Aboveground biomass density models for NASA’s Global Ecosystem Dynamics Investigation (GEDI) lidar mission

Laura Duncanson a,1, James R. Kellner b, John Armstrong a, Ralph Dubayah a, David M. Minor a, Steven Hancock c, Sean P. Healey d, Paul L. Patterson d, Svetlana Saarela e,f, Suzanne Marselis a, Carlos E. Silva a, Jamis Bruening c, Scott J. Goetz g, Hao Tang h,i, Michelle Hofton a, Bryan Blair h, Scott Luthcke h, Lola Fatoyinbo a, Katharine Abernethy j, Alfonso Alonso k, Hans-Erik Andersen l, Paul Aplin m, Timothy R. Baker n, Nicolas Barbier o, Jean Francois Bastin c, Peter Biber q, Pascal Boeckx r, Jan Bogaert s, Luigi Boschetti t, Peter Brehm Boucher t, Doreen S. Boyd t, David F.R.P. Burslem v, Sofia Calvo-Rodriguez w, Jérôme Chave x,y, Robin L. Chazdon z,aa, David B. Clark ab, Deborah A. Clark ab, Warren B. Cohen ac, David A. Coomes ad, Piermaria Corona ae,af, K.C. Cushman b,ag, Mark E.J. Cutler ah, James W. Dalling ag,ni, Michele Dalponte aj, Jonathan Dash ak, Sergio de-Miguel al,am, Songqiu Deng an, Peter Woods Ellis ap, Barend Erasmus bq,ar, Patrick A. Fekety as, Alfredo Fernandez-Landa as, Antonio Ferraz at,au, Rico Fischer av, Adrian G. Fisher aw,ax, Antonio García-Abri b y, Terje Gobakken b, Jorg M. Hacker az,ba, Marco Heurich bb,bc,bd, Ross A. Hill be, Chris Hopkinson bf, Huaping Huang bg, Stephen P. Hubbell ag,au, Andrew T. Hudak cw, Andreas Huth av,bb,bj, Benedikt Imbach bj, Kathryn J. Jeffery i, Masato Kato ab, Elizabeth Kearsey d, David Kenfack ag, Natascha Kljun bk, Nikolai Knapp av,cu, Kamil Král bl, Martin Krček bm, Nicolas Labrière c, Simon L. Lewis h,bn, Marcos Longo ai,bi, Richard M. Lucas bo, Russell Main ap,bp, Jose A. Manzanera bq, Rodolfo Vásquez Martínez br, Renaud Mathieu as,bs, Herve Memigaye bt,x, Victoria Meyer at,bu, Abel Monteagudo Mendoza br,bv, Alessandra Monerris bw, Paul Montesano h,bx, Felix Morsdorf by, Erik Näsett c, Laven Naïdoo bn, Reuben Nilus bs, Michael O’Brien ca, David A. Orwig cb, Konstantinos Papanastassiou cc, Geofffrey Parker cd, Christopher Philipson ce,cf, Oliver L. Phillips a, Jan Pisek sg, John R. Poulsen ch, Hans Pretzsch h, Christoph Rüdiger cl,cj, Sassan Saatchi at, Arturo Sanchez-Azofeifa w, Nuria Sanchez-Lopez z, Robert Scholes sp, Carlos A. Silva ck, Marc Simard at, Andrew Skidmore cl, Krzysztof Stereńczak cm, Mihai Tanase bw, Chiara Torresan ae,ca, Ruben Valbuena S,co, Hans Verbeeck t, Tomas Vrska t, Konrad Wessels sp, Joanne C. White eq, Lee J.T. White l,cr, Eliakimu Zahabu cs, Carlo Zgraggen bj

a University of Maryland, College Park, 2181 LeFrak Hall, College Park, Maryland 20742, USA
b Brown University, 75 Waterman St, Providence, RI 02912, USA
c University of Edinburgh, Drummond Street, Edinburgh EH8 9XP, UK
d USDA Forest Service, Rocky Mountain Research Station, 507 25th St., Ogden, UT 84401, USA
e Swedish University of Agricultural Sciences, SLU Skogsmarksgård 17, SE-901 83 Umeå, Sweden
f Norwegian University of Life Sciences, P.O. Box 5003, NMBU, 1432 As, Norway
g Norwegian University of Life Sciences, P.O. Box 5003, NMBU, 1432 As, Norway
h NASA Goddard Space Flight Center, 8800 Greenbelt Rd, Greenbelt, MD 20771, USA
i University of Stirling, University of Stirling, Stirling FK9 4LA, UK
j Institut de Recherche en Ecologie Tropicale, CENAREST, Gros Bouquet, Libreville, Gabon
k University of Maryland, College Park, 2181 LeFrak Hall, College Park, Maryland 20742, USA
l USDA Forest Service, Rocky Mountain Research Station, 507 25th St., Ogden, UT 84401, USA
m Swedish University of Agricultural Sciences, SLU Skogsmarksgård 17, SE-901 83 Umeå, Sweden
n Norwegian University of Life Sciences, P.O. Box 5003, NMBU, 1432 As, Norway
o Norwegian University of Life Sciences, P.O. Box 5003, NMBU, 1432 As, Norway
p NASA Goddard Space Flight Center, 8800 Greenbelt Rd, Greenbelt, MD 20771, USA
q University of Stirling, University of Stirling, Stirling FK9 4LA, UK
r Institut de Recherche en Ecologie Tropicale, CENAREST, Gros Bouquet, Libreville, Gabon
s USDA Forest Service, Rocky Mountain Research Station, 507 25th St., Ogden, UT 84401, USA
t Swedish University of Agricultural Sciences, SLU Skogsmarksgård 17, SE-901 83 Umeå, Sweden
u Norwegian University of Life Sciences, P.O. Box 5003, NMBU, 1432 As, Norway
v NASA Goddard Space Flight Center, 8800 Greenbelt Rd, Greenbelt, MD 20771, USA
w University of Stirling, University of Stirling, Stirling FK9 4LA, UK
x Institut de Recherche en Ecologie Tropicale, CENAREST, Gros Bouquet, Libreville, Gabon
y Institute of Geography, University of Copenhagen, Copenhagen, Denmark
z University of Maryland, College Park, 2181 LeFrak Hall, College Park, Maryland 20742, USA
a Corresponding author.
E-mail address: lduncans@umd.edu (L. Duncanson).

https://doi.org/10.1016/j.rse.2021.112845
Received 20 May 2021; Received in revised form 25 October 2021; Accepted 4 December 2021
Available online 7 January 2022
0034-4257/© 2021 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
1. Introduction

NASA’s Global Ecosystem Dynamics Investigation (GEDI) (Dubayah et al., 2020) mission has a primary science goal of mapping aboveground forest biomass across Earth’s temperate, subtropical, and tropical forests. Forests are a critical component of the global carbon cycle, with 2016–2020 deforestation emissions estimated to be ~2.9 Gt CO₂ per year, while remaining forests sequester ~2.5 Gt CO₂ per year (Tubiello et al., 2020). However, current estimates of carbon emissions from land use conversion and forest loss are highly variable, largely because of uncertainties in aboveground biomass (AGB [Mg]) estimates (Friedlingstein et al., 2019; Le Quéré et al., 2016), and depend on a combination of forest inventory, remote sensing data, and modeling efforts (Baccini et al., 2012; Houghton et al., 2012). Many remote sensing technologies have been used to quantify forest aboveground biomass density (AGB [Mg/ha]) at various scales, including both passive optical sensors such as Landsat (Foody et al., 2003) and active sensors such as Synthetic Aperture Radar (SAR) (Mitchard et al., 2009), airborne lidar (Coops et al., 2007; Næsset et al., 2013) and spaceborne lidar (Baccini et al., 2012; Saatchi et al., 2011). Each of these technologies has associated strengths and weaknesses for mapping AGBD, and are naturally synergistic.

Optical data provide the longest time series (Powell et al., 2010), and SAR data are particularly useful in areas with frequent cloud cover. Both optical and SAR instruments collect wall-to-wall imagery but optical reflectance and SAR backscatter have been demonstrated to saturate at relatively low AGB densities, although SAR signal saturation is wavelength dependent (Huete et al., 1997; Le Toan et al., 1992; Luckman et al., 1998; Rodriguez- Veiga et al., 2019). Many data fusion approaches focus on training wall-to-wall data with lidar samples (Healey et al., 2020; Potapov et al., 2021; Silva et al., 2021), because lidar measurements of vertical forest structure are sensitive to higher AGBD and have little or no saturation (Wulder et al., 2012). However, before NASA’s GEDI mission, no satellite-based lidar system had been specifically designed to study vegetation structure. Remote sensing does not directly measure aboveground woody biomass, which is a function of cumulative tree volume and the wood density of those trees. Therefore, while lidar may be the best technology for measuring 3D structure, and optical and SAR data can extend these measurements through space and time, all AGBD maps are modeled estimates that depend critically on the quality of field measurements of tree structure and wood density. Field estimates, in turn, are almost never direct measurements of biomass (Clark and Kellner, 2012) but typically rely on allometry to estimate AGBD from measurable attributes (e.g. Chave et al., 2014). This increases the challenge for any EO mission to accurately map AGBD.

