Algorithm Theoretical Basis Document (ATBD)
for
GEDI Level-4A (L4A) Footprint Level Aboveground Biomass Density

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

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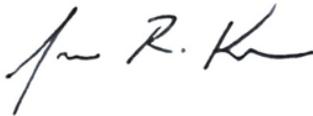[2] Department of Ecology, Evolution and Organismal Biology, Brown University, Providence RI
[3] Department of Geographical Sciences, University of Maryland, College Park MD

Version 1.0
Release date: August 11th, 2021
University of Maryland, College Park MD

Authors: Principal Investigator:

_____   _____

_____

_____

Abstract

The Global Ecosystem Dynamics Investigation (GEDI) lidar is a multibeam laser altimeter on the International Space Station. GEDI is the first spaceborne instrument designed specifically to measure vegetation structure and estimate 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 footprint-level AGBD (Mg · ha^{-1}) for individual GEDI footprints and the associated prediction uncertainty. GEDI04_A is a standalone data product, and GEDI04_A models are an input to the GEDI Level-4B (GEDI04_B) gridded AGBD data product. The GEDI04_A algorithm uses GEDI Level-2A (GEDI02_A) relative height metrics as input to parametric linear models to predict AGBD. GEDI04_A models were developed from a quality-filtered data set of GEDI footprint sized field plots paired with simulated GEDI waveforms across 21 countries and all continents within the GEDI domain (51.6 degrees N – S latitude). The models are stratified by combinations of world region and plant functional type (PFT). We describe the development of the GEDI04_A models and algorithm implementation for on-orbit prediction, including geographic transferability, elimination of GEDI02_A observations that do not meet requirements of the GEDI04_A algorithm, and quality flagging of GEDI04_A predictions. The GEDI04_A quality flag (l4_quality_flag) indicates the degree to which the input and output variables are representative of the conditions under which GEDI04_A models were developed.

Foreword

This document is the Algorithm Theoretical Basis Document for the GEDI Level-4A (L4A) Footprint Level Aboveground Biomass Density product. The GEDI Science Team assumes responsibility for this document and updates it, as required, as algorithms are refined. Reviews of this document are performed when appropriate and as needed updates to this document are made.

This document is a GEDI ATBD controlled document. Changes to this document require prior approval of the project. Proposed changes shall be noted in the change log, as well as incrementing the document version number.

Questions or comments concerning this document should be addressed to:

James R. Kellner
Department of Ecology, Evolution and Organismal Biology
Institute at Brown for Environment and Society
Brown University, Providence RI 02912
james_r_kellner@brown.edu
+1 (401) 863 5768

John Armston
2181 Lefrak Hall, Department of Geographical Sciences
University of Maryland, College Park MD 20742
armston@umd.edu
+1 (301) 405 8444

Laura Duncanson
2181 Lefrak Hall, Department of Geographical Sciences
University of Maryland, College Park MD 20742
lduncans@umd.edu
+1 (301) 405 3076

Ralph Dubayah
2181 Lefrak Hall, Department of Geographical Sciences
University of Maryland, College Park MD 20742
dubayah@umd.edu
+1 (301) 405 4069

Change History Log

Revision Level	Description of Change	Date Approved
1.0	Initial version	Aug 11, 2021

Table of Contents

List of Figures

List of Tables

1. INTRODUCTION

The Global Ecosystem Dynamics Investigation (GEDI) is producing measurements of vertical forest structure using a multibeam waveform lidar on the International Space Station (Dubayah et al., 2020a). 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 our understanding of the role of the land surface in the global carbon cycle (Canadell et al., 2007; Cox, 2019).

The sole GEDI observable is the received laser waveform (Dubayah et al., 2020a). The waveform represents the vertical distribution of intercepted surfaces within the extent of the illuminated laser footprint, and is available as the GEDI level-1B (GEDI01_B) data product (Dubayah et al., 2021). Estimating aboveground biomass density (AGBD) using GEDI data requires algorithms that process GEDI01_B waveforms to quantify relative height (RH) metrics. RH metrics define the percentage of the received laser waveform intensity that is less than a given height, where height is computed relative to the elevation of the lowest mode in the waveform. RH metrics are estimated in the presence of measurement uncertainty, including varying atmospheric conditions and solar background radiation. This uncertainty propagates through GEDI level-2A (GEDI02_A) RH metrics that are used to predict AGBD (Dubayah et al., 2020b).

The first release of the GEDI level-4A (GEDI04_A) data product is based on Version 1 of GEDI02_A (Dubayah et al., 2020b), and uses one of six algorithm setting groups to interpret the received waveform and identify the elevation of the lowest mode (Hofton and Blair, 2020). The Version 1 of GEDI04_A uses linear statistical models selected from an ensemble of candidates that predict AGBD as a function of one or more RH metrics. GEDI04_A models are a required input to the 1 km GEDI level-4B (GEDI04_B) AGBD data product (Patterson et al., 2019).

2. HISTORICAL PERSPECTIVE

Estimating AGBD using remote sensing requires aboveground biomass, M_i, for a sample of trees that has been computed using an allometric model in a fixed area, such as a field-inventory plot or lidar footprint. Summing the M_i over all individuals in the plot or footprint and expressing it per unit ground area produces an estimate of AGBD. Coincident remote sensing data are used to develop an empirical relationship between AGBD and a remotely sensed measurement. This relationship can then be used to predict AGBD using remotely sensed data (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 systems (Mitchard et al., 2009; Saatchi et al., 2011), discrete return airborne lidar (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 spaceborne waveform lidar (Boudreau et al., 2008; Lefsky et al., 2005; Rosette et al., 2013). Passive optical and SAR backscatter-based techniques typically saturate with increasing AGBD, with the degree of saturation depending on SAR wavelength (Huete et al., 1997; Luckman et al., 1998). Lidar consistently produces models

with the best performance in comparison to 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). GEDI faces a different challenge: it needs to develop 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 specifically to ensure transferability beyond the extent of calibration measurements (Friedl et al., 2002; Ploton et al., 2020).

3. STATISTICAL MODEL DEVELOPMENT

Models to produce GEDI04_A were developed using field estimates of AGBD colocated with simulated GEDI waveforms derived from discrete-return airborne lidar (Blair and 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 used to train GEDI models. Because GEDI is a sampling mission and most field plots are small, GEDI data will not intersect most of these locations during the mission life. The GEDI approach to developing footprint AGBD models considers multiple candidates stratified by world region and plant functional type (PFT; Fig. 1) with different functional forms.

