics in the model. Mean residual error and RMSE were computed using geographically cross validated predictions, where mean residual error was:

$$\frac{1}{n}\sum_{i=1}^{n}|\text{Observed}_i - \text{Expected}_i| \quad (1)$$

Observed_i is the estimate of AGBD from field data in footprint i, and Expected_i is the predicted value. The percentage RMSE was computed according to:

$$\text{RMSE} = 100 \times \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\text{Observed}_i - \text{Expected}_{i,j})^2} \Big/ \frac{1}{n}\sum_{i=1}^{n}\text{Observed}_i \quad (2)$$

RMSE was rounded down to the nearest 5% when ranking models to avoid trivial differences in RMSE driving model rankings. The model selection favors candidates that contain higher percentile RH metrics when all other terms are equivalent. This approach favors higher percentile RH metrics because RH metrics closer to the ground may be more sensitive to differences between simulated and real GEDI waveforms than RH metrics higher in the canopy (Hancock et al., 2019). Reducing simulator error for lower percentile RH metrics using the on-orbit transmit pulse shape and characteristics of recorded GEDI noise will be addressed in a subsequent release of GEDI04_A. Models with fewer coefficients and fewer RH metrics are preferred based on parsimony. The number of coefficients is not equivalent to the number of RH metrics because candidate models contain RH metric products. For example, a model that contains the product of RH98 and RH50 as a single predictor contains two coefficients (one slope and one intercept) and two RH metrics. A model that contains only RH98 and RH50 as independent variables contains three coefficients (two slopes and one intercept) and two RH metrics.

4. Algorithm Description

The GEDI04_A algorithm ingests GEDI02_A data and external input variables (Figure 3, Table 4; Dubayah et al., 2021). A prediction is generated for every GEDI02_A measurement for which it is possible to initiate the GEDI04_A algorithm. This is determined by the following six tests associated with variables in the GEDI04_A

Table 4
Input Variables Required to Run the GEDI04_A Algorithm

Input variable	Source	Description
algorithm_run_flag	GEDI04_A	Flag = 1 when the GEDI04_A algorithm is run. This occurs when rx_algrunflag = 1, rx_assess/quality_flag = 1, zcross >0, toploc >0, sensitivity >0 and sensitivity <1
bias_correction_name	GEDI04_A	Back-transform bias correction method
bias_correction_value	GEDI04_A	Back-transform bias correction value
dof	GEDI04_A	Degrees of freedom of the model used to predict agbd
landsat_water_persistence	GEDI02_A	Landsat permanent water bodies
leaf_off_flag	GEDI02_A	Flag indicating whether the observation was recorded during leaf-off conditions in DBT and DNT prediction strata (1 = leaf-off and 0 = leaf-on)
l2_quality_flag	GEDI04_A	Flag = 1 when algorithm_run_flag = 1, surface_flag = 1, stale_return_flag = 0, sensitivity >0.9, and rx_maxamp >8 × sd_corrected
urban_proportion	TanDEM-X	The proportion of land area within urban_focal_window_size that is urban land cover
par	GEDI04_A	Linear model parameters to predict agbd
predict_stratum	GEDI04_A	Character ID of the prediction stratum name for the 1 km cell that contains the footprint (e.g., DBT_Af = deciduous broadleaf trees in Africa)
rh_index	GEDI04_A	Index of the RH metrics used as predictors
rse	GEDI04_A	Residual standard error of the model used to predict AGBD
xvar	GEDI04_A	RH metric predictor variables using the optimal algorithm setting (transform and offset have been applied)
rx_algrunflag	GEDI02_A	Flag that indicates error run of the received waveform algorithm using selected settings (0 = good)
rx_assess/quality_flag	GEDI02_A	Flag that indicates a good waveform based on assess parameters (0 = good)
rx_maxamp	GEDI02_A	Maximum amplitude of the rxwaveform relative to the mean noise level
sd_corrected	GEDI01_B	Noise standard deviation
stale_return_flag	GEDI02_A	Flag = 0 when the pulse detection algorithm detects a return signal > the detection threshold within the search window
surface_flag	GEDI02_A	Flag = 1 when elev_lowestmode is within 300 m of the TanDEM-X 90 DEM or mean sea surface
vcov	GEDI04_A	Variance-covariance matrix of model parameters in transformed units (square root or natural logarithm)
xvar_aN	GEDI04_A	RH metric predictor variables using algorithm setting N (transform and offset have been applied)
x_transform	GEDI04_A	Transformation applied to the predictor variables (square root, natural logarithm, or none)
y_transform	GEDI04_A	Transformation applied to the response variable (square root or natural logarithm)
zcross	GEDI01_B	Sample number of the bin of the center the lowest mode above noise level

Note. AGBD, aboveground biomass density. These variables are available for every footprint in the GEDI04_A data product.

data product: rx_algrunflag = 1, rx_assess/quality_flag = 1, zcross > 0, toploc > 0, sensitivity > 0 and sensitivity < 1. Beam sensitivity is a measure of signal-to-noise that is related to the maximum canopy cover that can be penetrated by a waveform (Hofton & Blair, 2020). When these conditions are met, the GEDI04_A variable called algorithm_run_flag is set equal to 1. The algorithm looks up the PFT, world region, and algorithm selection setting, then applies the selected model to scaled and transformed GEDI02_A RH metrics. Additional checks are performed to determine whether the GEDI04_A prediction is valid, and ancillary data are computed (Table 5).

After a prediction is generated, the algorithm determines the value of two quality flags: l2_quality_flag and l4_quality_flag. The l2_quality_flag indicates whether GEDI02_A input data passed minimum quality standards

Table 5
Output Variables Produced by GEDI04_A Algorithm

Output variable	Units	Description
agbd	Mg/ha	Predicted AGBD using the optimal algorithm setting
agbd_aN	Mg/ha	Predicted AGBD using algorithm setting N
agbd_pi_lower	Mg/ha	Lower prediction interval for agbd, given alpha
agbd_pi_lower_aN	Mg/ha	Lower prediction interval for agbd_aN, given alpha
agbd_pi_upper	Mg/ha	Upper prediction interval for agbd, given alpha
agbd_pi_upper_aN	Mg/ha	Upper prediction interval for agbd_aN, given alpha
agbd_se	Mg/ha	The standard error of the agbd prediction
agbd_se_aN	Mg/ha	The standard error of the agbd_aN prediction using algorithm setting N
agbd_t	–	Predicted AGBD in transformed units (square root or natural logarithm)
agbd_t_aN	–	Predicted AGBD in transformed units (square root or natural logarithm) using algorithm setting N
agbd_t_se	–	Standard error of the agbt_t prediction in transformed units
agbd_t_se_aN	–	Standard error of the agbt_t prediction in transformed units using algorithm setting N
alpha	probability	Significance level used for calculation of prediction intervals
l2_quality_flag	–	Flag = 1 when algorithm_run_flag = 1, surface_flag = 1, stale_return_flag = 0, sensitivity >0.9, and rx_maxamp >8 × sd_corrected
l4_quality_flag	–	Flag = 1 when l2_quality_flag = 1, sensitivity >0.95, landsat_water_persistence <10, and urban_proportion <50. In DBT and DNT strata, l4_quality_flag also requires leaf_off_flag = 0.
predictor_limit_flag	–	Flag that indicates whether any of xvar are outside the range observed in training data for the given model using the optimal algorithm setting (0 = in bounds, 1 = below, 2 = above)
predictor_limit_flag_aN	–	Flag that indicates whether any of xvar_aN are outside the range observed in training data for the given model using algorithm setting N (0 = in bounds, 1 = below, 2 = above)
response_limit_flag	–	Flag that indicates whether agbd is outside the range observed in training data for the given model using the optimal algorithm setting (0 = in bounds, 1 = below, 2 = above)
response_limit_flag_aN	–	Flag that indicates whether agbd_aN is outside the range observed in training data for the given model using the algorithm setting N (0 = in bounds, 1 = below, 2 = above)

Note. AGBD, aboveground biomass density. These variables are available for every footprint in the GEDI04_A data product.

for AGBD estimation. The l2_quality_flag = 1 when the footprint passes five tests: algorithm_run_flag = 1, surface_flag = 1, stale_return_flag = 0, sensitivity >0.9, and rx_maxamp >8 × sd_corrected. The surface_flag = 1 when elev_lowestmode is within 300 m of the TanDEM-X 90 m digital elevation model (DEM) or mean sea surface. The stale_return_flag = 0 when the pulse detection algorithm detects a return signal > the detection threshold within the search window. The variable rx_maxamp is the maximum amplitude of the received waveform relative to the mean noise level, and sd_corrected is the corrected standard deviation of the waveform noise. The components of l4_qualtiy_flag depend on PFT. Within deciduous strata (DBT and DNT), l4_quality_flag = 1 when the footprint passes five tests: l2_quality_flag = 1, sensitivity >0.95, landsat_water_persistence <10, leaf_off_flag = 0, and urban_proportion <50. The variable landsat_water_persistence is the annual water percentage and is used to identify permanent open water bodies (Pickens et al., 2020). leaf_off_flag indicates whether the footprint was collected under leaf-off or leaf-on conditions and was derived for a 1 km grid using the VIIRS land surface phenology product VNP22Q2 (Zhang et al., 2016). The variable urban_proportion is from a 25 m global urban mask developed by the GEDI Science Team using the TerraSAR-X and TanDEM-X global urban footprint data product (Esch et al., 2013). In all remaining strata (EBT, ENT, and GSW), l4_quality_flag does not consider leaf_off_flag.

