Journal
Global Change Biology

Type
Research Article

Title
Toward a forest biomass reference measurement system for remote sensing applications

Running head
Forest Biomass Reference Measurement System

Abstract
Forests contribute to climate change mitigation through carbon storage and uptake, but the extent to which this carbon pool varies in space and time is still poorly known. Several Earth Observation missions have been specifically designed to address this issue, e.g. NASA's GEDI, NASA-ISRO's NISAR and ESA's BIOMASS. Yet, all these missions' products require independent and consistent validation. A permanent, global, *in situ*, site-based forest biomass reference measurement system relying on ground data of the highest possible quality is therefore needed. Here, we have assembled a list of almost two hundred high-quality sites through an in-depth review of the literature and expert knowledge. In this study, we explore how representative these sites are in terms of their coverage of environmental conditions, geographical space and biomass-related forest structure, compared to those experienced by forests worldwide. This work also aims at identifying which sites are the most representative, and where to invest to improve the representativeness of the proposed system. We show that the environmental coverage of the system does not seem to improve after at least the 175 most representative sites are included, but geographical and structural coverages continue to improve as more sites are added. We highlight areas of poor environmental, geographical or structural coverage, including, but not limited to, Canada, the western half of the USA, Mexico, Patagonia, Angola, Zambia, eastern Russia, tropical and subtropical highlands (e.g. in Colombia, the Himalayas, Borneo, Papua). For the proposed system to succeed, we stress that (1) data must be collected and processed applying the same standards across all countries and continents; (2) system establishment and management must be inclusive and equitable, with careful consideration of working conditions; (3) training and site partner involvement in downstream activities should be mandatory.

Keywords
Earth Observation – Forest vegetation – Aboveground biomass – Carbon – Permanent plots – Representativeness – Validation

Authors & Institutions
Nicolas Labrière[1], Stuart J. Davies[2], Mathias I. Disney[3,4], Laura I. Duncanson[5], Martin Herold[6], Simon L. Lewis[3,7], Oliver L. Phillips[7], Shaun Quegan[8], Sassan S Saatchi[9], Dmitry G. Schepaschenko[10,11], Klaus Scipal[12], Plinio Sist[13], Jérôme Chave[1]

[1] Laboratoire Evolution et Diversité Biologique (EDB), Toulouse, France

[2] Forest Global Earth Observatory, Smithsonian Tropical Research Institute, Washington, DC, USA
[3] University College London (UCL), London, UK
[4] NERC National Centre for Earth Observation (NCEO), London, UK
[5] University of Maryland, College Park, Maryland, USA
[6] GFZ German Research Centre for Geosciences, Potsdam, Brandenburg, Deutschland
[7] School of Geography, Leeds, UK
[8] University of Sheffield, Sheffield, UK
[9] Jet Propulsion Laboratory (JPL), California Institute of Technology, Pasadena, CA, USA
[10] International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria
[11] Peoples' Friendship University of Russia (RUDN), Moscow, Russia
[12] European Space Agency (ESA), Frascati, Italy
[13] Forests and Societies Research Unit, Montpellier, France

Reviewers

First name	Last name	Email	Institution	Preference
Gilberto	Câmara	gilberto.camara@inpe.br	Brazil's National Institute for Space Research (INPE)	Recommend
Nuno	Carvalhais	ncarvalhais@bgc-jena.mpg.de	Max Planck Institute for Biogeochemistry Jena	Recommend
Inge	Jonckheere	inge.jonckheere@fao.org	Food and Agriculture Organization of the United Nations (FAO)	Recommend
Anssi	Pekkarinen	Anssi.Pekkarinen@fao.org	Food and Agriculture Organization of the United Nations (FAO)	Recommend
Rasmus	Fensholt	rf@ign.ku.dk	University of Copenhagen	Recommend
Roberta	Martin	roberta.martin@asu.edu	University of Arizona	Recommend

Funding

Author	Funder name	Grant / Award number
NL	European Space Agency (ESA)	ESA ESRIN/Contract No. 4000133843/21/I/NB
NL, JC	French Agence Nationale de la Recherche	CEBA, ref. ANR-10-LABX-25-01
NL, JC	French Agence Nationale de la Recherche	TULIP, ref. ANR-10-LABX-0041
DS	European Space Agency (ESA)	IFBN, 4000114425/15/NL/FF/gp
DS	Russian Science Foundation	Project No. 19-77-30015

MH	European Space Agency (ESA)	Forest Carbon Monitoring and Biomass CCI (phase 2) project
SD	National Science Foundation (NSF), USA	Award #2020424
SS	National Aeronautics and Space Administration (NASA), USA	Terrestrial Ecology and Carbon Cycle program, 80NM0018F0590
OP	European Space Agency (ESA)	ForestScan

Acknowledgments

NL is grateful to John Armston for early exchanges on GEDI data. NL was supported by the European Space Agency (ESA) as part of the Climate Change Initiative (CCI) fellowship (ESA ESRIN/Contract No. 4000133843/21/I/NB). NL and JC benefited from 'Investissement d'Avenir' grants managed by the French Agence Nationale de la Recherche (CEBA, ref. ANR-10-LABX-25-01; TULIP, ref. ANR-10-LABX-0041), and funding from ESA CCI-Biomass (4000123662/18/I-NB) and CNES. The Forest Observation System (FOS) initiative is funded by ESA under contract No. 4000114425/15/NL/FF/gp. The Russian plot data preparation and pre-processing were financially supported by the Russian Science Foundation (Project No. 19-77-30015). The work of MH is supported by the ESA Forest Carbon Monitoring and Biomass CCI (phase 2) project. The work of SS is carried out under a grant by NASA's Terrestrial Ecology and Carbon Cycle program (80NM0018F0590). The TmFO network and the work of PS are supported by the Ministry of Foreign Affairs of France.

Conflict of interest

The authors declare no conflicts of interest.

Data availability statement

Land cover data are available from the Climate Data Store (CDS) of the Copernicus Climate Change Service (C3S; https://cds.climate.copernicus.eu/#!/home). Climatic data were obtained from Abatzoglou et al. (2018) (https://doi.org/10.1038/sdata.2017.191). Topographic data were downloaded from the EarthEnv project (http://www.earthenv.org/). Edaphic data are available from SoilGrids 2.0 (https://doi.org/10.5194/soil-7-217-2021). Canopy height information was obtained from the GEDI L3 Gridded Land Surface Metrics, Version 2 dataset (https://doi.org/10.3334/ORNLDAAC/1952). Tree cover fraction data were obtained from Version 3.0.1 of the global land cover maps distributed by the Copernicus Global Land Service (https://doi.org/10.3390/rs12061044). Realm borders are available from Dinerstein et al. (https://doi.org/10.1093/biosci/bix014). Aboveground carbon density (AGCD) estimates were obtained from Spawn et al. (2020) (https://doi.org/10.1038/s41597-020-0444-4).

Cover letter

* What scientific question is addressed in this manuscript?
=> This study evaluates how representative a system consisting of high-quality permanent *in situ* forest biomass reference measurement sites is in terms of its coverage of three key biomass-related dimensions, i.e. environmental, geographical and structural, in the context of forests worldwide.

* What is/are the key finding(s) that answer this question?
=> Environmental coverage does not improve after at least the 175 most representative sites are included in the system, but geographical and structural coverages exhibit continuous although light improvement as more sites are added. Areas of poor environmental, geographical or structural coverage include Canada, western USA, Mexico, Patagonia, Angola, Zambia, eastern Russia and tropical / subtropical highlands.

* Why is this work important and timely?
=> Forests contribute to climate change mitigation through carbon storage and uptake. To what extent this carbon pool varies in space and time is still poorly known. Several Earth Observation missions have been specifically designed to address this issue. The proposed system will make a major contribution to the essential independent validation of the biomass products provided by all these missions.

* Describe how your paper fits within the scope of GCB; What biological AND global change aspects does it address?
=> This study focuses on forests worldwide and describes a biomass reference measurement system designed to help validate estimates of their contribution to climate change mitigation through carbon storage and uptake from dedicated Earth Observation missions.

* What are the three most recently published papers that are relevant to this manuscript?
=> Chave et al. (2019) Sur Geophys; https://doi.org/10.1007/s10712-019-09528-w
Duncanson et al. (2019) Surv Geophys; https://doi.org/10.1007/s10712-019-09538-8
Schepaschenko et al. (2019) Sci Data; https://doi.org/10.1038/s41597-019-0196-1

If you listed non-preferred reviewers, please provide a justification for each.
=> NA

If your manuscript does not conform to author or formatting guidelines (e.g. exceeding word limit), please provide a justification.
=> NA

Introduction

Plants store about 80% of the Earth's biomass carbon (Bar-On et al., 2018), with forests constituting by far the largest plant carbon pool (ca. 80%; Pan et al., 2013). However, estimates of the spatial distribution and temporal variation of this carbon pool are still imprecise (Harris et al., 2021; Santoro et al., 2021). While forests are vulnerable to global change (Brienen et al., 2020; McDowell et al., 2020; Schimel et al., 2015), they currently provide a carbon sink (e.g. Pan et al., 2011; van Marle et al., 2022) and could contribute further to mitigating climate change given the large potential of intact and regenerating forests for carbon uptake and storage (Chazdon et al., 2016; Requena Suarez et al., 2019). Understanding the nature and distribution of forest carbon fluxes due to land use change and other processes depends critically on mapping the current distribution of vegetation biomass. Moreover, a key factor in projecting how and where forest regeneration or restoration projects would be most effective is detailed, spatially explicit knowledge of local biomass storage potential (see e.g. Heinrich et al., 2021).

The remote sensing community has made substantial investments to address the global challenge of mapping forest carbon stores, fluxes, and their sequestration potential. Several ongoing and upcoming Earth Observation (EO) missions are designed to measure key structural parameters of the world's forests, their carbon stores and their carbon fluxes, e.g. NASA's GEDI (Dubayah et al., 2020), NASA-ISRO's NISAR (NISAR, 2018) and ESA's BIOMASS (Quegan et al., 2019). Each is expected to deliver biomass maps with associated uncertainty. Their coverage, spatial resolution and range depend on mission specifications (e.g. coverage of Earth's surface between 51.6° N and 51.6° S for GEDI, biomass up to 100 Mg/ha for NISAR). Although these missions offer novel approaches to mapping forest carbon, their products require validation using standard procedures to bolster their uptake for a broad range of uses, including climate modelling, national reporting, and land use management (Duncanson et al., 2019). Only if the accuracy and uncertainty of biomass maps are comprehensively assessed and quantified will they meet the needs of the user communities.

How should this be done? We argue that given the wide range of users, instrument sensors, platforms, often limited lifetimes, and pace of technological change, validation strategies need a clear long-term ground vision. This means developing a consistent approach that covers the world's forests and is built to last. It requires designing and maintaining a permanent, global, *in situ*, site-based forest biomass reference measurement (henceforth, FBRM) system to enable independent validation of biomass products and proper quantification of associated uncertainty. Building and sustaining this high-quality distributed system of FBRM sites needs to be an integral part of all EO missions aimed at mapping forest biomass.

In compliance with the good practices protocol for the validation of aboveground woody biomass products (Duncanson et al., 2021), the design of the FBRM system needs to follow a number of principles: (1) Ground data should be of the highest possible quality, with large permanent sampling plots (at least 1 ha in size, 10 ha minimum in total), and airborne LiDAR coverage (at least 1000 ha) plus complementary terrestrial LiDAR acquisitions. The procedures for data acquisition and database compilation should be standardized by following established protocols, and all data should be collected as synchronously as

possible with EO measurements; (2) The system should cover the broadest possible range of environmental, geographical and structural conditions, so as to maximize the robustness of validation activities; (3) The selection of sites should be pragmatic, i.e. focusing on sites where previous expertise and capacity have been built and future operation is highly likely. Establishing and maintaining multiple, high-quality permanent plots is challenging, especially in the tropics (Davies et al., 2021; ForestPlots.net et al., 2021). Therefore, it is strategically sensible while building a potential FBRM system to leverage the experience, knowledge and investment of all stakeholders engaged in long-term permanent plot networks, from data originators (e.g. forest workers) to data curators. And for any such system to be fair and sustainable, the needs of data contributors should be of pivotal concern (de Lima et al., 2022).

Previous experience with the validation of EO products demonstrates the value of highly integrated FBRM sites compared to widely distributed small forest samples as established by most national forest inventories. This is because validation of EO-derived biomass maps depends strongly on accurate spatial registration of the ground plots, and because biomass estimates from individual plots are informative for calibration/validation only if the plots are large enough (Réjou-Méchain et al., 2014). All the aforementioned conditions for the inclusion of sites in a global monitoring system are difficult to meet, and for the moment, validation efforts for each individual EO mission have been based on a handful of sites.

