Remote Sensing of Environment 270 (2022) 112845

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

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

Laura Duncanson [a,*], James R. Kellner [b], John Armston [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 [a], Scott J. Goetz [g], Hao Tang [a,ct], Michelle Hofton [a], Bryan Blair [h], Scott Luthcke [h], Lola Fatoyinbo [h], Katharine Abernethy [i,j], Alfonso Alonso [k], Hans-Erik Andersen [l], Paul Aplin [m], Timothy R. Baker [n], Nicolas Barbier [o], Jean Francois Bastin [cv], Peter Biber [q], Pascal Boeckx [r], Jan Bogaert [p], Luigi Boschetti [s], Peter Brehm Boucher [t], Doreen S. Boyd [u], 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,ai], Michele Dalponte [aj], Jonathan Dash [ak], Sergio de-Miguel [al,am], Songqiu Deng [an], Peter Woods Ellis [ao], Barend Erasmus [ap,aq], Patrick A. Fekety [ar], Alfredo Fernandez-Landa [as], Antonio Ferraz [at,au], Rico Fischer [av], Adrian G. Fisher [aw,ax], Antonio García-Abril [ay], Terje Gobakken [f], Jorg M. Hacker [az,ba], Marco Heurich [bb,bc,bd], Ross A. Hill [be], Chris Hopkinson [bf], Huabing Huang [bg], Stephen P. Hubbell [ag,au], Andrew T. Hudak [cw], Andreas Huth [av,bh,bi], Benedikt Imbach [bj], Kathryn J. Jeffery [i], Masato Katoh [an], Elizabeth Kearsley [r], David Kenfack [ag], Natascha Kljun [bk], Nikolai Knapp [av,cu], Kamil Král [bl], Martin Krůček [bm], Nicolas Labrière [x], Simon L. Lewis [n,bm], Marcos Longo [at,bn], Richard M. Lucas [bo], Russell Main [aq,bp], Jose A. Manzanera [bq], Rodolfo Vásquez Martínez [br], Renaud Mathieu [aq,bs], Herve Memiaghe [bt,x], Victoria Meyer [at,bu], Abel Monteagudo Mendoza [br,bv], Alessandra Monerris [bw], Paul Montesano [h,bx], Felix Morsdorf [by], Erik Næsset [f], Laven Naidoo [bp], Reuben Nilus [bz], Michael O'Brien [ca], David A. Orwig [cb], Konstantinos Papathanassiou [cc], Geoffrey Parker [cd], Christopher Philipson [ce,cf], Oliver L. Phillips [n], Jan Pisek [cg], John R. Poulsen [ch], Hans Pretzsch [q], Christoph Rüdiger [ci,cj], Sassan Saatchi [at], Arturo Sanchez-Azofeifa [w], Nuria Sanchez-Lopez [s], Robert Scholes [ap], Carlos A. Silva [ck], Marc Simard [at], Andrew Skidmore [cl], Krzysztof Stereńczak [cm], Mihai Tanase [bw], Chiara Torresan [ae,cn], Ruben Valbuena [e,co], Hans Verbeeck [r], Tomas Vrska [bl], Konrad Wessels [cp], Joanne C. White [cq], Lee J.T. White [i,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 Skogsmarksgränd 17, SE-901 83 Umeå, Sweden
[f] Norwegian University of Life Sciences, P.O. Box 5003, NMBU, 1432 Ås, Norway
[g] Northern Arizona University, 1295 S. Knoles Dr., Flagstaff, AZ 86011, USA
[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

* 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

[k] Smithsonian Conservation Biology Institute, PO Box 37012, MRC 5516, Washington, DC 20013, USA
[l] USDA Forest Service, Pacific Northwest Research Station,University of Washington, Box 352100, Seattle, WA 98195-2100, USA
[m] Edge Hill University, St Helens Road, Ormskirk, Lancashire L394QP, UK
[n] University of Leeds, Woodhouse Lane, Leeds LS2 9JT, UK
[o] AMAP, Univ Montpellier, IRD, CNRS, INRAE, CIRAD, 34398 Montpellier cedex 5, France
[p] University of Liege, Passage des Déportés 2, B-5030 Gembloux, Belgium
[q] Technical University Munich, Arcisstraße 21, D-80333 Munich, Germany
[r] Ghent University, Coupure Links 653, 9000 Gent, Belgium
[s] University of Idaho, 875 Perimeter Dr., MS 1133, Moscow, ID 83844, USA
[t] Harvard University, 26 Oxford Street Cambridge, MA 02138, USA
[u] University of Nottingham, University Park, Nottingham NG7 2RD, UK
[v] University of Aberdeen, Cruickshank Building, St Marchar Drive, Aberdeen AB24 3UU, Scotland, UK
[w] University of Alberta, Edmonton, AB T6G 2E3, Canada
[x] Laboratoire Évolution et Diversité Biologique (EDB), UMR 5174 (CNRS/IRD/UPS), 118 route de Narbonne, 31062 Toulouse Cedex 9, France
[y] Université Toulouse, 41 Allées Jules Guesde - CS 61321, 31013 TOULOUSE - CEDEX 6, France
[z] University of Connecticut, 75 North Eagleville Road, Storrs, CT 06269-2043, USA
[aa] University of the Sunshine Coast, Sippy Downs, Queensland 4556, Australia
[ab] University of Missouri-St. Louis, 223 Research Building, One University Boulevard, St. Louis, MO 63121-4499, USA
[ac] USDA Forest Service, Pacific Northwest Research Station, 3200 SW Jefferson Way, Corvallis, OR 97331, USA
[ad] University of Cambridge, Downing Street, Cambridge CB2 3EA, UK
[ae] Council for Agricultural Research and Economics, viale Santa Margherita 80, 52100 Arezzo, Italy
[af] University of Tuscia, via San Camillo de Lellis, 01100 Viterbo, Italy
[ag] Smithsonian Tropical Research Institute, PO Box 0843-03092, Balboa, Ancón, Panama
[ah] University of Dundee, Nethergate, Dundee DD1 4HN, UK
[ai] University of Illinois at Urbana-Champaign, 286 Morrill Hall, 505 S Goodwin Ave, Urbana, IL 61801, USA
[aj] Research and Innovation Centre, Fondazione Edmund Mach, via E. Mach 1, 38098 San Michele all'Adige (TN), Italy
[ak] Scion New Zealand Forest Research Institute, New Zealand
[al] University of Lleida, Av. Alcalde Rovira Roure, 191, E-25198 Lleida, Spain
[am] Joint Research Unit CTFC - AGROTECNIO - CERCA, Crta. de St. Llorenç de Morunys a Port del Comte, km 2, E- 25280 Solsona, Spain
[an] Shinshu University, 8304, Minamiminowa-Vill., Kamiina-County, Nagano 399-4598, Japan
[ao] The Nature Conservancy, 4245 North Fairfax Drive, Suite 100, Arlington, VA 22203, USA
[ap] University of the Witwatersrand, 1 Jan Smuts Avenue, Braamfontein, 2000, Johannesburg, South Africa
[aq] University of Pretoria, Lynnwood Rd, Hatfield, Pretoria 0002, South Africa
[ar] Colorado State University, 1062 Campus Delivery, Fort Collins, CO 80523-1062, USA
[as] Agresta Sociedad Cooperativa, St. Numancia 1, E-42001 Soria, Spain
[at] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove, Pasadena, CA 91109, USA
[au] University of California, Los Angeles, LaKretz Hall, 619 Charles E Young Dr E #300, Los Angeles, CA 90024, USA
[av] Helmholtz Centre for Environmental Research - UFZ, Permoserstr. 15, 04318 Leipzig, Germany
[aw] University of New South Wales, Sydney, NSW 2052, Australia
[ax] University of Queensland, Brisbane, QLD 4072, Australia
[ay] Universidad Politécnica de Madrid (UPM), Ciudad Universitaria s/n, 28040 Madrid, Spain
[az] Airborne Research Australia, Hangar 60, Dakota Drive, Parafield Airport, 5106, Australia
[ba] Flinders University, 182 Victoria Square, Adelaide, SA, 5000, Australia
[bb] Bavarian Forest National Park, Freyunger Straße, 94481 Grafenau, Germany
[bc] University of Freiburg, Tennenbacher Straße 2, 79106 Freiburg, Germany
[bd] Inland Norway University of Applied Sciences, 2480 Koppang, Norway
[be] Bournemouth University, Talbot Campus, Poole, Dorset BH12 5BB, UK
[bf] University of Lethbridge, 4401 University Drive, Lethbridge, Alberta T1K 3M4, Canada
[bg] School of Geospatial Engineering and Science, Sun Yat-Sen University, Guangzhou 510275, P. R. China
[bh] University of Osnabrück, Barbarastr 12, D - 49076 Osnabrück, Germany
[bi] German Centre for Integrative Biodiversity Research (iDiv), Puschstrasse 4, 04103 Leipzig, Germany
[bj] Aeroscout, Hengstrain 14, 6280, Hochdorf, Switzerland
[bk] Lund University, Centre for Environmental and Climate Science, Sölvegatan 37, 223 62 Lund, Sweden
[bl] The Silva Tarouca Research Institute, Lidická 25/27, 602 00 Brno, Czech Republic
[bm] University College London, London WC1E 6BT, UK
[bn] Embrapa Agricultural Informatics, Av. André Tosello 209, 13083-886 Campinas, SP, Brazil
[bo] Aberystwyth University, Penglais Campus, Aberystwyth, Ceredigion SY23 3DY, UK
[bp] Council for Scientific and Industrial Research, PO BOX 395, Pretoria 0001, South Africa
[bq] Universidad Politécnica de Madrid, Ciudad Universitaria, 28040 Madrid, Spain
[br] Jardín Botánico de Missouri, Prolongación Bolognesi Mz.E-6, Peru
[bs] International Rice Research Institute, Pili Drive, Los Baños, Laguna 4031, Philippines
[bt] University of Oregon, 1585 E 13th Ave, Eugene, OR 97403, USA
[bu] Terraformation, Kamuela, HI 96743, USA
[bv] Universidad Nacional de San Antonio Abad del Cusco, Av. de La Cultura 773, Cusco 08000, Peru
[bw] University of Melbourne, Grattan Street, Parkville, Victoria, Australia
[bx] Science Systems and Applications, Inc. (SSAI), 10210 Greenbelt Road, Lanham, MD 20706, USA
[by] Department of Geography, University of Zürich, Winterthurerstr. 190, 8057 Zürich, Switzerland
[bz] Sabah Forestry Department, P.O.Box 1407, 90715 Sandakan, Sabah, Malaysia
[ca] Universidad Rey Juan Carlos, c/Tulipán s/n, E-28933 Móstoles, Spain
[cb] Harvard University, Harvard Forest, 324 North Main Street, Petersham, MA 01366, USA
[cc] DLR, Königswinterer Str. 522-524, D-53227 Bonn, Germany
[cd] Smithsonian Environmental Research Center, 647 Contees Wharf Road. Edgewater, MD 21037, USA
[ce] ETH Zürich, Universitätstrasse 16, 8092 Zürich, Switzerland
[cf] Permian Global, Savoy Hill House, 7-10 Savoy Hill, WC2R 0BU, London
[cg] University of Tartu, Tartu Observatory, Observatooriumi 1, Toravere 61602, Estonia
[ch] Duke University, PO Box 90328, USA
[ci] Monash University, Department of Civil Engineering, 23 College Walk, Clayton, VIC 3800, Australia
[cj] Bureau of Meteorology, 700 Collins St, Docklands, VIC 3008, Australia
[ck] University of Florida, 342 Newins-Ziegler Hall, PO Box 110410, Gainesville, FL, USA
[cl] University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands

[cm] Forest Research Institute, Braci Leśnej 3 Street, Sękocin Stary, 05-090 Raszyn, Poland
[cn] Institute of BioEconomy - National Research Council of Italy, via Biasi, 75, 38098 San Michele all'Adige (TN), Italy
[co] Bangor University, Today building, Deiniol Road, LL57 2UW, Bangor, UK
[cp] George Mason University, 4400 University Drive, MSN 6C3, Fairfax, VA 22030, USA
[cq] Canadian Forest Service, Natural Resources Canada, 506 West Burnside Road, Victoria, BC V8Z 1M5, Canada
[cr] Ministry of Forests, Sea, the Environment and Climate Change, Boulevard Triomphal Omar BONGO, Libreville, P.O. Box: 199, Gabon
[cs] The Sokoine University of Agriculture, P.O. Box 3000, Chuo Kikuu, Morogoro, Tanzania
[ct] Department of Geography, National University of Singapore, 117570, Singapore
[cu] Thünen Institute of Forest Ecosystems, Alfred-Möller-Straße 1, 16225 Eberswalde, Germany
[cv] TERRA Teaching and Research Centre, Gembloux Agro-Bio Tech, University of Liège, Gembloux, Belgium
[cw] USDA Forest Service, Rocky Mountain Research Station, 1221 South Main St., Moscow, ID 83843, USA

ARTICLE INFO

Editor name: Jing M. Chen

Keywords:
LiDAR
GEDI
Waveform
Forest
Aboveground biomass
Modeling

ABSTRACT

NASA's Global Ecosystem Dynamics Investigation (GEDI) is collecting spaceborne full waveform lidar data with a primary science goal of producing accurate estimates of forest aboveground biomass density (AGBD). This paper presents the development of the models used to create GEDI's footprint-level (~25 m) AGBD (GEDI04_A) product, including a description of the datasets used and the procedure for final model selection. The data used to fit our models are from a compilation of globally distributed spatially and temporally coincident field and airborne lidar datasets, whereby we simulated GEDI-like waveforms from airborne lidar to build a calibration database. We used this database to expand the geographic extent of past waveform lidar studies, and divided the globe into four broad strata by Plant Functional Type (PFT) and six geographic regions. GEDI's waveform-to-biomass models take the form of parametric Ordinary Least Squares (OLS) models with simulated Relative Height (RH) metrics as predictor variables. From an exhaustive set of candidate models, we selected the best input predictor variables, and data transformations for each geographic stratum in the GEDI domain to produce a set of comprehensive predictive footprint-level models. We found that model selection frequently favored combinations of RH metrics at the 98th, 90th, 50th, and 10th height above ground-level percentiles (RH98, RH90, RH50, and RH10, respectively), but that inclusion of lower RH metrics (e.g. RH10) did not markedly improve model performance. Second, forced inclusion of RH98 in all models was important and did not degrade model performance, and the best performing models were parsimonious, typically having only 1-3 predictors. Third, stratification by geographic domain (PFT, geographic region) improved model performance in comparison to global models without stratification. Fourth, for the vast majority of strata, the best performing models were fit using square root transformation of field AGBD and/or height metrics. There was considerable variability in model performance across geographic strata, and areas with sparse training data and/or high AGBD values had the poorest performance. These models are used to produce global predictions of AGBD, but will be improved in the future as more and better training data become available.

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 $GtCO_2$ per year, while remaining forests sequester ~2.5 $GtCO_2$ 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 (AGBD [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; Rodríguez-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;

Ranson and Sun, 2000). 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. Hurtt 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 GEDI04_A 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 GEDI_04A 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

Table 1
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.

Previous study	LiDAR type/ Additional datasets	Data acquisition date	Geographic extent	Modeling type	Predictor(s) for best AGBD model	Accuracy of best AGBD model	Plot size	Number of plots
Boreal								
Næsset and Gobakken (2008)	Discrete Return LiDAR	1998 - 2006	Norway (regional)	nonlinear regression model	Height Metrics, Density Metrics	$R^2 = 0.82$ RMSE(%) = 25	0.02-0.04 ha	1395
Andersen et al. (2011)	Discrete Return LiDAR	June 2009	Tanana Valley, Alaska (Local)	OLS	Height Metrics, Density Metrics	$R^2 = 0.74$	0.033 ha	79
Margolis et al. (2015)	GLAS	2005-2006	Canadian boreal (regional)	kNN	Waveform metrics	$R^2 = 0.66$ RMSE = 27.2 Mg/ha	40 x 60 m lidar plots	565
Temperate								
Zhao et al. (2009)	Discrete Return LiDAR	March 2004	Eastern Texas, USA	functional regression models	Height Metrics	$R^2 = 0.938\%$ RMSE = 14.6	1 ha	2000 (synthesized)
	Discrete Return LiDAR (Optech Gemini ALS & G-LiHT)	2008, 2011, 2012	Teakettle, California; Parker Tract, North Carolina; SERC, Maryland (Local)	PLS	Height Metrics, Density Decile Metrics	$R^2 = 0.84\%$ RMSE = 6	0.81 ha	16
Lefsky et al. (2002)	SLICER	September 1995, July 1996	Temperate Coniferous Forest, Temperate Deciduous Forest, Boreal Coniferous Forest (Local)	Stepwise multiple regression	Canopy Height Metrics, Canopy Cover	$R^2 = 0.91$	0.25 ha	22
Gleason and Im (2012)	Discrete Return LiDAR	August 2010	Heiberg Memorial Forest, NY (Local)	SVR	Height Metrics, LAI, Canopy Volume, Crown Area	$R^2 = 0.93$ RMSE (%) = 13.6	0.038 ha	18
Hernando et al. (2019)	Discrete Return LiDAR/ Multispectral data	July 2006	Spain (Local)	k-MSN	Height Metrics, NDVI Metrics	$R^2 = 0.64\%$ RMSE = 16.7	0.126 ha	37
Ferraz et al. (2016)	Discrete Return Lidar	2008	Agueda, Portugal (local)	OLS	Height Metrics	R^2=0.72% RMSE=23.3	0.04 ha	39
Tropical								
	LVIS	1998, 2005	Costa Rica (Local)	OLS	Height Metrics	$R^2 = 0.65$ RSE = 10.5 Mg/ha	0.5 ha, 1 ha	20
Drake et al. (2003)	LVIS	March 1998	Panama, Costa Rica (Local)	OLS	Height Metrics, HOME	$R^2 = 0.89\%$ RMSE = 14.06	0.5 ha, 1 ha	71
Ene et al. (2016)	Discrete Return LiDAR	February - June 2012	Liwale district, Tanzania (Local)	OLS	Height Metrics, Density Metrics	%RMSE = 47.4	0.07 ha	513
Hansen et al. (2015)	Discrete Return LiDAR	January - February 2012	East Usambara Mountains, Tanzania (Local)	OLS	Height Metrics, Density Metrics	$R^2 = 0.70\%$ RMSE=32.3	0.1 ha	153
Labriere et al. (2018)	Discrete Return LiDAR	2009 - 2016	French Guiana, Gabon (Local)	OLS	Median Height of CHM	$R^2 = 0.79\%$ RMSE = 14.3	1 ha	183
Laurin et al. (2014)	Discrete Return LiDAR/ Hyperspectral data	March 2012	Sierra Leone (Local)	PLS	Height Metrics, Hyperspectral Bands	$R^2 = 0.7$ RMSE = 61.7 Mg/ha	0.125 ha	600
Naidoo et al. (2015)	Discrete Return Lidar	April-May 2012	Savannah, Kruger National Park, SA (local)	OLS	Height Metrics, Canopy Cover	R^2=0.63 RMSE = 19.2 Mg/ha	0.0625 ha	152
Xu et al. (2017)	Discrete Return Lidar	2012	DRC (Regional)	Power-law	Canopy height metrics	$R^2 = 0.75$ RMSE = 59.9 Mg/ha	1 ha	92
Coomes et al. (2017)	Discrete Return LiDAR	November 2014	Sabah,Malaysia (Local)	Power-law model	Height Metrics, Gap Fraction	%RMSE = 13	1 ha	36
Ferraz et al. (2018)	Discrete Return Lidar	2014	Kalimantan, Indonesia (Regional)	power-law	Canopy height metrics	Drylands $R^2 = 0.81\%$ RMSE = 20 Wetlands $R^2 = 0.79\%$ RMSE = 9	0.1 - 0.25 ha	82 (drylands) + 22 (wetlands)
Meyer et al. (2018)	Discrete Return LiDAR	2009 - 2012	Neotropics	Jackknife regression	Large Canopy Area, Wood density	$R^2 = 0.78$ RMSE = 46.0 Mg/ha	0.25 ha, 1 ha	291
Lucas et al. (2008)	Discrete Return LiDAR/ Hyperspectral data	August 2000/ September 2000	Queensland, Australia (Local)	Jackknife regression	Height Metrics, Canopy Cover	$R^2 = 0.90$ RMSE = 11.8 Mg/ha	0.25 ha	31
Pantropical								

(*continued on next page*)

Table 1 (*continued*)

Previous study	LiDAR type/ Additional datasets	Data acquisition date	Geographic extent	Modeling type	Predictor(s) for best AGBD model	Accuracy of best AGBD model	Plot size	Number of plots
Asner and Mascaro (2014)	Discrete Return LiDAR	-	Pantropic	Power-law model	Height Metrics, Basal Area, Wood density	$R^2 = 0.92$ RMSE = 17.1 Mg/ha	0.1 - 1 ha	804
Saatchi et al. (2011)	GLAS/MODIS, SRTM, QSCAT	2003-2004	Pantropic	Maximum Entropy	Height Metrics	$R^2 = 0.80\%$ RMSE = 23.8	0.25 - 1 ha	493
Baccini et al. (2012)	GLAS/MODIS	2017-2018	Pantropic	OLS & RF	HOME, Height Metrics, Canopy energy	$R^2 = 0.83$ RMSE = 22.6 Mg/ha	0.16 ha	283
Simard et al. (2018)	GLAS/SRTM/ Landsat-derived maps	2000	Mangroves (Global)	OLS	Height metric	$R^2 = 0.67$ RMSE = 84.2 Mg/ha	various	331

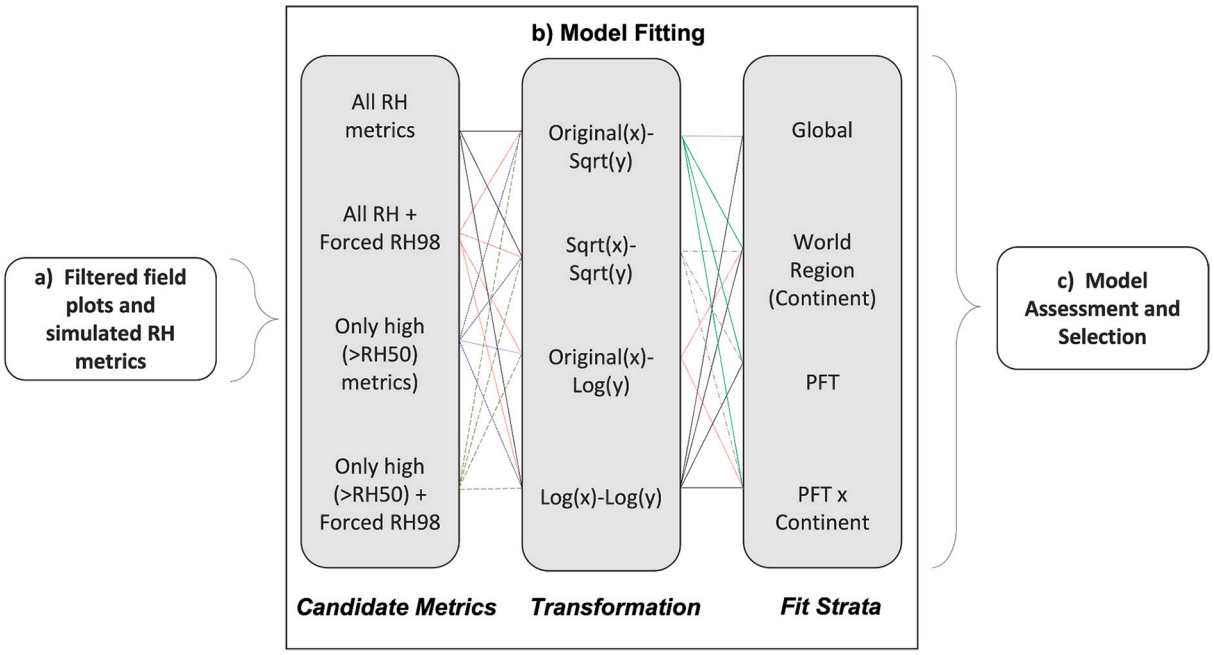

Fig. 1. Flow chart of the GEDI04_A modeling process. (a) Field estimates of AGBD and simulated GEDI RH metrics were used to (b) fit models considering an exhaustive set of RH metric selections, transformations, and geographic strata to produce thousands of candidate models. (c) These were then assessed by geographic cross-validation performance and used for final model selection.