The predictor_limit_flag and response_limit_flag indicate whether xvar or agbd are outside the range of training data for the given GEDI04_A model (Table 5). The variables xvar and agbd are the scaled and transformed predictor values, and predicted AGBD in original units, respectively. The predictor_limit_flag and response_limit_flag have a value of 0 when the data are inside the range, a value of 1 when outside the lower bound, and a value of 2 when outside the upper bound. For predictor_limit_flag, values of 1 or 2 are triggered when at least one predictor is outside the range of training data for the given model name (Table 5). The values of predictor_limit_flag and response_limit_flag do not impact l4_quality_flag.

4.1. Scientific Theory

The mass of a given individual tree, M_i, is derived using allometric-scaling models developed from destructive harvesting and weighing of trees. In their most general form, these models assume the mass of individual i, M_i is a power function of size, X_i:

$$M_i \propto aX_i^b \tag{3}$$

Numerous investigations have demonstrated that M_i is related to tree size under a wide range of conditions (Beets et al., 2011; Brown, 1997; Chave et al., 2014; Forrester et al., 2017; Jenkins et al., 2003; Moore, 2010; Muukkonen, 2007; Paul et al., 2016; Roxburgh et al., 2019; Ung et al., 2008).

In practice, the variable indicated by X_i is usually the diameter of the stem at breast height (DBH), defined as 1.3 m aboveground. Some allometric models use tree height in addition to DBH, and others incorporate wood specific gravity, defined as oven-dried mass divided by wet volume (Williamson & Wiemann, 2010). Because it is typically not possible to weigh an entire tree, wood samples are heated in an oven until mass stabilizes. The oven-dry mass per unit wet volume is computed from the samples, and measurements of DBH and height are used to compute total wood volume. Wood volume is multiplied by wood specific gravity to obtain M_i. Repeating this process for many individual trees results in data that are used to estimate the parameters of Equation 3.

For every tree-record in the FSBD an allometric model appropriate to the given PFT and world region was applied to predict M_i. When there was more than one model that could be used for a given tree, locally developed models were favored over regional ones, as long as locally developed models were not site-specific. The approach also favored models with finer taxonomic resolution and models developed using larger sample sizes. In Australia AGBD is from eight allometric models developed by Paul et al. (2016) and Roxburgh et al. (2019). In New Zealand AGBD is from the model developed by Moore (2010) for *Pinus radiata*, and the model of Beets et al. (2011) for all other species. In North America we selected the models of Jenkins et al. (2003) in the continental United States and the models of Ung et al. (2008) in Canada. In Europe AGBD is from the allometric models of Forrester et al. (2017). Throughout the tropics of South America, Africa and Asia, we selected the allometric model developed by Chave et al. (2014).

In some situations there was more than one candidate allometric model to predict M_i that met GEDI04_A requirements. In these situations no claim is made about the superiority of the selected allometric models in comparison to alternatives. For example, the models of Brown (1997) and Chave et al. (2014) have been used to predict M_i in Central American forests. The models of Muukkonen (2007) and Forrester et al. (2017) have been used in deciduous broadleaf and evergreen needleleaf forests of Europe. The models of Jenkins et al. (2003) and the component ratio method (CRM; Heath et al., 2009) have been used in North American forests. Choosing a model to predict M_i is important and has resulted in discrepancies in estimates of AGBD from spaceborne remote sensing (Duncanson et al., 2017; Mitchard et al., 2013). Studies are needed that compare predictions of M_i to harvested trees that have been dried and weighed to determine which allometric models have the best performance.

Some contributed data records for individual trees included measurements of height and DBH, but some records only contained DBH. When the allometric model required height as an input and a measured value was available, the measured value was used. When no height measurement was available height was predicted in one of three ways. If some records contained estimates of height and DBH, we developed a local height-diameter relationship using complete data records. If no height measurements were available, we used the published height-diameter allometry of Muukkonen (2007) in Europe, and the models of Feldpausch et al. (2011) elsewhere except the United States. In the United States, we developed a local height-diameter allometry using United States Department of

Agriculture Forest Inventory and Analysis data within the same county and used this locally developed model to predict tree height.

4.2. Scientific Assumptions

Allometric models are assumed to generate unbiased estimates of M_i when applied to non-harvested trees. This assumption is unlikely to true. Harvested trees used to develop allometric scaling relationships are usually not randomly sampled (Clark & Kellner, 2012), and validation studies that directly measure tree mass have demonstrated that allometric models systematically underestimate M_i for large trees (e.g., Gonzalez de Tanago et al., 2018). An important area for future research is the development of improved allometric scaling models or no-allometry methods based on terrestrial laser scanning or drone lidar (Calders et al., 2020; Disney et al., 2020; Kellner et al., 2019).

GEDI04_A models treat footprints as circular areas with a radius of 12.5 m. In model training, M_i is assigned to the footprint using stem positions or the mean AGBD associated with a given subplot that contains a simulated GEDI waveform. When mapped individual stems are available and the coordinates of a given stem are within the extent of the footprint, M_i, as defined by a given allometric model, is assigned to the footprint. When stem positions are unavailable, the mean AGBD in the square subplot is assigned to the footprint. Four assumptions underpin this approach. First, across-beam laser intensity follows a Gaussian distribution, but we ignore the impact of across-beam laser intensity on the relationship between RH metrics and AGBD (Hofton & Blair, 2020; Hyde et al., 2005). Second, because the across-beam laser intensity is Gaussian, intercepted surfaces >12.5 m from the footprint center contribute a small amount to the intensity of the returned laser waveform. For example, assuming the across-beam σ is 5.5 m (Hancock et al., 2019) about 2.3% of the returned laser energy on a uniform reflectance target orthogonal to the beam path is received from surfaces beyond the 12.5 m threshold. The third and fourth assumptions address the size of GEDI footprints relative to tree locations or subplots. A tree whose stem is outside the 12.5 m radius used to assign M_i to individual footprints could contribute to the simulated waveform if parts of the tree crown are inside the footprint. Similarly, a tree whose stem coordinates are inside the footprint has all of M_i assigned to the footprint, even though some branch or crown material (a portion of M_i) may be outside the extent of the simulated GEDI waveform. For non-stem-mapped plots, trees inside the subplot but outside the footprint contribute to the mean AGBD assigned to the footprint, even though they may be unobserved by the simulated waveform.

Error in AGBD associated with stem or crown positions and subplot geometry is likely to be most important for very large trees (Knapp et al., 2021). For example, the crown diameter of a single large tree can exceed 50 m, or two times the nominal GEDI footprint diameter (Martínez Cano et al., 2019). Very large trees can contribute disproportionally to AGBD. Clark and Clark (1996) estimated that the largest 2% of the stems accounted for 27% of AGBD in a lowland Neotropical rain forest.

Simulated waveforms used to develop the GEDI04_A models were generated in the absence of sensor noise, and where RH metrics are known exactly. RH metrics can be determined without error using simulated waveforms because they can be computed relative to ground elevation from ALS. Transferring these models from simulated to recorded GEDI data requires the assumption that ground-detection methods applied to recorded GEDI01_B waveforms are accurate, and that noise inherent to recorded GEDI data does not undermine the application of models developed on noiseless data. The first release of GEDI02_A RH metrics used a single algorithm to identify the elevation of the lowest mode, assumed to be ground elevation. This sometimes resulted in GEDI02_A ground elevation estimates that were biased high, and in turn RH metrics that were biased low. Release 2 of GEDI02_A addressed this issue by using one of six algorithm setting groups to interpret the received waveform, rather than one (Beck et al., 2021; Hofton & Blair, 2020).