How many observation sites would be necessary for global validation of biomass maps, and where should they be located? From a validation perspective, these sites should ideally span a wide range of biomass, and should encompass a variety of forest structures for any given level of biomass. But from an ecological point of view, the sites should cover an extensive range of bioclimatic and biogeographic conditions, as well as contrasting topographies, soil types and geological substrates, and be exposed to varying levels and types of anthropogenic pressures or natural disturbances. Given the enormous extent and diversity of forests globally, the replication of high-quality observation sites at thousands of locations is unrealistic, so the theoretical challenge in allocating limited resources to locations involves maximizing their distance from each other along key dimensions, to ensure an optimized coverage of conditions experienced by forests around the world. However, because these sites should ideally already be established (Chave et al., 2019), the problem of site selection is constrained by what is available. Here, we have assembled a list of almost two hundred potential FBRM sites through an in-depth review of the literature and expert knowledge. The aim of this study is to evaluate how representative these sites are in terms of their coverage of three key biomass-related dimensions, i.e. environmental, geographical and structural, in the context of forests worldwide.

Specifically, we ask the following research questions: (1) how well does a selection of existing forest sites represent environmental conditions, geographical space and forest structure globally?; (2) which combination of sites best represents each of the three biomass-related dimensions over global forested areas, for any given number of sites?; (3) how does a combination of potential FBRM sites compare in terms of representativeness with an equivalent number of forested locations randomly selected over the globe?; (4) where should efforts be invested to improve the environmental, geographical and structural coverage of the proposed FBRM system, possibly going beyond existing plots?

Materials and methods

1) Potential forest biomass reference measurement sites

We assembled a list of sites meeting all or most of the quality criteria required to become part of the FBRM system (e.g. plot size, likeliness to be revisited). We screened the following continental to global-scale forest plot networks for potential sites of interest: AfriTRON (Hubau et al., 2020), ForestGEO (Davies et al., 2021), IIASA (Schepaschenko et al., 2017), NEON (Metzger et al., 2019), RAINFOR (ForestPlots.net et al., 2021), SEOSAW (The SEOSAW partnership, 2020), TERN (Cleverly et al., 2019) and TmFO (Sist et al., 2015). Peer-reviewed and grey literature were also searched, and expert knowledge mobilized through consultation with key stakeholders, such as EO mission research scientists, space agencies and national forest/forestry departments. We tried to be as thorough and exhaustive as possible but some high-quality plots and networks might have escaped our notice and readers are encouraged to contact the corresponding author to notify us of this.

The screening resulted in a list of 195 potential FBRM sites (Table S1). Among these, plot cumulative area ranged from 0.5 ha for several of the Siberian sites to 125 ha at Paracou, French Guiana. About two thirds of the sites had a plot cumulative area ≥ 10 ha (n = 132), with about half of those that did not located in the Palearctic (n = 30). Potential FBRM sites were present in every forested biome, *sensu* Whittaker (1975), yet the coverage of annual precipitation and mean temperature gradients was uneven (Figure S1). About three-quarters were affiliated to (at least) one of the eight large-scale networks. The rest were usually monitored by research institutes, universities or national forest/forestry departments.

We use the terminology of “potential” FBRM sites, mindful that this list is likely to change in the future for various reasons. One is that most of the sites have not formally agreed to join the proposed system of FBRM sites (and many have probably not heard about the concept yet). Plus, some sites may in the end prove unsuitable, and others may join the initiative. However, the fairly large sample of sites represented in the list reported here is a useful step to test this study’s research questions.

2) Geographic information and study area

All spatial data were reprojected using a global equal-area map projection to reflect the respective and relative area contributions of realms and continents. EASE-Grid 2.0 (epsg:6933), version 2 of the Equal-Area Scalable Earth Grid (Brodzik et al., 2012), is commonly used for satellite-based data distribution (see e.g. GEDI; Dubayah et al., 2021). This projection is preferable to the longitude-latitude coordinate reference system (epsg:4326), that is neither equal-area nor conformal. The coarsest spatial resolution of all spatial datasets used in this study (2.5 arc-minute, which is about 5 km at the equator, for the TerraClimate dataset; Abatzoglou et al., 2018) was chosen, and all datasets were resampled accordingly. Following reprojection and resampling, gridded data were generated over 2,920 rows and 6,940 columns, that is 20,264,800 cells in total.

To restrict our analysis to forests, we built a forest mask using land cover data for 2020 from the ESA CCI Land Cover project. The original dataset (300 m spatial resolution; epsg:4326) was reprojected and resampled to 5 km (mode retained). The mask included cells with tree-dominated land cover classes (see Supporting Information for more details), for a total of 1,728,368 cells (that is, around 43 million km^2). Non-tree dominated land cover classes such as shrubland, grassland and cropland are also pools of carbon, but were not considered here.

3) Environmental space

Climatic, topographic and edaphic variables are widely used to investigate the influence of the environmental space on forest structure, composition and functioning (see e.g. Anderson-Teixeira et al., 2015; Sullivan et al., 2020).

Temperature and precipitation are key climatic factors influencing vegetation patterns (Holdridge, 1947; Whittaker, 1975), together with their seasonality (Mucina, 2019). So is solar radiation (Cox et al., 2016). Annual mean temperature (°C), temperature seasonality (% coefficient of variation CV), annual precipitation (mm), precipitation seasonality (% CV) and solar radiation (W m^{-2}) were therefore selected for subsequent analysis. Data were taken from the TerraClimate dataset (original spatial resolution 5 km; Abatzoglou et al., 2018) directly, or could be computed from it following O'Donnell & Ignizio (2012).

Topographic variables and especially elevation also shape the spatial distribution of species and habitats (see altitudinal zonation; von Humboldt & Bonpland, 1805). Data on elevation above sea level (m) were obtained from the EarthEnv project (http://www.earthenv.org/) (Amatulli et al., 2018).

Soil physico-chemical properties have a direct influence on vegetation, as they partly determine water and nutrient availability (Hulshof & Spasojevic, 2020). Estimated edaphic data were obtained from SoilGrids 2.0 (original spatial resolution 250 m; Poggio et al., 2021). Depth-weighted averaged values over the three topmost soil layers (i.e. 0–5 cm, 5–15 cm and 15–30 cm) were computed for each of the eleven variables provided. As in Sullivan et al. (2020), we selected variables representing both soil physical ("texture") and chemical ("fertility") properties. More specifically, we retained coarse fragment content (% volume), sand fraction (% mass), cation exchange capacity (cmol kg^{-1}) and pH (H_2O) (unitless). We favored sand fraction over clay fraction (commonly retained in similar analyses), as the latter was modelled less accurately (Poggio et al., 2021).

Some edaphic variables were found to be strongly correlated with climatic ones, like cation exchange capacity and annual mean temperature (Spearman's rank correlation coefficient $\rho < -0.75$). This may be because edaphic variables are modelled using other variables, including climatic ones (Poggio et al., 2021). Despite some strong pairwise correlations between the ten variables selected (5 climatic, 1 topographic and 4 edaphic), we kept them all as indicators of the environmental space as each bears relevant information. Correlation is unlikely to distort results from the analysis of network representativeness described below. Previous studies, e.g. Anderson-Texeira et al. (2015) or Hoffman et al. (2013), ran the same analysis using an even bigger number of variables (n = 17 and n = 37, respectively) without considering correlation.

4) Geographical space

We also explored whether the potential sites were sufficiently distant from each other to cover the entire forested area of the world. Since floristic composition varies greatly across continents, maximizing geographical distance across sites and minimizing the occurrence of geographical gaps is desirable in the optimal design of a reference measurement system.

5) Structural space

Canopy height and tree cover fraction are two structural variables commonly used to describe forest structure. Both can be estimated by spaceborne instruments. Canopy height (H) information was obtained from the GEDI L3 Gridded Land Surface Metrics, Version 2 dataset (Dubayah et al., 2021). Gridded data at 1 km spatial resolution (mean RH100, i.e. the 100th percentile of waveform energy relative to the ground, computed from individual waveforms collected between April 18th 2019 and April 14th 2021) were averaged to 5 km. We kept 5 km cells only when at least half of their area overlapped with non-empty 1 km cells. Due to GEDI discrete sampling and ISS-orbit limited spatial coverage (± 51.6° latitude), only about 60% of the potential FBRM sites (n = 118) and half of the forested cells (n = 829,256) had canopy height information available from GEDI first two years of data collection.

Tree cover fraction (TC) was also used, based on the PROBA-V satellite acquisitions for 2019. These data were obtained from Version 3.0.1 of the global land cover maps distributed by the Copernicus Global Land Service (Buchhorn et al., 2020). Original data at 100 m spatial resolution were reprojected and averaged to 5 km.

6) Analysis of network representativeness

To assess how well a network of observation sites represents environmental, geographical and structural conditions of forested areas globally, we performed a point-based “representativeness of network” analysis (Anderson-Teixeira et al., 2015; Hoffman et al., 2013). The principle of this analysis is as follows.

For each site, distances were computed between values at that site and those at any cell of the map included in the forest mask. More precisely, we computed Euclidean distances on standardized variables (after z-score normalization) for the environmental and structural spaces, and great-circle distance (i.e. the shortest distance between two points on the Earth surface, represented here by a sphere) computed using the haversine formula for the geographical space. This resulted in site-specific environmental, geographical and structural distance maps. Site-specific maps referring to the same space were then stacked, and the minimum value retained for each cell to produce environmental, geographical and structural dissimilarity maps. Lastly, maximum environmental, geographical and structural distances were searched for (see Supporting Information for more details) and relative dissimilarity was mapped as a percentage of the normalized value.

The representativeness of network analysis was performed for various sets of contributing sites (i.e. those included in the stack from which minimum values were selected) based on the following selection strategies: all potential FBRM sites, only those with a plot cumulative area ≥ 10 ha, the n most representative potential FBRM sites, n randomly selected potential FBRM sites, the n most representative virtual sites (i.e. cells with no potential FBRM site identified for the time being) over global forested areas, and n randomly selected virtual sites (for n ranging from 5 to 118 or 195 depending on selection strategy and space).

To identify the most representative FBRM or virtual sites for a given number of sites, n, we performed a partitioning around medoids (PAM) analysis. This clustering technique is suited for our purpose as clusters are built around actual objects (the so-called “medoids”, here potential FBRM sites or cells) and not “centroids” as in the k-means algorithm (Kaufman & Rousseeuw, 1990). Despite often being regarded as deterministic (see e.g. Reynolds et al., 2006), there might be ties in some cases e.g. during medoid selection when choosing between two objects that may give the same reduction in the cost function, i.e. the sum of dissimilarities. In this case, selecting one object over the other would depend on the order in which these two objects were presented to the algorithm. To address this problem, we ran the original PAM algorithm a hundred times for each number, n, of potential FBRM sites of interest, each time reshuffling the input dataset, and retained the most frequent combination to serve as the n most representative sites. For most representative virtual site selection, the “fasterPAM” algorithm was used on a subset of 20,000 cells geographically spanning global forested areas to reduce the computational burden (Schubert & Rousseeuw, 2021).

Finally, we selected potential FBRM sites randomly and ran the representativeness of network analysis. This operation was repeated 200 times, and median relative dissimilarity values retained for each cell of the study area to produce relative dissimilarity maps. The whole process was also performed for virtual sites selected randomly over global forested areas, with only 5 repetitions in this case because of computational cost. Only for n = 100 were random virtual site selection and subsequent representativeness of network analysis repeated 200 times (median retained for each cell of the forest mask), and the difference between random vs. most representative site resulting relative dissimilarity maps computed.

All analyses were conducted using the R statistical computing platform (R Core Team, 2021), and mainly packages ‘cluster’ (Maechler et al., 2021), 'data.table' (Dowle & Srinivasan, 2021), ‘gdalUtils’ (Greenberg & Mattiuzzi, 2020) and 'raster' (Hijmans, 2021).

Results

1) Representativeness of a potential forest biomass reference measurement system with all pre-existing sites currently identified

Environmental conditions were well represented (defined here as relative dissimilarity < 10%, that is ca. a third of maximum dissimilarity) by the system of potential FBRM sites in most lowland tropical rainforests, the eastern part of Canada and the United States of America (USA), northern Europe and the west and central parts of Russia (Figure 1, top). Among forested areas noticeably lacking sufficient coverage of environmental conditions (relative dissimilarity > 10%) were the western half of North America (incl. Mexico), Patagonia, Angola/Zambia and eastern Russia. Overall, the geographical and structural spaces benefited from a better representation by the potential FBRM sites than environmental space (Figure 1, center and bottom respectively). In the main, only Patagonia, the easternmost part of Siberia and New Zealand were poorly represented in geographical space (relative dissimilarity > 10%). Insufficient coverage of structural conditions (relative dissimilarity > 10%) mostly affected isolated cells present in limited areas such as the west coast of the USA, forested areas of the Himalayas and the Sunda Shelf (Sumatra, peninsular Malaysia, Borneo) (see Figure S2 for a close-up on portions of these three areas).