An important objective for GEDI04_A models is that they are transferable 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 robust in the presence of errors and uncertainties, including measurement errors associated with the GEDI02_A algorithm for derivation of RH metrics (Hofton and Blair, 2020).
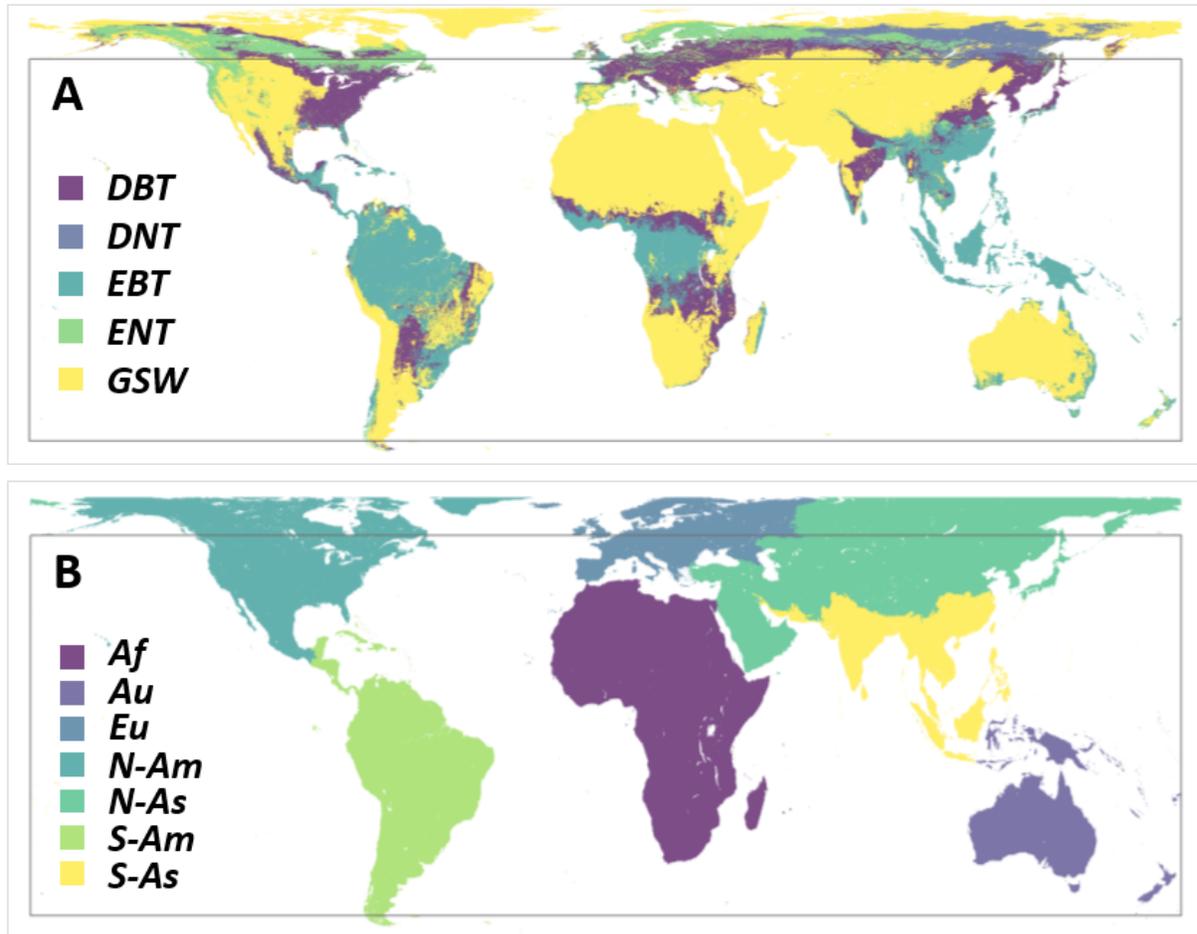
Figure 1. Global stratification by five combinations of error-corrected and infilled MODIS MCD12Q1 V006 PFT (A) and world region (B) to produce GEDI04_A models. The box inset is the GEDI observation domain of 51.6 degrees N to S latitude. DBT (deciduous broadleaf trees), DNT (deciduous needleleaf trees), EBT (evergreen broadleaf trees), ENT (evergreen needleleaf trees), GSW (grasses, shrubs and woodlands). Af (Africa), Au (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).

GEDI04_A models were developed using a quality-filtered calibration data set that contains simulated GEDI waveforms: the Forest Structure and Biomass Database (FSBD). This data set is one of the most exhaustive ever compiled for remote sensing of AGBD, but important regions are under-represented. These include the forests of continental Asia, the evergreen broadleaf forests throughout the islands of Southeast Asia and north of Australia, and the worldwide distribution of savannas and deciduous tropical forests (Table 1). To quantify geographic transferability, candidate models were evaluated within sets of 5-degree grid cells that contain simulated GEDI waveforms with coincident field data. This approach sets aside data from one grid cell for testing, and trains the model using data within the remaining grid cells. This model is used to predict AGBD within the held-out grid cell, and the process is repeated for all grid cells within each stratum for all models under consideration (Fig. 2).

Table 1. Numbers of simulated GEDI waveforms used for footprint model development and testing. 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).