GEDI04_A models were developed using training data collected under leaf-on conditions. The leaf_off_flag is used to identify GEDI waveforms that are likely to have been generated under leaf-on conditions in DBT and DNT forests. Some EBT forests and woodlands experience periods of partial deciduousness during which some percentage of crowns are leafless while the canopy as a whole is not. For example, a study across a rainfall gradient in Panama classified as EBT using MODIS data product MCD12Q1 found that 3.6%–19.1% of crown area was leafless at peak deciduousness (Condit et al., 2000). This indicates that some GEDI waveforms may represent partial leaf-off conditions in practice. An important assumption is that GEDI04_A training data are representative

of the variability introduced by partial leaf-off conditions, and that the impact of this variability is subsumed into the GEDI04_A model parameter uncertainty estimates.

A final assumption is that GEDI04_A models are representative of the geographic conditions to which they will be applied. Although the GEDI FSBD is comprehensive, important regions are under-represented or missing entirely (Table 1). Training data are lacking in continental Asia and throughout the GSW and DNT stratifications worldwide. In strata where training data are lacking, the need to select a model for on-orbit prediction necessitates the assumption that a model developed using data from a different location can be applied to that stratum to produce unbiased predictions of AGBD.

4.3. Statistical Theory

Because M_i is modeled as a power function of stem diameter and height, model functional forms that linearize the relationship between AGBD and RH metrics and minimize heteroskedasticity are necessary. GEDI04_A considers four functional forms: (a) a square-root transformation on the response, (b) a square-root transformation on the response and predictors, (c) a natural logarithm transformation on the response, and (d) a natural logarithm transformation on the response and predictors (Hansen et al., 2015). Back-transforming model predictions from the square-root or natural logarithm scale requires a back-transformation bias correction. Models using the natural logarithm transformation considered two bias corrections. The method originally developed by Baskerville (1972) transforms values from the natural logarithm scale to the original scale using:

$$\widehat{\mathrm{AGBD}}_{i,j} = exp\left(\mathbf{X}_i\beta_j + \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\log\left(\widehat{\mathrm{AGBD}}_i\right) - \mathbf{X}_i\beta_j\right)^2}{2}\right) \tag{4}$$

The term $\mathbf{X}_i\beta_j$ denotes predicted values for footprint i model j in natural logarithm units using matrix notation, where $\mathbf{X}_i$ is a row vector of predictor variables including a 1 for the intercept and β_j is column vector of coefficients.

Snowdon (1991) developed a ratio estimator for bias correction that is less sensitive to violations of the assumptions of logarithmic normality. The back-transformed value is:

$$\widehat{\mathrm{AGBD}}_{i,j} = C_j \times \exp\left(\mathbf{X}_i\beta_j\right) \tag{5}$$

C_j is a bias-correction coefficient:

$$C_j = \frac{\sum_{i=1}^{n}\mathrm{AGBD}_i}{\frac{1}{n}\sum_{i=1}^{n}\exp\left(\mathbf{X}_i\beta_j\right)} \tag{6}$$

For models with a square-root transformation on the response, we used the bias-correction of Snowdon (1991), where the bias-correction coefficient is:

$$C_j = \frac{\sum_{i=1}^{n}\mathrm{AGBD}_i}{\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\beta_j{}^2} \tag{7}$$

The back-transformed value for models with a square-root transformation on the response is:

$$\widehat{\mathrm{AGBD}}_{i,j} = C_j \times \left(\mathbf{X}_i\beta_j\right)^2 \tag{8}$$

We did not consider the square-root back-transformation correction of Gregoire et al. (2008) because it results in a positive bias for small values of AGBD.

The GEDI04_A data product contains footprint prediction intervals calculated using the associated standard error and alpha contained in the data product. This estimate of the standard error is provided so that users can produce arbitrary prediction intervals. The standard error of the prediction for GEDI footprint i in transformed units is available in the GEDI04_A data product as the variable agbd_t_se. This quantity is calculated according to:

$$\mathrm{SE}_i = \sqrt{\mathrm{MSE}_k + \mathbf{X}_i \,\mathrm{Cov}(\beta)\, \mathbf{X}_i^{\mathrm{T}}} \tag{9}$$

Here, MSE_k is the square of the residual standard error from the linear regression applied to prediction stratum k containing GEDI footprint i, $\mathbf{X}_i$ is the row vector of scaled and transformed RH metrics for GEDI footprint i, and Cov(β) is the variance-covariance matrix for the model parameters in transformed units (i.e., natural-logarithm, square-root, or none).

Prediction intervals are calculated for every predicted value of AGBD in transformed units according to:

$$\sqrt{\hat{\mathrm{AGBD}}_i} \pm t_{\left(1-\frac{\alpha}{2}, n-(k+1)\right)} \times \mathrm{SE}_i \tag{10}$$

The t-multiplier is the value from a t distribution with confidence level α and $n-(k+1)$ degrees of freedom, where k is the number of predictor variables. Users can compute prediction intervals for arbitrary values of α using the degrees of freedom within the model_data group of the GEDI04_A product and Equation 10. Back transforming these values to original units requires bias correction. All of the selected models for release 1 and release 2 use a square-root transformation on the response and the back-transformation correction in Equation 8. The parameter C_j is available in the GEDI04_A data product as the variable bias_correction_value.

4.4. Statistical Assumptions

Fitting linear models to transformed AGBD requires the assumption that transformations linearize the relationship between AGBD and RH metrics and reduce heteroskedasticity. Both of these assumptions underpin the methods used to compute the standard error of the estimate of AGBD at 1 km resolution (Dubayah, Armston, Healey, et al., 2022; Patterson et al., 2019; Ståhl et al., 2011). This approach assumes that a single bias-correction coefficient produces an unbiased estimate of AGBD after back-transformation across the range of AGBD. Flewelling and Pienaar (1981) demonstrated that this assumption can be violated at large values of predicted AGBD.

4.5. Algorithm Input Variables

The GEDI04_A algorithm requires GEDI02_A inputs, an error-corrected and infilled version of MODIS MCD12Q1 V006 PFT classification, a world region identifier, and linear models for all prediction strata within the GEDI domain (Table 4). Release 1 of the GEDI04_A product used release 1 GEDI02_A as input. However, we applied the algorithm setting group selection being implemented in release 2 of GEDI02_A to the release 1 GEDI02_A data on a per-footprint basis to generate release 1 of GEDI04_A. Release 2 of GEDI04_A is based on release 2 of GEDI02_A. The algorithm setting group used for each footprint is contained in the selected_algorithm variable in the root group of the GEDI04_A data product. Note that a selected_algorithm value of 10 indicates algorithm setting group 5 has been used, but that the lowest detected mode is likely a noise detection. When this occurs, a higher mode has been selected as the ground mode and used to calculate RH metrics (Hofton & Blair, 2020).

4.6. Algorithm Output Variables

The GEDI04_A algorithm outputs predicted AGBD in original (Mg/ha) and transformed units, associated prediction intervals, the standard error of the prediction, quality flags, and other ancillary information (Table 5). The algorithm produces these data for every selection setting and identifies the best selection setting for each waveform. A comprehensive list of input and output variables is in the GEDI04_A data dictionary (Dubayah, Armston, Kellner, et al., 2022).

Table 6
Quality Filtering of GEDI04_A Release 2 Generation 2 in Deciduous Broadleaf Trees Prediction Strata

Stratum	Beam	algorithm_run_flag = 1	l4_quality_flag = 1	Retained
Africa	coverage	133,344,523	10,698,029	8.0%
Africa	power	149,369,908	30,690,663	20.5%
Australia	coverage	17,357	1,188	6.8%
Australia	power	20,061	3,473	17.3%
Europe	coverage	226,494,254	26,824,934	11.8%
Europe	power	266,526,677	122,433,339	45.9%
N. America	coverage	231,851,095	36,552,050	15.8%
N. America	power	264,883,024	97,711,860	36.9%
N. Asia	coverage	184,857,796	19,740,237	10.7%
N. Asia	power	209,163,542	49,611,697	23.7%
S. America	coverage	73,671,901	10,600,806	14.4%
S. America	power	82,679,807	24,891,801	30.1%
S. Asia	coverage	60,044,292	4,662,244	7.8%
S. Asia	power	68,842,657	21,922,594	31.8%
Total	coverage	910,281,218	109,079,488	12.0%
	power	1,041,485,676	347,365,427	33.3%

Note. algorithm_run_flag = 1 when rx_algrunflag = 1, rx_assess/quality_flag = 1, zcross > 0, toploc > 0, sensitivity > 0 and sensitivity < 1. In DBT prediction strata, l4_quality_flag = 1 when l2_quality_flag = 1, sensitivity >0.95, landsat_water_persistence <10, leaf_off_flag = 0, and urban_proportion <50.