2) Maximum representativeness possible with different combinations of pre-existing sites currently identified

Comparing the distribution of relative dissimilarity values for environmental, geographical and structural conditions for various sets of potential FBRM sites, the spread (i.e. the variability of values) was highest for environmental space, whatever the set of sites under consideration (Figure 2). Representativeness was always maximized when all the potential FBRM sites were included. Conversely, the highest relative dissimilarity values were reached whatever the space when using the 132 sites with a plot cumulative area ≥ 10 ha, and not the 50 most representative ones. Consistent with Figure 1, only a low proportion of relative dissimilarity values exceeded 10% when considering geographical or structural conditions. Whatever the value n, the identity of the n most representative sites differed between spaces, notably because 40% of the sites from the initial pool did not have canopy height information and could therefore not be considered when studying site contribution to the representativeness of the structural space. For a given space, a site among the n most representative ones was not necessarily selected for higher values of n (Tables S1–S2, Figure S3).

3) Pre-existing site- vs. random location-based system

Less than half of global forested areas were better represented environmentally, geographically and structurally by the 100 most representative potential FBRM sites than a hundred random samples (proportion ranging from 39 to 48% depending on the space; Figure 3). This was particularly apparent for Canadian, Amazonian, Angolan/Zambian and Russian forested areas with respect to the environmental space (Figure 3, top).

Geographically, a better representation was achieved by the 100 most representative potential FBRM sites than by a hundred random ones in the vicinity of selected FBRM sites, creating island-like patterns (Figure 3, center). Regional patterns were less sharp for the representation of structural conditions, but North American and Asian forested areas appeared generally better represented by the 100 most representative potential FBRM sites than a hundred random samples, while South American and African forested areas showed the opposite (Figure 3, bottom).

4) Pre-existing site-based system improvement

Increasing plot cumulative area for all sites up to at least 10 ha would increase the number of potential FBRM sites meeting the CEOS requirements (Duncanson et al., 2021), and consequently improve the environmental, geographical and structural coverage of the resulting system (Figure 2). The more locations in the system, the lower the median relative dissimilarity values, whatever the space and location selection strategy (Figure 4). For example, as regards environmental coverage, median relative dissimilarity values decreased from 11.6 to 10.1 to 9.2% respectively when the 50, 100 and 150 most representative potential FBRM sites were selected. Selecting the n most representative cells over global forested areas always provided better environmental, geographical and structural coverage than other selection strategies. A system made up of random cells was more representative of the environmental and geographical spaces than its most representative pre-existing site-based counterpart, whenever at least 20 locations contributed to the system.

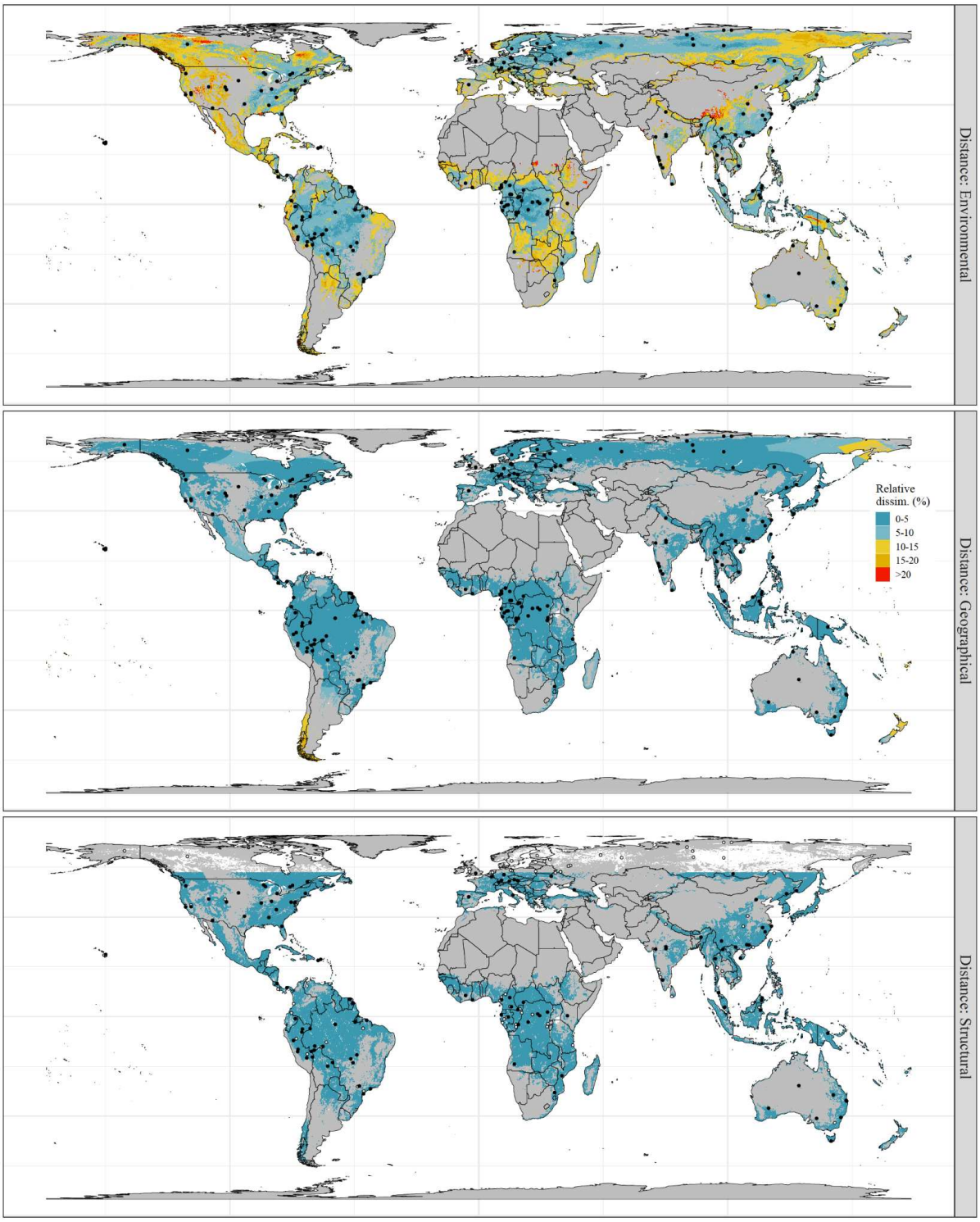

Figure 1. Relative environmental (top), geographical (center) and structural (bottom) dissimilarities (%) over global forested areas with respect to conditions covered by potential forest biomass reference measurement sites (n = 195, top and center; n = 118, bottom). Blank continental areas and hollow points (bottom) respectively correspond to forested areas and sites not sampled (yet, for those within ± 51.6° latitude) by GEDI. Relative dissimilarity was categorized for display purposes. Non-forested areas are in grey. The map projection is EASE-Grid 2.0 (epsg:6933), a global, equal-area protection, and spatial resolution is 5 km.

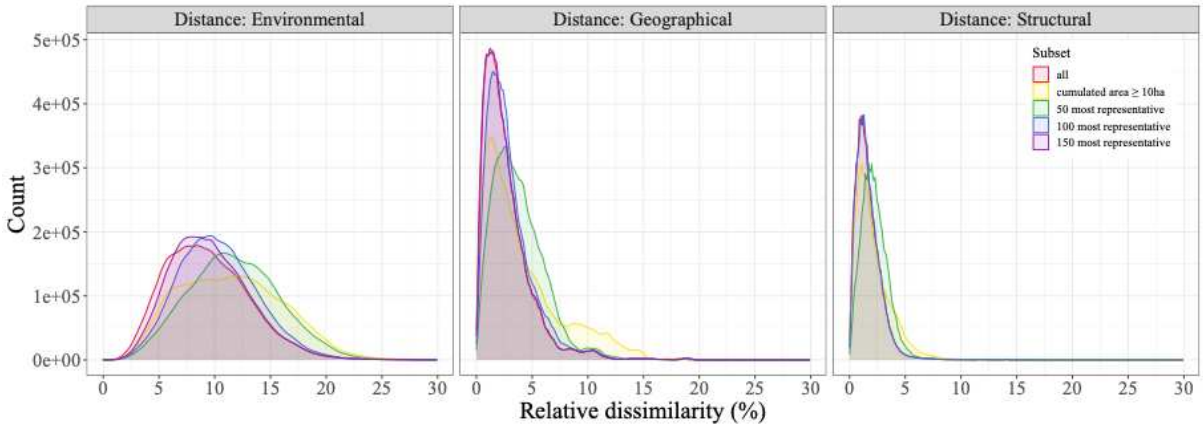

Figure 2. Relative dissimilarities for different types of distances and subsets of potential forest biomass reference measurement sites. There are 1,728,368 contributing cells (5 km spatial resolution) for the environmental (left) and geographical (center) density plots, and 829,256 for the structural (right) density plot because of GEDI discrete sampling and ISS-orbit limited spatial coverage (± 51.6° latitude). The X-axis was cropped to 30% of relative dissimilarity for display purposes, excluding ca. 0.045% of the overall data.

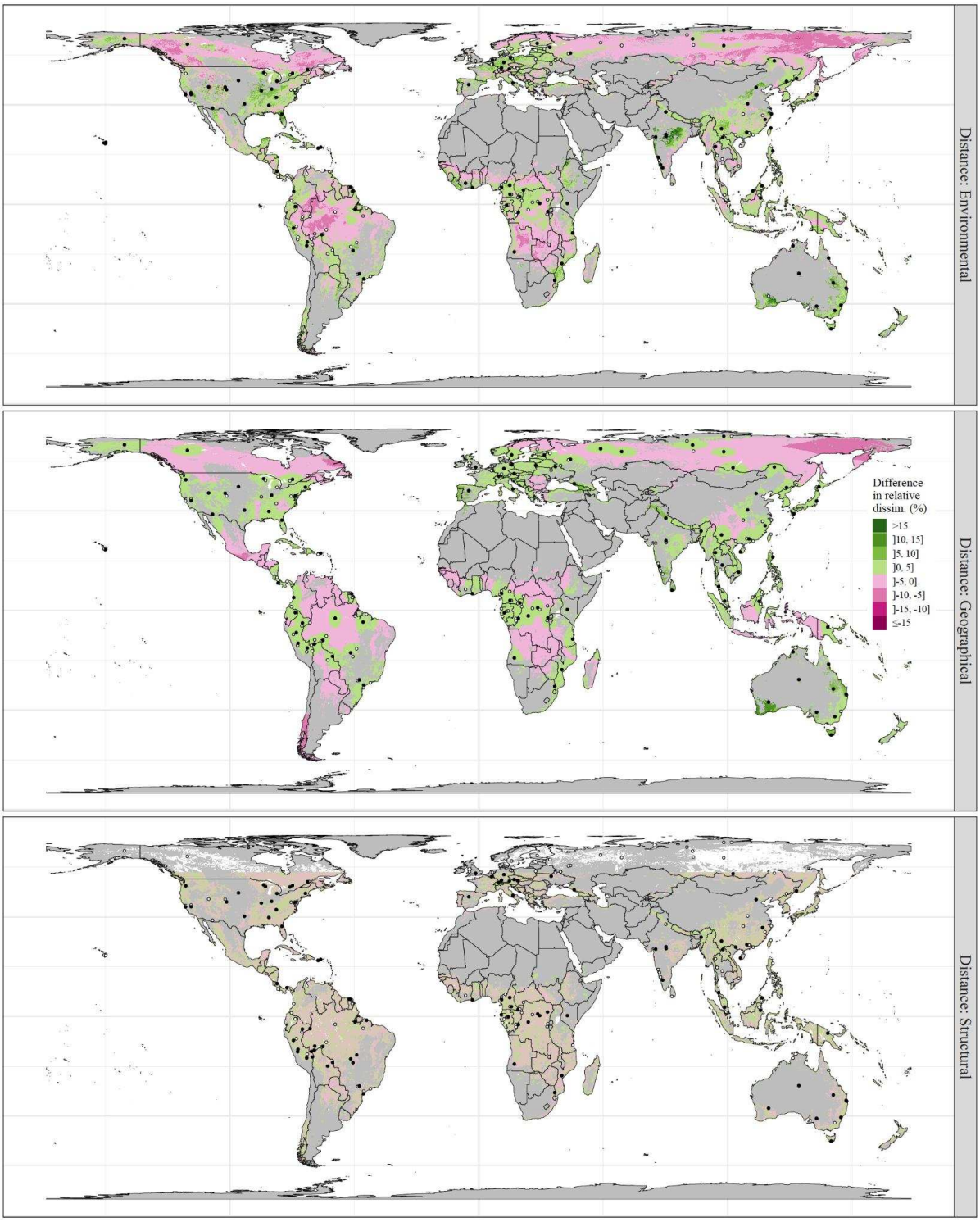

Figure 3. Difference in relative environmental (top), geographical (center) and structural (bottom) dissimilarities between a set of 100 randomly selected cells (median of 200 runs used) and the 100 most representative potential forest biomass reference measurement (FBRM) sites. A network made up of randomly selected cells is less representative of local conditions than one made up of the 100 most representative potential FBRM sites, wherever the difference in relative dissimilarity is positive. Difference in relative dissimilarity was categorized for display purposes. Non-forested areas are in grey. Blank continental areas within ± 51.6° latitude (bottom) correspond to areas not yet sampled by GEDI, and hollow points to sites not among the 100 most representative potential FBRM sites. The map

projection is EASE-Grid 2.0 (epsg:6933), a global, equal-area protection, and spatial resolution is 5 km.