	DBT	DNT	EBT	ENT	GSW	Total
Africa	490	0	834	0	6	**1,330**
Australia and Oceania	0	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	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**
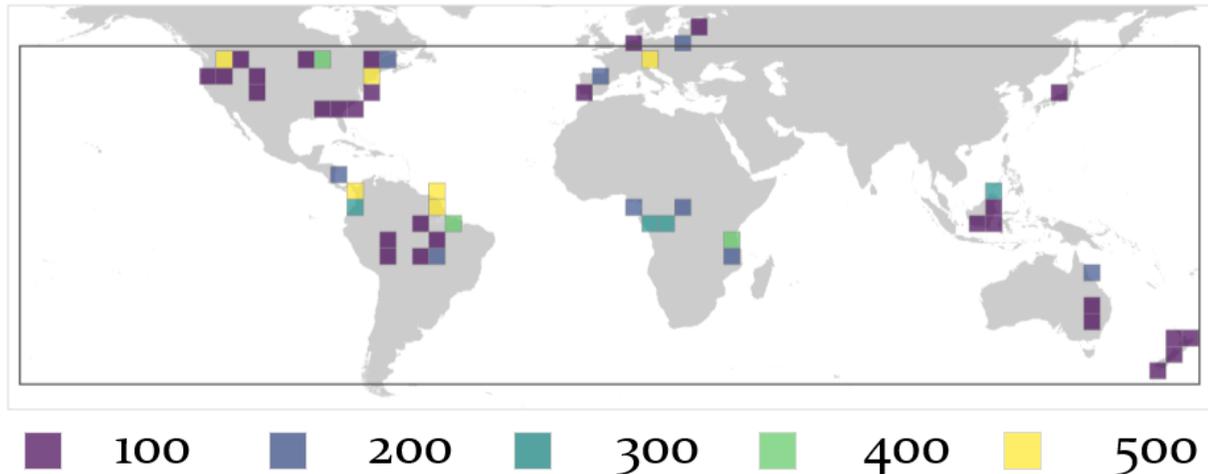
Figure 2. Geographic distribution of the number of simulated GEDI waveforms in the current version of the FSBD within 5 degree grid cells.

3.1. Stratification of GEDI04_A models

Building globally representative GEDI04_A models requires stratification. The models are stratified by world region and PFT (Fig. 1, Table 1). World regions are the geologically defined continents of Africa and Europe. 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, which defines the floral and faunal boundary between Australia and Asia during the Pleistocene (Mayr, 1944). 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. We divide the continent of Asia into north and south regions that approximately correspond to temperate and tropical forests (Fig 1.).

GEDI04_A models are stratified by combinations of PFT derived from an infilled and error-corrected version MODIS data product MCD12Q1 V006 (Friedl et al., 2010, 2002). 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; Fig. 1).

3.2 Quality-control filters

The FSBD is a living data archive that grows over time as new data sets are assimilated and improvements are made to existing records. The unfiltered database currently contains 31,414 simulated GEDI waveforms. After excluding incomplete projects and others that are inappropriate for GEDI (e.g., variable-radius plots), the unfiltered database contains 12,140 simulated GEDI waveforms. After applying quality-control filters, the database used to develop Version 1 of the GEDI04_A data product contains 8,587 simulated waveforms from 21 countries (Table 1). Below we indicate 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 FSBD.

3.2.1. Incongruent AGBD and height

We excluded footprints when there was an incongruence between field-estimated AGBD and simulated RH98. In particular, when AGBD was < 1 Mg $\cdot$ ha^{-1} and RH98 was > 5 m, the footprint was excluded (113 footprints, or 0.93% of the unfiltered database). When AGBD was > 150 Mg $\cdot$ ha^{-1} 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 CCF 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 with no elevation variation. 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 individual tree mass (M_i) require tree height. When the difference between measured or predicted height and RH98 was > 10 m, we excluded the footprint (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 domain over which predictions are valid. These 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 authors, the footprint was excluded (640 footprints, or 5.27%).

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. 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 from the FSBD (129 footprints, or 1.06%).

3.2.6. Large sample size

Data in the FSBD were contributed by numerous researchers without whom the development of comprehensive GEDI04_A models would not be possible. 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, stem-mapped plots contribute disproportionately to the total number of observations in the FSBD. 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 analysis-ready data. When the number of footprints in a plot was < 200, we accepted all footprints that passed other filters. When the number of footprints in a plot was > 200 after applying other filters, we collected a stratified random sample of 200 footprints, 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 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. Stratification sample size, variable selection, and multicollinearity

We developed models 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. We considered models 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 RH metrics in increments of 10% and RH98 in addition to all possible two-way interactions.

We refer to combinations of PFT and world region as prediction strata. When there are training data to develop a model for a given prediction stratum, we use that model within the given combination of PFT and world region. However, there are 25 out of 35 prediction strata represented by < 50 footprints in the filtered GEDI FSBD. In these cases, we use the corresponding PFT model in 18 cases and an alternative model stratified by PFT and world region in 7 cases. These 7 cases represent two EBT strata, two DBT strata, and three DNT strata.

In the EBT prediction stratum within Europe and North America we use the corresponding DBT by world region models. We assume 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 current version of the filtered FSBD, EBT samples are exclusively tropical or *Eucalyptus*, and thus not representative of EBT in North America or Europe. In the DBT stratum within the South America world region and the country of Australia, we use the corresponding EBT model for the associated world region. In Australia, we assume the DBT classification is an error in MCD12Q1, because Australia lacks deciduous broadleaf forests. In the South America world region, DBT forests are likely to be tropical moist or dry forests that are similar to EBT of South America. Finally, we lack training data in DNT globally. We use a corresponding ENT by world region model for three of these strata (Australia, Europe, and North America). In the remaining DNT strata we lack a corresponding ENT by world region model and use the global ENT model.

We computed all possible 1, 2, 3, and 4 variable predictor matrices. Two-way interactions were permitted in the absence of main effects. For example, we considered models that contained the interaction between RH50 and RH98, even when RH50 and RH98 were not included in the model as independent variables. We then eliminated predictor matrices that were multicollinear. We quantified multicollinearity by computing the Pearson correlation matrix for the candidate predictor matrix. If the maximum absolute value of the correlation coefficient was > 0.9, the candidate predictor matrix was excluded. We also computed the variance inflation factor (VIF). When VIF was > 10, the predictor matrix was excluded. For all candidates that passed both multicollinearity tests, we fit a weighted linear model by regressing the transformation of AGBD on the predictors. Weights were inversely proportional to the number of simulated footprints in each 5 degree grid cell used to quantify geographic transferability 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, smallest percentage root mean squared error, 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:

$$\frac{1}{5}\sum_{j=1}^{5}\frac{1}{n_j}\sum_{i=1}^{n_j} Observed_{i,j} - Expected_{i,j} \tag{1}$$

The interior sum is the mean within bin j, where j is one of five quantile bins computed using predicted AGBD. Computing the mean residual error within quantile bins favors models that perform well across the range of AGBD. The percentage RMSE was computed according to:

$$RMSE = 100 \times \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Observed_i - Expected_{i,j}\right)^2} \Big/ \frac{1}{n}\sum_{i=1}^{n} Observed_i \tag{2}$$

The model selection favors candidates that contain larger valued RH metrics over models with similar mean residual error and RMSE than models that contain smaller valued RH metrics. This is because RH metrics closer to the ground are more sensitive to differences between simulated and real GEDI waveforms than RH metrics higher in the canopy. Reducing simulator error for smaller valued RH metrics using the on-orbit transmit pulse shape and characteristics of recorded GEDI noise will be addressed in a subsequent version of GEDI04_A. Models with fewer coefficients and fewer RH metrics are preferred based on parsimony. The number of coefficients is not directly proportional to the number of RH metrics because candidate models contain interactions. For example, a model that contains the interaction between RH98 and RH50 as a single predictor contains two coefficients and two RH metrics. A model that contains only RH98 and RH50 as main effects contains three coefficients and two RH metrics.