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

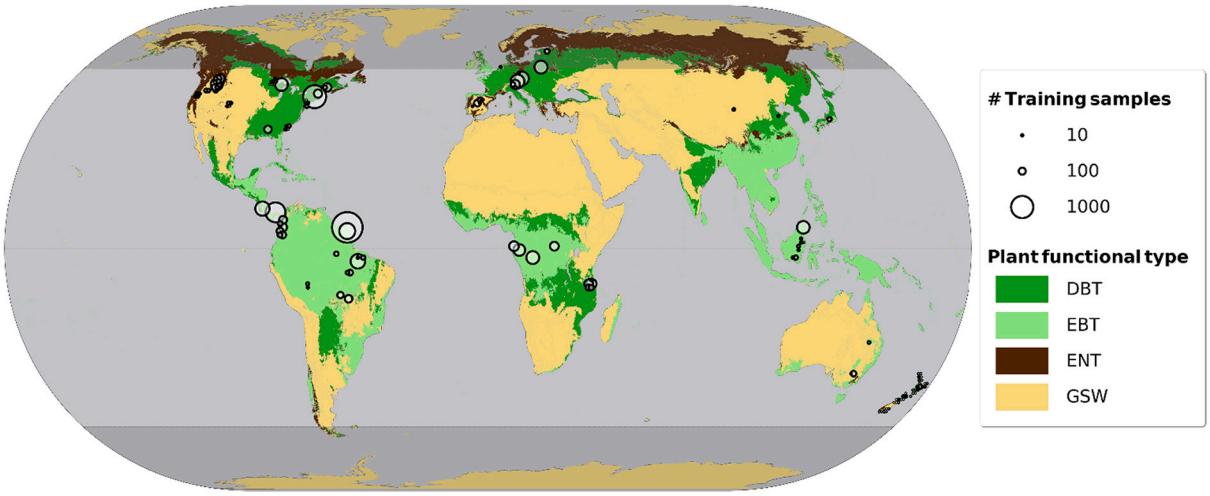

Fig. 2. The field and airborne lidar data used to fit the models in this paper are represented by circles indicating the number of samples per site. These are distributed across geographic fit strata (PFT and geographic region). Geographic regions were roughly continental. PFTs included Deciduous Broadleaf Trees (DBT), Evergreen Broadleaf Trees (EBT), Evergreen Needleleaf Trees (ENT), and Grasslands/Shrublands/Woodlands (GSW). PFTs displayed here were derived from the 2019 MODIS product MCD12Q1 V006 (Friedl et al., 2010). The light gray area highlights the geographical domain of the GEDI observations.

(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

$$h(AGBD) = \sum_{j=1}^{p} B_j f(x_j) + \epsilon \quad (1)$$

$$\widehat{AGBD} = h^* \left(\sum_{j=1}^{p} \widehat{B}_j f(x_j) + \epsilon \right) \quad (2)$$

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 ϵ 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 10th 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

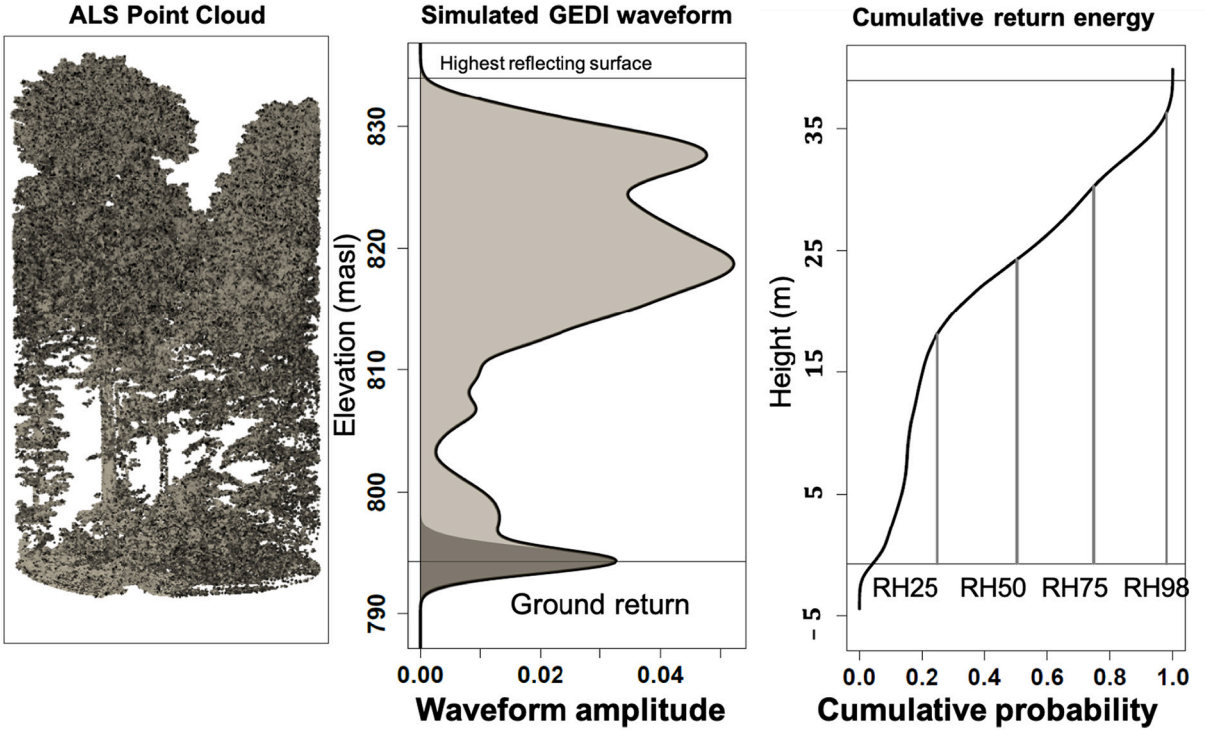

Fig. 3. Relative height (RH) metrics were calculated as the height relative to ground elevation under which a certain percentage of waveform energy has been returned. RH50, for example, is the height relative to the ground elevation below which 50% of waveform energy has been returned. Note that in cases of wide ground signals and/or sparse vegetation it is possible for RH metrics to be negative. This example is a simulated waveform in a temperate deciduous forest at Žofín, Czech Republic.

(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) ≤ 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; Friis 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

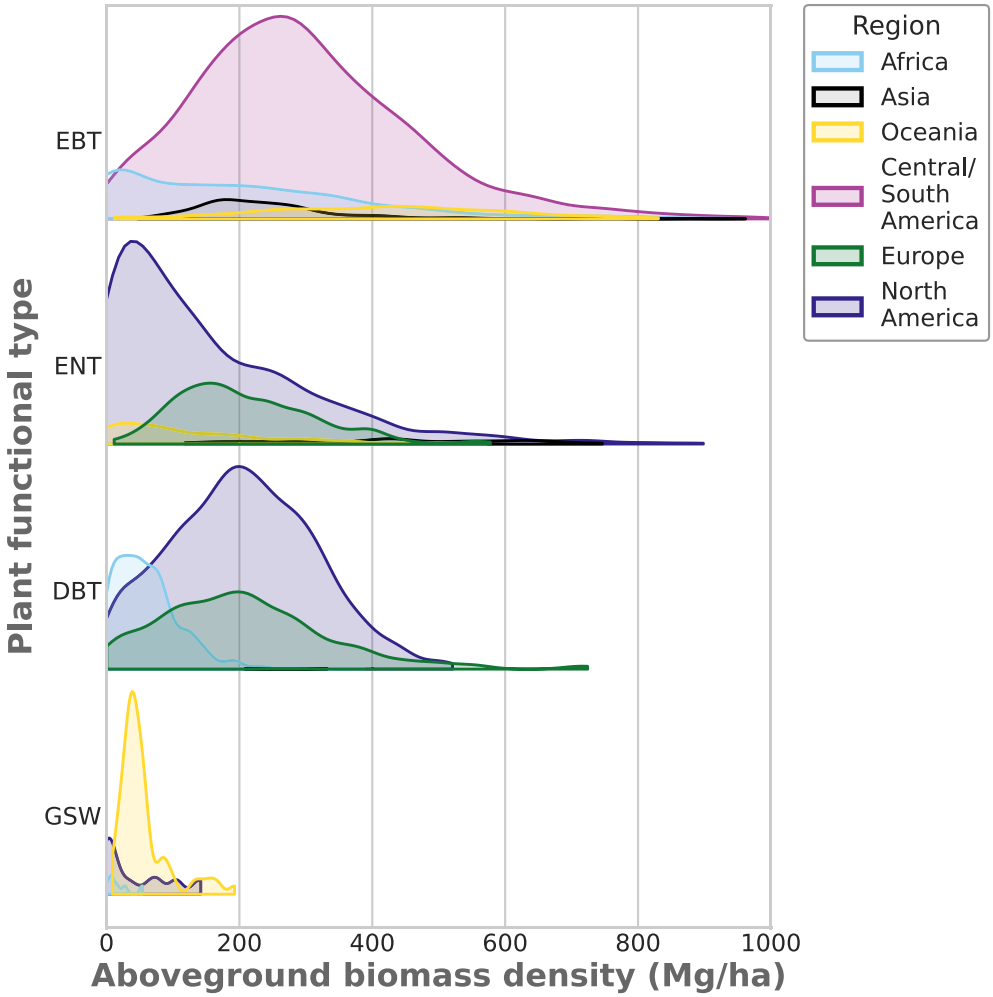

Fig. 4. Frequency distributions of AGBD for each PFT (Evergreen Broadleaf Trees (EBT), Evergreen Needleleaf Trees (ENT), Deciduous Broadleaf Trees (DBT), and Grassland, Shrub and Woodland (GSW)), and geographic region class show that different strata have different amounts of training data, with data-rich 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), typically 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. 7h, i). These occurred across the full range of heights, and across all PFTs. While these differences were relatively rare, where they occured 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

Table 2
The model for each PFT by geographic region stratum used for footprint-level prediction in the GEDI04_A product. R^2, %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.

Stratum	R^2	%RMSE	MRE (Mg/ha)	Transform	Predictors	AGBD (Mg/ha)			# Training samples
						min	mean	max	
DBT Africa	0.63	57.24	8.41	sqrt-sqrt	RH50, RH98	0.00	63.10	386.76	490
DBT Europe	0.66	47.27	21.52	sqrt-sqrt	RH70, RH98	0.52	206.13	724.06	333
DBT North America	0.66	38.08	22.81	sqrt-sqrt	RH50, RH98	0.00	205.22	520.78	873
EBT Africa	0.64	66.89	15.32	sqrt-sqrt	RH50, RH98	0.00	216.59	1489.70	834
EBT Asia	0.36	78.94	121.15	original-log	RH98	47.31	245.64	961.80	326
EBT Oceania	0.61	28.66	8.17	sqrt-sqrt	RH70, RH98	12.24	416.71	830.58	213
EBT South America	0.66	42.2	10.4	sqrt-sqrt	RH50, RH98	0.00	299.96	1578.00	3441
ENT Oceania	0.54	63.43	14.33	sqrt-sqrt	RH98	0.00	133.93	571.72	142
ENT Europe	0.68	35.02	14.93	sqrt-sqrt	RH70, RH98	11.46	208.39	577.72	417
ENT North America	0.69	65.47	16.22	sqrt-sqrt	RH70, RH98	0.00	157.47	1768.70	1391
GSW Oceania	0.9	58.60	11.41	sqrt-sqrt	RH50, RH80, RH98	9.15	54.83	192.56	65

low RH metrics, and constrained) allowed suites of models to be considered that had advantages beyond the accuracy of model fits. While allowing all predictors increased model performance in some strata, e.g., GSW Oceania, EBT Asia (Fig. 8, Table 3), forcing RH98 into models may have yielded higher sensitivity to biomass in low canopy covers or emergent trees, while removing low (<RH50) predictors yielded models that are theoretically more transferable to on-orbit GEDI data (Hancock et al., 2019). Constrained sets of predictors rarely impacted the accuracy of models (Fig. 9). In some strata (EBT Asia, GSW Oceania, ENT Oceania), applying both constraints increased the %RMSE more than 5% (Table 4), and in other strata (DBT Europe, DBT North America) there was an increase in mean bias by more than 10 Mg/ha. However, for most strata all four scenarios yielded similar results. The same was true when fitting models at the more broadly stratified continent, PFT, or global level, although certain models in some strata performed more poorly when removing low RH metrics (e.g., Oceania and Asia, Fig. SI 3). We therefore selected the best performing model in the most constrained scenario, with forced inclusion of RH98 and without any RH predictor lower than RH50.