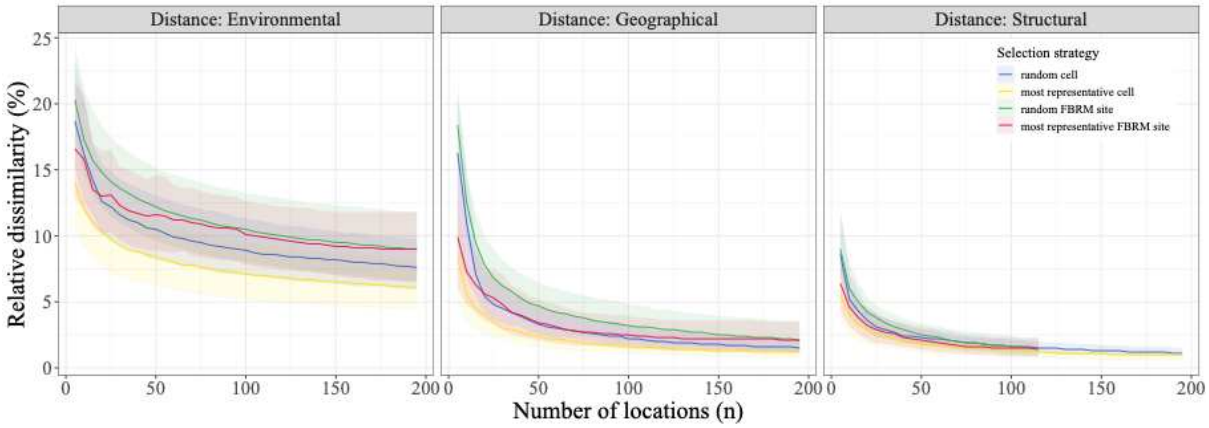

Figure 4. Relative dissimilarities vs. number of locations for different types of distances and selection strategies. Only numbers of locations, n, which are multiples of 5 are used here. Lines and shaded areas correspond to the median and interquartile range of relative dissimilarity values over global forested areas, respectively.

Discussion

1) Guaranteeing and improving system representativeness

Various ways were identified to guarantee and further improve the representativeness of the proposed system of FBRM sites. First and foremost, efforts (discussed extensively later) should be made to ensure that every single potential FBRM site identified in this study joins the proposed system. The environmental coverage of the system does not seem to improve after at least the 175 most representative potential FBRM sites are included, but geographical and structural coverages showed a continuous although slight improvement (Figure 4).

Second, plot cumulative area should be increased to at least 10 ha at each site wherever this is not the case to comply with CEOS recommendations (Duncanson et al., 2021). This would clearly improve the environmental, geographical and structural coverage of the system (Figure 2), if we were to consider that sites where plots do not cover at least 10 ha overall should consistently be dismissed. Apart from plot cumulative area, ancillary data will likely need to be acquired, updated or upgraded, including more accurate location of plot corners (using differential global navigation satellite systems), soil samples to determine characterize local soil physico-chemical properties, and airborne and terrestrial LiDAR acquisitions. While the FBRM system is being formed, the “representativeness of network" analysis developed in this study can help prioritize sites for main and ancillary data acquisition (Table S2, Figure S3).

Third, efforts should be made to identify pre-existing sites in areas of poor environmental, geographical or structural coverage (Figure 1, Figure 3). These include, but are not limited to, Canada, the western half of the USA, Mexico, Patagonia, Angola, Zambia, eastern Russia, tropical and subtropical highlands (e.g. in Colombia, the Himalayas, Borneo, Papua). It should be noted here that in some of these areas, forest inventory data are already collected but with designs suboptimal to have been identified as potential FBRM sites and included in this study. Nonetheless, there will ideally be opportunities to expand on some key locations.

Fourth, given the obvious coverage gaps in these areas, new sites should be established if none already exist. The manifold added values of long-term permanent plots compared to newly-established ones include good knowledge of site history, the availability of ancillary and recensus inventory data, and the fact that plot remeasurement is cheaper than establishment. Yet relative dissimilarities are minimized whatever the space when most representative virtual sites (i.e. cells) instead of most representative potential FBRM sites are used (Figure 4). This likely arises from the fact that potential FBRM sites are not located randomly. Individual plot networks were usually built with certain criteria in mind, for example to study well-defined geographical areas (e.g. Australia for TERN; Cleverly et al., 2019) and/or to answer specific research questions (e.g. what are the long-term effects of logging on tropical forests for TmFO?; Sist et al., 2015). However, their aggregation does not guarantee a satisfactory representativeness of the environmental, geographical and structural spaces covered by global forested areas. Within a given biome or ecoregion, plot location might also be biased due to, e.g., logistical considerations like accessibility. Such could be the case over Amazonia, where a recent study suggested that plots were

preferentially located in areas of high ancient human impact, potentially slanting our understanding of Amazonian forest dynamics (McMichael et al., 2017).

Last, in order to improve the system representativeness and avoid presenting a potentially distorted picture of its performances regionally (e.g. over-optimistic in the tropics?; see Figure 1), other spaces could be considered, such as biogeographical and disturbance (both exogenous and anthropogenic) spaces. The former could include, among other information, layers of global tree species α and β diversity (Keil & Chase, 2019). The latter could encompass map-based information on, e.g., forest integrity with respect to anthropogenic pressures (Grantham et al., 2020) or susceptibility to natural disturbances (windstorms, wildfires, etc.). Concurrently, integration to the FBRM system of long-term permanent plot networks focused on the study of secondary forests such as 2ndFOR (Poorter et al., 2021) should be favored to keep increasing the heterogeneity of forest conditions and successional stages covered by ground data.

2) Relationship between forest structure and aboveground biomass

Environmental conditions are used to model potential (i.e. theoretical) aboveground biomass (Prentice et al., 2011). Differences between potential and actual biomass stocks are hypothesized to originate from human disturbances (Pan et al., 2013). Structural conditions were represented in this study using remote sensing data (tree cover fraction and canopy height derived from PROBA-V and GEDI data, respectively) acquired during the last 2-3 years. Their contemporaneity is an asset to keep track of biomass stocks in a rapidly changing world.

Aboveground biomass is commonly estimated from structural attributes across various scales, using e.g. tree height and diameter at the individual tree scale (Chave et al., 2014) and top-of-canopy height at the (sub-)hectare scale (Labrière et al., 2018). At larger scale, previous exploratory work showed that spatial variations in the product of tree cover fraction (TC) and canopy height closely corresponded to those of LiDAR-derived aboveground biomass carbon density (AGCD) maps (see “CCI Biomass Product Validation and Algorithm Selection Report” 1 and 2; https://climate.esa.int/en/projects/biomass/key-documents/). We tested how well TC×H correlated with AGCD at the 5 km cell scale over global forested areas and for the subset of cells bearing potential FBRM sites. AGCD estimates were obtained from Spawn et al. (2020), after original data at 300 m spatial resolution were reprojected and averaged to 5 km. We found that AGCD was strongly correlated with TC×H over global forested areas (n = 829,256, Spearman’s ρ = 0.77, $p < 0.001$) (Figure S4). Root-mean-square error (RMSE), coefficient of correlation (R^2) and bias were 26.7 MgC ha^{-1}, 0.85 and 4.1 MgC ha^{-1}, respectively. Similar statistics were obtained with potential FBRM site-bearing cells only (n = 118): Spearman’s ρ = 0.74 ($p < 0.001$), RMSE = 30.5 MgC ha^{-1}, R^2 = 0.82 and bias = 3.5 MgC ha^{-1}. This confirmed that structural attributes are important predictors of aboveground biomass.

Nonetheless, local information may be essential to reduce uncertainties in AGB due to locally variable parameters such as community wood density (Phillips et al., 2019) inferred using tree-by-tree identity information that at present can only be provided by *in situ* data. The pivotal role of *in situ* data was recently exemplified in the case of GEDI waveform data. Accurately predicting AGCD from GEDI waveforms alone was shown to be suboptimal as

two forest stands with similar waveforms can have very different AGCD (Bruening et al., 2021), and allometries heavily rely on *in situ* training data (Duncanson et al., 2022). Beyond such direct use, tree-by-tree identity information can also be mobilized to calibrate and validate hyperspectral data (Draper et al., 2019; Jucker et al., 2018), which can in turn improve forest stratification and the use of the most relevant structure metrics-based allometries. In this study, structural coverage was represented by two of the most meaningful variables that can be remotely sensed over global forested areas: tree cover fraction and canopy height. Including other structure-related variables, such as canopy height variability, could complement our understanding of how representative the proposed FBRM sites are of the structural space. This analysis will gain in completeness as new datasets, and new versions of the ones we used, are released. The current GEDI L3 gridded dataset (Version 2) is still patchy, especially in the tropics, and coverage should keep improving with following versions. In addition, plant area index (PAI) and vertical foliage profile, two variables that have already proven useful to distinguish vegetation types (see e.g. Marselis et al., 2018), should be part of the next releases. Also, as boreal forests are barely sampled by GEDI due to ISS-orbit limited spatial coverage, incorporating canopy height information from NASA's Ice, Cloud and Land Elevation Satellite-2 (ICESat-2) mission (ALT18; Markus et al., 2017) will help fill a major gap in structure data.

3) On the uniqueness of tropical forests

The proposed system of FBRM sites should encompass a wide variety of forest conditions (incl. old-growth, regenerating, managed) and soil types (incl. well-drained, nutrient-poor, seasonally flooded, swampy). Adequate coverage of the three main forest biomes (tropical, temperate and boreal) is also essential. But how should this “adequate” coverage be established? Areal forest biome proportions of global forested areas are close to 50, 20 and 30% for tropical, temperate and boreal forest biomes, respectively (Pan et al., 2013). Based on areal considerations only, this would mean that half of the potential FBRM sites should be located in the tropics, a fifth in temperate regions and the rest (about a third) in boreal ones. This condition is satisfied for most representative virtual sites (i.e. cells), whatever the value n of cells and for both environmental and geographical distances, but not for most representative potential FBRM sites (Figure S5). This is likely due to the different balance of forest biome proportions in the list of potential FBRM sites (ca. 60, 35 and 5% for tropical, temperate and boreal forest biomes, respectively) compared to global forested areas. Regarding structural coverage, forest biome proportions are most probably influenced by the truncated coverage of boreal forests (see above). In terms of aboveground biomass instead of area, forest biome proportions would be 65, 20, 15% for tropical, temperate and boreal forest biomes, respectively (using data from Spawn et al., 2020). Focusing on either gross or net primary productivity (GPP and NPP, respectively) also shows the disproportionate contribution of tropical forests compared to their area (more than two thirds; Pan et al., 2013), which is even more apparent when emphasizing on gross forest emissions (almost four fifths over the years 2001-2019; Harris et al., 2021). Tropical sites should consequently be the cornerstone of the FBRM system, reasonably representing 65-70% of all the potential FBRM sites. This is all the more relevant because 80-95% of all known tree species in each continent were sampled in their tropical region (Cazzolla Gatti et al., 2022), a hyperdiversity further complicating community wood density determination (Phillips et al., 2019). While the PAM algorithm does not, either in its original or most recent form, include weighting options, other clustering techniques could be envisioned that would allow weighting existing or virtual

potential FBRM sites depending on a cell's AGB, GPP, NPP, tree diversity or a combination of some or all of these.

4) Practical implementation of a forest biomass reference measurement system and final considerations

The proposed FBRM system will provide a framework within which a diverse community of stakeholders (e.g. Earth Observation agencies, individual countries, forest organizations) can make a lasting contribution to (and of course benefit from) a comprehensive and sustained system of high-quality biomass reference data. This system also has to be recognized and supported as an opportunity to train the next generation of researchers with expertise at the confluence of forest science and remote sensing, leveraging investments made by the forest science community. Funding the FBRM system will require significant investment. However this investment, even on a global scale, is a fraction of the cost of a single space mission. Plus, this cost is likely to be largely offset by the resulting widespread, consistent and effective use of the EO-derived biomass maps. Two possible funding mechanisms could be imagined, one where funding bodies collaborate with long-term permanent plot networks and another where funders collaborate directly with individual plot principal investigators. Whatever the funding scheme favored, for the FBRM concept to succeed, plot networks must collect and process the data applying the same standards across all countries and continents, and subsequently share the derived data products with the global community, for example through the Forest Observation System (FOS; Schepaschenko et al., 2019). Protocol harmonization and standardization are key to ensuring high quality of the data generated and maximizing interoperability across all FBRM sites, and should be conducted for all the necessary steps from fieldwork (e.g. plot shape, tree diameter measurement) to post-field data processing (e.g. allometric equations, error propagation scheme). It must be stressed that the proposed system needs to be established and managed inclusively, with careful consideration of working conditions. Training and site partner involvement in downstream activities should be mandatory. Only this would allow for proper recognition of the disadvantaged social, economic and historical context in which most staff involved in forest research activities operate, which is overwhelmingly true in tropical nations.