4. ALGORITHM DESCRIPTION

The GEDI04_A data product is AGBD ($Mg \cdot ha^{-1}$) for individual GEDI footprints and the associated prediction uncertainty. The GEDI04_A algorithm ingests GEDI02_A data and external input variables (Fig. 2, Table 2). 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: 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 and Blair, 2020). For more information about waveform processing, see the ATBD for GEDI transmit and receive waveform processing (Hofton and Blair, 2020). When these conditions are met, the GEDI04_A algorithm_run_flag = 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 3).

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 metrics met minimum quality standards 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 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 l4_quality_flag indicates whether each GEDI observation is representative of the conditions under which GEDI04_A models were developed. The 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_percentage < 50. The variable landsat_water_persistence indicates permanent water bodies. The leaf_off_flag indicates whether the footprint was collected under leaf-off or leaf-on conditions, and was derived for a 1 km EASE 2.0 grid using the VIIRS land surface phenology product VNP22Q2 (Zhang et al., 2016). leaf_off_flag = 1 when the footprint was collected after the onset of maximum greenness and before the midpoint of the senescence phase for the given 1 km grid cell.

Some GEDI04_A models use exclusively RH98 to predict AGBD. When this is the case, the l4_quality_flag does not consider leaf_off_flag. This is because the impact of leaf-off conditions is assumed to be minimal for RH98. The variable urban_percentage is from a 25 m global urban mask developed by the GEDI Science Team using the TerraSAR-X and TanDEM-X global urban footprint (GUF) data product (Esch et al., 2013).

Two diagnostic flags are provided independent of l4_quality_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 3). These flags 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.
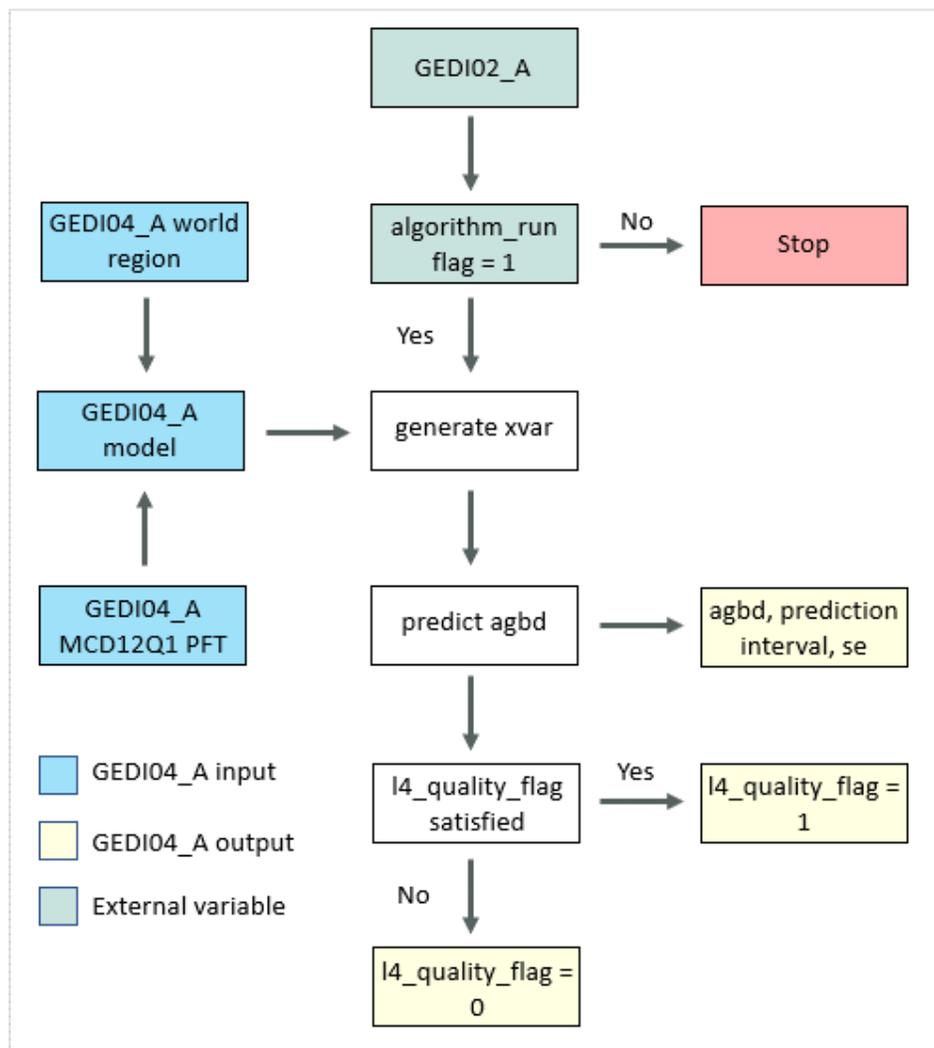
Figure 3. GEDI04_A algorithm flow. The GEDI04_A algorithm assimilates external data from GEDI02_A and other sources. A prediction is generated for every GEDI shot where algorithm_run_flag = 1. The algorithm looks up the GEDI04_A model using a world region grid and error-corrected and infilled MODIS MCD12Q1 PFT (Fig. 1). 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.