Lidar applications for forest AGBD mapping have been developed over a range of spatial resolutions, scales and ecosystems, and with differing lidar instruments, statistical approaches, and model accuracies. GEDI’s algorithm development builds on this wealth of research, including studies using NASA’s Land Vegetation and Ice Sensor (LVIS), a large footprint full waveform airborne instrument which is used as a simulator for GEDI (Blair and Hofton, 1999). LVIS data have proved effective for mapping forest structure and biomass in tropical (Drake et al., 2002b, 2003; Tang et al., 2012) and temperate systems (Andersen et al., 2006; Lefsky et al., 2002), confirming the assertions of previous modeling studies that have explored the theoretical utility of waveform lidar for forest structure mapping using radiative transfer model inversion (Hancock et al., 2011, 2015; Koetz et al., 2006; North et al., 2010;
Remote Sensing of Environment 270 (2022) 112845

4


Biomass mapping with waveform lidar has also been expanded to spaceborne sensors through studies exploring the utility of ICESat Geoscience Laser Altimeter System (GLAS) waveforms for biomass estimation (Boudreau et al., 2008; Duncanson et al., 2010; Lefsky et al., 2007; Sun et al., 2008) and global vegetation height (Los et al., 2012; Simard et al., 2011). Most currently existing continental or global-scale forest biomass products use data from GLAS, often in combination with radar and/or optical data (Avitabile et al., 2016; Baccini et al., 2012; Saatchi et al., 2011; Simard et al., 2018). Similar to ICESat’s GLAS, GEDI uses full waveform near infrared 1064 nm lasers, but with a smaller footprint size (~25 m vs 65 m diameter), and collects denser spatial coverage with 8 ground tracks and 60-m along-track spacing (compared to a single track with 172 m spacing for GLAS).

An exhaustive meta-analysis of previous biomass modeling efforts, such as in Zolkos et al. (2013) is beyond our scope here, but a sample of previous studies (Table 1) helps to illustrate the breadth of geographic domain, ecosystems, statistical algorithms, sensors, and associated accuracies. This diversity of models, data, and methods highlights the enormous task of creating a parsimonious set of calibration models appropriate for global biomass predictions using GEDI. As we examined these previous studies, a few key points emerged that served to inform our efforts. Model performance is generally shown to improve with larger training plots where edge effects and geolocation impacts are minimized (Labriere et al., 2018; Næsset et al., 2015; Réjou-Méchain et al., 2019; Zolkos et al., 2013), and over smaller geographic domains (e.g. sub-national compared to pantropical) (Healey et al., 2020). Most of these previous calibration studies have been conducted over local areas, generally the result of the limited spatial extent of associated airborne lidar surveys. Larger domain studies were almost exclusively based on GLAS data, which presented their own challenge of being spatially sparse and having a large footprint size that blended together reflectance from many trees of varying heights and the underlying terrain, leading to height estimation errors over slopes (Duncanson et al., 2010; Mahoney et al., 2014).

In general, local studies (e.g., project-level airborne lidar collection) tend to include more predictor variables, and have higher reported accuracies (e.g. see Zolkos et al., 2013 and Table 1). In contrast, regional or larger area studies (continental, pantropical, global) have lower reported accuracies due to broader domains that cover an enormous range of edaphic, topographic, floristic and climatic gradients that can create local variations in canopy structure, and hence aboveground biomass (Ferraz et al., 2018; Meyer et al., 2019; Xu et al., 2017). In addition, such studies often include multiple ancillary datasets (e.g., soils), some of which may only be weakly correlated with biomass (Ploton et al., 2020).

In terms of modeling methods, both parametric and non-parametric (e.g., machine learning) approaches have been widely used, but with similar accuracies, although local studies sometimes show higher accuracy using machine learning approaches (Corte et al., 2020; Hudak et al., 2016; Silva et al., 2021), while broader area studies show essentially comparable performance for both types of approaches (Tang et al., 2021). As we scale to broader spatial extents, more forest structural variability occurs and models converge to generalities that appear to be equally well captured by parametric and machine learning approaches.

1.1. GEDI mission overview

The previous efforts outlined in Table 1 from more than 20 years of research underscore the complexity of the task any space mission must face towards developing calibration models for biomass. The GEDI mission has a primary science goal of mapping aboveground woody biomass across Earth’s temperate, subtropical, and tropical forests. A complete description of the mission, its goals, objectives and data products can be found in Dubayah et al. (2020). GEDI uses a full waveform lidar system that operates from the International Space Station (ISS), and produces lidar measurements of forest structure within approximately ± 51.6 degrees latitude. GEDI was launched to the ISS on December 5, 2018, and started collecting science data in April 2019. It was projected to collect ~10 billion cloud-free land surface observations over a nominal two-year mission length, but at the time of writing has already surpassed this data collection goal, and has been extended until at least January, 2023. Data from GEDI were designed to produce gridded AGBD maps with higher accuracies than previously possible. These maps will aid climate change mitigation programs such as the United Nation’s Reduced Emissions from Deforestation and forest Degradation (REDD+) (Corbera and Schroeder, 2011; Gibbs et al., 2007; Goetz et al., 2015), and help constrain global climate and carbon cycle models (e.g. Hurr et al., 2002).

1.2. GEDI biomass modeling challenges

An exhaustive global scale field campaign to collect reference samples co-located with GEDI footprints was logistically infeasible and would be hindered by GEDI’s ~10 m geolocation uncertainty for footprint locations. For relatively small (~25 m) plots that already have ~5 m nominal geolocation uncertainty under dense canopies (Réjou-Méchain et al., 2019), matching plots to footprints is problematic.

To overcome these limitations, we adopted a ‘crowd-sourced calibration’ approach to develop biomass algorithms. As described in Methods, this involved associating existing field plots coincident with airborne lidar datasets, then simulating GEDI waveforms from the airborne lidar data to produce a global training dataset for AGBD models.

In addition to constraints related to training data availability, GEDI’s biomass algorithm development also required adoption of a statistical framework compatible with GEDI’s approach to gridded biomass estimation (the L4B product). A major limitation of most existing biomass maps generated from remotely sensed data is their ad hoc or poorly defined uncertainty estimation framework. GEDI has a mission requirement to provide a 1-km gridded product (L4B) with a standard error that is 20% of the mean AGBD in at least 80% of 1 km cells. This in turn places specific accuracy requirements on GEDI4A footprint-level waveform-to-biomass models, which should be parametric models (Patterson et al., 2019).

In this paper, we used a training dataset to produce predictive footprint-level models that addressed three model development questions. First, what predictors should these models use to provide sufficient explanatory power while preventing inclusion of too many independent variables that may lead to overfitting (Valbuena et al., 2017)? Further (and relatedly), what data transformations linearizes the relationship between predictors and AGBD without compromising model performance? Finally, what level of geographic stratification optimizes model performance while maintaining sufficient training data? This paper presents GEDI’s conceptual approach to footprint-level AGBD modeling, and the version 1 GEDI footprint biomass (GEDI04_A) models. These models were fit per geographic stratum, and have differing input predictor variables and data transformations. The models developed here are applied to the entire archive of observations in the GEDI04A product (2019 - 2021). These models may change for future versions of this product, as more training data become available and we learn about the relative strengths and weaknesses of the first products in different geographic regions.

2. Methods

The GEDI AGBD model development relied on a compilation of existing field and airborne lidar datasets gathered through international partnerships within the GEDI domain (Fig. 1a, Fig. 2). Airborne lidar point clouds were processed through a GEDI waveform simulator (Hancock et al., 2019, Fig 1b) to produce GEDI-like waveforms and derived metrics commensurate with field plot data. An exhaustive set of models was fit to predict field AGBD as a function of simulated RH metrics, with permutations in candidate predictor metrics (all possible
Examples of previous lidar biomass studies in forested ecosystems provide context for the unique geographic extent and spatial resolution of GEDI footprint-level biomass (GEDI04_A) models. Models are listed from local to pantropical studies, using a range of input data (discrete return lidar, airborne full waveform (LVIS and SLICER)), and spaceborne waveform lidar (GLAS). Modeling types include Ordinary Least Squares (OLS), Support Vector Regression (SVR), Random Forest (RF), Partial Least Square regression (PLS), k-Nearest Neighbours (kNN), among others listed. The accuracies reported here are from the respective papers.