Bibliography

Abatzoglou, J. T., Dobrowski, S. Z., Parks, S. A., & Hegewisch, K. C. (2018). TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. *Scientific Data*, *5*(1), 170191. https://doi.org/10.1038/sdata.2017.191

Amatulli, G., Domisch, S., Tuanmu, M.-N., Parmentier, B., Ranipeta, A., Malczyk, J., & Jetz, W. (2018). A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. *Scientific Data*, *5*(1), 180040. https://doi.org/10.1038/sdata.2018.40

Anderson-Teixeira, K. J., Davies, S. J., Bennett, A. C., Gonzalez-Akre, E. B., Muller-Landau, H. C., Joseph Wright, S., Abu Salim, K., Almeyda Zambrano, A. M., Alonso, A., Baltzer, J. L., Basset, Y., Bourg, N. A., Broadbent, E. N., Brockelman, W. Y., Bunyavejchewin, S., Burslem, D. F. R. P., Butt, N., Cao, M., Cardenas, D., … Zimmerman, J. (2015). CTFS-ForestGEO: a worldwide network monitoring forests in an era of global change. *Global Change Biology*, *21*(2), 528–549. https://doi.org/10.1111/gcb.12712

Bar-On, Y. M., Phillips, R., & Milo, R. (2018). The biomass distribution on Earth. *Proceedings of the National Academy of Sciences*, *115*(25), 6506. https://doi.org/10.1073/pnas.1711842115

Brienen, R. J. W., Caldwell, L., Duchesne, L., Voelker, S., Barichivich, J., Baliva, M., Ceccantini, G., Di Filippo, A., Helama, S., Locosselli, G. M., Lopez, L., Piovesan, G., Schöngart, J., Villalba, R., & Gloor, E. (2020). Forest carbon sink neutralized by pervasive growth-lifespan trade-offs. *Nature Communications*, *11*(1), 4241. https://doi.org/10.1038/s41467-020-17966-z

Brodzik, M. J., Billingsley, B., Haran, T., Raup, B., & Savoie, M. H. (2012). EASE-Grid 2.0: Incremental but Significant Improvements for Earth-Gridded Data Sets. *ISPRS International Journal of Geo-Information*, *1*(1). https://doi.org/10.3390/ijgi1010032

Bruening, J. M., Fischer, R., Bohn, F. J., Armston, J., Armstrong, A. H., Knapp, N., Tang, H.,

Huth, A., & Dubayah, R. (2021). Challenges to aboveground biomass prediction from waveform lidar. *Environmental Research Letters*, *16*(12), 125013. https://doi.org/10.1088/1748-9326/ac3cec

Buchhorn, M., Lesiv, M., Tsendbazar, N.-E., Herold, M., Bertels, L., & Smets, B. (2020). Copernicus Global Land Cover Layers—Collection 2. *Remote Sensing*, *12*(6). https://doi.org/10.3390/rs12061044

Cazzolla Gatti, R., Reich, P. B., Gamarra, J. G. P., Crowther, T., Hui, C., Morera, A., Bastin, J.-F., de-Miguel, S., Nabuurs, G.-J., Svenning, J.-C., Serra-Diaz, J. M., Merow, C., Enquist, B., Kamenetsky, M., Lee, J., Zhu, J., Fang, J., Jacobs, D. F., Pijanowski, B., … Liang, J. (2022). The number of tree species on Earth. *Proceedings of the National Academy of Sciences*, *119*(6), e2115329119. https://doi.org/10.1073/pnas.2115329119

Chave, J., Davies, S. J., Phillips, O. L., Lewis, S. L., Sist, P., Schepaschenko, D., Armston, J., Baker, T. R., Coomes, D., Disney, M., Duncanson, L., Hérault, B., Labrière, N., Meyer, V., Réjou-Méchain, M., Scipal, K., & Saatchi, S. (2019). Ground Data are Essential for Biomass Remote Sensing Missions. *Surveys in Geophysics*, *40*(4), 863–880. https://doi.org/10.1007/s10712-019-09528-w

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., … Vieilledent, G. (2014). Improved allometric models to estimate the aboveground biomass of tropical trees. *Global Change Biology*, *20*, 3177–3190. http://dx.doi.org/10.1111/gcb.12629

Chazdon, R. L., Broadbent, E. N., Rozendaal, D. M. A., Bongers, F., Zambrano, A. M. A., Aide, T. M., Balvanera, P., Becknell, J. M., Boukili, V., Brancalion, P. H. S., Craven, D., Almeida-Cortez, J. S., Cabral, G. A. L., de Jong, B., Denslow, J. S., Dent, D. H., DeWalt, S. J., Dupuy, J. M., Durán, S. M., … Poorter, L. (2016). Carbon sequestration potential of second-growth forest regeneration in the Latin American

tropics. *Science Advances*, *2*(5), e1501639. https://doi.org/10.1126/sciadv.1501639

Cleverly, J., Eamus, D., Edwards, W., Grant, M., Grundy, M. J., Held, A., Karan, M., Lowe, A. J., Prober, S. M., Sparrow, B., & Morris, B. (2019). TERN, Australia's land observatory: Addressing the global challenge of forecasting ecosystem responses to climate variability and change. *Environmental Research Letters*, *14*(9), 095004. https://doi.org/10.1088/1748-9326/ab33cb

Cox, C. B., Moore, P. D., & Ladle, R. J. (2016). *Biogeography: An ecological and evolutionary approach*. John Wiley & Sons.

Davies, S. J., Abiem, I., Abu Salim, K., Aguilar, S., Allen, D., Alonso, A., Anderson-Teixeira, K., Andrade, A., Arellano, G., Ashton, P. S., Baker, P. J., Baker, M. E., Baltzer, J. L., Basset, Y., Bissiengou, P., Bohlman, S., Bourg, N. A., Brockelman, W. Y., Bunyavejchewin, S., … Zuleta, D. (2021). ForestGEO: Understanding forest diversity and dynamics through a global observatory network. *Biological Conservation*, *253*, 108907. https://doi.org/10.1016/j.biocon.2020.108907

de Lima, R. A. F., Phillips, O. L., Duque, A., Tello, J. S., Davies, S. J., de Oliveira, A. A., Muller, S., Honorio Coronado, E. N., Vilanova, E., Cuni-Sanchez, A., Baker, T. R., Ryan, C. M., Malizia, A., Lewis, S. L., ter Steege, H., Ferreira, J., Marimon, B. S., Luu, H. T., Imani, G., … Vásquez, R. (2022). Making forest data fair and open. *Nature Ecology & Evolution*. https://doi.org/10.1038/s41559-022-01738-7

Dinerstein, E., Olson, D., Joshi, A., Vynne, C., Burgess, N. D., Wikramanayake, E., Hahn, N., Palminteri, S., Hedao, P., Noss, R., Hansen, M., Locke, H., Ellis, E. C., Jones, B., Barber, C. V., Hayes, R., Kormos, C., Martin, V., Crist, E., … Saleem, M. (2017). An Ecoregion-Based Approach to Protecting Half the Terrestrial Realm. *BioScience*, *67*(6), 534–545. https://doi.org/10.1093/biosci/bix014

Dowle, M., & Srinivasan, A. (2021). *Data.table: Extension of `data.frame`. R package version 1.14.2. Https://CRAN.R-project.org/package=data.table*.

Draper, F. C., Baraloto, C., Brodrick, P. G., Phillips, O. L., Martinez, R. V., Honorio Coronado, E. N., Baker, T. R., Zárate Gómez, R., Amasifuen Guerra, C. A., Flores,

M., Garcia Villacorta, R., V. A. Fine, P., Freitas, L., Monteagudo-Mendoza, A., J. W Brienen, R., & Asner, G. P. (2019). Imaging spectroscopy predicts variable distance decay across contrasting Amazonian tree communities. *Journal of Ecology*, *107*(2), 696–710. https://doi.org/10.1111/1365-2745.13067

Dubayah, R. O., 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. *Science of Remote Sensing*, *1*, 100002. https://doi.org/10.1016/j.srs.2020.100002

Dubayah, R. O., Luthcke, S. B., Sabaka, T. J., Nicholas, J. B., Preaux, S., & Hofton, M. A. (2021). *GEDI L3 Gridded Land Surface Metrics, Version 2*. https://doi.org/10.3334/ORNLDAAC/1952

Duncanson, L., Armston, J., Disney, M., Avitabile, V., Barbier, N., Calders, K., Carter, S., Chave, J., Herold, M., Crowther, T. W., Falkowski, M., Kellner, J. R., Labrière, N., Lucas, R., MacBean, N., McRoberts, R. E., Meyer, V., Næsset, E., Nickeson, J. E., … Williams, M. (2019). The Importance of Consistent Global Forest Aboveground Biomass Product Validation. *Surveys in Geophysics*, *40*(4), 979–999. https://doi.org/10.1007/s10712-019-09538-8

Duncanson, L., Armston, J., Disney, M., Avitabile, V., Barbier, N., Calders, K., Carter, S., Chave, J., Herold, M., MacBean, N., McRoberts, R., Minor, D., Paul, K., Réjou-Méchain, M., Roxburgh, S., Williams, M., Albinet, C., Baker, T., Bartholomeus, H., … Margolis, H. (2021). *Aboveground Woody Biomass Product Validation Good Practices Protocol. Version 1.0.* (L. Duncanson, M. Disney, J. Armston, J. Nickeson, D. Minor, and F. Camacho (Eds.)). Good Practices for Satellite Derived Land Product Validation, (p. 236): Land Product Validation Subgroup (WGCV/CEOS), doi:10.5067/doc/ceoswgcv/lpv/agb.001.

Duncanson, L., Kellner, J. R., Armston, J., Dubayah, R., Minor, D. M., Hancock, S., Healey,

S. P., Patterson, P. L., Saarela, S., Marselis, S., Silva, C. E., Bruening, J., Goetz, S. J., Tang, H., Hofton, M., Blair, B., Luthcke, S., Fatoyinbo, L., Abernethy, K., … Zgraggen, C. (2022). Aboveground biomass density models for NASA's Global Ecosystem Dynamics Investigation (GEDI) lidar mission. *Remote Sensing of Environment*, *270*, 112845. https://doi.org/10.1016/j.rse.2021.112845

ForestPlots.net, Blundo, C., Carilla, J., Grau, R., Malizia, A., Malizia, L., Osinaga-Acosta, O., Bird, M., Bradford, M., Catchpole, D., Ford, A., Graham, A., Hilbert, D., Kemp, J., Laurance, S., Laurance, W., Ishida, F. Y., Marshall, A., Waite, C., … Tran, H. D. (2021). Taking the pulse of Earth's tropical forests using networks of highly distributed plots. *Biological Conservation*, *260*, 108849. https://doi.org/10.1016/j.biocon.2020.108849

Grantham, H. S., Duncan, A., Evans, T. D., Jones, K. R., Beyer, H. L., Schuster, R., Walston, J., Ray, J. C., Robinson, J. G., Callow, M., Clements, T., Costa, H. M., DeGemmis, A., Elsen, P. R., Ervin, J., Franco, P., Goldman, E., Goetz, S., Hansen, A., … Watson, J. E. M. (2020). Anthropogenic modification of forests means only 40% of remaining forests have high ecosystem integrity. *Nature Communications*, *11*(1), 5978. https://doi.org/10.1038/s41467-020-19493-3

Greenberg, J. A., & Mattiuzzi, M. (2020). *GdalUtils: Wrappers for the Geospatial Data Abstraction Library (GDAL) Utilities. R package version 2.0.3.2. Https://CRAN.R-project.org/package=gdalUtils*.

Harris, N. L., Gibbs, D. A., Baccini, A., Birdsey, R. A., de Bruin, S., Farina, M., Fatoyinbo, L., Hansen, M. C., Herold, M., Houghton, R. A., Potapov, P. V., Suarez, D. R., Roman-Cuesta, R. M., Saatchi, S. S., Slay, C. M., Turubanova, S. A., & Tyukavina, A. (2021). Global maps of twenty-first century forest carbon fluxes. *Nature Climate Change*, *11*(3), 234–240. https://doi.org/10.1038/s41558-020-00976-6

Heinrich, V. H. A., Dalagnol, R., Cassol, H. L. G., Rosan, T. M., de Almeida, C. T., Silva Junior, C. H. L., Campanharo, W. A., House, J. I., Sitch, S., Hales, T. C., Adami, M., Anderson, L. O., & Aragão, L. E. O. C. (2021). Large carbon sink potential of

secondary forests in the Brazilian Amazon to mitigate climate change. *Nature Communications*, *12*(1), 1785. https://doi.org/10.1038/s41467-021-22050-1

Hijmans, R. J. (2021). *Raster: Geographic Data Analysis and Modeling. R package version 3.5-2. Https://CRAN.R-project.org/package=raster*.