4.1. Scientific theory

All remote sensing estimates of AGBD rely on field measurements of individual tree structure (Clark and Kellner, 2012; Labriere et al., 2018). 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 that M_i is a power function of trunk diameter, X_i:

$$M_i \propto aX_i^b \qquad\qquad (3)$$

Numerous investigations including hundreds of species have demonstrated that M_i is a power function of 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. Some models use tree height in addition to DBH, and others incorporate wood specific gravity, defined as oven-dried mass divided by wet volume (Williamson and Wiemann, 2010). Because it is not always 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 we used an allometric model appropriate to the given PFT and world region to predict M_i. When there was more than one model that could be used for a given tree, we favored locally developed models over regional ones, as long as locally-developed models were not site-specific. We also favored models with finer taxonomic resolution. In Australia we use eight allometric models developed by Paul et al. (2016) and Roxburgh et al. (2019). In New Zealand we use 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 use the models of Jenkins et al. (2003) in the continental United States and the models of Ung et al. (2008) in Canada. In Europe we use the allometric models of Forrester et al. (2017). Throughout the tropics of South America, Africa and Asia we use the model of Chave et al. (2014).

In some situations there was more than one candidate model to predict M_i that met GEDI04_A requirements. For example, the models of Brown (1997) and Chave et al. (2014) have been used to predict M_i in Central American evergreen broadleaf 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 of the United States of America. There is evidence that the CRM underestimates AGBD, especially in deciduous broadleaf forests of the eastern United States (Duncanson et al., 2017; Radtke et al., 2017). However, the CRM produces values that are greater than Jenkins et al. (2003) in forests taller than 50 m. Choosing a model to predict M_i is important and has resulted in large

17

discrepancies in estimates of AGBD from spaceborne remote sensing (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. Subsequent versions of GEDI04_A may use different allometric models.

Some contributed data records for individual trees included measurements of height and diameter, but some measurements only contained diameter. When the allometric model required height as an input and a measured value was available, we used the measured value. When no height measurement was available, we predicted height in one of three ways. If some records contained estimates of height and diameter, 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 (FIA) data within the same county. We then applied this locally developed model to predict tree height.

4.2. Scientific assumptions

We assume that allometric models generate accurate estimates of M_i when applied to non-harvested trees. Whether this assumption is true has been debated. Harvested trees used to develop allometric scaling relationships are usually not randomly sampled (Clark and Kellner, 2012), and validation studies that directly measure tree mass have demonstrated that scaling 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; 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. When the coordinates of a given stem are within the extent of the footprint, M_i, as defined by a given allometric scaling model, is assigned to the footprint. In practice, there are four potential violations of this assumption. First, across-beam laser intensity follows a Gaussian distribution (Blair and Hofton, 1999). This means that intercepted surfaces near the center of the footprint have a stronger impact on the returned laser waveform than objects near the edges (Hyde et al., 2005). Second, because the across-beam laser intensity is Gaussian, the extent of the footprint is infinite. 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 with constant elevation is received from surfaces beyond the 12.5 m threshold. The third and fourth assumptions address the size of GEDI footprints relative to the trees within them. 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.

Collectively, the third and fourth assumptions are called the point-mass assumption, because they assume that M_i is associated with a point location defined by the stem position. Violation of the point-mass assumption is likely to be important for very large trees that contribute to large values of AGBD (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).

Simulated waveforms used to develop the GEDI04_A models were generated in the absence of sensor noise, where RH metrics are known exactly. RH metrics can be determined without error on simulated waveforms because they can be computed relative to true ground. Transferring these models from simulated to recorded GEDI data requires the assumption that ground-finding 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 resulted in GEDI02_A ground elevation estimates that were biased high, and therefore in RH metrics that were biased low. Version 2 of GEDI02_A addressed this issue by using one of six algorithm settings to interpret the received waveform, rather than one (Hofton and Blair, 2020). Optimal settings for every combination of the GEDI04_A modified MCD12Q1 PFT and world region have been identified by the GEDI Science Team using a comprehensive dataset of GEDI-ALS crossovers. GEDI-ALS crossovers are locations where recorded GEDI data intersects discrete-return airborne lidar. At these locations we can remove systematic geolocation error in recorded GEDI data and compare GEDI waveforms to simulated waveforms developed using discrete-return lidar data. These comparisons enable selection of optimal algorithm settings by comparison to true ground.

GEDI04_A models were developed using training data collected under leaf-on conditions. We use leaf_off_flag to identify GEDI waveforms that are likely to be under leaf-on conditions. However the use of this flag in drought deciduous tropical forests may be problematic. This is because some EBT forests experience periods of partial deciduousness during which some percentage of crowns are leafless while the canopy as a whole is green. For example, a study across a rainfall gradient in Panama found that 3.6 – 19.1% of crown area was leafless at peak deciduousness (Condit et al., 2000). All of the areas in this study are classified as EBT using MCD12Q1. This indicates that some GEDI footprints with leaf_off_flag = 0 may represent partial leaf-off conditions in practice. We assume that the GEDI04_A training data are representative of the variability introduced by such partial leaf-off conditions and that the impact of this variability is subsumed into the GEDI04_A model parameter uncertainty estimates.

As noted in above, 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). For example, we lack training data in continental Asia and throughout the GSW and DNT stratifications worldwide. In strata where training data are lacking, we assume that a model developed for a different stratum can be applied to that stratum to produce unbiased predictions of AGBD.

4.3. Mathematical 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: (i) a square-root transformation on the response, (ii) a square-root transformation on the response and predictors, (iii) a natural logarithm transformation on the response, and (iv) 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. For models using the natural logarithm transformation, we considered two bias corrections. The method originally developed by Baskerville (1972) transforms values from the natural logarithm scale to the original scale using:

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

The term $\mathbf{X}_i\boldsymbol{\beta}_j$ 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 $\boldsymbol{\beta}_j$ is column vector of coefficients. $AGBD_i$ is the natural logarithm of AGBD from field data within the simulated GEDI footprint.

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:

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

C_j is a bias-correction coefficient:

$$C_j = \frac{\sum_{i=1}^{n} AGBD_i}{\frac{1}{n}\sum_{i=1}^{n} exp(\mathbf{X}_i\boldsymbol{\beta}_j)} \tag{6}$$

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

$$C_j = \frac{\sum_{i=1}^{n} AGBD_i}{\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i\boldsymbol{\beta}_j{}^2} \tag{7}$$

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

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

We provide prediction intervals and estimates of the standard error of the prediction for every predicted value of AGBD. The standard error of the prediction for GEDI footprint h is:

$$SE_h = \sqrt{MSE_k + \mathbf{X}_h \, \text{Cov}(\boldsymbol{\beta}) \, \mathbf{X}_h^{\text{T}}} \tag{9}$$

Here, MSE_k is the square of the residual standard error from the linear regression applied to prediction stratum k containing GEDI footprint h, $\mathbf{X}_h$ is the row vector of scaled and transformed RH metrics for GEDI footprint h, and $\text{Cov}(\boldsymbol{\beta})$ 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 according to:

$$\widehat{AGBD}_h \pm t_{\left(1-\frac{\alpha}{2}, n-2\right)} \times SE_h \tag{10}$$

The t multiplier is the value from a t distribution with confidence level α and $n-2$ degrees of freedom. 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.