5. Algorithm Implementation and Usage Constraints

The software that generates the GEDI04_A product was implemented at the GEDI Science Office at the Department of Geographical Sciences, University of Maryland, College Park, in collaboration with the GEDI Science Data Processing System at the NASA Goddard Space Flight Center in Greenbelt, Maryland and the Institute at Brown for Environment and Society at Brown University.

The GEDI04_A algorithms were developed for prediction of AGBD using GEDI data. Although the approach developed here could be replicated for other sensors, the GEDI04_A models should not be directly applied to alternative sensor data. For example, Duncanson et al. (2020) applied the GEDI04_A model framework to simulated ICEsat-2 data. This required the development of alternative statistical models. The models in Duncanson et al. (2020) were developed specifically to accommodate the instrument response and spatial resolution of ICESat-2.

6. Performance Assessment

The performance of the GEDI04_A algorithm was evaluated by quantifying the frequency of observations that were excluded by quality filters in every prediction stratum for coverage and power lasers, and by disaggregating the impact of variables that can trigger l4_quality_flag = 0. This performance assessment determines the percentages of GEDI shots that are flagged as low-quality observations in every prediction stratum relative to the number of observations where algorithm_run_flag = 1, and it identifies the causes of the low-quality trigger. The analysis is based on mission weeks 16–153 of release 2, generation 2 of the GEDI04_A data product.

Quality filtering results in data losses when it is possible to implement the GEDI04_A algorithm. These data losses are expected, because most conditions under which it is possible to run the GEDI04_A algorithm do not meet minimum quality standards. Estimates of AGBD are provided for low-quality waveforms in addition to those where l4_quality_flag = 1 because some users may want access to GEDI04_A AGBD estimates under a wide range of conditions, and because AGBD under varying conditions can help to characterize algorithm performance.

In DBT forests worldwide, application of l4_quality_flag retained 12.0% of coverage observations and 33.3% of power observations, with variation among world regions (Table 6). Data losses are less substantial in EBT and ENT strata. For example, in EBT forests application of l4_quality flag retains 51.5% of coverage observations and 73.7% of power observations worldwide (Table 7). The GEDI04_B algorithm applies a more stringent sensitivity threshold of 0.98 in EBT forests of Africa, South America, and South Asia, which retains 10.4%–22.0% and 32.2%–44.1% of coverage and power observations in these strata, respectively (Table 7). Worldwide, beam sensitivity, l2_quality_flag, and leaf_off_flag drive data losses in DBT strata (Table 8), and beam sensitivity and l2_quality_flag drive data losses in global EBT forests (Table 9). The other two variables that contribute to l4_qualtiy_flag are landsat_water_persistence and urban_proportion, both of which are comparatively small contributors to data losses. Similar conclusions characterize remaining PFT strata (Tables S1–S6 in Supporting Information S1).

7. Improvement of GEDI04_A Algorithms

7.1. Geographic Representation of Training Data

Although the training data used to develop GEDI04_A models is comprehensive, important areas are underrepresented. There are 24 out of 32 prediction strata where it was not possible to develop or test a candidate model, including many stratifications that contain forested aboveground carbon density (Pan et al., 2011). These strata represent all of North Asia, the dry savannas and woodlands of South America and South Asia, and DNT and

Table 7
Quality Filtering of GEDI04_A Release 2 Generation 2 in Evergreen Broadleaf Trees Prediction Strata

Stratum	Beam	algorithm_run_flag = 1	l4_quality_flag = 1	Retained	Sensitivity >0.98	Retained
Africa	coverage	104,132,309	50,035,427	48.0%	10,822,440	10.4%
Africa	power	127,811,969	88,546,666	69.3%	41,196,324	32.2%
Australia	coverage	67,322,240	25,761,816	38.3%	–	–
Australia	power	77,982,343	57,909,260	74.3%	–	–
Europe	coverage	22,502,234	5,020,769	22.3%	–	–
Europe	power	26,123,805	16,041,229	61.4%	–	–
N. America	coverage	16,327,053	7,408,796	45.4%	–	–
N. America	power	18,875,801	14,085,387	74.6%	–	–
N. Asia	coverage	6,418,298	3,134,591	48.8%	–	–
N. Asia	power	7,807,837	5,022,501	64.3%	–	–
S. America	coverage	255,471,862	155,680,698	60.9%	56,198,576	22.0%
S. America	power	298,794,469	241,406,525	80.8%	131,704,665	44.1%
S. Asia	coverage	127,909,960	62,141,220	48.6%	20,117,226	15.7%
S. Asia	power	157,433,639	103,783,449	65.9%	55,789,270	35.4%
Total	coverage	600,083,956	309,183,317	51.5%	–	–
	power	714,829,863	526,795,017	73.7%	–	–

Note. algorithm_run_flag = 1 when rx_algrunflag = 1, rx_assess/quality_flag = 1, zcross > 0, toploc > 0, sensitivity > 0 and sensitivity < 1. In EBT prediction strata, l4_quality_flag = 1 when l2_quality_flag = 1, sensitivity >0.95, landsat_water_persistence < 10, and urban_proportion < 50. GEDI04_B processing applies a more stringent sensitivity threshold of 0.98 in EBT × Africa, EBT × South America and EBT × South Asia.

Table 8
Conditions Where l4_quality_flag = 0 in Deciduous Broadleaf Trees Prediction Strata

Stratum	Beam	Sensitivity	l2_quality_flag	landsat_water_pers	leaf_off_flag	urban_proportion
Africa	coverage	75.5%	37.5%	0.3%	31.9%	0.1%
Africa	power	32.3%	16.0%	0.3%	32.9%	0.1%
Australia	coverage	84.3%	65.5%	3.2%	76.7%	0.0%
Australia	power	63.2%	52.6%	1.6%	77.3%	0.1%
Europe	coverage	78.9%	36.8%	0.9%	77.0%	3.7%
Europe	power	29.7%	19.5%	0.9%	76.1%	3.8%
N. America	coverage	65.7%	24.3%	2.9%	52.4%	1.4%
N. America	power	20.2%	15.4%	3.0%	51.6%	1.4%
N. Asia	coverage	64.2%	25.6%	2.8%	37.7%	1.9%
N. Asia	power	22.8%	13.3%	2.7%	38.5%	1.9%
S. America	coverage	60.3%	21.1%	0.8%	39.0%	0.2%
S. America	power	15.5%	12.6%	0.8%	39.1%	0.2%
S. Asia	coverage	79.5%	51.1%	1.4%	70.9%	1.1%
S. Asia	power	41.2%	18.2%	1.4%	69.3%	1.2%
Mean	coverage	72.6%	37.4%	1.8%	55.1%	1.2%
	power	32.1%	21.1%	1.5%	55.0%	1.2%

Note. Numbers are the percentage of shots where algorithm_run_flag = 1 and the given variable caused l4_quality_flag = 0. Some shots were flagged by multiple variables, such that rows do not sum to 100%. Data are from generation 2 release 2 of GEDI04_A. In DBT prediction strata, l4_quality_flag = 0 when sensitivity ≤ 0.95 or l2_quality_flag = 0 or landsat_water_persistence ≥ 10 or leaf_off_flag = 1 or urban_proportion ≥ 50.

Table 9
Conditions Where l4_quality_flag = 0 in Evergreen Broadleaf Trees Prediction Strata

Stratum	Beam	Sensitivity	l2_quality_flag	landsat_water_pers	leaf_off_flag	urban_proportion
Africa	coverage	51.3%	25.6%	1.3%	–	0.3%
Africa	power	24.6%	14.7%	1.2%	–	0.3%
Australia	coverage	60.9%	24.3%	4.0%	–	0.5%
Australia	power	19.6%	13.4%	3.9%	–	0.5%
Europe	coverage	76.1%	33.5%	3.6%	–	7.0%
Europe	power	27.2%	19.7%	3.8%	–	7.1%
N. America	coverage	53.5%	19.1%	3.3%	–	2.0%
N. America	power	16.0%	13.6%	3.3%	–	1.9%
N. Asia	coverage	49.2%	24.8%	5.7%	–	11.4%
N. Asia	power	21.1%	17.5%	5.6%	–	10.7%
S. America	coverage	38.3%	14.4%	2.2%	–	0.4%
S. America	power	12.7%	10.6%	2.2%	–	0.4%
S. Asia	coverage	50.1%	27.1%	4.7%	–	2.6%
S. Asia	power	23.3%	17.9%	4.5%	–	2.5%
Mean	coverage	54.2%	24.1%	3.5%	–	3.4%
	power	20.6%	15.4%	3.5%	–	3.4%

Note. Numbers are the percentage of shots where algorithm_run_flag = 1 and the given variable caused l4_quality_flag = 0. Some shots were flagged by multiple variables, such that rows do not sum to 100%. Data are from generation 2 release 2 of GEDI04_A. In EBT prediction strata, l4_quality_flag = 0 when sensitivity ≤ 0.95 or l2_quality_flag = 0 or landsat_water_persistence ≥ 10 or urban_proportion ≥ 50.