Hoffman, F. M., Kumar, J., Mills, R. T., & Hargrove, W. W. (2013). Representativeness-based sampling network design for the State of Alaska. *Landscape Ecology*, *28*(8), 1567–1586. https://doi.org/10.1007/s10980-013-9902-0

Holdridge, L. R. (1947). Determination of World Plant Formations From Simple Climatic Data. *Science*, *105*(2727), 367–368. https://doi.org/10.1126/science.105.2727.367

Hubau, W., Lewis, S. L., Phillips, O. L., Affum-Baffoe, K., Beeckman, H., Cuní-Sanchez, A., Daniels, A. K., Ewango, C. E. N., Fauset, S., Mukinzi, J. M., Sheil, D., Sonké, B., Sullivan, M. J. P., Sunderland, T. C. H., Taedoumg, H., Thomas, S. C., White, L. J. T., Abernethy, K. A., Adu-Bredu, S., … Zemagho, L. (2020). Asynchronous carbon sink saturation in African and Amazonian tropical forests. *Nature*, *579*(7797), 80–87. https://doi.org/10.1038/s41586-020-2035-0

Hulshof, C. M., & Spasojevic, M. J. (2020). The edaphic control of plant diversity. *Global Ecology and Biogeography*, *29*(10), 1634–1650. https://doi.org/10.1111/geb.13151

Jucker, T., Bongalov, B., Burslem, D. F. R. P., Nilus, R., Dalponte, M., Lewis, S. L., Phillips, O. L., Qie, L., & Coomes, D. A. (2018). Topography shapes the structure, composition and function of tropical forest landscapes. *Ecology Letters*, *21*(7), 989–1000. https://doi.org/10.1111/ele.12964

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. John Wiley & Sons.

Keil, P., & Chase, J. M. (2019). Global patterns and drivers of tree diversity integrated across a continuum of spatial grains. *Nature Ecology & Evolution*, *3*(3), 390–399. https://doi.org/10.1038/s41559-019-0799-0

Labrière, N., Tao, S., Chave, J., Scipal, K., Toan, T. L., Abernethy, K., Alonso, A., Barbier, N., Bissiengou, P., Casal, T., Davies, S. J., Ferraz, A., Hérault, B., Jaouen, G.,

Jeffery, K. J., Kenfack, D., Korte, L., Lewis, S. L., Malhi, Y., … Saatchi, S. (2018). In Situ Reference Datasets From the TropiSAR and AfriSAR Campaigns in Support of Upcoming Spaceborne Biomass Missions. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *11*(10), 3617–3627. https://doi.org/10.1109/JSTARS.2018.2851606

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2021). *Cluster: Cluster Analysis Basics and Extensions. R package version 2.1.2.*

Markus, T., Neumann, T., Martino, A., Abdalati, W., Brunt, K., Csatho, B., Farrell, S., Fricker, H., Gardner, A., Harding, D., Jasinski, M., Kwok, R., Magruder, L., Lubin, D., Luthcke, S., Morison, J., Nelson, R., Neuenschwander, A., Palm, S., … Zwally, J. (2017). The Ice, Cloud, and land Elevation Satellite-2 (ICESat-2): Science requirements, concept, and implementation. *Remote Sensing of Environment*, *190*, 260–273. https://doi.org/10.1016/j.rse.2016.12.029

Marselis, S. M., Tang, H., Armston, J. D., Calders, K., Labrière, N., & Dubayah, R. (2018). Distinguishing vegetation types with airborne waveform lidar data in a tropical forest-savanna mosaic: A case study in Lopé National Park, Gabon. *Remote Sensing of Environment*, *216*, 626–634. https://doi.org/10.1016/j.rse.2018.07.023

McDowell, N. G., Allen, C. D., Anderson-Teixeira, K., Aukema, B. H., Bond-Lamberty, B., Chini, L., Clark, J. S., Dietze, M., Grossiord, C., Hanbury-Brown, A., Hurtt, G. C., Jackson, R. B., Johnson, D. J., Kueppers, L., Lichstein, J. W., Ogle, K., Poulter, B., Pugh, T. A. M., Seidl, R., … Xu, C. (2020). Pervasive shifts in forest dynamics in a changing world. *Science*, *368*(6494), eaaz9463. https://doi.org/10.1126/science.aaz9463

McMichael, C. N. H., Matthews-Bird, F., Farfan-Rios, W., & Feeley, K. J. (2017). Ancient human disturbances may be skewing our understanding of Amazonian forests. *Proceedings of the National Academy of Sciences*, *114*(3), 522. https://doi.org/10.1073/pnas.1614577114

Metzger, S., Ayres, E., Durden, D., Florian, C., Lee, R., Lunch, C., Luo, H., Pingintha-

Durden, N., Roberti, J. A., SanClements, M., Sturtevant, C., Xu, K., & Zulueta, R. C. (2019). From NEON Field Sites to Data Portal: A Community Resource for Surface–Atmosphere Research Comes Online. *Bulletin of the American Meteorological Society*, *100*(11), 2305–2325. https://doi.org/10.1175/BAMS-D-17-0307.1

Mucina, L. (2019). Biome: Evolution of a crucial ecological and biogeographical concept. *New Phytologist*, *222*(1), 97–114. https://doi.org/10.1111/nph.15609

NISAR. (2018). *NASA-ISRO SAR (NISAR) Mission Science Users' Handbook. NASA Jet Propulsion Laboratory. 261 pp*.

O'Donnell, M. S., & Ignizio, D. A. (2012). *Bioclimatic predictors for supporting ecological applications in the conterminous United States: U.S. Geological Survey Data Series 691, 10 p.* https://pubs.usgs.gov/ds/691/ds691.pdf

Pan, Y., Birdsey, R. A., Fang, J. Y., Houghton, R. A., Kauppi, P. E., Kurz, W. A., Phillips, O. L., Shvidenko, A., Lewis, S. L., Canadell, J. G., Ciais, P., Jackson, R. B., Pacala, S. W., McGuire, A. D., Piao, S. L., Rautiainen, A., Sitch, S., & Hayes, D. (2011). A Large and Persistent Carbon Sink in the World's Forests. *Science*, *333*, 988–993. http://dx.doi.org/10.1126/science.1201609

Pan, Y., Birdsey, R. A., Phillips, O. L., & Jackson, R. B. (2013). The Structure, Distribution, and Biomass of the World's Forests. *Annual Review of Ecology, Evolution, and Systematics*, *44*(1), 593–622. https://doi.org/10.1146/annurev-ecolsys-110512-135914

Phillips, O. L., Sullivan, M. J. P., Baker, T. R., Monteagudo Mendoza, A., Vargas, P. N., & Vásquez, R. (2019). Species Matter: Wood Density Influences Tropical Forest Biomass at Multiple Scales. *Surveys in Geophysics*, *40*(4), 913–935. https://doi.org/10.1007/s10712-019-09540-0

Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., & Rossiter, D. (2021). SoilGrids 2.0: Producing soil information for the globe with quantified spatial uncertainty. *SOIL*, *7*(1), 217–240. https://doi.org/10.5194/soil-7-217-2021

Poorter, L., Craven, D., Jakovac, C. C., van der Sande, M. T., Amissah, L., Bongers, F., Chazdon, R. L., Farrior, C. E., Kambach, S., Meave, J. A., Muñoz, R., Norden, N., Rüger, N., van Breugel, M., Almeyda Zambrano, A. M., Amani, B., Andrade, J. L., Brancalion, P. H. S., Broadbent, E. N., … Hérault, B. (2021). Multidimensional tropical forest recovery. *Science*, *374*(6573), 1370–1376. https://doi.org/10.1126/science.abh3629

Prentice, I. C., Harrison, S. P., & Bartlein, P. J. (2011). Global vegetation and terrestrial carbon cycle changes after the last ice age. *New Phytologist*, *189*(4), 988–998. https://doi.org/10.1111/j.1469-8137.2010.03620.x

Quegan, S., Le Toan, T., Chave, J., Dall, J., Exbrayat, J.-F., Minh, D. H. T., Lomas, M., D'Alessandro, M. M., Paillou, P., Papathanassiou, K., Rocca, F., Saatchi, S., Scipal, K., Shugart, H., Smallman, T. L., Soja, M. J., Tebaldini, S., Ulander, L., Villard, L., & Williams, M. (2019). The European Space Agency BIOMASS mission: Measuring forest above-ground biomass from space. *Remote Sensing of Environment*, *227*, 44–60. https://doi.org/10.1016/j.rse.2019.03.032

R Core Team. (2021). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/*.

Réjou-Méchain, M., Muller-Landau, H. C., Detto, M., Thomas, S. C., Le Toan, T., Saatchi, S. S., Barreto-Silva, J. S., Bourg, N. A., Bunyavejchewin, S., Butt, N., Brockelman, W. Y., Cao, M., Cárdenas, D., Chiang, J. M., Chuyong, G. B., Clay, K., Condit, R., Dattaraja, H. S., Davies, S. J., … Chave, J. (2014). Local spatial structure of forest biomass and its consequences for remote sensing of carbon stocks. *Biogeosciences*, *11*, 6827–6840. https://doi.org/10.5194/bg-11-6827-2014

Requena Suarez, D., Rozendaal, D. M. A., De Sy, V., Phillips, O. L., Alvarez-Dávila, E., Anderson-Teixeira, K., Araujo-Murakami, A., Arroyo, L., Baker, T. R., Bongers, F., Brienen, R. J. W., Carter, S., Cook-Patton, S. C., Feldpausch, T. R., Griscom, B. W., Harris, N., Hérault, B., Honorio Coronado, E. N., Leavitt, S. M., … Herold, M. (2019). Estimating aboveground net biomass change for tropical and subtropical forests:

Refinement of IPCC default rates using forest plot data. *Global Change Biology*, *25*(11), 3609–3624. https://doi.org/10.1111/gcb.14767

Reynolds, A. P., Richards, G., de la Iglesia, B., & Rayward-Smith, V. J. (2006). Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms. *Journal of Mathematical Modelling and Algorithms*, *5*(4), 475–504. https://doi.org/10.1007/s10852-005-9022-1

Santoro, M., Cartus, O., Carvalhais, N., Rozendaal, D. M. A., Avitabile, V., Araza, A., de Bruin, S., Herold, M., Quegan, S., Rodríguez-Veiga, P., Balzter, H., Carreiras, J., Schepaschenko, D., Korets, M., Shimada, M., Itoh, T., Moreno Martínez, Á., Cavlovic, J., Cazzolla Gatti, R., … Willcock, S. (2021). The global forest above-ground biomass pool for 2010 estimated from high-resolution satellite observations. *Earth System Science Data*, *13*(8), 3927–3950. https://doi.org/10.5194/essd-13-3927-2021

Schepaschenko, D., Chave, J., Phillips, O. L., Lewis, S. L., Davies, S. J., Réjou-Méchain, M., Sist, P., Scipal, K., Perger, C., Herault, B., Labrière, N., Hofhansl, F., Affum-Baffoe, K., Aleinikov, A., Alonso, A., Amani, C., Araujo-Murakami, A., Armston, J., Arroyo, L., … Zo-Bi, I. C. (2019). The Forest Observation System, building a global reference dataset for remote sensing of forest biomass. *Scientific Data*, *6*(1), 198. https://doi.org/10.1038/s41597-019-0196-1

Schepaschenko, D., Shvidenko, A., Usoltsev, V., Lakyda, P., Luo, Y., Vasylyshyn, R., Lakyda, I., Myklush, Y., See, L., McCallum, I., Fritz, S., Kraxner, F., & Obersteiner, M. (2017). A dataset of forest biomass structure for Eurasia. *Scientific Data*, *4*(1), 170070. https://doi.org/10.1038/sdata.2017.70

Schimel, D., Stephens, B. B., & Fisher, J. B. (2015). Effect of increasing CO_2 on the terrestrial carbon cycle. *Proceedings of the National Academy of Sciences*, *112*(2), 436. https://doi.org/10.1073/pnas.1407302112

Schubert, E., & Rousseeuw, P. J. (2021). Fast and eager k-medoids clustering: O(k) runtime improvement of the PAM, CLARA, and CLARANS algorithms. *Information Systems*,

101, 101804. https://doi.org/10.1016/j.is.2021.101804

Sist, P., Rutishauser, E., Peña-Claros, M., Shenkin, A., Hérault, B., Blanc, L., Baraloto, C., Baya, F., Benedet, F., da Silva, K. E., Descroix, L., Ferreira, J. N., Gourlet-Fleury, S., Guedes, M. C., Bin Harun, I., Jalonen, R., Kanashiro, M., Krisnawati, H., Kshatriya, M., … Yamada, T. (2015). The Tropical managed Forests Observatory: A research network addressing the future of tropical logged forests. *Applied Vegetation Science*, *18*(1), 171–174. https://doi.org/10.1111/avsc.12125

Spawn, S. A., Sullivan, C. C., Lark, T. J., & Gibbs, H. K. (2020). Harmonized global maps of above and belowground biomass carbon density in the year 2010. *Scientific Data*, *7*(1), 112. https://doi.org/10.1038/s41597-020-0444-4

Sullivan, M. J. P., Lewis, S. L., Affum-Baffoe, K., Castilho, C., Costa, F., Sanchez, A. C., Ewango, C. E. N., Hubau, W., Marimon, B., Monteagudo-Mendoza, A., Qie, L., Sonké, B., Martinez, R. V., Baker, T. R., Brienen, R. J. W., Feldpausch, T. R., Galbraith, D., Gloor, M., Malhi, Y., … Phillips, O. L. (2020). Long-term thermal sensitivity of Earth's tropical forests. *Science*, *368*(6493), 869. https://doi.org/10.1126/science.aaw7578

The SEOSAW partnership. (2020). A network to understand the changing socio-ecology of the southern African woodlands (SEOSAW): Challenges, benefits, and methods. *Plants, People, Planet*. https://doi.org/10.1002/ppp3.10168

van Marle, M. J. E., van Wees, D., Houghton, R. A., Field, R. D., Verbesselt, J., & van der Werf, Guido. R. (2022). New land-use-change emissions indicate a declining CO2 airborne fraction. *Nature*, *603*(7901), 450–454. https://doi.org/10.1038/s41586-021-04376-4

von Humboldt, A., & Bonpland, A. (1805). *Essai sur la géographie des plantes; accompagné d'un tableau physique des régions équinoxiales*. Paris, France: Levrault, Schoell et Compagnie.