<table>
<thead>
<tr>
<th>Previous study</th>
<th>LiDAR type/ Additional datasets</th>
<th>Data acquisition date</th>
<th>Geographic extent</th>
<th>Modeling type</th>
<th>Predictor(s) for best AGBD model</th>
<th>Accuracy of best AGBD model</th>
<th>Plot size</th>
<th>Number of plots</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boreal</td>
<td>Discrete Return LiDAR</td>
<td>1998 - 2006</td>
<td>Norway (regional)</td>
<td>nonlinear regression model</td>
<td>Height Metrics, Density Metrics</td>
<td>R² = 0.82, RMSE(%) = 25</td>
<td>0.02-0.04 ha</td>
<td>1395</td>
</tr>
<tr>
<td>Næsset and Gobølken (2008)</td>
<td>Discrete Return LiDAR</td>
<td>June 2009</td>
<td>Tanana Valley, Alaska (Local)</td>
<td>OLS</td>
<td>Height Metrics, Density Metrics</td>
<td>R² = 0.74</td>
<td>0.033 ha</td>
<td>79</td>
</tr>
<tr>
<td>Margolis et al. (2015)</td>
<td>GLAS</td>
<td>2005-2006</td>
<td>Canadian boreal (regional)</td>
<td>kNN</td>
<td>Waveform metrics</td>
<td>R² = 0.66, RMSE = 27.2 Mg/ha</td>
<td>40 x 60 m lidar plots</td>
<td>565</td>
</tr>
<tr>
<td>Temperate</td>
<td>Discrete Return LiDAR</td>
<td>March 2004</td>
<td>Eastern Texas, USA</td>
<td>functional regression models</td>
<td>Height Metrics</td>
<td>R² = 0.938%, RMSE = 14.6</td>
<td>1 ha</td>
<td>2000 (synthesized)</td>
</tr>
<tr>
<td>Zhao et al. (2009)</td>
<td>Discrete Return LiDAR</td>
<td>2008, 2011, 2012</td>
<td>Teakettle, California; Parker Tract, North Carolina; SERC, Maryland (Local)</td>
<td>PLS</td>
<td>Height Metrics, Density Decile Metrics</td>
<td>R² = 0.84%, RMSE = 6</td>
<td>0.81 ha</td>
<td>16</td>
</tr>
<tr>
<td>LeSkiy et al. (2002)</td>
<td>Slicer</td>
<td>September 1995, July 1996</td>
<td>Temperate Coniferous Forest, Temperate Deciduous Forest, Boreal Coniferous Forest (Local)</td>
<td>Stepwise multiple regression</td>
<td>Canopy Height Metrics, Canopy Cover</td>
<td>R² = 0.91</td>
<td>0.25 ha</td>
<td>22</td>
</tr>
<tr>
<td>Gleason and Im (2012)</td>
<td>Discrete Return LiDAR</td>
<td>August 2010</td>
<td>Heilberg Memorial Forest, NY (Local)</td>
<td>SVR</td>
<td>Height Metrics, LAI, Canopy Volume, Crown Area</td>
<td>R² = 0.93, RMSE(%) = 13.6</td>
<td>0.038 ha</td>
<td>18</td>
</tr>
<tr>
<td>Hernando et al. (2019)</td>
<td>Discrete Return LiDAR/ Multispectral data</td>
<td>July 2006</td>
<td>Spain (Local)</td>
<td>k-MSN</td>
<td>Height Metrics, NDVI Metrics</td>
<td>R² = 0.64%, RMSE = 16.7</td>
<td>0.126 ha</td>
<td>37</td>
</tr>
<tr>
<td>Ferraz et al. (2016)</td>
<td>Discrete Return Lidar</td>
<td>2008</td>
<td>Agueda, Portugal (local)</td>
<td>OLS</td>
<td>Height Metrics</td>
<td>R² = 0.72%, RMSE=23.3</td>
<td>0.04 ha</td>
<td>39</td>
</tr>
<tr>
<td>Tropical</td>
<td>LVIS</td>
<td>1998, 2005</td>
<td>Costa Rica (Local)</td>
<td>OLS</td>
<td>Height Metrics</td>
<td>R² = 0.65, RMSE = 10.5 Mg/ha</td>
<td>0.5 ha, 1 ha</td>
<td>20</td>
</tr>
<tr>
<td>Drake et al. (2003)</td>
<td>LVIS</td>
<td>March 1998</td>
<td>Panama, Costa Rica (Local)</td>
<td>OLS</td>
<td>Height Metrics, HOME</td>
<td>R² = 0.89%, RMSE = 14.06 %RMSE = 47.4</td>
<td>0.5 ha, 1 ha</td>
<td>71</td>
</tr>
<tr>
<td>Ene et al. (2016)</td>
<td>Discrete Return LiDAR</td>
<td>February - June 2012</td>
<td>Liwale district, Tanzania (Local)</td>
<td>OLS</td>
<td>Height Metrics, Density Metrics</td>
<td>R² = 0.70%, RMSE = 32.3</td>
<td>0.07 ha</td>
<td>513</td>
</tr>
<tr>
<td>Hansen et al. (2015)</td>
<td>Discrete Return LiDAR</td>
<td>January - February 2012</td>
<td>East Usambara Mountains, Tanzania (Local)</td>
<td>OLS</td>
<td>Height Metrics, Density Metrics</td>
<td>R² = 0.79%, RMSE = 14.3</td>
<td>0.125 ha</td>
<td>600</td>
</tr>
<tr>
<td>Labriere et al. (2018)</td>
<td>Discrete Return LiDAR</td>
<td>2009 - 2016</td>
<td>French Guiana, Gabon (Local)</td>
<td>OLS</td>
<td>Median Height of CHM, Height Metrics, Hyperspectral Bands</td>
<td>R² = 0.7, RMSE = 61.7 Mg/ha</td>
<td>1 ha</td>
<td>183</td>
</tr>
<tr>
<td>Laurin et al. (2014)</td>
<td>Discrete Return LiDAR/ Hyperspectral data</td>
<td>March 2012</td>
<td>Sierra Leone (Local)</td>
<td>PLS</td>
<td>Canopy height metrics, Hyperspectral Bands</td>
<td>R² = 0.63, RMSE = 19.2 %RMSE = 19.2</td>
<td>0.0625 ha</td>
<td>152</td>
</tr>
<tr>
<td>Naidoo et al. (2015)</td>
<td>Discrete Return Lidar</td>
<td>April-May 2012</td>
<td>Savannah, Kruger National Park, SA (local)</td>
<td>OLS</td>
<td>Canopy height metrics, Canopy Cover</td>
<td>R² = 0.75, RMSE = 59.9 Mg/ha</td>
<td>1 ha</td>
<td>92</td>
</tr>
<tr>
<td>Xu et al. (2017)</td>
<td>Discrete Return Lidar</td>
<td>2012</td>
<td>DRC (Regional)</td>
<td>Power-law</td>
<td>Canopy height metrics, Gap Fraction</td>
<td>R² = 0.81%, RMSE = 20 Wetlands, R² = 0.79%, RMSE = 9</td>
<td>0.1 - 0.25 ha</td>
<td>82 (drylands) + 22 (wetlands)</td>
</tr>
<tr>
<td>Coomes et al. (2017)</td>
<td>Discrete Return LiDAR</td>
<td>November 2014</td>
<td>Sabah, Malaysia (Local)</td>
<td>Power-law model</td>
<td>Height Metrics, Gap Fraction</td>
<td>0.25 ha</td>
<td>291</td>
<td></td>
</tr>
<tr>
<td>Ferraz et al. (2018)</td>
<td>Discrete Return Lidar</td>
<td>2014</td>
<td>Kalimantan, Indonesia (Regional)</td>
<td>Power-law</td>
<td>Canopy height metrics, Gap Fraction</td>
<td>0.25 ha</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>Meyer et al. (2018)</td>
<td>Discrete Return LiDAR</td>
<td>2009 - 2012</td>
<td>Neotropics</td>
<td>Jackknife regression</td>
<td>Large Canopy Area, Wood density</td>
<td>R² = 0.78, RMSE = 46.0 Mg/ha</td>
<td>0.25 ha</td>
<td>1 ha</td>
</tr>
<tr>
<td>Lucas et al. (2008)</td>
<td>Discrete Return LiDAR/ Hyperspectral data</td>
<td>August 2000/ September 2000</td>
<td>Queensland, Australia (Local)</td>
<td>Jackknife regression</td>
<td>Height Metrics, Canopy Cover</td>
<td>R² = 0.90, RMSE = 11.8 Mg/ha</td>
<td>0.25 ha</td>
<td>1395</td>
</tr>
</tbody>
</table>
combinations of 1, 2, 3, and 4 predictors from suites of RH metrics and their two-way interaction terms), transformations, and geographic stratifications (Fig. 1b, Fig. 2). Each of these models was evaluated based on its model fit and geographic cross-validation performance (that is, how well a model developed in one location worked in a different location within a strata), and ranked accordingly (Fig. 1).

2.1. Data

2.1.1. Field data

The field data used in the development of the GEDI04_A data product were from 74 sites (Fig. 2), taken from a total of 142 sites or projects that contributed data to this research (Table S1). These datasets were assembled by an international consortium of researchers and represent both publicly available and privately managed datasets (Fig. S1). A wide range of AGB densities was covered on every continent and PFT, although some geographic regions, such as continental Asia, were relatively data-sparse (Fig. 2). All datasets were collected with different protocols. The GEDI Forest Structure and Biomass Database (FSBD) follows the framework developed for the Australian Terrestrial Ecosystem Research Network (TERN) Biomass Plot Library (Auscover, 2016) and is a harmonization of these projects to a common set of georeferenced plots and, where available, tree-level observations. Individual tree measurements including diameter at breast height (DBH) or above basal deformities, tree height and species (as available) and plot geometries were used to match simulated GEDI footprints to field plots. We predicted individual tree AGB from DBH, and wood density based on taxonomic information, as well as height for tropical datasets, where available, using available broadly applicable allometric models (Table S1). While these allometric models are known to have high uncertainties (Vorster et al., 2020), e.g. for estimation of large tree biomass (Disney et al., 2020), the set of allometric models adopted for GEDI04_A was the most generalized available for the geographic scale of the GEDI04_A models. The degree to which species information and tree height were required depended on which allometric model was applied. For the tropics, if sufficient taxonomic information was not available to estimate wood density, the plot data were not included.