GSW strata worldwide. Even within strata that are more densely sampled, data used for model training may not be representative of relationships between RH metrics and AGBD throughout the prediction stratum. For example, training data in the EBT × Australia and Oceania world region are from two projects in the Australian state of Queensland. This prediction stratum also includes the Island of New Guinea, and other islands of the Malay Archipelago east of the Wallace line. The degree to which existing training data represent an entire prediction stratum underpins the quality of GEDI04_A and GEDI04_B data products. Two important priorities are documenting representativeness of training data within prediction strata, and the addition of new training data in under-represented locations. Both of these actions can also support the activities of other current and forthcoming space missions, including ICESat-2, NISAR and ESA BIOMASS (Duncanson et al., 2019, 2021; Narine et al., 2020; Scipal et al., 2010; Siqueira et al., 2021).

7.2. Algorithm Sensitivity to GEDI Data Waveform Properties

Training data must be representative of GEDI waveform properties. The release 1 and release 2 GEDI04_A models were developed using simulated noiseless waveforms, where ground elevation and canopy height were known from discrete-return ALS. In practice, identifying the ground-return and the top location in recorded GEDI data can be challenging under conditions of dense canopy cover, complex terrain, and low-altitude clouds or fog. The GEDI02_A data product includes results for six algorithm setting groups to interpret the received waveform and identify the signal start, end, and elevation of the lowest mode (Hofton & Blair, 2020), which is treated as ground elevation. Errors in waveform measurement will propagate through RH metrics and impact AGBD predictions. A priority for subsequent releases of the GEDI04_A data product is examining the frequency of waveform processing errors and the impact of these errors on estimates of AGBD. Subsequent releases of the GEDI04_A data product will be updated using models trained on simulated waveform data that has been parameterized with on-orbit measurements of the transmitted pulse and sensor noise.

GEDI04_B processing applies filtering criteria that excludes some footprints where l4_quality_flag = 1. These filters are necessary to avoid violating assumptions of GEDI04_B hybrid estimators (Dubayah, Armston, Healey,

Bruening, et al., 2022; Healey et al., 2022; Patterson et al., 2019), and to remove measurement artifacts identified after the generation of release 1 and release 2 of the GEDI04_A data products. In particular, GEDI04_B processing identifies and removes footprints that are likely to be impacted by low-altitude clouds or fog and shots with degradation of geolocation performance. GEDI04_B processing also applies a more stringent beam sensitivity threshold in EBT forests of Africa, South America and South Asia. Subsequent releases of the GEDI04_A data product will implement GEDI04_B quality filtering and will contain additional flags that standardize measurement quality between GEDI04_A and GEDI04_B data products.

7.3. Strengthening Allometric Models

All remote sensing of AGBD depends on field measurements of individual trees and allometric models to estimate individual tree mass. Recent work has demonstrated that allometric scaling equations generate biased estimates of individual tree mass in some conditions. An important area for future research is the development of improved allometric models or alternative methods based on terrestrial laser scanning or drone lidar (Brede et al., 2019; Calders et al., 2020; Disney et al., 2019; Duncanson et al., 2021; Kellner et al., 2019; Trochta et al., 2017). Subsequent releases of the GEDI04_A data product may use different allometric scaling equations.

7.4. Alternative Approaches to AGBD Prediction

The hybrid-inference framework selected for GEDI04_B requires footprint models that produce a covariance matrix that describes relationships among model parameters. This covariance matrix enables a closed-form estimate of uncertainty when GEDI04_A predictions are aggregated to large areas (e.g., the 1 km GEDI04_B grid, or other arbitrary regions; Dubayah, Armston, Healey, et al., 2022; Dubayah, Armston, Kellner, et al., 2022; Dubayah, Armston, Healey, Bruening, et al., 2022; Patterson et al., 2019; Ståhl et al., 2016). This requirement ruled out non-parametric methods and some approaches based on machine learning in the development of GEDI04_A models (e.g., ensemble-based decision trees and neural networks; Lang et al., 2021). Whether alternative specifications based on machine learning or other methods can improve the quality of GEDI04_A predictions is not known. Reducing uncertainty in GEDI04_A independent of GEDI04_B would improve the GEDI04_A data product and support investigations that require footprint-level resolution. This includes integration of footprint AGBD with Landsat forest cover loss (Hansen et al., 2010; Healey et al., 2020), fusion of GEDI04_A with TanDEM-X to produce gridded AGBD at a finer resolution than the 1 km GEDI04_B data product (Qi et al., 2019), and simulations from prognostic ecosystem model outputs (Ma et al., 2019; Medvigy et al., 2010), all of which are GEDI demonstrative products that require footprint AGBD. It may also be possible to identify or engineer features using machine learning that are linearly related to footprint AGBD. If successful, this could facilitate the use of machine learning in a way that is compliant with the hybrid-inference framework.

It may also be possible to simplify or reduce the number of GEDI04_A models used for global prediction or to make changes to GEDI04_A stratification. Examination of the models used to develop release 1 and release 2 of the GEDI04_A data product (Table 3) indicates broad similarity among coefficients and selected variables among prediction strata. Some of this is by design. For example, we required the inclusion of RH98, and excluded RH metrics < RH50 in all strata but one (ENT × Australia). However, no selected model used a natural logarithm transformation, and almost all models selected one additional RH metric close to RH50 in addition to RH98. In one case we selected a model with two additional RH metrics (GSW × Australia and Oceania). Of the 11 remaining models, five selected RH50, two selected RH60 and four selected RH70. Among all selected models there was a large negative intercept and all remaining coefficients were positive. No selected model contained a term that included a product of RH metric pairs. This suggests generality in the relationship between AGBD and RH metrics that could be exploited. Evaluating the impact of classification errors in PFT used to stratify GEDI04_A models is a research priority. Future versions of the GEDI04_A data product may consider alternative stratifications.

Data Availability Statement

The GEDI04_A version 2 generation 2 data product is publicly available through the Oak Ridge National Laboratory Distributed Active Archive Center (https://doi.org/10.3334/ORNLDAAC/2056; Dubayah, Armston, Kellner, et al., 2022).