4.4. Mathematical 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 propagate model parameter uncertainty in GEDI04_A models to the 1 km GEDI04_B AGBD data product (Ståhl et al., 2011). We also assume 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 35 prediction strata. The Version 1 GEDI04_A product uses Version 1 GEDI02_A as input. However, we applied the algorithm setting group selection being implemented in Version 2 GEDI02_A to the release 1 GEDI02_A data on a per-footprint basis. The algorithm setting group used for each footprint is contained in the selected_algorithm variable in the root group of the GEDI04_A 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 used to calculate RH metrics (Hofton and Blair, 2020).

Table 2. GEDI04_A input variables. Input variables are required to run the GEDI04_A algorithm. These variables are available for every footprint in the GEDI04_A data product.

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 (Snowdon or Baskerville)
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 deciduous needleleaf or deciduous broadleaf forests (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

4.6. Algorithm output variables

The GEDI04_A algorithm outputs predicted AGBD in original (Mg · ha^{-1}) and transformed units, associated prediction intervals, the standard error of the prediction, quality flags, and other ancillary information (Table 3). The algorithm produces these data for every

algorithm selection setting and identifies the best selection setting for each waveform. For models that contain exclusively RH98 as a predictor, the l4a_quality_flag = 1 (good) when l2_quality_flag = 1, sensitivity > 0.95, landsat_water_persistence < 10, and urban_percentage < 50. When there is more than one predictor, the l4_quality_flag also requires leaf_off_flag = 0, which indicates leaf-on conditions.

Table 3. GEDI04_A output variables. Output variables are produced by GEDI04_A algorithm. These variables are available for every footprint in the GEDI04_A data product, except alpha which is an attribute of the agbd_prediction group for the beam.

Output variable	Units	Description
agbd	$Mg \cdot ha^{-1}$	Predicted AGBD using the optimal algorithm setting
agbd_aN	$Mg \cdot ha^{-1}$	Predicted AGBD using algorithm setting N
agbd_pi_lower	$Mg \cdot ha^{-1}$	Lower prediction interval for agbd, given alpha
agbd_pi_lower_aN	$Mg \cdot ha^{-1}$	Lower prediction interval for agbd_aN, given alpha
agbd_pi_upper	$Mg \cdot ha^{-1}$	Upper prediction interval for agbd, given alpha
agbd_pi_upper_aN	$Mg \cdot ha^{-1}$	Upper prediction interval for agbd_aN, given alpha
agbd_se	$Mg \cdot ha^{-1}$	The standard error of the agbd prediction
agbd_se_aN	$Mg \cdot ha^{-1}$	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, leaf_off_flag = 0, and urban_percentage < 50
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)

23

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)

5. ALGORITHM IMPLEMENTATIONS

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 (UMD), in collaboration with the GEDI Science Data Processing System at the NASA Goddard Space Flight Center (GSFC) in Greenbelt, Maryland and the Institute at Brown for Environment and Society (IBES) at Brown University.

6. ALGORITHM USAGE CONSTRAINTS

There are no algorithm usage constraints.

7. PERFORMANCE ASSESSMENT VALIDATION

7.1. Performance assessment validation methods

The performance of GEDI04_A models is being assessed using a comprehensive database of GEDI-ALS crossovers and other GEDI data. By removing systematic geolocation error in GEDI data, transferability of GEDI04_A models from simulated to recorded GEDI data can be tested.

7.2. Performance assessment validation uncertainties

Coincident GEDI waveforms and discrete-return airborne lidar exist for a sample of locations worldwide, but do not provide a systematic or random sample of the land surface within the GEDI domain.

8. DATA ACCESS

8.1. Data access input data

Input predictors and selected model data are publicly available and contained with the GEDI04_A data product, which is accessible through the Oak Ridge National Laboratory Distributed Archive Center (ORNL DAAC).

8.2. Data access output data

The GEDI04_A data product is publicly available and accessible through ORNL DAAC.

8.3. Data access related URLs

Level-4 GEDI data products are available through the ORNL DAAC: https://daac.ornl.gov

10. REFERENCES

Baskerville, G.L., 1972. Use of Logarithmic Regression in the Estimation of Plant Biomass. Canadian Journal of Forest Research. https://doi.org/10.1139/x72-009

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 177–189.

Blair, J.B., Hofton, M.A., 1999. Modeling laser altimeter return waveforms over complex vegetation using high-resolution elevation data. Geophys. Res. Lett. 26, 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, 3876–3890. https://doi.org/10.1016/j.rse.2008.06.003

Brown, S., 1997. Estimating biomass and biomass change of tropical forests: a primer (FAO Forestry Paper - 134).

Calders, K., Adams, J., Armston, J., Bartholomeus, H., Bauwens, S., Bentley, L.P., Chave, J., Danson, F.M., Demol, M., Disney, M., Gaulton, R., Krishna Moorthy, S.M., Levick, S.R., Saarinen, N., Schaaf, C., Stovall, A., Terryn, L., Wilkes, P., Verbeeck, H., 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

Canadell, J.G., Quéré, C.L., Raupach, M.R., Field, C.B., Buitenhuis, E.T., Ciais, P., Conway, T.J., Gillett, N.P., Houghton, R.A., Marland, G., 2007. Contributions to accelerating atmospheric CO2 growth from economic activity, carbon intensity, and efficiency of natural sinks. PNAS 104, 18866–18870. https://doi.org/10.1073/pnas.0702737104

Chave, J., Réjou-Méchain, M., Búrquez, A., Chidumayo, E., Colgan, M.S., Delitti, W.B.C., Duque, A., Eid, T., Fearnside, P.M., Goodman, R.C., Henry, M., Martínez-Yrízar, A., Mugasha, W.A., Muller-Landau, H.C., Mencuccini, M., Nelson, B.W., Ngomanda, A., Nogueira, E.M., Ortiz-Malavassi, E., Pélissier, R., Ploton, P., Ryan, C.M., Saldarriaga, J.G., Vieilledent, G., 2014. Improved allometric models to estimate the aboveground biomass of tropical trees. Global Change Biology 20, 3177–3190. https://doi.org/10.1111/gcb.12629