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.

Table 3
The model for each broad stratum (PFT or regionally aggregated) applied in the GEDI04_A product when a more specified model (Table 2) is not available. R^2(Rsq), %RMSE, and Mean Rresidual Error (MRE) were all calculated from geographic cross validation. MRE was the absolute mean binned residual error, expressed in Mg/ha.

Strata	R^2	% RMSE	MRE (Mg/ha)	Transform	Predictors
DBT	0.58	49.21	21.62	sqrt-sqrt	RH60, RH98
EBT	0.64	52.14	17.2	sqrt-sqrt	RH50, RH98
ENT	0.70	59.26	19.69	sqrt-sqrt	RH60, RH98
GSW	0.86	55.39	8.85	sqrt-sqrt	RH98
North America	0.67	55.72	18.76	sqrt-sqrt	RH50, RH98
South America	0.65	43	12.38	sqrt-sqrt	RH60*RH70, RH98
Oceania	0.62	69.75	73.73	sqrt-sqrt	RH98
Asia	0.46	98.83	142.98	log-log	RH50, RH98
Europe	0.64	40.78	13.78	sqrt-sqrt	RH98
Africa	0.71	76.12	11.56	sqrt-sqrt	RH98

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 R^2). 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 R^2 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 R^2, 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

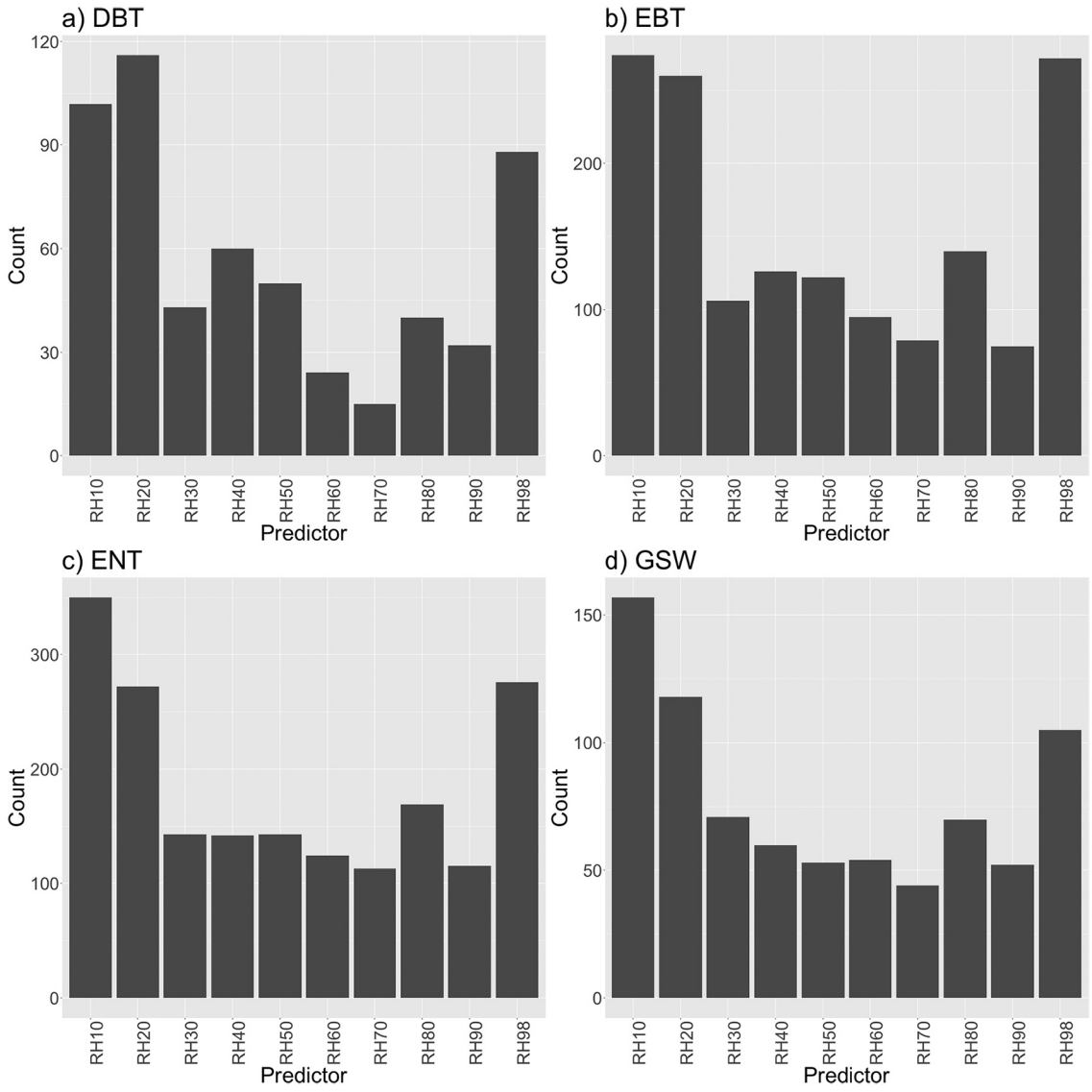

Fig. 5. The number of times predictor variables werer included in the 20 top ranked models per fit stratum, aggregated by PFT, when allowing all variables as predictors. Each variable in an interaction term (e.g. RH50 x RH98) was counted separately.

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 geolocation uncertainties, and the increase in AGBD variance with decreasing plot size (Chave et al., 2004; Labriere et al., 2018; Mauya et al., 2015; Réjou-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

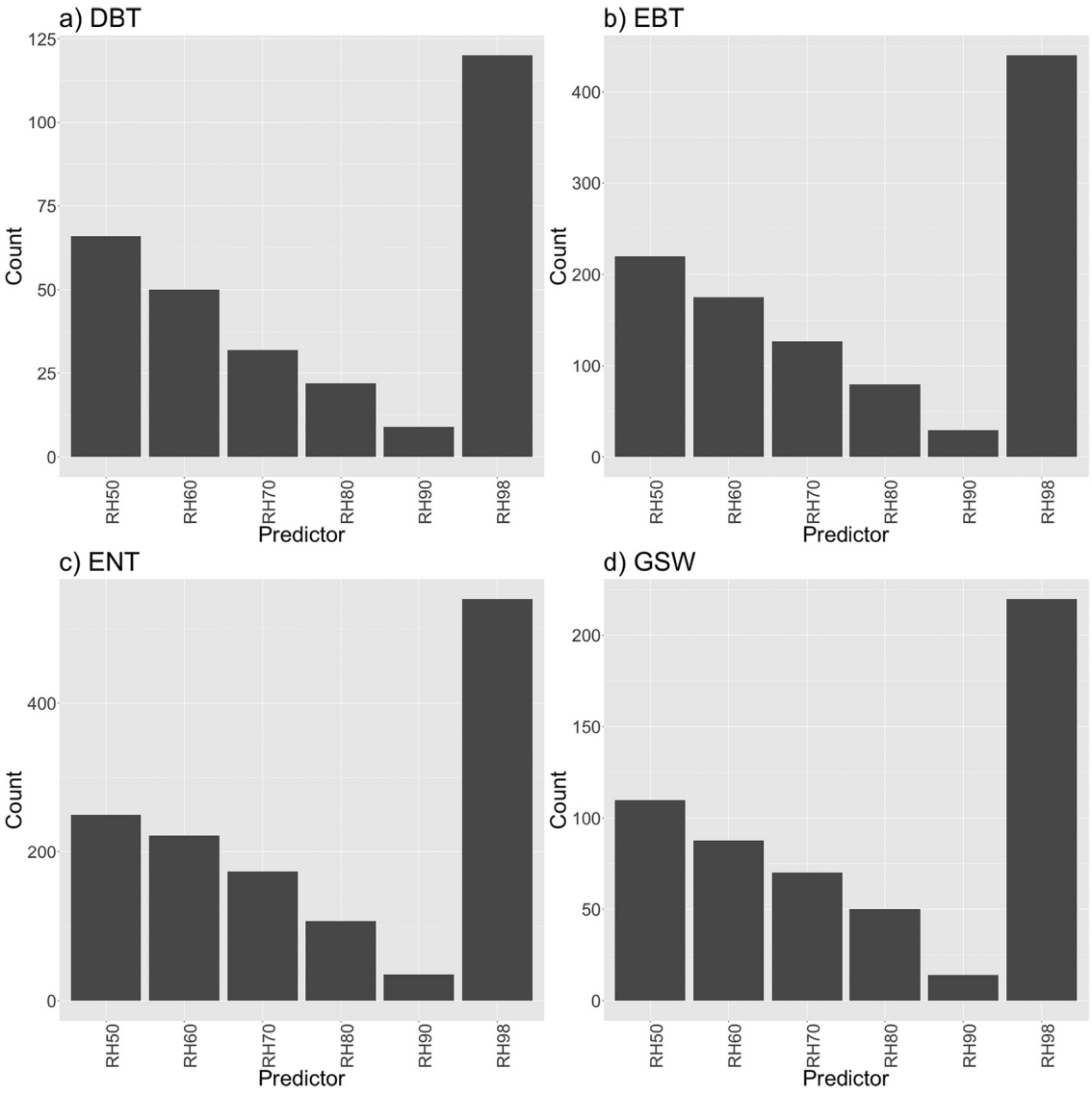

Fig. 6. The number of times predictor variables were included in the top 20 models per fit stratum, aggregated by PFT, when forcing RH98 and removing low RH metrics as predictors (constrained models). Each variable in an interaction term (e.g. RH50 x RH98) was counted separately.

stratum includes an exceptionally wide range of AGBD, and is 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