Whittaker, R. H. (1975). *Communities and Ecosystems*. Macmillan.

Supplementary information

Study area

The forest mask was built using land cover data for 2020 from the ESA CCI Land Cover project. Original data were downloaded from the Climate Data Store (CDS) of the Copernicus Climate Change Service (C3S; https://cds.climate.copernicus.eu/#!/home). The entire dataset is called "Land cover classification gridded maps from 1992 to present derived from satellite observations", and data can be downloaded on a yearly basis. To build the forest mask, only the following classes were retained: "Tree cover, broadleaved, evergreen, closed to open (>15%)" (class **50**), "Tree cover, broadleaved, deciduous, closed to open (>15%)" (class **60**), "Tree cover, broadleaved, deciduous, closed (>40%)" (class **61**), "Tree cover, broadleaved, deciduous, open (15-40%)" (class **62**), "Tree cover, needleleaved, evergreen, closed to open (>15%)" (class **70**), "Tree cover, needleleaved, evergreen, closed (>40%)" (class **71**), "Tree cover, needleleaved, evergreen, open (15-40%)" (class **72**), "Tree cover, needleleaved, deciduous, closed to open (>15%)" (class **80**), "Tree cover, needleleaved, deciduous, closed (>40%)" (class **81**), "Tree cover, needleleaved, deciduous, open (15-40%)" (class **82**), "Tree cover, mixed leaf type (broadleaved and needleleaved)" (class **90**), "Tree cover, flooded, fresh or brakish water" (class **160**), and "Tree cover, flooded, saline water" (class **170**).

Maximum environmental, geographical and structural distances

We computed Euclidean distances on ten environmental and two structural variables. Considering for each set of variables minimum and maximum values reached over global forested areas would allow to compute theoretical extreme distances. For structural variables, i.e. canopy height and tree cover fraction, while it is plausible that a single cell has maxima for both variables and another cell has both minima, this is unlikely for environmental variables where multiple climatic, topographic and edaphic variables are under consideration. Therefore, we searched for realized extreme, i.e. minimum and maximum, distances instead of theoretical ones. First, we selected 1,000 cells geographically spanning the study area (different sampling was used for the search of environmental and structural extremes, as the pool of cells was different given GEDI discrete sampling and ISS-orbit limited spatial coverage). Second, for each of those we computed their distance from each other cell over global forested areas, and retained extreme values along with the cell for which each extreme was obtained. Third, we built occurrence tables of those cells involved in extreme distances (4 occurrence tables in total, for minimum/maximum values for environmental/structural conditions). Last, we repeated step 2 for the 15 most common cells identified in each occurrence table and final realized extreme distances were identified consequently. For geographical space, minimum and maximum distances were set to 0 km and half a great circle (ca. 20,037.51 km), respectively.

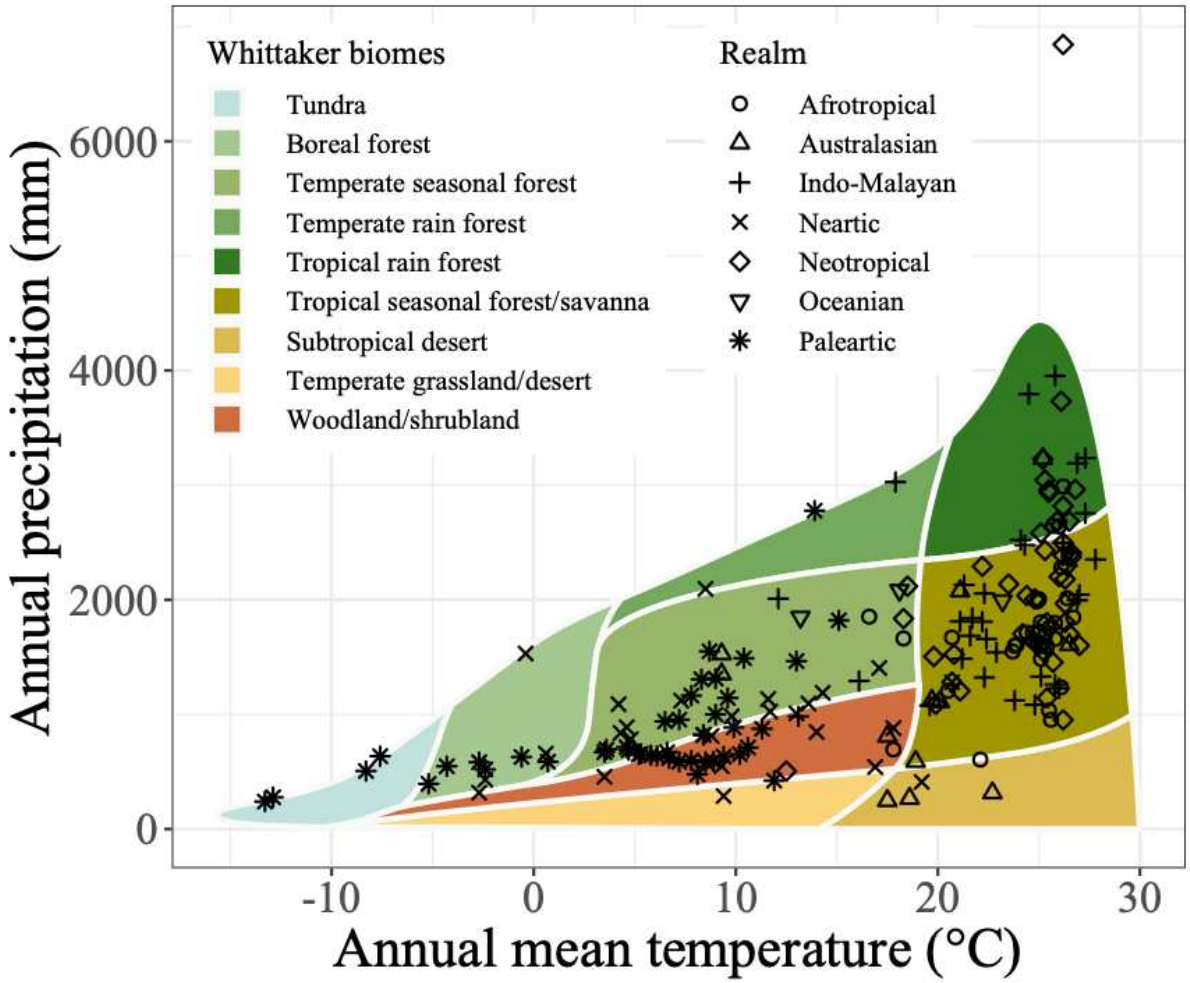

Figure S1. Whittaker diagram showing biome classification and delineation *sensu* Whittaker (1975) as a function of annual precipitation and mean temperature. All the forest biomass reference measurement sites (n = 195) are also displayed, along with information on their respective realm (realm borders obtained from Dinerstein et al., 2017). Note that one potential forest biomass reference measurement site is located in the Colombian part of the "Chocó–Darién moist forests" ecoregion, one of the wettest regions of Earth (annual precipitation > 6,000 mm).

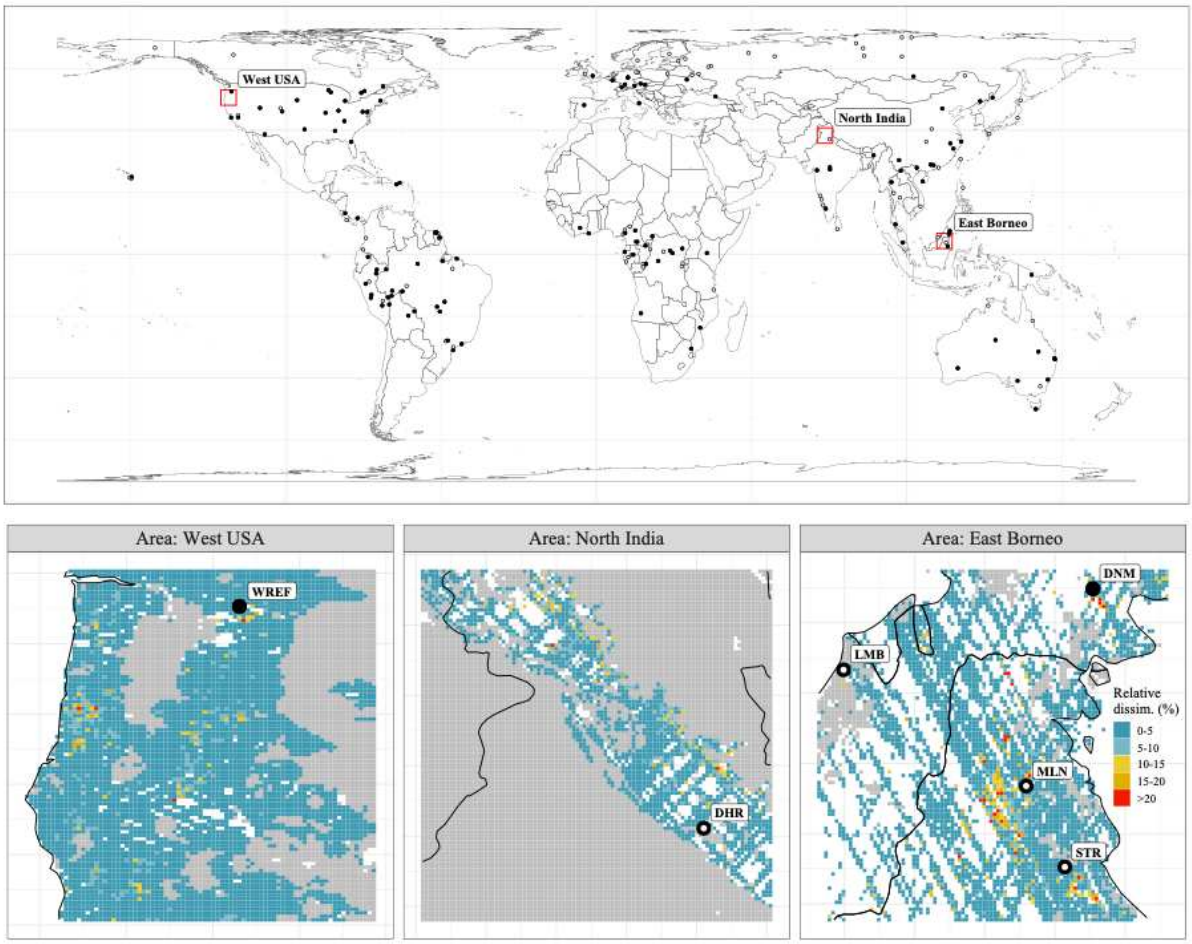

Figure S2. Location (top) and relative structural dissimilarity (bottom) for a subset of forested areas with insufficient coverage (relative dissimilarity > 10%, bottom) with respect to conditions covered by potential forest biomass reference measurement (FBRM) sites. Blank continental areas and hollow points (bottom) respectively correspond to forested areas and sites not sampled (yet) by GEDI, and each facet displays a 500 × 500 km area. Relative dissimilarity was categorized for display purposes. Non-forested areas are in grey. The map projection is EASE-Grid 2.0 (epsg:6933), a global, equal-area protection, and spatial resolution is 5 km.