Each of the projects/sites included in the training database (Table S1) had its own unique characteristics. For example, in the United States six datasets were collected under a NASA Carbon Monitoring
System (CMS) project, each including 50 ~GEDI footprint-sized (25 m diameter) plots providing independent sampling units across a range of geographic conditions. Conversely, large stem-mapped plots (>1 ha, e.g. Robson Creek site in Australia or Barro Colorado Island in Panama) were divided into several GEDI footprint-sized plots placed side-by-side in a tighter range of biogeographic conditions. Plot-level AGBD was calculated by summing all individual tree AGB (above a minimum DBH threshold, see Table S1) in a GEDI footprint-sized plot, but this summation took two general forms: 1) summing all trees when the plot was approximately GEDI sized (~25 m diameter), or 2) dividing stem-mapped plots into GEDI footprint-sized plots prior to summing all trees. Total plot-level AGB was then divided by plot area to produce estimates of AGBD.

2.1.2. Airborne Laser Scanning (ALS) data

Only field data with spatially coincident airborne lidar data collected during the leaf-on season within two years of field data acquisition were used for GEDI’s footprint-level AGBD models. The lidar data were from a range of instruments (Table S1). Lidar datasets were filtered to ensure sufficient sampling density of returns were available (>4 pulses m⁻²) to simulate GEDI waveforms (Hancock et al., 2019).

2.1.3. GEDI waveform simulator

A complete description of the waveform simulator is found in Hancock et al. (2019) and is briefly summarized here. To simulate GEDI waveforms, airborne lidar returns were spatially extracted over GEDI footprint-sized plots, binned vertically and weighted by a Gaussian distribution based on their distance from the footprint centroid, with a Gaussian width set to a sigma of 5.5 m to match GEDI’s footprint width. They were then convolved with the GEDI pulse shape (full width at half maximum, FWHM = 15.6 ns) to produce realistic simulations. Relative Height (RH) metrics (Blair and Hofton, 1999) were calculated with respect to ground elevation, and these metrics become the candidate predictor variables for the AGBD models. There were several possible algorithms for estimating ground elevation from GEDI waveforms, which are available in GEDI’s footprint level height and elevation (L2A) product with the associated sets of RH metrics. For GEDI04_A model development, we used the center of gravity from the ALS points classified as ground returns, to preclude uncertainties related to waveform ground finding algorithms in AGBD model fitting. While the GEDI simulator was capable of simulating metrics from both GEDI’s coverage and power beams acquired with different noise conditions and ground finding algorithms, we used only one set of simulated waveform metrics for model fitting (noise-free, with ground elevation as detected with ALS). We assumed that on-orbit waveforms with erroneous RH metrics due to low signal to noise ratio, e.g. coverage beams acquired during the day, would be flagged in L2A. Thus, our objective was to select a single set of predictor metrics and model parameters applicable to all quality on-orbit GEDI data. RH metrics from the waveform simulator have previously been validated against LVIS data and shown to be unbiased (<25 cm), with an RMSE of 5.7 m (Hancock et al., 2019).

2.1.4. Data filtering

Prior to model fitting, several filters were applied to identify erroneous outliers and ensure the training dataset (Fig. 2, Table S1) was not biased toward large plots. To prevent model training being weighted too heavily to plots with a larger number of footprints, we fit OLS models with a weighting factor based on the number of simulated footprints per plot (i.e. every plot will have the same influence on model fits, regardless of the plot size). We also subsampled large stem-mapped plots that have >200 simulated footprints in a single plot (i.e. in stem-mapped plots). To create this subsample, we divided the footprints into 20 equally-spaced AGBD bins, and randomly sampled 200 footprints, where the per-footprint probability of inclusion was inversely proportional to the number of footprints per bin. This resulted in a sample that was random relative to the collection of footprints in each large plot, while ensuring that the observed range of AGBD was covered. In addition to sampling, we applied several filters to remove erroneous training points. Most notably, we filtered data where field estimates of maximum tree height differed by more than 10 m from airborne lidar estimated maximum canopy height. The filters applied are detailed in the Supplementary Information section on data filtering, and generally addressed a) poor geolocation of the field data, b) temporal changes (e.g. growth, disturbance) between the field and lidar collection, c) measurement or transcription error in the field data, d) in the case of modeled heights, an inappropriate diameter-to-height allometric model and e) tree sizes that were outside the range of allometric model training.

2.2. Model formulation and data transformation

For GEDI’s footprint-level AGBD algorithms, it was necessary to adopt parametric model forms to satisfy assumptions of hybrid as well as generalized hierarchical model-based (GHMB) estimators used in
GEDI’s L4B gridded biomass product (Patterson et al., 2019; Saarela et al., 2018). A widely used parametric model for AGBD modeling is an OLS regression model with possible transforms of AGBD and predictors, which can be written as

\[ h(\text{AGBD}) = \sum_{j=1}^{p} B_j f(x_j) + \epsilon \]  

(1)

\[ \text{AGBD} = h^{-1} \left( \sum_{j=1}^{p} \hat{B}_j f(x_j) + \epsilon \right) \]  

(2)

\( B_j \) are the regression coefficients and \( p \) predictors \((x_j)\), \( f() \) is a transformation function (identity, log or square root) and \( h() \) is a back-transformation function (identity, exponential function or second power), and \( \epsilon \) is a normally distributed error term with a mean of zero. GEDI’s hybrid and GHMB estimators assume the model expressed by Eqn. 1 is a biased predictor for the true AGBD, and Eqn. 2 is an unbiased predictor of transformed AGBD, thus we applied a bias correction factor after back transformation (Snowdon, 1991). We also assume that the model has been properly specified (Patterson et al., 2019). Therefore, errors due to model misspecification must be minimized. The allometric models used to generate the field estimates of AGB, usually as a function of stem diameter and/or height, and species/wood density, are typically nonlinear. Transformation of predictor variables is often used to linearize relationships between height predictors and AGBD. Prior to applying OLS models, we therefore explored square root (sqrt) and log transformations of predictor variables, since these have proved useful in previous AGBD modelling studies (e.g. Hansen et al., 2017). We also transformed the response variable, AGBD, both for linearizing relationships between AGBD and the predictors, and to satisfy the assumption that errors were random observations from a normal distribution with zero mean and constant variance (homoscedasticity).

While the GEDI04_A algorithm focused on parametric modeling with OLS, we also conducted a comparison between OLS and three other popular modeling approaches, Partial Least Squares (PLS) regression, Random Forests, and Support Vector Regression to test whether OLS models performed comparably to these other approaches. Our analysis demonstrated that these alternative approaches did not increase the performance of candidate GEDI04_A models (detailed in Supplementary Information).

### 2.3. Candidate predictor variables

A suite of GEDI waveform metrics was derived from each simulated waveform including RH metrics (Dubayah et al., 2020, Fig. 3) representing the height above ground elevation below which a given percentage of waveform energy has been returned. RH50 is equivalent to the height of median energy (HOME, Drake et al., 2002a), and has been used in other AGBD modeling studies (Baccini et al., 2008). We explored the use of the 10\textsuperscript{th} percentile divisions of RH metrics between RH10 and RH90, as well as RH98 which we used as our maximum height metric as this is a more stable height metric than RH100 (see Blair and Hofton, 1999). We also considered interaction terms between these RH metrics (e.g. RH50 x RH98) as potential predictors of AGBD. While the suite of candidate predictors were often correlated (e.g. RH90 and RH98), our model fitting procedure (described below) removed candidate models with highly correlated predictors.

We considered four levels of predictor variable constraints in our model development: 1) No constraints (hereafter referred to as unconstrained models), 2) Forced inclusion of maximum height (RH98), 3) No RH metrics below RH50, and 4) Both forced inclusion of RH98 and no metrics < RH50. This fourth category is hereafter referred to as our constrained model set. The scenarios with forced inclusion of RH98 were justified in terms of the theoretical importance of maximum height for biomass modeling. The scenarios omitting the lower (<50) RH metrics were introduced to account for the sensitivity of low RH metrics to potential differences between simulated and observed GEDI waveforms that were not included in the waveform simulations here. On-orbit measurements of the GEDI pulse shape are not perfectly Gaussian.
Remote Sensing of Environment 270 (2022) 112845

Dubayah et al., 2020) and early received waveforms sometimes exhibited a long trailing edge, similar to that documented by Hancock et al. (2019) in LVIS data. A constant canopy/ground reflectance ratio was also assumed for simulations, whereas some degree of variation is expected for measured waveforms (Tang et al., 2019). The implication of these differences for prediction of footprint level biomass is the subject of current research by the GEDI Science Team. Constrained (no low RH metrics, forced RH98) and unconstrained model performance for each stratum was considered for final model selection.