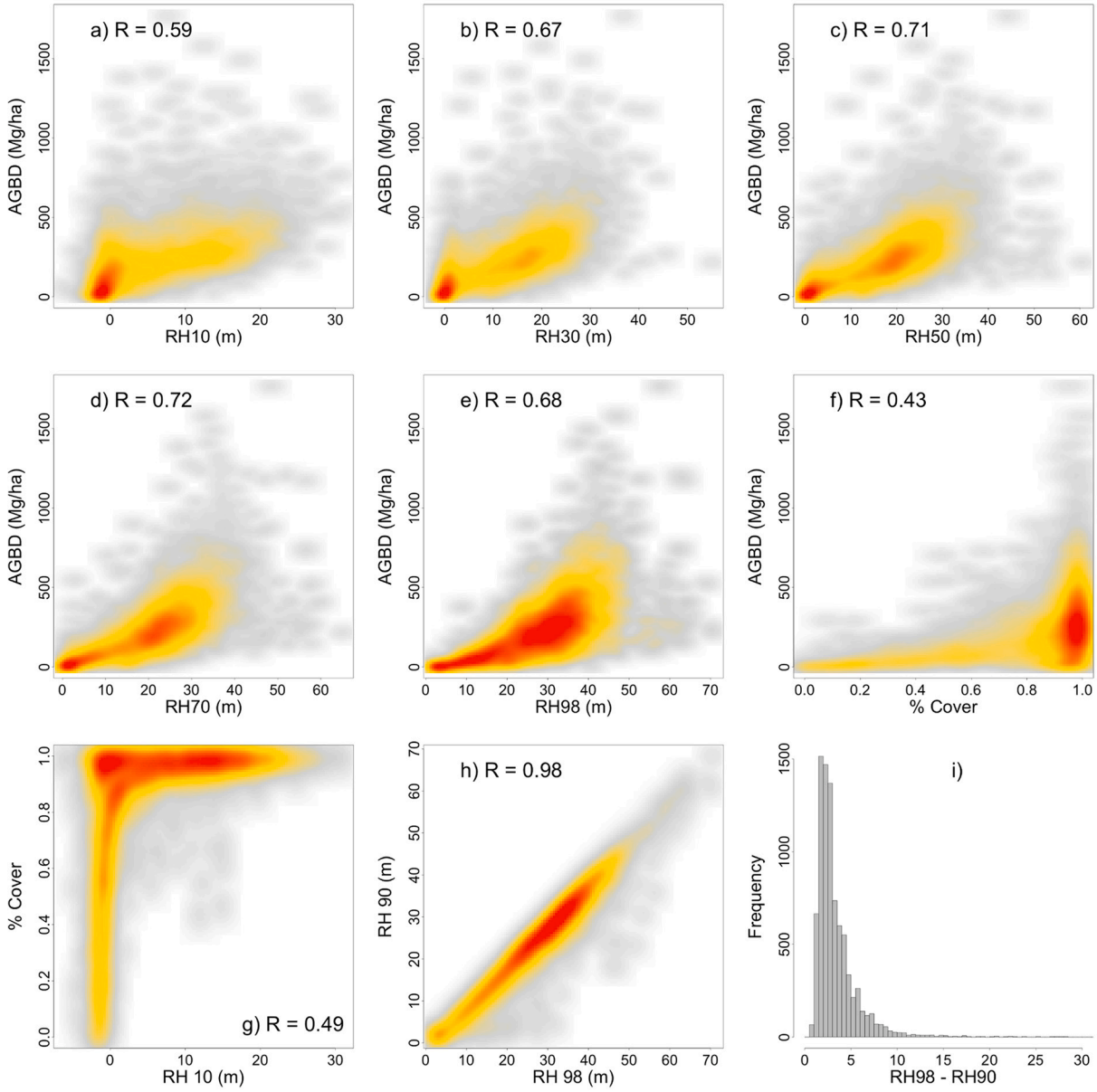

Fig. 7. Relationships between RH metrics (predictors), AGBD, and cover, coloured by density of data points (grey = few, red = most). All RH metrics were correlated with AGBD (a-e), and RH10 had the most variance in high canopy cover forests (g). While RH98 and RH90 were highly correlated (h), in many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

GEDI04_A models included RH98.

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. 7g). 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.

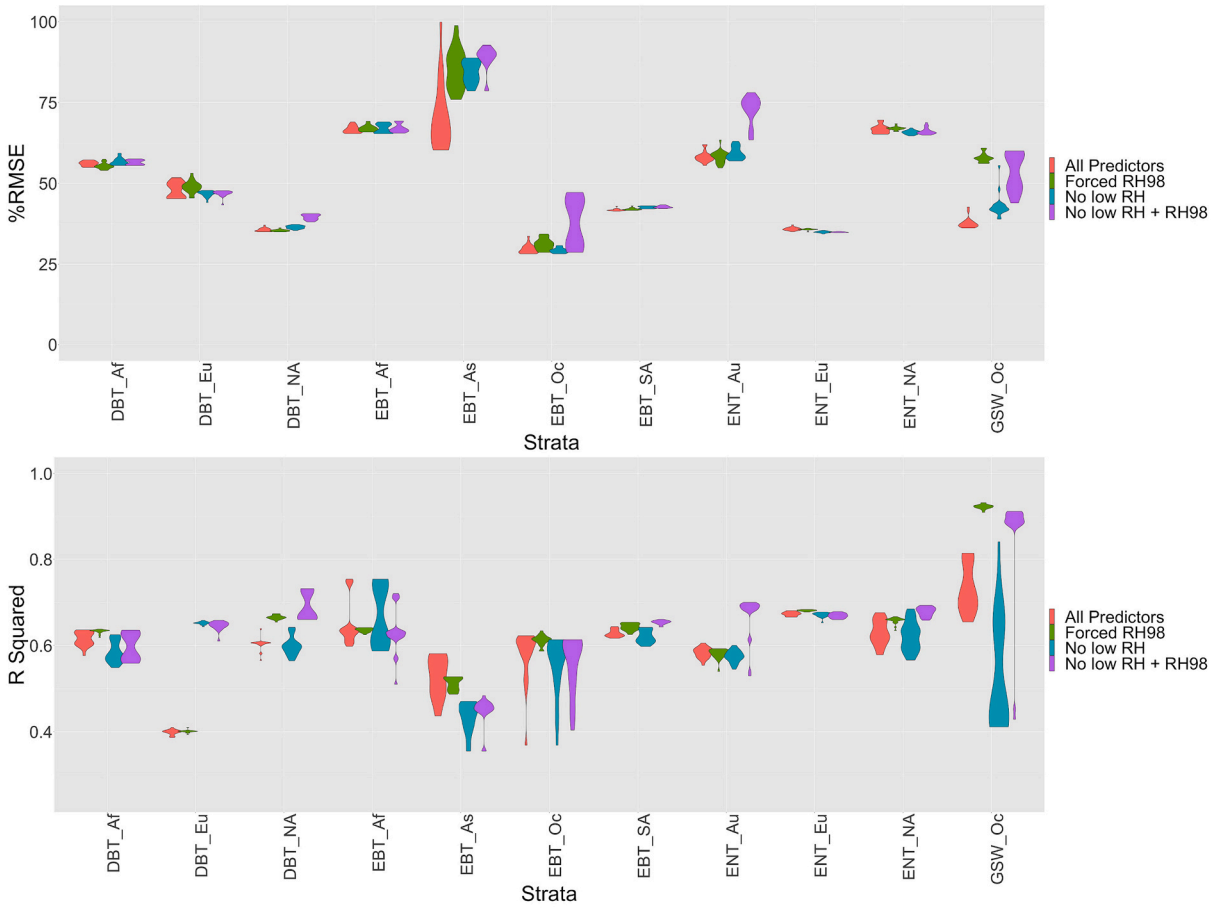

Fig. 8. Violin plots of the distributions of R^2 and %RMSE from four different variable selection scenarios; allowing any variables to be selected, forcing maximum height (RH98) as a predictor, removing low RH metrics (>RH50), and enforcing RH98 while removing low RH metrics. Metrics were assessed by geographic cross validation for models stratified by PFT and geographic region. Violin plots show the top performing 20 models per geographic stratum in each scenario.

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 even 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 to 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

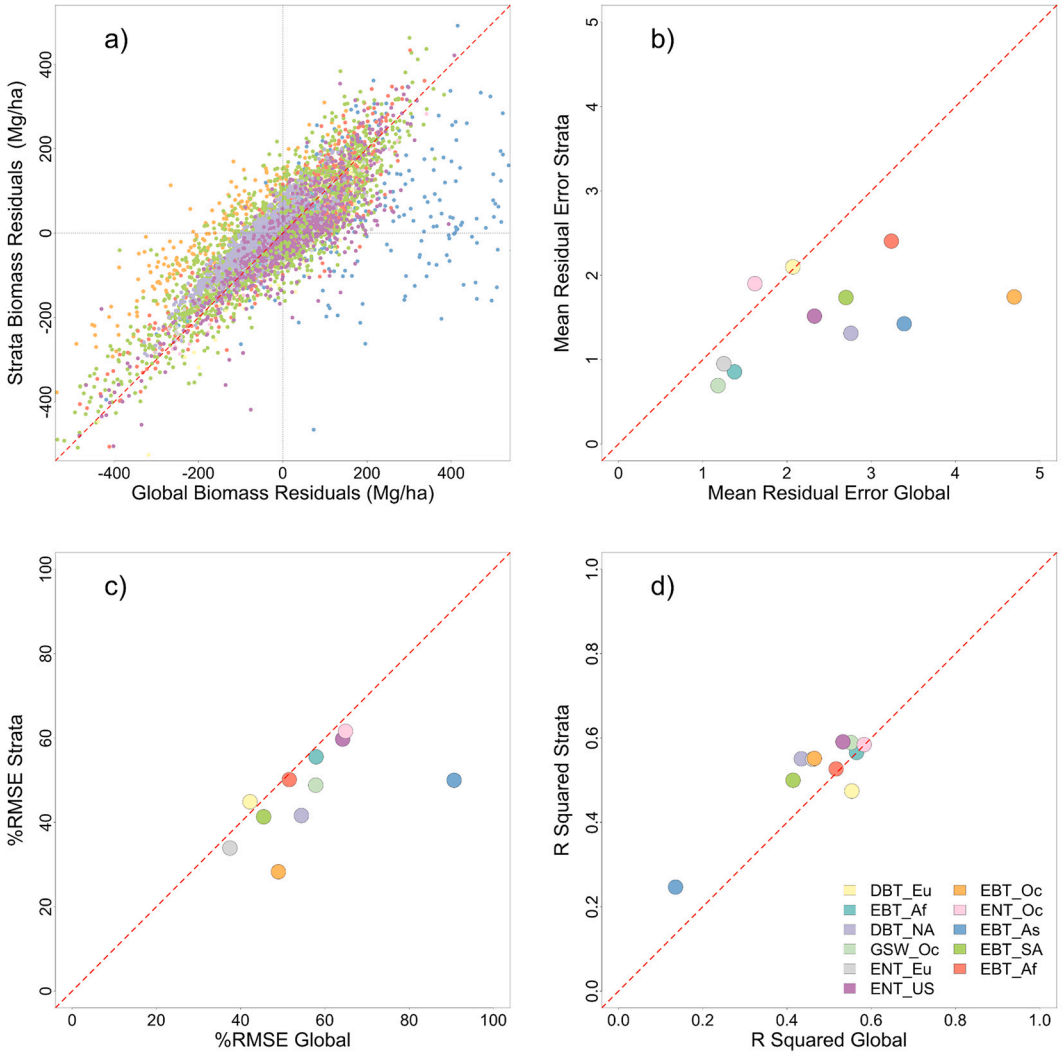

Fig. 9. Stratified models (PFT by geographic region) were equal or better than a single global model applied to each strata in terms of mean residual error (Mg/ha) (b), %RMSE (c) and R^2 (d), with the exception of DBT_Eu.

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.

	Unconstrained		Constrained		Difference (U-C)	
Strata	% RMSE	MRE	% RMSE	MRE	% RMSE	MRE
DBT Africa	54.97	5.97	56.88	7.81	-1.91	-1.84
DBT Europe	48.86	3.84	47.27	21.52	1.59	**-17.68**
DBT North America	35.05	7.60	38.08	22.81	-3.03	**-15.21**
EBT Africa	68.02	11.31	66.89	15.32	1.13	-4.01
EBT Asia	60.31	75.75	78.94	121.15	**-18.63**	**-45.4**
EBT Oceania	28.66	8.17	28.66	8.17	0	0
EBT South America	41.66	5.56	42.20	10.40	-0.54	-4.84
ENT Oceania	58.06	11.21	63.43	14.33	**-5.37**	-3.12
ENT Europe	35.72	11.77	35.02	14.93	0.7	-3.16
ENT North America	69.46	9.77	66.44	16.71	3.02	-6.94
GSW Oceania	37.45	2.62	44.53	8.70	**-7.08**	-6.08

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

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 Armston (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

Armston, Kellner, Hancock, and Dubayah were supported by NASA Contract #NNL 15AA03C 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 80HQTR18T0016, and 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

under cooperative agreement by Battelle. We also thank the National Science and Engineering Research Council of Canada (NSERC), Discovery Grant Program (PI Sanchez-Azofeifa). We also thank the Spanish institutions and programs Instituto Geográfico Nacional, Organismo Autónomo de Parques Nacionales and Inventario Forestal Nacional for supporting this science with open data. The Council for Scientific and Industrial Research (CSIR) project "National Woody Vegetation Monitoring System for Ecosystem and Value-added Services" contributed to the collection of South African ALS and field data. We also thank the Sabie Sand Wildtuin, South African National Parks (SANPARKS), the Wits Rural Knowledge Hub and the Bushbuckridge Municipality in South Africa, for support in the South African field data collection. Additional Australian data were collected as part of the SMAPEx project funded by an Australian Research Council Discovery Project (DP0984586). We thank Shell Gabon and the Smithsonian Conservation Biology Institute for funding the Rabi plot in Gabon, which is contribution No. 204 of the Gabon Biodiversity Program. We also acknowledge funding in French Guiana from CNES and "Investissement d'Avenir" grants managed by Agence Nationale de la Recherche (CEBA, ref. ANR-10-LABX-25-01). We thank the Project LIFE+ ForBioSensing PL "Comprehensive monitoring of stand dynamics in Białowieża Forest supported with remote sensing techniques" co-funded by Life Plus (contract number LIFE13 ENV/PL/000048) and Poland's National Fund for Environmental Protection and Water Management (contract number 485/2014/WN10/OP-NM-LF/D) for funding the collection of the Polish data, and Rafał Sadkowski for helping with data preparation from the ForBioSensing project. We also thank The Silva Tarouca Research Institute (Czech Republic) for collecting and providing field reference data under an INTER-ACTION project (LTAUSA18200). We also thank the former NERC Airborne Research Facility for their support with airborne data collection, and funding for airborne Lidar data provided by the Australian Department of Agriculture, Fisheries, and Forestry (DAFF). We also thank the Norwegian Agency for Development Cooperation (Norad), although the views expressed in this publication do not necessarily reflect the views of Norad. We also acknowledge DfID and UK Natural Environment Research Council (NE/P004806/1) for collection of field data. The Tanzanian field work for this study was carried out as part of the project "Enhancing the measuring, reporting and verification (MRV) of forests in Tanzania through the application of advanced remote sensing techniques", funded by the Royal Norwegian Embassy in Tanzania as part of the Norwegian International Climate and Forest Initiative. Finally, data from RAINFOR plots were supported by the Moore Foundation, and SERNANP (Peru) granted research permissions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rse.2021.112845.