Clark, D.B., Kellner, J.R., 2012. Tropical forest biomass estimation and the fallacy of misplaced concreteness. J Veg Sci 23, 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. J. Veg. Sci. 11, 649–658. https://doi.org/10.2307/3236572

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

Cox, P.M., 2019. Emergent Constraints on Climate-Carbon Cycle Feedbacks. Curr Clim Change Rep 5, 275–281. https://doi.org/10.1007/s40641-019-00141-y

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

Dubayah, R., Blair, J.B., Goetz, S., Fatoyinbo, L., Hansen, M., Healey, S., Hofton, M., Hurtt, G., Kellner, J., Luthcke, S., Armston, J., Tang, H., Duncanson, L., Hancock, S., Jantz, P.,

Marselis, S., Patterson, P.L., Qi, W., Silva, C., 2020a. 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., 2020b. GEDI L2A Elevation and Height Metrics Data Global Footprint Level V001.

Dubayah, R., Luthcke, S., Blair, J., Hofton, M., Armston, J., Tang, H., 2021. GEDI L1B Geolocated Waveform Data Global Footprint Level V002. https://doi.org/10.5067/GEDI/GEDI01_B.001

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: Biogeosciences 115, n/a-n/a. https://doi.org/10.1029/2009JG000933

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 Manage 12, 18. https://doi.org/10.1186/s13021-017-0086-9

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, 1465–1475. https://doi.org/10.1111/geb.12371

Esch, T., Marconcini, M., Felbier, A., Roth, A., Heldens, W., Huber, M., Schwinger, M., Taubenböck, H., Müller, A., Dech, S., 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, 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., Affum-Baffoe, K., Arets, E.J.M.M., Berry, N.J., Bird, M., Brondizio, E.S., de Camargo, P., Chave, J., Djagbletey, G., Domingues, T.F., Drescher, M., Fearnside, P.M., França, M.B., Fyllas, N.M., Lopez-Gonzalez, G., Hladik, A., Higuchi, N., Hunter, M.O., Iida, Y., Salim, K.A., Kassim, A.R., Keller, M., Kemp, J., King, D.A., Lovett, J.C., Marimon, B.S., Marimon-Junior, B.H., Lenza, E., Marshall, A.R., Metcalfe, D.J., Mitchard, E.T.A., Moran, E.F., Nelson, B.W., Nilus, R., Nogueira, E.M., Palace, M., Patiño, S., Peh, K.S.-H., Raventos, M.T., Reitsma, J.M., Saiz, G., Schrodt, F., Sonké, B., Taedoumg, H.E., Tan, S., White, L., Wöll, H., Lloyd, J., 2011. Height-diameter allometry of tropical forest trees. Biogeosciences 8, 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, 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, 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., Stark, H., Vacchiano, G., Zlatanov, T., Chakraborty, T., Saha, S., Sileshi, G.W., 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., Woodcock, C.E., Gopal, S., Schneider, A., Cooper, A., 2002. Global land cover mapping from MODIS: algorithms and early results. Remote sensing of Environment 83, 287–302.

Gonzalez de Tanago, J., Lau, A., Bartholomeus, H., Herold, M., Avitabile, V., Raumonen, P., Martius, C., Goodman, R.C., Disney, M., Manuri, S., Burt, A., Calders, K., 2018. Estimation of above-ground biomass of large tropical trees with terrestrial LiDAR. Methods in Ecology and Evolution 9, 223–234. https://doi.org/10.1111/2041-210X.12904

Hancock, S., Armston, J., Hofton, M., Sun, X., Tang, H., Duncanson, L.I., Kellner, J.R., Dubayah, R., 2019. The GEDI Simulator: A Large-Footprint Waveform Lidar Simulator for Calibration and Validation of Spaceborne Missions. Earth and Space Science 6, 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, 788–807. https://doi.org/10.3390/rs70100788

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: McWilliams, Will; Moisen, Gretchen; Czaplewski, Ray, comps. Forest Inventory and Analysis (FIA) Symposium 2008; October 21-23, 2008; Park City, UT. Proc. RMRS-P-56CD. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 26 p. 056.

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, 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., Knox, R., Walker, W., 2005. Mapping forest structure for wildlife habitat analysis using waveform lidar: Validation of montane ecosystems. Remote Sensing of Environment 96, 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, 12–35.

Kellner, J.R., Armston, J., Birrer, M., Cushman, K.C., Duncanson, L., Eck, C., Falleger, C., Imbach, B., Král, K., Krůček, M., Trochta, J., Vrška, T., Zgraggen, C., 2019. New Opportunities for Forest Remote Sensing Through Ultra-High-Density Drone Lidar. Surv Geophys 40, 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, 1592. https://doi.org/10.3390/rs13081592

Labriere, N., Tao, S., Chave, J., Scipal, K., Toan, T.L., Abernethy, K., Alonso, A., Barbier, N., Bissiengou, P., Casal, T., Davies, S.J., Ferraz, A., Herault, B., Jaouen, G., Jeffery, K.J., Kenfack, D., Korte, L., Lewis, S.L., Malhi, Y., Memiaghe, H.R., Poulsen, J.R., Rejou-Mechain, M., Villard, L., Vincent, G., White, L.J.T., Saatchi, S., 2018. In Situ Reference Datasets From the TropiSAR and AfriSAR Campaigns in Support of Upcoming Spaceborne Biomass Missions. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 1–11. https://doi.org/10.1109/JSTARS.2018.2851606

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, 393–399.

Lefsky, M.A., Harding, D.J., Keller, M., Cohen, W.B., Carabajal, C.C., Del Bom Espirito-Santo, F., Hunter, M.O., de Oliveira, R., 2005. Estimates of forest canopy height and aboveground biomass using ICESat: ICESAT ESTIMATES OF CANOPY HEIGHT. Geophysical Research Letters 32, n/a-n/a. https://doi.org/10.1029/2005GL023971

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, 126–139. https://doi.org/10.1016/S0034-4257(97)00133-8

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, 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, 1–14.