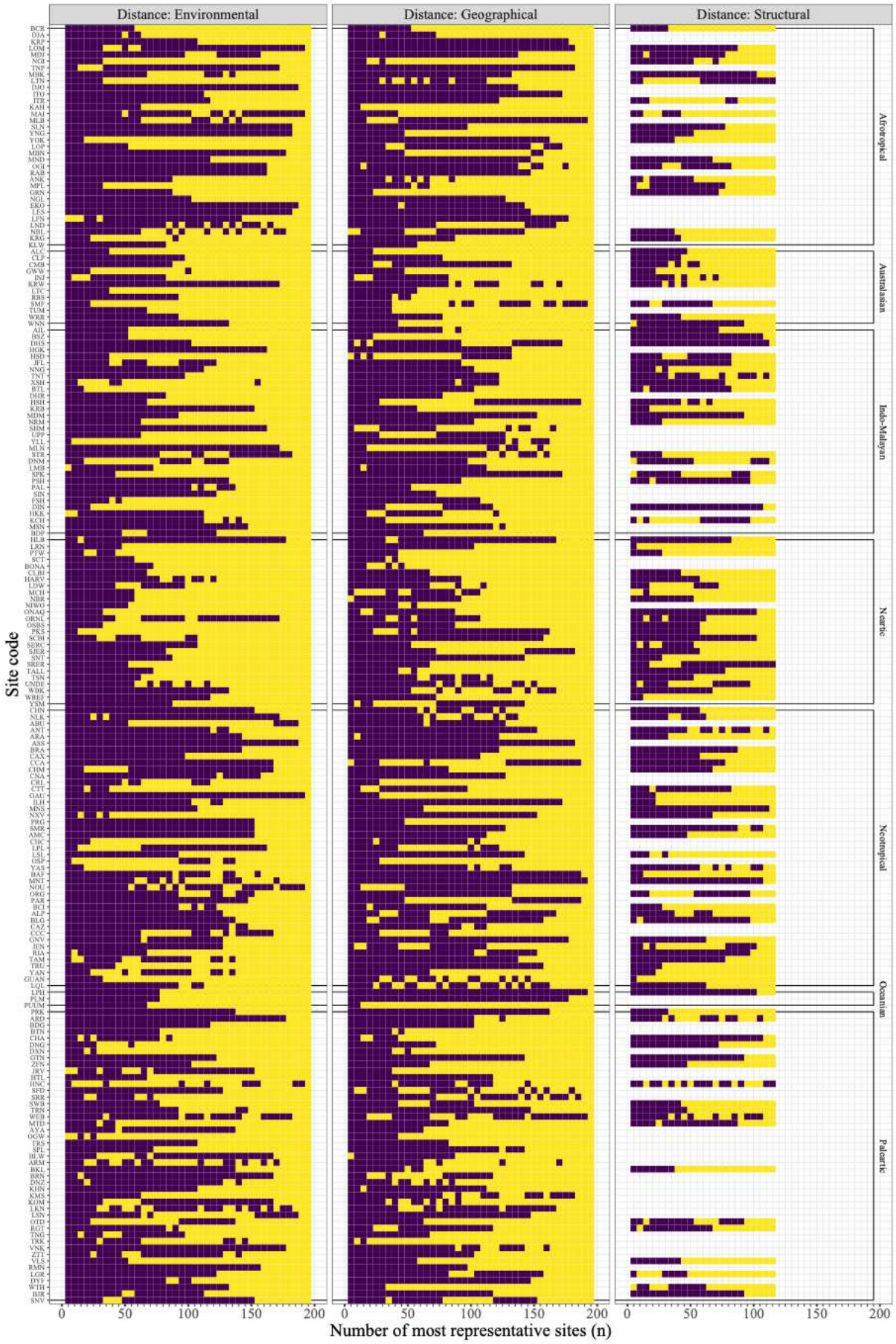
Distance: Environmental
Distance: Geographical
Distance: Structural
Site code
Afrotropical
Australasian
Indo-Malayan
N.eartic
Neotropical
Oceanian
Palearctic
0
50
100
150
200
Number of most representative sites (n)

Figure S3. Most representative potential forest biomass reference measurement sites for different types of distances. Selected and non-selected sites are in light and dark colors, respectively. For display purposes, only numbers, n, of most representative sites that are multiples of 5 are used here. Sites are ordered by biome (biome borders obtained from Dinerstein et al., 2017). The correspondence between site codes and names can be found in Table S1.

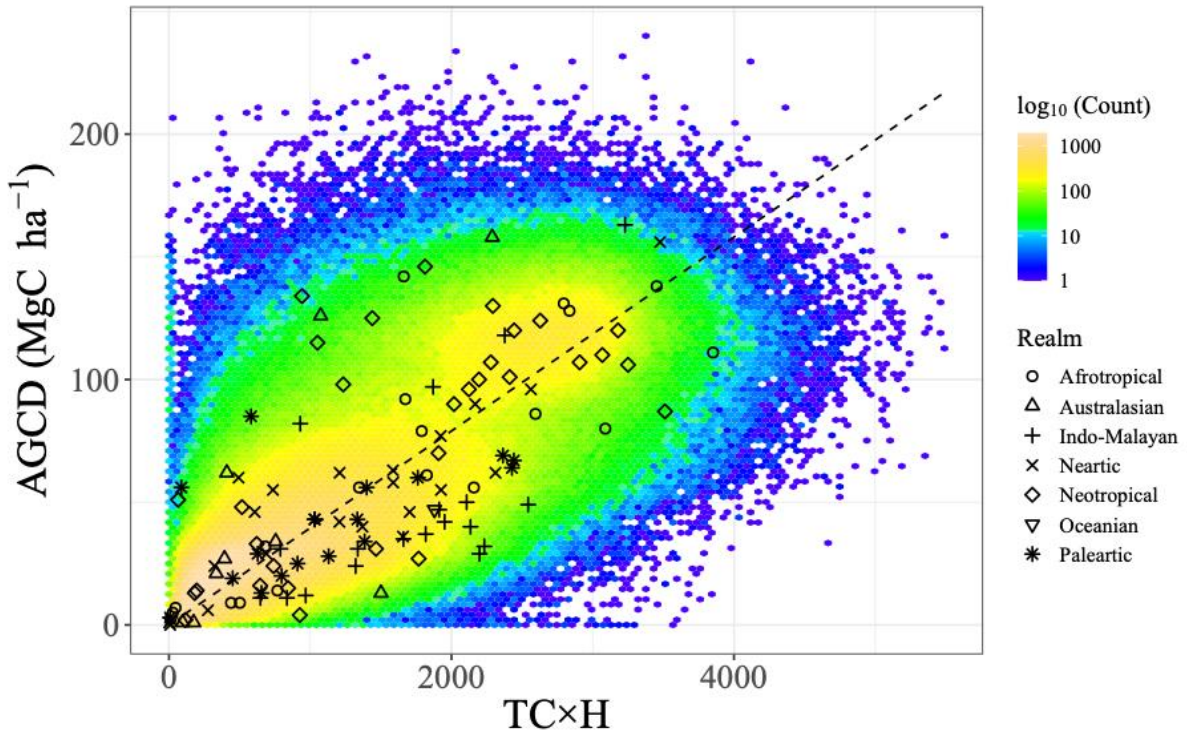

Figure S4. Aboveground biomass carbon density (AGCD; MgC ha^{-1}) vs. tree cover fraction (TC; %) times canopy height (H; m). AGCD estimates were obtained from Spawn et al. (2020). TC and H were derived from Proba-V and GEDI observations, respectively. Data from the 829,256 contributing cells (5 km spatial resolution) were binned together for display purposes (100 bins on both axes). The dashed line represents a linear regression forced through zero. All the potential biomass reference measurement sites with GEDI information (n = 118) are also displayed, along with information on their respective realm (realm borders obtained from Dinerstein et al., 2017).

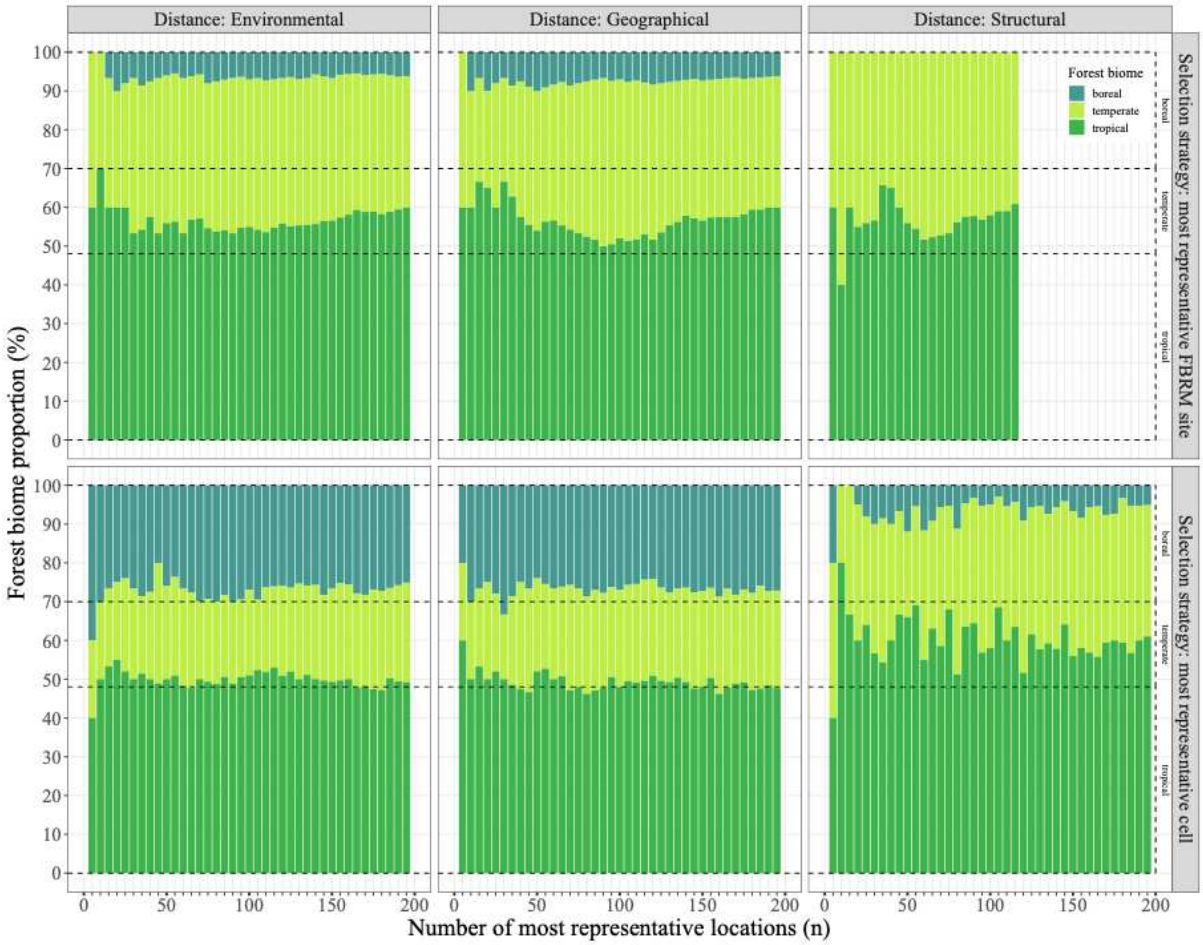

Figure S5. Forest biome proportion (%) vs. number of most representative locations for different types of distances and selection strategies. For display purposes, only numbers n of most representative locations multiple of 5 were used here. Forest biomes *sensu* Dinerstein et al. (2017) were classified as tropical, temperate or boreal. Dashed lines indicate areal forest biome proportions of global forested areas (48, 22 and 30% for tropical, temperate and boreal forest biomes, respectively).

Table S1. Potential forest biomass reference measurement site coordinator, location, plot cumulative area, structural and environmental attributes, and inclusion (+) or not (-) in the set of the 150, 100 and 50 most representative sites. Attribute values were extracted at each site location from the corresponding layers produced at 5 km spatial resolution. Cum. area = plot cumulative area (ha), Year = established year of (first) plot establishment at the site (yr), H = canopy height (m), TC = tree cover fraction (%), AGCD = aboveground biomass carbon density (MgC ha^{-1}; [range]), AMT = annual mean temperature (°C), TSE = temperature seasonality (% coefficient of variation CV), APR = annual precipitation (mm), PSE = precipitation seasonality (% CV), SRAD = downward surface shortwave radiation (W m^{-2}), Elevation = elevation above sea level (m), CFVO = coarse fragments (% volume), Sand = sand fraction (% mass), CEC = cation exchange capacity (cmol kg^{-1}), pH (in H_2O) (unitless).

Table S2. Partitioning of potential forest biomass reference measurement sites depending on number of sites, n, and space (environmental, geographical, structural). Only numbers of sites which are multiples of 5 are used here. Codes correspond to partition medoids.
For a given column, all rows with the same code belong to the same partition.