References

Andersen, H.-E., Reutebuch, S.E., McGaughey, R.J., 2006. A rigorous assessment of tree height measurements obtained using airborne lidar and conventional field methods. Can. J. Remote. Sens. https://doi.org/10.5589/m06-030.

Andersen, H.-E., Strunk, J., Temesgen, H., Atwood, D., Winterberger, K., 2011. Using multilevel remote sensing and ground data to estimate forest biomass resources in remote regions: a case study in the boreal forests of interior Alaska. Can. J. Remote. Sens. 37 (6), 1–16.

Asner, G.P., Mascaro, J., 2014. Mapping tropical forest carbon: calibrating plot estimates to a simple LiDAR metric. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2013.09.023.

Auscover, T., 2016. Biomass Plot Library-National collation of tree and shrub inventory data, allometric model predictions of above and below-ground biomass, Australia.

Avitabile, V., Herold, M., Heuvelink, G.B.M., Lewis, S.L., Phillips, O.L., Asner, G.P., Armston, J., Ashton, P.S., Banin, L., Bayol, N., Berry, N.J., Boeckx, P., de Jong, B.H. J., DeVries, B., Girardin, C.A.J., Kearsley, E., Lindsell, J.A., Lopez-Gonzalez, G., Lucas, R., Malhi, Y., Morel, A., Mitchard, E.T.A., Nagy, L., Qie, L., Quinones, M.J., Ryan, C.M., Ferry, S.J.W., Sunderland, T., Laurin, G.V., Gatti, R.C., Valentini, R., Verbeeck, H., Wijaya, A., Willcock, S., 2016. An integrated pan-tropical biomass map using multiple reference datasets. Glob. Chang. Biol. 22, 1406–1420.

Baccini, A., Laporte, N., Goetz, S.J., Sun, M., Dong, H., 2008. A first map of tropical Africa's above-ground biomass derived from satellite imagery. Environ. Res. Lett. https://doi.org/10.1088/1748-9326/3/4/045011.

Baccini, A., Goetz, S.J., Walker, W.S., Laporte, N.T., Sun, M., Sulla-Menashe, D., Hackler, J., Beck, P.S.A., Dubayah, R., Friedl, M.A., Samanta, S., Houghton, R.A., 2012. Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps. Nat. Clim. Chang. https://doi.org/10.1038/nclimate1354.

Baccini, A., Walker, W., Carvalho, L., Farina, M., Sulla-Menashe, D., Houghton, R.A., 2017. Tropical forests are a net carbon source based on aboveground measurements of gain and loss. Science 358, 230–234.

Banin, L., Lewis, S.L., Lopez-Gonzalez, G., Baker, T.R., Quesada, C.A., Chao, K.-J., Burslem, D.F.R.P., Nilus, R., Abu Salim, K., Keeling, H.C., Tan, S., Davies, S.J., Monteagudo Mendoza, A., Vásquez, R., Lloyd, J., Neill, D.A., Pitman, N., Phillips, O. L., 2014. Tropical forest wood production: a cross-continental comparison. J. Ecol. 102, 1025–1037.

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

Carlucci, M.B., Seger, G.D.S., Sheil, D., Amaral, I.L., Chuyong, G.B., Ferreira, L.V., Galatti, U., Hurtado, J., Kenfack, D., Leal, D.C., Lewis, S.L., Lovett, J.C., Marshall, A. R., Martin, E., Mugerwa, B., Munishi, P., Oliveira, Á.C.A., Razafimahaimodison, J.C., Rovero, F., Sainge, M.N., Thomas, D., Pillar, V.D., Duarte, L.D.S., 2017. Phylogenetic composition and structure of tree communities shed light on historical processes influencing tropical rainforest diversity. Ecography. https://doi.org/10.1111/ecog.02104.

Chave, J., Condit, R., Aguilar, S., Hernandez, A., Lao, S., Perez, R., 2004. Error propagation and scaling for tropical forest biomass estimates. Philos. Trans. R. Soc. Lond. B Biol. Sci. 359, 409–420.

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. Glob. Chang. Biol. 20, 3177–3190.

Clark, D.B., Kellner, J.R., 2012. Tropical forest biomass estimation and the fallacy of misplaced concreteness. J. Veg. Sci. 23, 1191–1196.

Coomes, D.A., Dalponte, M., Jucker, T., Asner, G.P., Banin, L.F., David, F.R., Lewis, S.L., Nilus, R., Phillips, O.L., Phua, M.-H., Qie, L., 2017. Area-based vs tree-centric approaches to mapping forest carbon in Southeast Asian forests from airborne laser scanning data. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2017.03.017.

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

Corbera, E., Schroeder, H., 2011. Governing and implementing REDD+. Environ. Sci. Pol. 14, 89–99.

Corlett, R.T., Primack, R.B., 2006. Tropical rainforests and the need for cross-continental comparisons. Trends Ecol. Evol. 21, 104–110.

Corte, A.P.D., Souza, D.V., Rex, F.E., Sanquetta, C.R., Mohan, M., Silva, C.A., Zambrano, A.M.A., Prata, G., de Almeida, D.R.A., Trautenmüller, J.W., Klauberg, C., de Moraes, A., Sanquetta, M.N., Wilkinson, B., Broadbent, E.N., 2020. Forest inventory with high-density UAV-Lidar: machine learning approaches for predicting individual tree attributes. Comput. Electron. Agric. https://doi.org/10.1016/j.compag.2020.105815.

Diaz, S., Cabido, M., 1997. Plant functional types and ecosystem function in relation to global change. J. Veg. Sci. https://doi.org/10.1111/j.1654-1103.1997.tb00842.x.

Disney, M., Burt, A., Wilkes, P., Armston, J., Duncanson, L., 2020. New 3D measurements of large redwood trees for biomass and structure. Sci. Rep. 10, 16721.

Drake, J.B., Dubayah, R.O., Clark, D.B., Knox, R.G., Blair, J.B., Hofton, M.A., Chazdon, R. L., Weishampel, J.F., Prince, S., 2002a. Estimation of tropical forest structural characteristics using large-footprint lidar. Remote Sens. Environ. 79, 305–319.

Drake, J.B., Dubayah, R.O., Clark, D.B., Knox, R.G., Bryan Blair, J., Hofton, M.A., Chazdon, R.L., Weishampel, J.F., Prince, S., 2002b. Estimation of tropical forest structural characteristics using large-footprint lidar. Remote Sens. Environ. https://doi.org/10.1016/s0034-4257(01)00281-4.

Drake, J.B., Knox, R.G., Dubayah, R.O., Clark, D.B., Condit, R., Bryan Blair, J., Hofton, M., 2003. Above-ground biomass estimation in closed canopy Neotropical forests using lidar remote sensing: factors affecting the generality of relationships. Glob. Ecol. Biogeogr. https://doi.org/10.1046/j.1466-822x.2003.00010.x.

Dubayah, R., Blair, J.B., Goetz, S., Fatoyinbo, L., Hansen, M., Healey, S., Hofton, M., Hurtt, G., Kellner, J., Luthcke, S., Armston, J., Tang, H., Duncanson, L., Hancock, S., Jantz, P., Marselis, S., Patterson, P.L., Qi, W., Silva, C., 2020. The global ecosystem dynamics investigation: high-resolution laser ranging of the Earth's forests and topography. Sci. Remote Sens. https://doi.org/10.1016/j.srs.2020.100002.

Duncanson, L.I., Niemann, K.O., Wulder, M.A., 2010. Estimating forest canopy height and terrain relief from GLAS waveform metrics. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2009.08.018.

Duncanson, L., Neuenschwander, A., Hancock, S., Thomas, N., Fatoyinbo, T., Simard, M., Silva, C.A., Armston, J., Luthcke, S.B., Hofton, M., Kellner, J.R., Dubayah, R., 2020. Biomass estimation from simulated GEDI, ICESat-2 and NISAR across environmental gradients in Sonoma County, California. Remote Sens. Environ. 242, 111779. https://doi.org/10.1016/j.rse.2020.111779.

Duncanson, L., Armston, J., Disney, M., Avitabile, V., Barbier, N., Calders, K., Carter, S., Chave, J., Herold, M., MacBean, N., et al., 2021. Aboveground Woody Biomass Product Validation Good Practices Protocol..

Ene, L.T., Næsset, E., Gobakken, T., Mauya, E.W., Bollandsås, O.M., Gregoire, T.G., Ståhl, G., Zahabu, E., 2016. Large-scale estimation of aboveground biomass in miombo woodlands using airborne laser scanning and national forest inventory data. Remote Sens. Environ. 186, 626–636.

Enquist, B.J., West, G.B., Brown, J.H., 2009. Extensions and evaluations of a general quantitative theory of forest structure and dynamics. Proc. Natl. Acad. Sci. U. S. A. 106, 7046–7051.

Fatoyinbo, T., Armston, J., Simard, M., Saatchi, S., Denbina, M., Lavalle, M., Hofton, M., Tang, H., Marselis, S., Pinto, N., Hancock, S., Hawkins, B., Duncanson, L., Blair, B., Hansen, C., Lou, Y., Dubayah, R., Hensley, S., Silva, C., Poulsen, J.R., Labrière, N., Barbier, N., Jeffery, K., Kenfack, D., Herve, M., Bissiengou, P., Alonso, A., Moussavou, G., White, L.T.J., Lewis, S., Hibbard, K., 2021. The NASA AfriSAR campaign: airborne SAR and lidar measurements of tropical forest structure and biomass in support of current and future space missions. Remote Sens. Environ. 264, 112533.

Feldpausch, T.R., Lloyd, J., Lewis, S.L., Brienen, R.J.W., Gloor, M., Monteagudo Mendoza, A., Lopez-Gonzalez, G., Banin, L., Abu Salim, K., Affum-Baffoe, K., Alexiades, M., Almeida, S., Amaral, I., Andrade, A., Aragao, L.E.O., Araujo Murakami, A., Arets, E.J.M.M., Arroyo, L., Aymard, C.G.A., Baker, T.R., Banki, O.S., Berry, N.J., Cardozo, N., Chave, J., Comiskey, J.A., Alvarez, E., de Oliveira, A., Di Fiore, A., Djagbletey, G., Domingues, T.F., Erwin, T.L., Fearnside, P.M., Franca, M.B., Freitas, M.A., Higuchi, N., Honorio, C.E., Iida, Y., Jimenez, E., Kassim, A.R., Killeen, T.J., Laurance, W.F., Lovett, J.C., Malhi, Y., Marimon, B.S., Marimon-Junior, B.H., Lenza, E., Marshall, A.R., Mendoza, C., Metcalfe, D.J., Mitchard, E.T.A., Neill, D.A., Nelson, B.W., Nilus, R., Nogueira, E.M., Parada, A., Peh, K.S.-H., Pena Cruz, A., Penuela, M.C., Pitman, N.C.A., Prieto, A., Quesada, C.A., Ramırez, F., Ramırez-Angulo, H., Reitsma, J.M., Rudas, A., Saiz, G., Salomao, R.P., Schwarz, M., Silva, N., Silva-Espejo, J.E., Silveira, M., Sonke, B., Stropp, J., Taedoumg, H.E., Tan, S., ter Steege, H., Terborgh, J., Torello-Raventos, M., van der Heijden, G.M.F., Vasquez, R., Vilanova, E., Vos, V.A., White, L., Willcock, S., Woell, H., Phillips, O.L., 2012. Tree height integrated into pantropical forest biomass estimates. Biogeosciences 9, 3381–3403.