In OLS modeling, excessive multicollinearity can cause model parameters to be sensitive to changes in fit data and/or potentially misleading evaluation metrics (e.g. %RMSE) (Wood, 2017). Through simulating the effects of collinearity amongst predictor variables, we determined that a Pearson’s correlation coefficient $r < 0.9$ between any two predictors was a sufficient threshold for minimizing the effects of multicollinearity on model fitting (Saarela et al., 2021). We filtered our results to only consider models that use combinations of predictors deemed sufficiently independent ($r < 0.9$), with a Variable Inflation Factor (VIF) $\leq 10$. Note that because RH metrics can be negative (cf. Section 3.2.2), we added an offset of 100.0 m to all RH metrics prior to model fitting to ensure that all RH predictors will be positive. This offset was applied prior to computing interaction terms.

We fit OLS models for every combination of 1-4 predictor variables from RH metrics (both constrained and unconstrained predictor sets) and associated two-way interaction terms for original-sqrt, original-log, sqrt-sqrt, and log-log transformations.

2.4. Candidate stratifications

We explored three levels of geographic stratification to produce globally representative models. For vegetation type classification we used PFT, a broadly adopted classification of ecosystem structure and function (Diaz and Cabido, 1997) commonly used for Earth System modeling (Poulter et al., 2011). The relationships between height metrics and AGBD might be expected to vary by PFT considering the breadth of relationships between lidar and AGBD (i.e., as expressed by the lidar-to-AGBD model) in different ecosystem types (e.g., Table 1). We also stratified our database by geographic region as it has been well documented that forest composition and structure both vary within and between continents (Carlucci et al., 2017; Corlett and Primack, 2006; Feldpausch et al., 2012; Frits and Balslev, 2005).

We considered model stratification by a) geographic region, b) PFT, and c) PFT within a given geographic region. The four considered PFTs are Evergreen Broadleaf Trees (EBT), Evergreen Needleleaf Trees (ENT), Deciduous Broadleaf Trees (DBT), and Grasslands/Shrublands/Woodlands (GSW). The most geographically specific model was therefore for a single PFT in a single geographic region, e.g. Evergreen Broadleaf Trees in South America. However, we were limited by the availability of field and airborne lidar datasets within each of these strata – in a stratum where no calibration data were available, we were unable to develop a

Fig. 4. Frequency distributions of AGBD for each PFT (Evergreen Broadleaf Trees (EBT), Evergreen Needleleaf Trees (ENT), Deciduous Broadleaf Trees (DBT), and Grassland, Shrub and Woodland (GSW)), and geographic region class show that different strata have different amounts of training data, with data-rich areas in North America, South America and Europe, and relative paucities in Africa and Asia. AGBD in training plots ranges from 0 to over 1000 Mg/ha at the footprint level.
model and thus applied a more coarsely stratified model. For example, if no training data were available for Deciduous Broadleaf Trees PFT in Asia, we had the choice of applying either a Deciduous Broadleaf Trees model calibrated from other continents, or an Asia-wide model calibrated for other ecosystems.

As the gridded AGBD algorithm is at a 1-km resolution, we adopted a 1-km stratification for assigning training plots to a PFT, as available from the MODIS product MCD12Q1 V006 (Friedl et al., 2010). For model fitting, we extracted the MCD12Q1 PFT classification (Type 5) for the corresponding pixel to each field plot. Our database was primarily representative of three PFT classes: EBT, ENT, and DBT. We also aggregated plots classified as Shrubs, Grasslands or Barren (<10% vegetation) into a fourth PFT-class that we called Grassland, Shrub and Woodland (GSW, Figs. 2, 4). The MODIS PFT classes were checked and corrected with field data. Where field data did not report PFT, MCD12Q1 was used, which potentially introduced spatial and temporal matching issues as well as the possibility of MODIS classification errors. For example, a few field plots in a coastal Gabonese tropical forest (Akanda National Park) were classified as water due to their proximity to the ocean, and were manually corrected. A wide range of AGBD values existed within each stratum (Fig. 4). For predictions in the GEDI04_A product, every 1-km cell on the GEDI grid will be assigned to one of the PFT and geographic region IDs represented in the model fit strata, as described in Kellner et al. (2021).

2.5. Model assessment and ranking

A large number of models were fit with permutations of stratification, transformation, and predictor variable selection. Models were fit to each stratification level (global, per-PFT, geographic region and PFT within geographic region), each transformation, and every combination of predictors that passed our multicollinearity filters. Model performance was evaluated both on the model fit metrics and via geographic transferability cross-validation, where models were applied to geographic areas that were outside the training dataset but inside the fit stratum. Thus, models were evaluated based on their ability to predict AGBD in a different geographic region within the same stratum. The latter cross-validation provided an assessment of geographic transferability, a reasonable estimate of expected performance for a given model, insofar as the training dataset could inform. This approach also minimized the influence of spatial autocorrelation, which can lead to overestimation of model predictive capability (Ploton et al., 2020).

Predictions from each model were back-transformed prior to model assessment. We applied a ratio method for back-transformation bias correction to predictions following (Snowdon, 1991). Models were sorted by mean residual error (MRE) and relative root mean square error (%RMSE) under geographic transferability cross-validation, wherein RMSE was calculated relative to the observed (calibration data), and %RMSE was calculated as RMSE / mean(observed). MRE was calculated by taking the average of the absolute mean residual error (predicted minus observed) in each of five quantile bins of AGBD. This MRE metric ideally should be zero for every AGBD bin, and was used as a systematic prediction error measure, and was selected to account for mean residual error for different biomass magnitudes (i.e., where overestimation at the low end and underestimation at the high end balance out). Models were ranked by cross-region MRE in addition to the %RMSE rounded down to the nearest 5% bin, the largest RH metric in the model (i.e., preference to models that include RH98, RH90 etc. over low RH metrics), the number of fitted coefficients, and the number of RH metrics in that order. The final predictive models selected for application to on-orbit GEDI RH metrics were the top performing models in each geographic stratum.

3. Results

We present summaries of model fits from an exhaustive suite of candidate OLS models, focusing results on presenting the accuracies of the highest ranked models (Section 3.1), selection of predictor variables (Section 3.2), selection of data transformation (Section 3.3) and the implications of geographic stratification (Section 3.4).

3.1. Summary of top candidate model performance

The total number of models fit per stratum depended on the number that passed the various filters described in the Methods section (typically 2,000-10,000 models per stratum). Summing across all strata and fit scenarios, we fit >100,000 models. The statistics from the top model for each PFT by geographic region stratum and for each broad stratum are described in Table 2 and 3 respectively. All reported accuracy statistics were calculated by geographic cross validation.

The model geographic cross validation %RMSE ranged from ~28-66% for the top 20 models per stratum, with the majority of models having an %RMSE ~30-50%. MRE was one of the metrics used for model ranking, and the MRE for all strata was <25 Mg/ha, with the notable exception of the EBT x Asia stratum (Table 2).

3.2. Variable selection

When allowing full flexibility of variable selection, with only multicollinearity constraints (unconstrained models, see Methods), both high and low RH metrics were selected (Fig. 5). RH10 was the most common predictor in the top 20 models per strata, either independently or within an interaction term. Relative Height metrics RH90 and RH98 were also frequently selected. When constraining variable selection by forcing inclusion of RH98, a similar frequency of selected predictors was apparent, but RH98 was the most selected (by force), which reduced the frequency of the selection of RH90, likely due to high correlation between the two. In the model set that allowed low RH metrics, but forced RH98, RH10 was the second most frequently selected metric. Frequencies of variable selection did not differ substantially by PFT (Figs 5, 6). Note that forcing the inclusion of RH98 and removing low RH metrics (constrained models) typically simplified the models in that fewer predictors were included (Table 2, Table S2). Interaction terms were more commonly used as predictors than single RH metrics (Table S2), when all predictors were allowed, while fewer interaction terms were included in the constrained models.

3.2.1. Relationships between RH metrics and AGBD

All RH metrics were correlated with AGBD, but there were markedly different relationships among them (Fig. 7a-e). While RH98 had a nonlinear relationship with AGBD, RH50 was more linearly related. As expected, RH10 was sensitive to canopy cover. In areas of low cover, RH10 was typically negative (Fig. 7g), becoming positive at approximately 80% canopy cover in the simulated waveform data. When RH10 is negative it indicates that > 20% of waveform energy is within the ground return. The two relationships seen between RH10 and AGBD are distinguished by canopy cover, representing roughly a low cover (<80%) and high cover (>80%) relationship.

Maximum height was directly related to the AGB of trees and therefore RH98 usually had a relationship with AGBD at the plot level, where the biomass may be a product of many smaller trees or a few large trees at the size of GEDI footprints (Fig. 7e). However, the 90th percentile lidar height is often used in the literature instead of maximum height for AGBD prediction (Table 1). We compared RH98 and RH90 across geographic regions, and found that while they were highly correlated, there were often large differences between the two metrics (Fig. 7i, j). These occurred across the full range of heights, and across all PFTs. While these differences were relatively rare, where they occurred they led to underestimations of AGBD when predicting with RH90 instead of RH98, most notably in the low biomass range.

3.2.2. Implications of candidate variable selection on model performance

The four candidate predictor sets (unconstrained, forced RH98, no
The model for each broad stratum (PFT or regionally aggregated) applied in the GEDI04_A product. R², %RMSE, and Mean Residual Error (MRE) were all calculated from geographic cross validation. MRE was the absolute mean binned residual error, expressed in Mg/ha. The minimum, mean, and maximum aboveground biomass density of training samples for each stratum are reported in Mg/ha, along with the total number of training samples used from each stratum.