Acknowledgments
This work was funded by NASA contract NNL 15AA03C to the University of Maryland for the development and execution of the GEDI mission. We thank our colleagues on the GEDI Mission Science Team, including research technicians and graduate students whose work directly supported the mission: Bryan Blair, Jamis Bruening, Patrick Burns, Ralph Dubayah, Temilola Fatoyinbo, Scott J. Goetz, Steve Hancock, Matt Hansen, Sean P. Healey, Michelle Hofton, George Hurtt, Scott Luthcke, Suzanne Marselis, David M. Minor, Paul Patterson, Jim Pontius, and Hao Tang. We are grateful to the NASA Terrestrial Ecology Program, Hank Margolis and Michael Falkowski for supporting the GEDI mission, and the University of Maryland for providing independent financial support. We thank Mike Wulder and Andy Hudak for reviewing draft versions of the ATBD. Many individuals contributed data that allowed comprehensive GEDI04_A models to be developed: Katharine Abernethy, Hans-Erik Andersen, Paul Aplin, Timothy R. Baker, Nicolas Barbier, Jean Francois Bastin, Pascal Boeckx, Jan Bogaert, Luigi Boschetti, Peter Brehm Boucher, Doreen S. Boyd, David F.R.P. Burslem, Sofia Calvo-Rodriguez, Jérôme Chave, Robin L. Chazdon, David B. Clark, Deborah A. Clark, Warren B. Cohen, David A. Coomes, Piermaria Corona, K. C. Cushman, Mark E. J. Cutler, James William Dalling, Michele Dalponte, Sergio de-Miguel, Songqiu Deng, Peter Woods Ellis, Barend Erasmus, Michael Falkowski, Patrick A. Fekety, Alfredo Fernández-Landa, Antonio Ferraz, Rico Fischer, Adrian G. Fisher, Antonio García-Abril, Terje Gobakken, Jonathan A. Greenberg, Jorg M. Hacker, Marco Heurich, Ross A. Hill, Sören Holm, Chris Hopkinson, Chengquan Huang, Huabing Huang, Stephen P. Hubbell, Andrew T. Hudak, Andreas Huth, Benedikt Imbach, Patrick Jantz, Kathryn Jeffery, Masato Katoh, Elizabeth Kearsley, Natascha Kljun, Nikolai Knapp, Kamil Král, Martin Krůček, Nicolas Labrière, Seung-kuk Lee, Simon L. Lewis, Marcos Longo, Richard M. Lucas, Russell Main, Jose A. Manzanera, Rodolfo Vásquez Martínez, Renaud Mathieu, Victoria Meyer, Paul Montesano, Felix Morsdorf, Erik Næsset, Laven Naidoo, Reuben Nilus, Michael J. O'Brien, David A. Orwig, Geoffrey Parker, Christopher Philipson, Oliver L. Phillips, Jan Pisek, John R. Poulsen, Wenlu Qi, Christoph Rüdiger, Svetlana Saarela, Sassan Saatchi, Arturo Sanchez-Azofeifa, Nuria Sanchez-Lopez, Crystal B. Schaff, Marc Simard, Andrew Kerr Skidmore, Göran Ståhl, Krzysztof Stereńczak, Chiara Torresan, Rubén Valbuena, Hans Verbeeck, Tomas Vrska, Konrad Wessels, Joanne C. White, and Carlo Zgraggen.

References

Baskerville, G. L. (1972). Use of logarithmic regression in the estimation of plant biomass. *Canadian Journal of Forest Research*, *2*(1), 49–53. https://doi.org/10.1139/x72-009

Beck, J., Wirt, B., Luthcke, S., Hofton, M., & Armston, J. (2021). *Global ecosystem dynamics investigation (GEDI) level 02 user guide version 2.0. U.S. Geological survey*. Earth Resources Observation and Science Center.

Beets, P. N., Kimberley, M. O., Paul, T. S. H., & Garrett, L. G. (2011). Planted forest carbon monitoring system–forest carbon model validation study for Pinus radiata. *New Zealand Journal of Forestry Science*, *41*, 177–189.

Blair, J. B., & Hofton, M. A. (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

Boudreau, J., Nelson, R., Margolis, H., Beaudoin, A., Guindon, L., & Kimes, D. (2008). Regional aboveground forest biomass using airborne and spaceborne LiDAR in Québec. *Remote Sensing of Environment*, *112*(10), 3876–3890. https://doi.org/10.1016/j.rse.2008.06.003

Brede, B., Calders, K., Lau, A., Raumonen, P., Bartholomeus, H. M., Herold, M., & Kooistra, L. (2019). Non-destructive tree volume estimation through quantitative structure modelling: Comparing UAV laser scanning with terrestrial LIDAR. *Remote Sensing of Environment*, *233*, 111355. https://doi.org/10.1016/j.rse.2019.111355

Brown, S. (1997). Estimating biomass and biomass change of tropical forests: A primer (FAO forestry paper - 134). Retrieved from http://www.fao.org/docrep/w4095e/w4095e00.HTM

Calders, K., Adams, J., Armston, J., Bartholomeus, H., Bauwens, S., Bentley, L. P., et al. (2020). Terrestrial laser scanning in forest ecology: Expanding the horizon. *Remote Sensing of Environment*, *251*, 112102. https://doi.org/10.1016/j.rse.2020.112102

Chave, J., Réjou-Méchain, M., Búrquez, A., Chidumayo, E., Colgan, M. S., Delitti, W. B. C., et al. (2014). Improved allometric models to estimate the aboveground biomass of tropical trees. *Global Change Biology*, *20*(10), 3177–3190. https://doi.org/10.1111/gcb.12629

Clark, D. B., & Clark, D. A. (1996). Abundance, growth and mortality of very large trees in neotropical lowland rain forest. https://doi.org/10.1016/0378-1127(95)03607-5

Clark, D. B., & Kellner, J. R. (2012). Tropical forest biomass estimation and the fallacy of misplaced concreteness. *Journal of Vegetation Science*, *23*(6), 1191–1196. https://doi.org/10.1111/j.1654-1103.2012.01471.x

Condit, R., Watts, K., Bohlman, S. A., Perez, R., Foster, R. B., & Hubbell, S. P. (2000). Quantifying the deciduousness of tropical forest canopies under varying climates. *Journal of Vegetation Science*, *11*(5), 649–658. https://doi.org/10.2307/3236572

Coops, N. C., Hilker, T., Wulder, M. A., St-Onge, B., Newnham, G., Siggins, A., et al. (2007). Estimating canopy structure of Douglas-fir forest stands from discrete-return LiDAR. *Trees*, *21*(3), 295–310. https://doi.org/10.1007/s00468-006-0119-6

Disney, M., Burt, A., Calders, K., Schaaf, C., & Stovall, A. (2019). Innovations in ground and airborne technologies as reference and for training and validation: Terrestrial laser scanning (TLS). *Surveys in Geophysics*, *40*(4), 937–958. https://doi.org/10.1007/s10712-019-09527-x

Disney, M., Burt, A., Wilkes, P., Armston, J., & Duncanson, L. (2020). New 3D measurements of large redwood trees for biomass and structure. *Scientific Reports*, *10*(1), 16721. https://doi.org/10.1038/s41598-020-73733-6

Drake, J., Dubayah, R., Clark, D. B., Knox, R. G., Hofton, M. A., Chazdon, R. L., et al. (2002). Estimation of tropical forest structural characteristics using large-footprint lidar. https://doi.org/10.1016/S0034-4257(01)00281-4

Dubayah, R., Armston, J., Healey, S. P., Bruening, J. M., Patterson, P. L., Kellner, J. R., et al. (2022). GEDI launches a new era of biomass inference from space. *Environmental Research Letters*, *17*(9), 095001. https://doi.org/10.1088/1748-9326/ac8694

Dubayah, R., Blair, J. B., Goetz, S., Fatoyinbo, L., Hansen, M., Healey, S., et al. (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

Dubayah, R., Hofton, M., Blair, J. B., Armston, J., Tang, H., & Luthcke, S. (2021). *GEDI L2A elevation and height metrics data global footprint level V002*. NASA EOSDIS Land Processes DAAC. https://doi.org/10.5067/GEDI/GEDI02_A.002

Dubayah, R. O., Armston, J., Healey, S. P., Yang, Z., Patterson, P. L., Saarela, S., et al. (2022). *GEDI L4B gridded aboveground biomass density, version 2*. ORNL DAAC. https://doi.org/10.3334/ORNLDAAC/2017

Dubayah, R. O., Armston, J., Kellner, J. R., Duncanson, L., Healey, S. P., Patterson, P. L., et al. (2022). *GEDI L4A footprint level aboveground biomass density, version 2.1*. ORNL DAAC. https://doi.org/10.3334/ORNLDAAC/2056

Dubayah, R. O., Sheldon, S. L., Clark, D. B., Hofton, M. A., Blair, J. B., Hurtt, G. C., & Chazdon, R. L. (2010). Estimation of tropical forest height and biomass dynamics using lidar remote sensing at La Selva, Costa Rica: Forest dynamics using lidar. *Journal of Geophysical Research*, *115*(G2). https://doi.org/10.1029/2009JG000933

Duncanson, L., Armston, J., Disney, M., Avitabile, V., Barbier, N., Calders, K., et al. (2019). The importance of consistent global forest aboveground biomass product validation. *Surveys in Geophysics*, *40*(4), 979–999. https://doi.org/10.1007/s10712-019-09538-8

Duncanson, L., Armston, J., Disney, M., Avitabile, V., Barbier, N., Calders, K., et al. (2021). Aboveground woody biomass product validation good practices protocol. In *Good practices for satellite derived land product validation* (p. 236). Land Product Validation Subgroup (WGCV/CEOS).