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, 10. https://doi.org/10.1186/1750-0680-8-10

Mitchard, E.T.A., Saatchi, S.S., Woodhouse, I.H., Nangendo, G., Ribeiro, N.S., Williams, M., Ryan, C.M., Lewis, S.L., Feldpausch, T.R., Meir, P., 2009. Using satellite radar backscatter to predict above-ground woody biomass: A consistent relationship across four different African landscapes. Geophys. Res. Lett. 36, L23401. https://doi.org/10.1029/2009GL040692

Moore, J.R., 2010. Allometric equations to predict the total above-ground biomass of radiata pine trees. Ann. For. Sci. 67, 806–806. 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, 157–166.

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

Patterson, P.L., Healey, S.P., Ståhl, G., Saarela, S., Holm, S., Andersen, H.-E., Dubayah, R.O., Duncanson, L., Hancock, S., Armston, J., Kellner, J.R., Cohen, W.B., Yang, Z., 2019. Statistical properties of hybrid estimators proposed for GEDI—NASA's global ecosystem dynamics investigation. Environ. Res. Lett. 14, 065007. https://doi.org/10.1088/1748-9326/ab18df

Paul, K.I., Roxburgh, S.H., Chave, J., England, J.R., Zerihun, A., Specht, A., Lewis, T., Bennett, L.T., Baker, T.G., Adams, M.A., Huxtable, D., Montagu, K.D., Falster, D.S., Feller, M., Sochacki, S., Ritson, P., Bastin, G., Bartle, J., Wildy, D., Hobbs, T., Larmour, J., Waterworth, R., Stewart, H.T.L., Jonson, J., Forrester, D.I., Applegate, G., Mendham, D., Bradford, M., O'Grady, A., Green, D., Sudmeyer, R., Rance, S.J., Turner, J., Barton, C., Wenk, E.H., Grove, T., Attiwill, P.M., Pinkard, E., Butler, D., Brooksbank, K., Spencer,

B., Snowdon, P., O'Brien, N., Battaglia, M., Cameron, D.M., Hamilton, S., McAuthur, G., Sinclair, J., 2016. Testing the generality of above-ground biomass allometry across plant functional types at the continent scale. Glob Change Biol 22, 2106–2124. https://doi.org/10.1111/gcb.13201

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., Pélissier, R., 2020. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nature Communications 11, 4540. https://doi.org/10.1038/s41467-020-18321-y

Radtke, P., Walker, D., Frank, J., Weiskittel, A., DeYoung, C., MacFarlane, D., Domke, G., Woodall, C., Coulston, J., Westfall, J., 2017. Improved accuracy of aboveground biomass and carbon estimates for live trees in forests of the eastern United States. Forestry: An International Journal of Forest Research 90, 32–46. https://doi.org/10.1093/forestry/cpw047

Rejou-Mechain, M., Muller-Landau, H.C., Detto, M., Thomas, S.C., Le Toan, T., Saatchi, S.S., Barreto-Silva, J.S., Bourg, N.A., Bunyavejchewin, S., Butt, N., Brockelman, W.Y., Cao, M., Cardenas, D., Chiang, J.-M., Chuyong, G.B., Clay, K., Condit, R., Dattaraja, H.S., Davies, S.J., Duque, A., Esufali, S., Ewango, C., Fernando, R.H.S., Fletcher, C.D., Gunatilleke, I. a. U.N., Hao, Z., Harms, K.E., Hart, T.B., Herault, B., Howe, R.W., Hubbell, S.P., Johnson, D.J., Kenfack, D., Larson, A.J., Lin, L., Lin, Y., Lutz, J.A., Makana, J.-R., Malhi, Y., Marthews, T.R., McEwan, R.W., McMahon, S.M., McShea, W.J., Muscarella, R., Nathalang, A., Noor, N.S.M., Nytch, C.J., Oliveira, A.A., Phillips, R.P., Pongpattananurak, N., Punchi-Manage, R., Salim, R., Schurman, J., Sukumar, R., Suresh, H.S., Suwanvecho, U., Thomas, D.W., Thompson, J., Uriarte, M., Valencia, R., Vicentini, A., Wolf, A.T., Yap, S., Yuan, Z., Zartman, C.E., Zimmerman, J.K., Chave, J., 2014. Local spatial structure of forest biomass and its consequences for remote sensing of carbon stocks. Biogeosciences 11, 6827–6840. https://doi.org/10.5194/bg-11-6827-2014

Rosette, J., North, P.R.J., Rubio-Gil, J., Cook, B., Los, S., Suarez, J., Sun, G., Ranson, J., Blair, J.B., 2013. Evaluating Prospects for Improved Forest Parameter Retrieval From Satellite LiDAR Using a Physically-Based Radiative Transfer Model. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. 6, 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

Saarela, S., Holm, S., Grafström, A., Schnell, S., Næsset, E., Gregoire, T.G., Nelson, R.F., Ståhl, G., 2016. Hierarchical model-based inference for forest inventory utilizing three sources of information. Annals of Forest Science 73, 895–910. https://doi.org/10.1007/s13595-016-0590-1

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, DESDynI VEG-3D Special Issue 115, 2836–2849. https://doi.org/10.1016/j.rse.2010.07.015

Snowdon, P., 1991. A ratio estimator for bias correction in logarithmic regressions. Canadian Journal of Forest Research. 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. Can. J. For. Res.-Rev. Can. Rech. For. 41, 96–107. https://doi.org/10.1139/X10-161

Ståhl, G., Saarela, S., Schnell, S., Holm, S., Breidenbach, J., Healey, S.P., Patterson, P.L., Magnussen, S., Næsset, E., McRoberts, R.E., Gregoire, T.G., 2016. Use of models in large-area forest surveys: comparing model-assisted, model-based and hybrid estimation. Forest Ecosystems 3, 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, 2917–2930. https://doi.org/10.1016/j.rse.2010.08.027

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. https://doi.org/10.1139/X07-224

Williamson, G.B., Wiemann, M.C., 2010. Measuring wood specific gravity...Correctly. Am. J. Bot. 97, 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

12. ACKNOWLEDGMENTS

Azofeifa, Nuria Sanchez-Lopez, Crystal B. Schaff, Marc Simard, Andrew Kerr Skidmore, Göran Ståhl, Krzysztof Stereńczak, Hao Tang, Chiara Torresan, Rubén Valbuena, Hans Verbeeck, Tomas Vrska, Konrad Wessels, Joanne C. White, and Carlo Zgraggen.