<table>
<thead>
<tr>
<th>Stratum</th>
<th>R²</th>
<th>%RMSE</th>
<th>MRE (Mg/ha)</th>
<th>Transform</th>
<th>Predictors</th>
<th>AGBD (Mg/ha)</th>
<th># Training samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBT Africa</td>
<td>0.63</td>
<td>55.24</td>
<td>8.41</td>
<td>sqrt-sqrt</td>
<td>RH50, RH98</td>
<td>0.00</td>
<td>63.10 386.76 490</td>
</tr>
<tr>
<td>DBT Europe</td>
<td>0.66</td>
<td>47.27</td>
<td>21.52</td>
<td>sqrt-sqrt</td>
<td>RH70, RH98</td>
<td>0.52</td>
<td>206.13 724.06 333</td>
</tr>
<tr>
<td>DBT North America</td>
<td>0.66</td>
<td>38.08</td>
<td>22.81</td>
<td>sqrt-sqrt</td>
<td>RH50, RH98</td>
<td>0.00</td>
<td>205.22 520.78 873</td>
</tr>
<tr>
<td>EBT Africa</td>
<td>0.64</td>
<td>66.89</td>
<td>15.32</td>
<td>sqrt-sqrt</td>
<td>RH50, RH98</td>
<td>0.00</td>
<td>216.59 1489.70 834</td>
</tr>
<tr>
<td>EBT Asia</td>
<td>0.36</td>
<td>78.94</td>
<td>121.15</td>
<td>original-log</td>
<td>RH50, RH98</td>
<td>0.00</td>
<td>245.64 961.80 326</td>
</tr>
<tr>
<td>EBT Oceania</td>
<td>0.61</td>
<td>28.66</td>
<td>8.17</td>
<td>sqrt-sqrt</td>
<td>RH70, RH98</td>
<td>12.24</td>
<td>416.71 830.58 213</td>
</tr>
<tr>
<td>EBT South America</td>
<td>0.66</td>
<td>42.2</td>
<td>10.4</td>
<td>sqrt-sqrt</td>
<td>RH50, RH98</td>
<td>0.00</td>
<td>299.96 1578.00 3441</td>
</tr>
<tr>
<td>ENT Europe</td>
<td>0.54</td>
<td>63.43</td>
<td>14.33</td>
<td>sqrt-sqrt</td>
<td>RH98</td>
<td>0.00</td>
<td>133.93 571.72 142</td>
</tr>
<tr>
<td>ENT Oceania</td>
<td>0.68</td>
<td>35.02</td>
<td>14.93</td>
<td>sqrt-sqrt</td>
<td>RH70, RH98</td>
<td>11.46</td>
<td>208.39 577.72 417</td>
</tr>
<tr>
<td>ENT North America</td>
<td>0.68</td>
<td>66.44</td>
<td>16.71</td>
<td>sqrt-sqrt</td>
<td>RH50, RH98</td>
<td>0.00</td>
<td>157.47 1768.70 1391</td>
</tr>
<tr>
<td>GSW Oceania</td>
<td>0.9</td>
<td>58.60</td>
<td>11.41</td>
<td>sqrt-sqrt</td>
<td>RH98</td>
<td>9.15</td>
<td>54.83 192.56 65</td>
</tr>
</tbody>
</table>

The model for each geographic region stratum used for footprint-level prediction in the GEDI04_A product. R², %RMSE, and mean residual error (MRE) were all calculated from geographic cross validation. MRE was the absolute mean binned residual error, expressed in Mg/ha. The minimum, mean, and maximum aboveground biomass density of training samples for each stratum are reported in Mg/ha, along with the total number of training samples used from each stratum.

<table>
<thead>
<tr>
<th>Stratum</th>
<th>R²</th>
<th>%RMSE</th>
<th>MRE (Mg/ha)</th>
<th>Transform</th>
<th>Predictors</th>
<th># Training samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>DBT</td>
<td>0.58</td>
<td>49.21</td>
<td>21.62</td>
<td>sqrt-sqrt</td>
<td>RH60, RH98</td>
<td>157.47 245.64 326</td>
</tr>
<tr>
<td>EBT</td>
<td>0.64</td>
<td>52.14</td>
<td>17.2</td>
<td>sqrt-sqrt</td>
<td>RH50, RH98</td>
<td>206.13 961.80 326</td>
</tr>
<tr>
<td>ENT</td>
<td>0.70</td>
<td>59.26</td>
<td>19.69</td>
<td>sqrt-sqrt</td>
<td>RH60, RH98</td>
<td>206.13 961.80 326</td>
</tr>
<tr>
<td>GSW</td>
<td>0.86</td>
<td>55.39</td>
<td>8.85</td>
<td>sqrt-sqrt</td>
<td>RH98</td>
<td>63.10 386.76 490</td>
</tr>
<tr>
<td>North</td>
<td>0.67</td>
<td>55.72</td>
<td>18.76</td>
<td>sqrt-sqrt</td>
<td>RH50, RH98</td>
<td>206.13 961.80 326</td>
</tr>
<tr>
<td>South</td>
<td>0.65</td>
<td>43</td>
<td>12.38</td>
<td>sqrt-sqrt</td>
<td>RH60*RH70, RH98</td>
<td>206.13 961.80 326</td>
</tr>
<tr>
<td>America</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oceania</td>
<td>0.62</td>
<td>69.75</td>
<td>73.73</td>
<td>sqrt-sqrt</td>
<td>RH98</td>
<td>63.10 386.76 490</td>
</tr>
<tr>
<td>Asia</td>
<td>0.46</td>
<td>98.83</td>
<td>142.98</td>
<td>log-log</td>
<td>RH50, RH98</td>
<td>206.13 961.80 326</td>
</tr>
<tr>
<td>Europe</td>
<td>0.64</td>
<td>40.78</td>
<td>13.78</td>
<td>sqrt-sqrt</td>
<td>RH98</td>
<td>63.10 386.76 490</td>
</tr>
<tr>
<td>Africa</td>
<td>0.71</td>
<td>76.12</td>
<td>11.56</td>
<td>sqrt-sqrt</td>
<td>RH98</td>
<td>63.10 386.76 490</td>
</tr>
</tbody>
</table>

3.3. Predictor and response variable transformation

Within each geographic stratum, we also allowed flexibility in the adopted transformation, enabling different model forms in structurally different forests rather than insisting on, for example, log-log linear fits in all forests. As aforementioned, transformations were desirable both to enable the application of OLS models to account for nonlinear relationships between predictor and response variables (Fig S7), but also to ensure the errors were normally distributed in model fit space. The majority of models adopted a square root transform on the response variable (Table 2, Table 3). The primary exceptions were for EBT Asia, and Asia.

### 3.4. Model stratification

Models were also fit at differing levels of geographic stratification. Models fit at both PFT by geographic region level (Table 2) typically performed better than models stratified by geographic region or PFT alone (Table 3) in terms of accuracy assessment (lower mean residual error, lower %RMSE, higher R²). When directly comparing estimates from the most refined PFT by geographic region models with estimates from a single global model fit (Fig. 9), the more stratified models were equal to or better than the global model in a given stratum with respect to %RMSE. The stratified models also had lower MRE values, and the R² values were similar between the two sets. Some strata did not benefit from a more stratified model, while others improved substantially.

### 4. Discussion

Global-scale lidar specifically designed for measuring forest structure has not been available at a footprint size of 25 m, so generating a set of globally representative models is a relatively novel endeavor (with the exception of the pantropical studies listed in Table 1). GEDI04_A models performed comparably to other wide area AGBD modeling efforts (Table 1), but generally did not produce as accurate results as local studies where models are specifically developed for the area of interest (Ploton et al., 2020). We found that accuracies varied considerably by geographic strata (Section 4.1), but that variable selection was fairly consistent and primarily used high (RH98, RH90) and low (RH10, RH20) height metrics (Section 4.2). Given the potential differences between simulated and on-orbit waveforms (Section 2.3), where performance was roughly equivalent (Table 4), the models without low RH metrics were selected. The degree of spatial stratification had a meaningful impact on accuracies (Section 4.3).

#### 4.1. Model performance

The highest accuracies were found when models were fit at the most detailed (PFT by geographic region) stratification level, and typically models in the more poorly performing geographic regions adopted training data from a broader domain. For example, the best EBT Asia model used training data across all EBT forests. We see this model exhibits high %RMSE, low R², has a non-zero mean residual error, and consistently overestimated low biomass and underestimated high biomass. In this example, very limited training data were available, with only a few training sites in particularly high biomass areas of Borneo,
that do not represent the composition of the broader PFT by geographic region (Banin et al., 2014). While more training data in this geographic stratum are highly desirable, in their absence we were faced with the decision either to apply a poorly performing model to any footprints in EBT Asia, or to apply a more generalized model (i.e., for all of Asia, or for all EBT forests). As the model performance was particularly unreliable (Table 2), a more broadly trained, PFT-wide model (Table 3) was selected for the GEDI04_A product.