Duncanson, L., Huang, W., Johnson, K., Swatantran, A., McRoberts, R. E., & Dubayah, R. (2017). Implications of allometric model selection for county-level biomass mapping. *Carbon Balance and Management*, *12*(1), 18. https://doi.org/10.1186/s13021-017-0086-9

Duncanson, L., Kellner, J. R., Armston, J., Dubayah, R., Minor, D. M., Hancock, S., et al. (2022). Aboveground biomass density models for NASA's Global Ecosystem Dynamics Investigation (GEDI) lidar mission. *Remote Sensing of Environment*, *270*, 112845. https://doi.org/10.1016/j.rse.2021.112845

Duncanson, L., Neuenschwander, A., Hancock, S., Thomas, N., Fatoyinbo, T., Simard, M., et al. (2020). Biomass estimation from simulated GEDI, ICESat-2 and NISAR across environmental gradients in Sonoma County, California. *Remote Sensing of Environment*, *242*, 111779. https://doi.org/10.1016/j.rse.2020.111779

Duncanson, L. I., Dubayah, R. O., & Enquist, B. J. (2015). Assessing the general patterns of forest structure: Quantifying tree and forest allometric scaling relationships in the United States: Forest allometric variability in the United States. *Global Ecology and Biogeography*, *24*(12), 1465–1475. https://doi.org/10.1111/geb.12371

Esch, T., Marconcini, M., Felbier, A., Roth, A., Heldens, W., Huber, M., et al. (2013). Urban footprint processor—Fully automated processing chain generating settlement masks from global data of the TanDEM-X mission. *IEEE Geoscience and Remote Sensing Letters*, *10*(6), 1617–1621. https://doi.org/10.1109/LGRS.2013.2272953

Feldpausch, T. R., Banin, L., Phillips, O. L., Baker, T. R., Lewis, S. L., Quesada, C. A., et al. (2011). Height-diameter allometry of tropical forest trees. *Biogeosciences*, *8*(5), 1081–1106. https://doi.org/10.5194/bg-8-1081-2011

Flewelling, J. W., & Pienaar, L. V. (1981). Multiplicative regression with lognormal errors. *Forest Science*, *27*(2), 281–289. https://doi.org/10.1093/forestscience/27.2.281

Foody, G. M., Boyd, D. S., & Cutler, M. E. J. (2003). Predictive relations of tropical forest biomass from Landsat TM data and their transferability between regions. *Remote Sensing of Environment*, *85*(4), 463–474. https://doi.org/10.1016/S0034-4257(03)00039-7

Forrester, D. I., Tachauer, I. H. H., Annighoefer, P., Barbeito, I., Pretzsch, H., Ruiz-Peinado, R., et al. (2017). Generalized biomass and leaf area allometric equations for European tree species incorporating stand structure, tree age and climate. *Forest Ecology and Management*, *396*, 160–175. https://doi.org/10.1016/j.foreco.2017.04.011

Friedl, M. A., McIver, D. K., Hodges, J. C., Zhang, X. Y., Muchoney, D., Strahler, A. H., et al. (2002). Global land cover mapping from MODIS: Algorithms and early results. *Remote Sensing of Environment*, *83*(1–2), 287–302. https://doi.org/10.1016/s0034-4257(02)00078-0

Friedl, M. A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A., & Huang, X. (2010). MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. *Remote Sensing of Environment*, *114*(1), 168–182. https://doi.org/10.1016/j.rse.2009.08.016

Gonzalez de Tanago, J., Lau, A., Bartholomeus, H., Herold, M., Avitabile, V., Raumonen, P., et al. (2018). Estimation of above-ground biomass of large tropical trees with terrestrial LiDAR. *Methods in Ecology and Evolution*, *9*(2), 223–234. https://doi.org/10.1111/2041-210X.12904

Gregoire, T. G., Lin, Q. F., Boudreau, J., & Nelson, R. (2008). Regression estimation following the square-root transformation of the response. *Forest Science*, *54*(6), 597–606. https://doi.org/10.1093/forestscience/54.6.597

Hancock, S., Armston, J., Hofton, M., Sun, X., Tang, H., Duncanson, L. I., et al. (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

Hansen, E. H., Gobakken, T., Bollandsås, O. M., Zahabu, E., & Næsset, E. (2015). Modeling aboveground biomass in dense tropical submontane rainforest using airborne laser scanner data. *Remote Sensing*, *7*(1), 788–807. https://doi.org/10.3390/rs70100788

Hansen, M. C., Stehman, S. V., & Potapov, P. V. (2010). Quantification of global gross forest cover loss. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(19), 8650–8655. https://doi.org/10.1073/pnas.0912668107

Healey, S. P., Patterson, P. L., & Armston, J. (2022). Algorithm theoretical basis document (ATBD) for GEDI level-4B gridded aboveground biomass density.

Healey, S. P., Yang, Z., Gorelick, N., & Ilyushchenko, S. (2020). Highly local model calibration with a new GEDI LiDAR asset on Google Earth engine reduces landsat forest height signal saturation. *Remote Sensing*, *12*(17), 2840. https://doi.org/10.3390/rs12172840

Heath, L. S., Hansen, M., Smith, J. E., & Miles, P. D. (2009). Investigation into calculating tree biomass and carbon in the FIADB using a biomass expansion factor approach. In W. McWilliams, G. Moisen, & R. C. Czaplewski (Eds.), *Forest inventory and analysis (FIA) symposium 2008; October 21-23, 2008; Park city, UT. Proc. RMRS-P-56CD* (Vol. 56, p. 26). U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. Retrieved from https://www.fs.usda.gov/treesearch/pubs/33351

Hofton, M. A., & Blair, J. B. (2020). Algorithm theoretical basis document (ATBD) for GEDI transmit and receive waveform processing for L1 and L2 products.

Huete, A. R., Liu, H. Q., Batchily, K., & Leeuwen, W. V. (1997). A comparison of vegetation indices over a global set of TM images for EOS-MODIS. *Remote Sensing of Environment*, *59*(3), 440–451. https://doi.org/10.1016/S0034-4257(96)00112-5

Hyde, P., Dubayah, R., Peterson, B., Blair, J. B., Hofton, M., Hunsaker, C., et al. (2005). Mapping forest structure for wildlife habitat analysis using waveform lidar: Validation of montane ecosystems. *Remote Sensing of Environment*, *96*(3), 427–437. https://doi.org/10.1016/j.rse.2005.03.005

Jenkins, J. C., Chojnacky, D. C., Heath, L. S., & Birdsey, R. A. (2003). National-scale biomass estimators for United States tree species. *Forest Science*, *49*(1), 12–35.

Kellner, J. R., Armston, J., Birrer, M., Cushman, K. C., Duncanson, L., Eck, C., et al. (2019). New opportunities for forest remote sensing through ultra-high-density drone lidar. *Surveys in Geophysics*, *40*(4), 959–977. https://doi.org/10.1007/s10712-019-09529-9

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

Lang, N., Kalischek, N., Armston, J., Schindler, K., Dubayah, R., & Wegner, J. D. (2021). 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

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

Lefsky, M. A., Harding, D. J., Keller, M., Cohen, W. B., Carabajal, C. C., Del Bom Espirito-Santo, F., et al. (2005). Estimates of forest canopy height and aboveground biomass using ICESat: ICESat estimates of canopy height. *Geophysical Research Letters*, *32*(22). https://doi.org/10.1029/2005GL023971

Le Rest, K., Pinaud, D., Monestiez, P., Chadoeuf, J., & Bretagnolle, V. (2014). Spatial leave-one-out cross-validation for variable selection in the presence of spatial autocorrelation. *Global Ecology and Biogeography*, *23*(7), 811–820. https://doi.org/10.1111/geb.12161

Luckman, A., Baker, J., Honzák, M., & Lucas, R. (1998). Tropical forest biomass density estimation using JERS-1 SAR: Seasonal variation, confidence limits, and application to image mosaics. *Remote Sensing of Environment*, *63*(2), 126–139. https://doi.org/10.1016/S0034-4257(97)00133-8

Ma, L., Hurtt, G. C., Ott, L. E., Sahajpal, R., Fisk, J., Flanagan, S., et al. (2019). Global ecosystem demography model (ED-global v1.0): Development, calibration and evaluation for NASA's global ecosystem dynamics investigation (GEDI). In *AGU fall meeting abstracts* (Vol. *11*). Retrieved from http://adsabs.harvard.edu/abs/2019AGUFM.B11E2372M

Martínez Cano, I., Muller-Landau, H. C., Wright, S. J., Bohlman, S. A., & Pacala, S. W. (2019). Tropical tree height and crown allometries for the Barro Colorado nature monument, Panama: A comparison of alternative hierarchical models incorporating interspecific variation in relation to life history traits. *Biogeosciences*, *16*(4), 847–862. https://doi.org/10.5194/bg-16-847-2019