Ferraz, A., Saatchi, S., Mallet, C., Jacquemoud, S., Gil, G., Silva, C.A., Soares, P., Tomé, M., Pereira, L., 2016. Airborne lidar estimation of aboveground forest biomass in the absence of field inventory. Remote Sens. 8, 653.

Ferraz, A., Saatchi, S., Xu, L., Hagen, S., Chave, J., Yu, Y., Meyer, V., Garcia, M., Silva, C., Roswintiart, O., Samboko, A., Sist, P., Walker, S., Pearson, T.R.H., Wijaya, A., Sullivan, F.B., Rutishauser, E., Hoekman, D., Ganguly, S., 2018. Carbon storage potential in degraded forests of Kalimantan, Indonesia. Environ. Res. Lett. 13, 095001.

Ferry, B., Morneau, F., Bontemps, J.-D., Blanc, L., Freycon, V., 2010. Higher treefall rates on slopes and waterlogged soils result in lower stand biomass and productivity in a tropical rain forest. J. Ecol. https://doi.org/10.1111/j.1365-2745.2009.01604.x.

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 Sens. Environ. https://doi.org/10.1016/s0034-4257(03)00039-7.

Friedl, M.A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A., Huang, X., 2010. MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2009.08.016.

Friedlingstein, P., Jones, M.W., O'Sullivan, M., Andrew, R.M., Hauck, J., Peters, G.P., Peters, W., Pongratz, J., Sitch, S., Le Quéré, C., Bakker, D.C.E., Canadell, J.G., Ciais, P., Jackson, R.B., Anthoni, P., Barbero, L., Bastos, A., Bastrikov, V., Becker, M., Bopp, L., Buitenhuis, E., Chandra, N., Chevallier, F., Chini, L.P., Currie, K.I., Feely, R.A., Gehlen, M., Gilfillan, D., Gkritzalis, T., Goll, D.S., Gruber, N., Gutekunst, S., Harris, I., Haverd, V., Houghton, R.A., Hurtt, G., Ilyina, T., Jain, A.K., Joetzjer, E., Kaplan, J.O., Kato, E., Klein Goldewijk, K., Korsbakken, J.I., Landschützer, P., Lauvset, S.K., Lefèvre, N., Lenton, A., Lienert, S., Lombardozzi, D., Marland, G., McGuire, P.C., Melton, J.R., Metzl, N., Munro, D.R., Nabel, J.E.M.S., Nakaoka, S.-I., Neill, C., Omar, A.M., Ono, T., Peregon, A., Pierrot, D., Poulter, B., Rehder, G., Resplandy, L., Robertson, E., Rödenbeck, C., Séférian, R., Schwinger, J., Smith, N., Tans, P.P., Tian, H., Tilbrook, B., Tubiello, F.N., Van der Werf, G.R., Wiltshire, A.J., Zaehle, S., 2019. Global carbon budget 2019. Earth Syst. Sci. Data 11, 1783–1838.

Friis, I., Balslev, H., 2005. Plant diversity and complexity patterns: local, regional, and global dimensions. Biologiske Skrifter 55.

Gibbs, H.K., Brown, S., Niles, J.O., Foley, J.A., 2007. Monitoring and estimating tropical forest carbon stocks: making REDD a reality. Environ. Res. Lett. https://doi.org/10.1088/1748-9326/2/4/045023.

Gleason, C.J., Im, J., 2012. Forest biomass estimation from airborne LiDAR data using machine learning approaches. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2012.07.006.

Goetz, S.J., Hansen, M., Houghton, R.A., Walker, W., Laporte, N., Busch, J., 2015. Measurement and monitoring needs, capabilities and potential for addressing reduced emissions from deforestation and forest degradation under REDD+. Environ. Res. Lett. 10, 123001.

Hancock, S., Disney, M., Muller, J.-P., Lewis, P., Foster, M., 2011. A threshold insensitive method for locating the forest canopy top with waveform lidar. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2011.07.012.

Hancock, S., Armston, J., Li, Z., Gaulton, R., Lewis, P., Disney, M., Mark Danson, F., Strahler, A., Schaaf, C., Anderson, K., Gaston, K.J., 2015. Waveform lidar over vegetation: an evaluation of inversion methods for estimating return energy. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2015.04.013.

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 Space Sci. 6, 294–310.

Hansen, E.H., Bollandsås, Gobakken Terje, Zahabu, E., Næsset, E., 2015. Modeling aboveground biomass in dense tropical submontane rainforest using airborne laser scanner data. Remote Sens. 7, 788–807.

Hansen, E., Ene, L., Mauya, E., Patočka, Z., Mikita, T., Gobakken, T., Næsset, E., 2017. Comparing empirical and semi-empirical approaches to forest biomass modelling in different biomes using airborne laser scanner data. Forests. https://doi.org/10.3390/f8050170.

Healey, S.P., Yang, Z., Gorelick, N., Ilyushchenko, S., 2020. Highly local model calibration with a new GEDI LiDAR asset on Google Earth Engine reduces landsat forest height signal saturation. Remote Sens. 12 (17), 2840.

Hernando, A., Puerto, L., Mola-Yudego, B., Manzanera, J.A., García-Abril, A., Maltamo, M., Valbuena, R., 2019. Estimation of forest biomass components using airborne LiDAR and multispectral sensors. iForest - Biogeosci. Forestry. https://doi.org/10.3832/ifor2735-012.

Houghton, R.A., House, J.I., Pongratz, J., Werf, G.R. van der, DeFries, R.S., Hansen, M.C., Le Quéré, C., Ramankutty, N., 2012. Carbon emissions from land use and land-cover change. Biogeosciences 9, 5125–5142.

Huang, W., Sun, G., Dubayah, R., Cook, B., Montesano, P., Ni, W., Zhang, Z., 2013. Mapping biomass change after forest disturbance: Applying LiDAR footprint-derived models at key map scales. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2013.03.017.

Hudak, A.T., Bright, B.C., Pokswinski, S.M., Louise Loudermilk, E., O'Brien, J.J., Hornsby, B.S., Klauberg, C., Silva, C.A., 2016. Mapping forest structure and composition from low-density LiDAR for informed forest, fuel, and fire management at Eglin Air Force Base, Florida, USA. Can. J. Remote. Sens. https://doi.org/10.1080/07038992.2016.1217482.

Huete, A., Liu, H.Q., Batchily, K., van Leeuwen, W., 1997. A comparison of vegetation indices over a global set of TM images for EOS-MODIS. Remote Sens. Environ. 59, 440–451.

Hurtt, G.C., Pacala, S.W., Moorcroft, P.R., Caspersen, J., Shevliakova, E., Houghton, R.A., Moore, B., 2002. Projecting the Future of the U.S. Carbon Sink. In: Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.012249999.

Jucker, T., Caspersen, J., Chave, J., Antin, C., Barbier, N., Bongers, F., Dalponte, M., van Ewijk, K.Y., Forrester, D.I., Haeni, M., Higgins, S.I., Holdaway, R.J., Iida, Y., Lorimer, C., Marshall, P.L., Momo, S., Moncrieff, G.R., Ploton, P., Poorter, L., Rahman, K.A., Schlund, M., Sonké, B., Sterck, F.J., Trugman, A.T., Usoltsev, V.A., Vanderwel, M.C., Waldner, P., Wedeux, B.M.M., Wirth, C., Wöll, H., Woods, M., Xiang, W., Zimmermann, N.E., Coomes, D.A., 2017. Allometric equations for integrating remote sensing imagery into forest monitoring programmes. Glob. Chang. Biol. 23, 177–190.

Kearsley, E., de Haulleville, T., Hufkens, K., Kidimbu, A., Toirambe, B., Baert, G., Huygens, D., Kebede, Y., Defourny, P., Bogaert, J., Beeckman, H., Steppe, K., Boeckx, P., Verbeeck, H., 2013. Conventional tree height-diameter relationships significantly overestimate aboveground carbon stocks in the Central Congo Basin. Nat. Commun. 4, 2269.

Kellner, J.R., Armston, J., Duncanson, L., 2021. Algorithm Theoretical Basis Document for GEDI Footprint Aboveground Biomass Density.

Koetz, B., Morsdorf, F., Sun, G., Ranson, K.J., Itten, K., Allgower, B., 2006. Inversion of a lidar waveform model for forest biophysical parameter estimation. IEEE Geosci. Remote Sens. Lett. https://doi.org/10.1109/lgrs.2005.856706.

Labriere, N., Tao, S., Chave, J., Scipal, K., Le Toan, T., 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 J. Select. Top. Appl. Earth Observ. Remote Sens. https://doi.org/10.1109/jstars.2018.2851606.

Lang, N., Schindler, K., Wegner, J.D., 2019. Country-wide high-resolution vegetation height mapping with Sentinel-2. Remote Sens. Environ. 233, 111347.

Laurin, G.V., Chen, Q., Lindsell, J.A., Coomes, D.A., Del Frate, F., Guerriero, L., Pirotti, F., Valentini, R., 2014. Above ground biomass estimation in an African tropical forest with lidar and hyperspectral data. ISPRS J. Photogramm. Remote Sens. https://doi.org/10.1016/j.isprsjprs.2014.01.001.

Le Quéré, C., Andrew, R., Canadell, J.G., Sitch, S., Korsbakken, J.I., Peters, G.P., Manning, A.C., Boden, T.A., Tans, P.P., Houghton, R.A., Keeling, R.F., Alin, S., Andrews, O.D., Anthoni, P., Barbero, L., Bopp, L., Chevallier, F., Chini, L.P., Ciais, P., Currie, K., Delire, C., Doney, S.C., Friedlingstein, P., Gkritzalis, T., Harris, I., Hauck, J., Haverd, V., Hoppema, M., Klein Goldewijk, K., Jain, A.K., Kato, E., Körtzinger, A., Landschutzer, P., Lefevre, N., Lenton, A., Lienert, S., Lombardozzi, D., Melton, J.R., Metzl, N., Millero, F., Monteiro, P.M.S., Munro, D.R., Nabel, J.E.M.S., Nakaoka, S.-I., O'Brien, K., Olsen, A., Omar, A., Pierrot, D., Ono, T., Poulter, B., Rödenbeck, C., Salisbury, J., Schuster, U., Schwinger, J., Seferian, R., Skjelvan, I., Stocker, B.D., Sutton, A.J., Takahashi, T., Tian, H., Tilbrook, B., Van Der Laan-Luijkx, I.T., Van Der Werf, G.R., Viovy, N., Walker, A.P., Wiltshire, A.J., Zaehle, S., 2016. Global Carbon Budget 2016. Earth Syst. Sci. Data 8, 605–649.

Le Toan, T., Beaudoin, A., Riom, J., Guyon, D., 1992. Relating forest biomass to SAR data. IEEE Trans. Geosci. Remote Sens. 30, 403–411.

Lefsky, M.A., Cohen, W.B., Acker, S.A., Parker, G.G., Spies, T.A., Harding, D., 1999. Lidar remote sensing of the canopy structure and biophysical properties of douglas-fir western hemlock forests. Remote Sens. Environ. https://doi.org/10.1016/s0034-4257(99)00052-8.

Lefsky, M.A., Cohen, W.B., Parker, G.G., Harding, D.J., 2002. Lidar remote sensing for ecosystem studies. BioScience. doi:10.1641/0006-3568(2002)052[0019,lrsfes]2.0.co;2.

Lefsky, M.A., Keller, M., Pang, Y., De Camargo, P.B., Hunter, M.O., 2007. Revised method for forest canopy height estimation from Geoscience Laser Altimeter System waveforms. J. Appl. Remote. Sens. 1, 013537.