While EBT Asia was the most challenging stratum for model fitting, high %RMSE values were also found in EBT Africa, and ENT North America. The former was also likely due to data limitations, as the EBT Africa data were constrained to two primary clusters; one being the AfriSAR sites in Gabon (Fatoyinbo et al., 2021), and the second being from sampled forest plots in the Democratic Republic of Congo (Kearsley et al., 2013). Conversely, the ENT North America model was one of the most data rich training strata in the database. The large %RMSE values here were likely related to the wide range of forest types represented in this class, with high biomass forests presenting estimation challenges while the mean biomass remains low, thus increasing the %RMSE rather than the RMSE. Uncertainties (both absolute and relative RMSE) in AGBD estimation generally increased with AGBD (Baccini et al., 2017; Duncanson et al., 2020; Zolkos et al., 2013), and the highest AGBD in the database were in the tall conifer forests of the Western United States (Fig. 4).

We anticipated greater uncertainties in areas of both high biomass and dense canopy cover. GEDI’s 25-meter footprint was designed partially to overcome blending of ground and canopy signals, particularly over slopes. Small plots are known to add uncertainties to AGBD predictions, partially from edge effects, larger consequences from geo-location uncertainties, and the increase in AGBD variance with decreasing plot size (Chave et al., 2004; Labriere et al., 2018; Mauya et al., 2015; Rejou-Méchain et al., 2019; Zolkos et al., 2013).

Our results reaffirm that biomass prediction with small plots and GEDI waveforms is most challenging in high biomass, closed canopy forests. While we found models fit in conifer dominated systems (ENT) had higher %RMSE values than models fit in broadleaf dominated systems (DBT and EBT), the absolute RMSE values in terms of Mg/ha were highest for the EBT strata. Because %RMSE values were dependent on the AGBD itself, high %RMSE values were seen in the GSW and ENT classes, despite high $R^2$ values (Fig. SI 3). The ENT North America forests included the highest plot-level AGBD values in the database, but this stratum includes an exceptionally wide range of AGBD, and is

---

**Fig. 5.** The number of times predictor variables were included in the 20 top ranked models per fit stratum, aggregated by PFT, when allowing all variables as predictors. Each variable in an interaction term (e.g. RH50 x RH98) was counted separately.
Remote Sensing of Environment 270 (2022) 112845

dominated by shorter trees, which sum to a lower AGBD, rather than the giant redwood stands in the Western United States (Fig. 2). The general model performance in these strata was as expected (Table 1), but there remains ample room for improvement, particularly in tropical EBT regions, and most notably in Asia. This may include the improvement of reference data (e.g. better allometric models, more training data), the inclusion of new predictors (e.g. GEDI02_B metrics), and/or advances in model development (e.g. from machine learning).

4.2. Selection of predictor variables

When all RH metrics were candidate predictors, low RH metrics (RH10, RH20) and maximum height (RH98) were most frequently selected. Note that the results presented in Table 2 show that when forcing RH98 into models and removing RH metrics below RH50, the top model from each stratum had a simple form using only RH98 alone, RH98 and RH70, or RH98 and RH50. These closely matched models published in previous studies, (e.g. Drake et al., 2002a), and are theoretically more transferable to on-orbit data than more complicated models including low RH metrics (Hancock et al., 2019).

4.2.1. Importance of maximum height in biomass estimation

Maximum canopy height has been used as a predictor for forest AGBD for over a century, dating back to early forest inventories in Norway, Sweden and the United States (Smith, 2002). Conceptually, if AGBD is linearly related to woody volume, and the volume of a tree is approximately height multiplied by basal area, then height and some metric related to basal area should be tightly coupled to AGBD (Asner and Mascaro, 2014; Coomes et al., 2017; Drake et al., 2003; Enquist et al., 2009; Jucker et al., 2017; Lefsky et al., 1999). Some metric representing forest maximum height is therefore expected to be a strong predictor of plot-level AGBD, and our unconstrained models frequently selected RH98 or RH90 as predictors. However, RH90 can be considerably shorter than maximum canopy height. Large deviations between RH98 and RH90 were found across the full range of heights in our training database, and in every PFT (Fig. 7h). Some of these deviations in short forests were likely related to canopy cover where in low cover environments RH90 may be within the ground return. In taller, denser forests, large deviations would be expected due to either sparse crowns (e.g., tall conifers) or emergent crowns in broadleaf systems where a footprint may only capture an emergent crown’s edge. Considering the advantages of including RH98 as a predictor, all of the selected version 1

Fig. 6. The number of times predictor variables were included in the top 20 models per fit stratum, aggregated by PFT, when forcing RH98 and removing low RH metrics as predictors (constrained models). Each variable in an interaction term (e.g. RH50 x RH98) was counted separately.
4.2.2. Importance of low RH metrics in biomass modeling

In contrast to the clear biophysical meaning of RH98 (maximum height), the frequently selected low RH metrics in the unconstrained models were more difficult to interpret with respect to their importance for AGBD estimation. Other waveform lidar-based studies have differed in the selection of RH metrics, where Huang et al., 2013; Sun et al., 2011 found RH50 and RH75 the most useful, Swatantran et al., 2011 used RH75, RH100 and canopy cover, Ni-Meister et al., 2010 used RH50, RH100, and canopy cover, Drake et al., 2003 used RH50 (HOME) alone. However, many lidar studies have demonstrated the importance of canopy cover ( Table 1 ), and low RH metrics may be particularly sensitive to canopy cover ( Fig. 7 g). RH10 and RH20 should also be sensitive to terrain slope, which may in turn be correlated with AGBD (Ferry et al., 2010). Future iterations of GEDI04_A models will consider other predictors from the L2B product (e.g., canopy cover, Plant Area Index, Foliage Height Diversity) after their generation in the GEDI waveform simulator has been validated. We anticipate that these cover-based metrics may play an important role; as to whether they provide enough independent information to improve biomass modeling relative to low RH metrics remains to be tested. However, inclusion of these lower RH metrics did not yield a substantial enough improvement in model performance ( Table 4 ) to overcome other considerations. For the strata where there was a substantial increase in MRE when removing models with low RH metrics (DBT North America, DBT Europe), the top candidate models that included low RH metrics included many more predictor metrics and several interaction terms (Table S2), thus there was a trade off between minimizing MRE and maximizing parsimony. After consideration, the selected version 1 GEDI04_A models ( Table 2, 3 ) did not include any RH metrics below RH50.

Fig. 7. Relationships between RH metrics (predictors), AGBD, and cover, coloured by density of data points (grey = few, red = most). All RH metrics were correlated with AGBD (a-e), and RH10 had the most variance in high canopy cover forests (g). While RH98 and RH90 were highly correlated (h), in many footprints there was a large (>5 m) difference (i).
4.3. Model stratification

Models stratified by both PFT and geographic region generally performed better than more broadly stratified (only PFT, only geographic region, or global) models, as expected (Fig. 9). However, more detailed strata often used training data outside of their stratum, suggesting either that there were insufficient training data within a stratum to fit a geographically transferable model (e.g., EBT Asia), or that certain strata are structurally similar across broader geographic domains, and thus broader training datasets were beneficial. The former is a hypothesis about data scarcity, while the latter presents a hypothesis about structural convergence across continents or PFTs. While our results from the more data rich strata (e.g. ENT North America) supports the second hypothesis, the degree to which either of these hypotheses holds can only be determined with new inputs of data, either to bolster stratification efforts or provide independent reference data.

The most noteworthy difference observed between a single global model fit and stratified models is the increased MRE associated with the global model fit (Fig. 9b). This suggests that when a global model and stratified model had comparable $R^2$ or %RMSE values (Fig 8), systematic errors were introduced when applying a model that was unrepresentative of the spatial domain to which it was applied. Note that the MRE term selected did not show the direction of error (it was the absolute mean residual error in bins of predicted AGBD), but in general models overestimated low biomass and underestimated high biomass (have systematic error), and this was particularly common when applying broadly stratified models.

4.3.1. Between-strata variation

Our results confirm that stratification is important when fitting simple OLS models to predict AGBD at a global scale. However, the degree of variation between model forms, variable selection, and accuracy may illuminate areas where we could improve or further refine the current stratification. For example, if two strata had nearly identical models, they may be good candidates for a new, merged stratum. Conversely, if a stratum had poor model performance and substantial within stratum variability, it may be a candidate for further stratification. We hypothesized that different geographic strata would have different relationships between waveform metrics and AGBD, and our analysis confirmed this is indeed the case (Table 2). Even with our relatively sparse samples of field and simulated waveform datasets, we observed clear discrepancies in model performance and variable selection between strata (Table 2, Table 3, Fig 8).

4.4. Alternative model forms

The framework for creating GEDI’s gridded 1-km biomass product (GEDI04_B) is enabled through the use of a parametric framework for footprint level biomass prediction. However machine learning modeling approaches may present an attractive alternative. For example, recent work has shown that convolutional neural nets may be trained directly from waveforms, with comparable or greater accuracy than existing waveform processing methods (Lang et al., 2019) to estimate RH metrics and ground elevations. This method could be used to bypass the use of RH metrics entirely and derive biomass only using the waveform (the GEDI01_B product). The implementation of such an approach to support
the GEDI04_B gridded product will depend on the development of theory to link machine learning methods with the statistical methods GEDI uses for gridded biomass, including hybrid estimation (Patterson et al., 2019). Linear mixed-effects models may also be useful alternatives to the stratification approach adopted in the first version of the GEDI04_A product and will be explored in future iterations.