Mayr, E. (1944). Wallace's line in the light of recent zoogeographic studies. *The Quarterly Review of Biology*, *19*(1), 1–14. https://doi.org/10.1086/394684

Medvigy, D., Wofsy, S. C., Munger, J. W., & Moorcroft, P. R. (2010). Responses of terrestrial ecosystems and carbon budgets to current and future environmental variability. *Proceedings of the National Academy of Sciences*, *107*(18), 8275–8280. https://doi.org/10.1073/pnas.0912032107

Mitchard, E. T., Saatchi, S. S., Baccini, A., Asner, G. P., Goetz, S. J., Harris, N. L., & Brown, S. (2013). Uncertainty in the spatial distribution of tropical forest biomass: A comparison of pan-tropical maps. *Carbon Balance and Management*, *8*(1), 10. https://doi.org/10.1186/1750-0680-8-10

Mitchard, E. T., Saatchi, S. S., White, L. J. T., Abernethy, K. A., Jeffery, K. J., Lewis, S. L., et al. (2012). Mapping tropical forest biomass with radar and spaceborne LiDAR in Lopé National park, Gabon: Overcoming problems of high biomass and persistent cloud. *Biogeosciences*, *9*(1), 179–191. https://doi.org/10.5194/bg-9-179-2012

Moore, J. R. (2010). Allometric equations to predict the total above-ground biomass of radiata pine trees. *Annals of Forest Science*, *67*(8), 806. https://doi.org/10.1051/forest/2010042

Muukkonen, P. (2007). Generalized allometric volume and biomass equations for some tree species in Europe. *European Journal of Forest Research*, *126*(2), 157–166. https://doi.org/10.1007/s10342-007-0168-4

Narine, L. L., Popescu, S. C., & Malambo, L. (2020). Using ICESat-2 to estimate and map forest aboveground biomass: A first example. *Remote Sensing*, *12*(11), 1824. https://doi.org/10.3390/rs12111824

Næsset, E., Gobakken, T., Bollandsås, O. M., Gregoire, T. G., Nelson, R., & Ståhl, G. (2013). Comparison of precision of biomass estimates in regional field sample surveys and airborne LiDAR-assisted surveys in Hedmark County, Norway. *Remote Sensing of Environment*, *130*, 108–120. https://doi.org/10.1016/j.rse.2012.11.010

Pan, Y., Birdsey, R. A., Houghton, R., Kauppi, P., Kurz, W., Phillips, O. L., et al. (2011). A large and persistent carbon sink in the world's forests. *Science*, *333*(6045), 984–988. https://doi.org/10.1126/science.1204588

Patterson, P. L., Healey, S. P., Ståhl, G., Saarela, S., Holm, S., Andersen, H.-E., et al. (2019). Statistical properties of hybrid estimators proposed for GEDI—NASA's global ecosystem dynamics investigation. *Environmental Research Letters*, *14*(6), 065007. https://doi.org/10.1088/1748-9326/ab18df

Paul, K. I., Roxburgh, S. H., Chave, J., England, J. R., Zerihun, A., Specht, A., et al. (2016). Testing the generality of above-ground biomass allometry across plant functional types at the continent scale. *Global Change Biology*, *22*(6), 2106–2124. https://doi.org/10.1111/gcb.13201

Pickens, A. H., Hansen, M. C., Hancher, M., Stehman, S. V., Tyukavina, A., Potapov, P., et al. (2020). Mapping and sampling to characterize global inland water dynamics from 1999 to 2018 with full Landsat time-series. *Remote Sensing of Environment*, *243*, 111792. https://doi.org/10.1016/j.rse.2020.111792

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., et al. (2020). Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, *11*(1), 4540. https://doi.org/10.1038/s41467-020-18321-y

Pohjankukka, J., Pahikkala, T., Nevalainen, P., & Heikkonen, J. (2017). Estimating the prediction performance of spatial models via spatial k-fold cross validation. *International Journal of Geographical Information Science*, *31*(10), 2001–2019. https://doi.org/10.1080/13658816.2017.1346255

Qi, W., Saarela, S., Armston, J., Ståhl, G., & Dubayah, R. (2019). Forest biomass estimation over three distinct forest types using TanDEM-X InSAR data and simulated GEDI lidar data. *Remote Sensing of Environment*, *232*, 111283. https://doi.org/10.1016/j.rse.2019.111283

Rejou-Mechain, M., Muller-Landau, H. C., Detto, M., Thomas, S. C., Le Toan, T., Saatchi, S. S., et al. (2014). Local spatial structure of forest biomass and its consequences for remote sensing of carbon stocks. *Biogeosciences*, *11*(23), 6827–6840. https://doi.org/10.5194/bg-11-6827-2014

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. https://doi.org/10.1111/ecog.02881

Rosette, J., North, P. R. J., Rubio-Gil, J., Cook, B., Los, S., Suarez, J., et al. (2013). Evaluating prospects for improved forest parameter retrieval from satellite LiDAR using a physically-based radiative transfer model. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *6*(1), 45–53. https://doi.org/10.1109/JSTARS.2013.2244199

Roxburgh, S. H., Karunaratne, S. B., Paul, K. I., Lucas, R. M., Armston, J. D., & Sun, J. (2019). A revised above-ground maximum biomass layer for the Australian continent. *Forest Ecology and Management*, *432*, 264–275. https://doi.org/10.1016/j.foreco.2018.09.011

Saatchi, S., Marlier, M., Chazdon, R. L., Clark, D. B., & Russell, A. E. (2011). Impact of spatial variability of tropical forest structure on radar estimation of aboveground biomass. *Remote Sensing of Environment*, *115*(11), 2836–2849. https://doi.org/10.1016/j.rse.2010.07.015

Scipal, K., Arcioni, M., Chave, J., Dall, J., Fois, F., LeToan, T., et al. (2010). The BIOMASS mission #x2014; an ESA Earth explorer candidate to measure the BIOMASS of the Earth's forests. In *2010 IEEE international geoscience and remote sensing symposium* (pp. 52–55). https://doi.org/10.1109/IGARSS.2010.5648979

Siqueira, P., Armston, J., Chapman, B., Das, A., Dubayah, R., Kellndorfer, J., et al. (2021). Ecosystem sciences with NISAR. In *2021 IEEE international geoscience and remote sensing symposium IGARSS* (pp. 547–549). https://doi.org/10.1109/IGARSS47720.2021.9553600

Snowdon, P. (1991). A ratio estimator for bias correction in logarithmic regressions. *Canadian Journal of Forest Research*, *21*(5), 720–724. https://doi.org/10.1139/x91-101

Ståhl, G., Holm, S., Gregoire, T. G., Gobakken, T., Naesset, E., & Nelson, R. (2011). Model-based inference for biomass estimation in a LiDAR sample survey in Hedmark County, Norway. *Canadian Journal of Forest Research-Revue Canadienne De Recherche Forestiere*, *41*(1), 96–107. https://doi.org/10.1139/X10-161

Ståhl, G., Saarela, S., Schnell, S., Holm, S., Breidenbach, J., Healey, S. P., et al. (2016). Use of models in large-area forest surveys: Comparing model-assisted, model-based and hybrid estimation. *Forest Ecosystems*, *3*(1), 5. https://doi.org/10.1186/s40663-016-0064-9

Swatantran, A., Dubayah, R., Roberts, D., Hofton, M., & Blair, J. B. (2011). Mapping biomass and stress in the Sierra Nevada using lidar and hyperspectral data fusion. *Remote Sensing of Environment*, *115*(11), 2917–2930. https://doi.org/10.1016/j.rse.2010.08.027

Trochta, J., Krůček, M., Vrška, T., & Král, K. (2017). 3D Forest: An application for descriptions of three-dimensional forest structures using terrestrial LiDAR. *PLoS One*, *12*(5), e0176871. https://doi.org/10.1371/journal.pone.0176871

Ung, C.-H. U.-H., Bernier, P. B., & Guo, X.-J. G.-J. (2008). Canadian national biomass equations: New parameter estimates that include British Columbia data. *Canadian Journal of Forest Research*, *38*(5), 1123–1132. https://doi.org/10.1139/X07-224

Williamson, G. B., & Wiemann, M. C. (2010). Measuring wood specific gravity correctly. *American Journal of Botany*, *97*(3), 519–524. https://doi.org/10.3732/ajb.0900243

Zhang, X., Friedl, M. A., & Henebry, G. M. (2016). Algorithm theoretical basis document: VIIRS land surface phenology product.

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