Los, S.O., Rosette, J.A.B., Kljun, N., North, P.R.J., Chasmer, L., Suárez, J.C., Hopkinson, C., Hill, R.A., van Gorsel, E., Mahoney, C., Berni, J.A.J., 2012. Vegetation height and cover fraction between 60° S and 60° N from ICESat GLAS data. Geosci. Model Dev. 5, 413–432.

Lucas, R.M., Lee, A.C., Bunting, P.J., 2008. Retrieving forest biomass through integration of CASI and LiDAR data. Int. J. Remote Sens. https://doi.org/10.1080/01431160701736497.

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 Sens. Environ. https://doi.org/10.1016/s0034-4257(97)00133-8.

Mahoney, C., Kljun, N., Los, S.O., Chasmer, L., Hacker, J.M., Hopkinson, C., North, P.R. J., Rosette, J.A.B., van Gorsel, E., 2014. Slope estimation from ICESat/GLAS. Remote Sens. 6, 10051–10069. https://doi.org/10.3390/rs61010051.

Margolis, H.A., Nelson, R.F., Montesano, P.M., Beaudoin, A., Sun, G., Andersen, H.-E., Wulder, M.A., 2015. Combining satellite lidar, airborne lidar, and ground plots to estimate the amount and distribution of aboveground biomass in the boreal forest of North America. Can. J. For. Res. 45, 838–855.

Mauya, E.W., Hansen, E.H., Gobakken, T., Bollandsås, O.M., Malimbwi, R.E., Næsset, E., 2015. Effects of field plot size on prediction accuracy of aboveground biomass in airborne laser scanning-assisted inventories in tropical rain forests of Tanzania. Carbon Balance Manag. https://doi.org/10.1186/s13021-015-0021-x.

Meyer, V., Saatchi, S., Clark, D.B., Keller, M., Grégoire, V., Ferraz, A., Espírito-Santo, F., d'Oliveira, M., Kaki, D., Chave, J., 2018. Canopy area of large trees explains aboveground biomass variations across neotropical forest landscapes. Biogeosciences 15, 3377–3390.

Meyer, V., Saatchi, S., Ferraz, A., Xu, L., Duque, A., García, M., Chave, J., 2019. Forest degradation and biomass loss along the Chocó region of Colombia. Carbon Balance Manag. 14, 2.

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. https://doi.org/10.1029/2009gl040692.

Næsset, E., Gobakken, T., 2008. Estimation of above- and below-ground biomass across regions of the boreal forest zone using airborne laser. Remote Sens. Environ. 112, 3079–3090.

Næsset, E., Bollandsås, O.M., Gobakken, T., Gregoire, T.G., Ståhl, G., 2013. Model-assisted estimation of change in forest biomass over an 11year period in a sample survey supported by airborne LiDAR: A case study with post-stratification to provide "activity data.". Remote Sens. Environ. https://doi.org/10.1016/j.rse.2012.10.008.

Næsset, E., Bollandsås, O.M., Gobakken, T., Solberg, S., McRoberts, R.E., 2015. The effects of field plot size on model-assisted estimation of aboveground biomass change using multitemporal interferometric SAR and airborne laser scanning data. Remote Sens. Environ. 168, 252–264.

Naidoo, L., Mathieu, R., Main, R., Kleynhans, W., Wessels, K., Asner, G., Leblon, B., 2015. Savannah woody structure modelling and mapping using multi-frequency (X-, C- and L-band) Synthetic Aperture Radar data. ISPRS J. Photogramm. Remote Sens. 105, 234–250.

Ni-Meister, W., Lee, S., Strahler, A.H., Woodcock, C.E., Schaaf, C., Yao, T., Jon Ranson, K., Sun, G., Bryan Blair, J., 2010. Assessing general relationships between aboveground biomass and vegetation structure parameters for improved carbon estimate from lidar remote sensing. J. Geophys. Res. Biogeosci. https://doi.org/10.1029/2009jg000936.

North, P.R.J., Rosette, J.A.B., Suárez, J.C., Los, S.O., 2010. A Monte Carlo radiative transfer model of satellite waveform LiDAR. Int. J. Remote Sens. https://doi.org/10.1080/01431160903380664.

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. https://doi.org/10.1088/1748-9326/ab18df.

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. Nat. Commun. 11, 4540.

Potapov, P., Li, X., Hernandez-Serna, A., Tyukavina, A., Hansen, M.C., Kommareddy, A., Pickens, A., Turubanova, S., Tang, H, Silva, C.E., Armston, J., Dubayah, R., Blair, J. B., Hofton, M., 2021. Mapping global forest canopy height through integration of GEDI and Landsat data. Remote Sens. Environ. 253, 112165.

Poulter, B., Ciais, P., Hodson, E., Lischke, H., Maignan, F., Plummer, S., Zimmermann, N. E., 2011. Plant functional type mapping for earth system models. Geosci. Model Dev. https://doi.org/10.5194/gmd-4-993-2011.

Powell, S.L., Cohen, W.B., Healey, S.P., Kennedy, R.E., Moisen, G.G., Pierce, K.B., Ohmann, J.L., 2010. Quantification of live aboveground forest biomass dynamics with Landsat time-series and field inventory data: a comparison of empirical modeling approaches. Remote Sens. Environ. 114, 1053–1068.

Ranson, K.J., Sun, G., 2000. Modeling lidar returns from forest canopies. IEEE Trans. Geosci. Remote Sens. https://doi.org/10.1109/36.885208.

Réjou-Méchain, M., Barbier, N., Couteron, P., Ploton, P., Vincent, G., Herold, M., Mermoz, S., Saatchi, S., Chave, J., de Boissieu, F., Féret, J.-B., Takoudjou, S.M., Pélissier, R., 2019. Upscaling forest biomass from field to satellite measurements: sources of errors and ways to reduce them. Surv. Geophys. https://doi.org/10.1007/s10712-019-09532-0.

Rodríguez-Veiga, P., Quegan, S., Carreiras, J., Persson, H.J., Fransson, J.E.S., Hoscilo, A., Ziółkowski, D., Stereńczak, K., Lohberger, S., Stängel, M., Berninger, A., Siegert, F., Avitabile, V., Herold, M., Mermoz, S., Bouvet, A., Le Toan, T., Carvalhais, N., Santoro, M., Cartus, O., Rauste, Y., Mathieu, R., Asner, G.P., Thiel, C., Pathe, C., Schmullius, C., Seifert, F.M., Tansey, K., Balzter, H., 2019. Forest biomass retrieval approaches from earth observation in different biomes. Int. J. Appl. Earth Obs. Geoinf. 77, 53–68.

Saarela, S., Holm, S., Healey, S., Andersen, H.-E., Petersson, H., Prentius, W., Patterson, P., Næsset, E., Gregoire, T., Ståhl, G., 2018. Generalized hierarchical model-based estimation for aboveground biomass assessment using GEDI and Landsat Data. Remote Sens. https://doi.org/10.3390/rs10111832.

Saarela, S., Holm, S., Healey, S.P., Patterson, P.L., Duncanson, L.I., Armston, J.D., Kellner, J.K., Gobakken, T., Næsset, E., Ekström, M., Ståhl, G., 2021. Effects of multicollinearity in model-based inference predicting aboveground biomass using NASA's GEDI and Landsat missions (research note). Silva Fennica. In press.

Saatchi, S.S., Harris, N.L., Brown, S., Lefsky, M., Mitchard, E.T.A., Salas, W., Zutta, B.R., Buermann, W., Lewis, S.L., Hagen, S., Petrova, S., White, L., Silman, M., Morel, A., 2011. Benchmark map of forest carbon stocks in tropical regions across three continents. Proc. Natl. Acad. Sci. https://doi.org/10.1073/pnas.1019576108.

Silva, C.A., Duncanson, L., Hancock, S., Neuenschwander, A., Thomas, N., Hofton, M., Fatoyinbo, L., Simard, M., Marshak, C.Z., Armston, J., Lutchke, S., Dubayah, R., 2021. Fusing simulated GEDI, ICESat-2 and NISAR data for regional aboveground biomass mapping. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2020.112234.

Simard, M., Pinto, N., Fisher, J.B., Baccini, A., 2011. Mapping forest canopy height globally with spaceborne lidar. J. Geophys. Res. 116 https://doi.org/10.1029/2011jg001708.

Simard, M., Fatoyinbo, L., Smetanka, C., Rivera-Monroy, V.H., Castañeda-Moya, E., Thomas, N., Van der Stocken, T., 2018. Mangrove canopy height globally related to precipitation, temperature and cyclone frequency. Nat. Geosci. 12, 40–45.

Smith, W.B., 2002. Forest inventory and analysis: a national inventory and monitoring program. Environ. Pollut. 116, S233–S242.

Snowdon, P., 1991. A ratio estimator for bias correction in logarithmic regressions. Can. J. For. Res. 21, 720–724.

Sun, G., Ranson, K., Kimes, D., Blair, J., Kovacs, K., 2008. Forest vertical structure from GLAS: An evaluation using LVIS and SRTM data. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2006.09.036.

Sun, G., Jon Ranson, K., Guo, Z., Zhang, Z., Montesano, P., Kimes, D., 2011. Forest biomass mapping from lidar and radar synergies. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2011.03.021.

Swatantran, A., Dubayah, R., Roberts, D., Hofton, M., Bryan Blair, J., 2011. Mapping biomass and stress in the Sierra Nevada using lidar and hyperspectral data fusion. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2010.08.027.

Tang, H., Armston, J., Hancock, S., Marselis, S., Goetz, S., Dubayah, R., 2019. Characterizing global forest canopy cover distribution using spaceborne lidar. Remote. Sens. Environ. 231, 111262.

Tang, H., Dubayah, R., Swatantran, A., Hofton, M., Sheldon, S., Clark, D.B., Blair, B., 2012. Retrieval of vertical LAI profiles over tropical rain forests using waveform lidar at La Selva, Costa Rica. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2012.05.005.

Tang, H., Ma, L., Lister, A., O'Neill-Dunne, J., Lu, J., Lamb, R.L., Dubayah, R., Hurtt, G., 2021. High-resolution forest carbon mapping for climate mitigation baselines over the RGGI region, USA. Environ. Res. Lett. 16, 035011.

Tubiello, F.N., Pekkarinen, A., Marklund, L., Wanner, N., Conchedda, G., Federici, S., Rossi, S., Grassi, G., 2020. Carbon emissions and removals by forests: New estimates 1990–2020. Earth Syst. Sci. Data. https://doi.org/10.5194/essd-2020-203.

Valbuena, R., Hernando, A., Manzanera, J.A., Görgens, E.B., Almeida, D.R.A., Mauro, F., García-Abril, A., Coomes, D.A., 2017. Enhancing of accuracy assessment for forest above-ground biomass estimates obtained from remote sensing via hypothesis testing and overfitting evaluation. Ecol. Model. 366, 15–26.

Vorster, A.G., Evangelista, P.H., Stovall, A.E.L., Ex, S., 2020. Variability and uncertainty in forest biomass estimates from the tree to landscape scale: the role of allometric equations. Carbon Balance Manag. 15, 8.

Wood, S.N., 2017. Generalized Additive Models: An Introduction with R, Second edition. CRC Press.

Wulder, M.A., White, J.C., Nelson, R.F., Næsset, E., Ørka, H.O., Coops, N.C., Hilker, T., Bater, C.W., Gobakken, T., 2012. Lidar sampling for large-area forest characterization: a review. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2012.02.001.

Xu, L., Saatchi, S.S., Shapiro, A., Meyer, V., Ferraz, A., Yang, Y., Bastin, J.-F., Banks, N., Boeckx, P., Verbeeck, H., Lewis, S.L., Muanza, E.T., Bongwele, E., Kayembe, F., Mbenza, D., Kalau, L., Mukendi, F., Ilunga, F., Ebuta, D., 2017. Spatial distribution of carbon stored in forests of the Democratic Republic of Congo. Sci. Rep. 7, 15030.

Zhao, K., Popescu, S., Nelson, R., 2009. Lidar remote sensing of forest biomass: a scale-invariant estimation approach using airborne lasers. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2008.09.009.

Zolkos, S.G., Goetz, S.J., Dubayah, R., 2013. A meta-analysis of terrestrial aboveground biomass estimation using lidar remote sensing. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2012.10.017.