4.5. Unreported sources of uncertainty

While we attempted to minimize uncertainties in our models, and therefore in the AGBD predictions on the GEDI04_A product, there are several sources of uncertainty that we were unable to estimate or that were beyond the scope of this work. Measurement errors in the field inventory data, measurement error in RH metrics derived from GEDI waveforms collected on-orbit, errors associated with the selection and application of allometric models, errors associated with any lack of representativeness of our training data (sampling errors), and errors associated with the transference of models using simulated RH metrics to on-orbit predictions were not included in the uncertainty estimates of the footprint-level GEDI04_A predictions.

We attempted to select models that would be applicable outside their geographic domain of model training. While we undertook checks to ensure the models showed no systematic lack of fit to the sample data, the assumption that the range of values in the training data was similar

Table 4

Differences in top model %RMSE and mean residual error (MRE) between the unconstrained (all RH metrics) models and the most constrained (no low RH metrics, forced RH98) models. Differences are unconstrained minus constrained. Thus, when the constrained model performed more poorly, differences are negative. Differences larger than 5% RMSE or 10 Mg/ha mean residual error are highlighted in bold.

<table>
<thead>
<tr>
<th>Strata</th>
<th>Unconstrained</th>
<th>Constrained</th>
<th>Difference (U-C)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>%RMSE</td>
<td>MRE %</td>
<td>%RMSE</td>
</tr>
<tr>
<td>DTB Africa</td>
<td>54.97</td>
<td>5.97</td>
<td>56.88</td>
</tr>
<tr>
<td>DTB Europe</td>
<td>48.86</td>
<td>3.84</td>
<td>47.27</td>
</tr>
<tr>
<td>DTB North America</td>
<td>35.05</td>
<td>7.60</td>
<td>38.08</td>
</tr>
<tr>
<td>EBT Africa</td>
<td>60.02</td>
<td>11.31</td>
<td>66.89</td>
</tr>
<tr>
<td>EBT Asia</td>
<td>60.31</td>
<td>75.75</td>
<td>78.94</td>
</tr>
<tr>
<td>EBT Oceania</td>
<td>28.66</td>
<td>8.17</td>
<td>28.66</td>
</tr>
<tr>
<td>EBT South America</td>
<td>41.66</td>
<td>5.56</td>
<td>42.20</td>
</tr>
<tr>
<td>ENT Oceania</td>
<td>58.06</td>
<td>11.21</td>
<td>63.43</td>
</tr>
<tr>
<td>ENT Europe</td>
<td>35.72</td>
<td>11.77</td>
<td>35.02</td>
</tr>
<tr>
<td>ENT North America</td>
<td>69.46</td>
<td>9.77</td>
<td>66.44</td>
</tr>
<tr>
<td>GSW Oceania</td>
<td>37.45</td>
<td>2.62</td>
<td>44.53</td>
</tr>
</tbody>
</table>

Fig. 9. Stratified models (PFT by geographic region) were equal or better than a single global model applied to each strata in terms of mean residual error (Mg/ha) (b), %RMSE (c) and R² (d), with the exception of DTB_Eu.
in the sample as in the population was more difficult to achieve because of the limited availability of high quality data across all prediction strata to which the models will be applied. The GEDI04_A data product includes the covariance matrix of the model parameters, which we assume conveys the uncertainty of footprint estimates of AGBD, and two quality flags that indicate whether the predictor variables or the predicted response are outside the range of values observed in the training data. As more training data and on-orbit GEDI data become available, these will contribute to a more complete uncertainty assessment of GEDI04_A predictions. A comprehensive and current discussion of the sources of uncertainty associated with generation of training data, parametric modeling of biomass using these data, and application of models for prediction is presented in Duncanson et al. (2021).

5. Conclusions

Here we have described GEDI’s footprint-level AGBD models using a data-informed approach for variable selection, data filtering and transformation, and model stratification. Most models perform well, and are consistent with results from previous studies. We found low RH metrics (e.g. RH10) and maximum canopy height (RH98) were primarily selected as predictors of AGBD, but constraining models by removing low RH metrics and forcing RH98 as a predictor did not substantially reduce model performance in the majority of strata, and where there was a reduction in performance, there was also a tradeoff with model parsimony (Table S2).

The models described herein are the set of models used in the version 1 and 2 of the GEDI04_A product, but future releases will incorporate improved versions of these models. Specifically, we anticipate improvements will come from the incorporation of more candidate predictors (L2B metrics, e.g. cover, and relevant ancillary data), and as more training data become available. The latter is particularly important given the lack of training data in some strata and associated poor model performance (e.g., Continental Asia). Our approach to model selection through geographic transferability cross-validation attempted to overcome the issue of sparse training data, but the effectiveness of this approach needs to be assessed, and that in turn requires validation data in these areas. As models improve, new versions of the GEDI04_A product will be produced, with details of this anticipated cadence presented in Kellner et al. (2021).

The development of global biomass maps from GEDI rests upon the development of robust, transparent, and reproducible calibration models with well-known error structures. GEDI’s small footprint size and geolocation accuracy, when coupled with randomly precessing orbits and limited pointing capability, preclude the implementation of any kind of post-hoc calibration strategy. Instead, GEDI’s approach to biomass modeling has been, from its inception, to use a pre-launch calibration strategy that exploits the array of existing ground plots for which associated airborne lidar exist. The collation of these data in GEDI’s Forest Structure and Biomass Database was, and continues to be, a laborious task and has resulted in a dataset that is unprecedented in its scope; yet it exists only because of the equally arduous efforts of many data collaborators to collect and process field data, and then to make these data available. The importance of these data cannot be stressed enough and their continued development should be encouraged and supported. As open data policies become more widely adopted, and forest biomass continues to be a priority for inventory or climate change mitigation efforts, we hope and expect that more funding for targeted reference data will become available, bolstering improved product development and thus improved science and applications. This will be greatly strengthened if the precarious situation of many potential data contributors is recognised, especially those working in tropical nations. This means adequately funding not just the narrow process of data collection, but developing non-extractive models in which training, career development, herbarium work, long-term data management and sustained research funding are all a core, directly funded part of the mission calibration and validation process, not as optional add-ons or after thoughts.

Based on the work presented here, GEDI will ultimately provide beyond 10 billion estimates of footprint biomass during its mission, dwarfing the existing archive of space-based lidar estimates. GEDI is the first mission that was developed to explicitly measure ecosystem structure, and whose statistical estimation framework was integrated into its mission design from the outset. This estimation framework has driven our approach to GEDI04_A biomass and the result will be maps of gridded biomass where, perhaps for the first time, the precision of those estimates is well understood. Thus, our efforts here are an important step towards a next generation of biomass products that may confidently be used by themselves, and in harmony with other data towards addressing the pressing environmental challenges GEDI was designed to meet.

Credit author statement

Laura Duncanson (LD), James Kellner (JK), and John Armstrong (JA) formulated the concept for the paper, conducted the analysis, and led the writing of the paper. Ralph Dubayah led the GEDI mission, guided the development and writing of paper writing. LD, JA, David Minor (DM), Suzanne Marselis, Carlos E Silva, and Jamis Bruening compiled and processed the training dataset. DM also created several figures for the paper and contributed to writing and editing the paper. Steven Hancock developed and applied the waveform simulator, and helped with the writing and theoretical development of the work. Sean Healey, Paul Patterson and Svetlana Saarela provided statistical guidance and contributed to the writing. Scott J. Goetz, Hao Tang, Michelle Hofton, Bryan Blair, Scott Luthcke, and Lola Fatoyinbo provided comments and content on the GEDI mission and associated past studies. All other authors provided training data to the study, through collection of the field data, and/or data processing and curation, as well as providing helpful comments and edits to the paper.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

Armstrong, Kellner, Hancock, and Dubayah were supported by NASA Contract #NNL15AA03C to the University of Maryland for the development and execution of the GEDI mission. Duncanson and Minor were supported by a NASA GEDI Science Team Grant NNH20ZDA001N and a NASA Post Doctoral Program fellowship. Saarela was supported through NASA Carbon Monitoring System Grant 80HQR18T0016, Healey and Patterson were funded by the GEDI mission through Interagency Agreement RPO201523. We thank the NASA Terrestrial Ecology program for continued support of the GEDI mission, and the University of Maryland for providing independent financial support of the GEDI mission. We also thank NASA for contributing to several lidar data collections used in this study, including from the NASA Carbon Monitoring System (Grant number NNH13AW621, to PI Cohen at the USFS Service). We also gratefully acknowledge the collection and provision of field and airborne data from a wide variety of other sources, including by the Sustainable Landscapes Brazil project supported by the Brazilian Agricultural Research Corporation (EMBRAPA), the US Forest Service, the National Science Foundation (DEB 0939907), Smithsonian Tropical Research Institute, USAID, and the US Department of State, among others. Additional data were acquired from the Terrestrial Ecosystem Research Network (TERN), an Australian Government NCRIS-enabled research infrastructure project, for provision of data used in this analysis, and from the National Ecological Observatory Network (NEON), a program sponsored by the National Science Foundation and operated.
Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.

References


