ENVIRONMENTAL RESEARCH
LETTERS

LETTER • OPEN ACCESS

Estimating aboveground biomass density using hybrid statistical inference with GEDI lidar data and Paraguay's national forest inventory

To cite this article: Eric L Bullock *et al* 2023 *Environ. Res. Lett.* **18** 085001

ENVIRONMENTAL RESEARCH
LETTERS

LETTER

Estimating aboveground biomass density using hybrid statistical inference with GEDI lidar data and Paraguay's national forest inventory

Eric L Bullock[1,*], **Sean P Healey**[1], **Zhiqiang Yang**[1], **Regino Acosta**[2], **Hermelinda Villalba**[2], **Katherin Patricia Insfrán**[2], **Joana B Melo**[3], **Sylvia Wilson**[4], **Laura Duncanson**[5], **Erik Næsset**[6], **John Armston**[5], **Svetlana Saarela**[6], **Göran Ståhl**[7], **Paul L Patterson**[8] **and Ralph Dubayah**[5]

[1] US Forest Service, Rocky Mountain Research Station, Riverdale, UT 84405, United States of America
[2] Instituto Forestal Nacional (INFONA), San Lorenzo, Paraguay
[3] Joint Research Centre, European Commission, Ispra 21027, Italy
[4] United States Geological Survey, Reston, VA 20192, United States of America
[5] Department of Geographical Sciences, University of Maryland, College Park, MD 20742, United States of America
[6] Faculty of Environmental Sciences and Natural Resource Management, Norwegian University of Life Sciences, Ås NO-1432, Norway
[7] Department of Forest Resource Management, Swedish University of Agricultural Sciences, Umeå, Sweden
[8] USDA Forest Service, Rocky Mountain Research Station, Fort Collins, CO 80526, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: eric.bullock@usda.gov

Keywords: lidar, GEDI, biomass, carbon stocks, greenhouse gas inventories, hybrid inference, forest inventories

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
9 January 2023

REVISED
19 May 2023

ACCEPTED FOR PUBLICATION
16 June 2023

PUBLISHED
11 July 2023

Abstract

Forests are widely recognized as critical to combating climate change due to their ability to sequester and store carbon in the form of biomass. In recent years, the combined use of data from ground-based forest inventories and remotely sensed data from light detection and ranging (lidar) has proven useful for large-scale assessment of forest biomass, but airborne lidar is expensive and data acquisition is infeasible for many countries. By contrast, the spaceborne Global Ecosystem Dynamics Investigation (GEDI) lidar instrument has collected freely available data for most of the world's temperate and tropical forests since 2019. GEDI's biomass products rely on models calibrated with a global network of field plots paired with GEDI waveforms simulated from airborne lidar to predict biomass. While this calibration strategy minimizes spatial and temporal offsets between field measurements and corresponding lidar returns, calibration data are sparse in many regions. Paraguay's forests are known to be poorly represented in GEDI's current calibration dataset, and here we demonstrate that local models calibrated opportunistically with on-orbit GEDI data and field surveys from Paraguay's national forest inventory can be used with GEDI's statistical estimators of aboveground biomass density (AGBD). We specify a protocol for opportunistically matching GEDI observations with field plots to calibrate a field-to-GEDI biomass model for use in GEDI's hybrid statistical framework. Country-specific calibration using on-orbit data resulted in relatively accurate and unbiased predictions of footprint-level biomass, and importantly, supported the assumption underlying model-based inference that the model must 'apply' to the area of interest. Using a locally calibrated biomass model, we estimate that the mean AGBD in Paraguay is 65.55 Mg ha^{-1}, which coincides well with the design-based approach employed by the national forest inventory. The GEDI estimates for individual forest strata range from 52.34 Mg ha^{-1} to 103.88 Mg ha^{-1}. On average, the standard errors are 47% lower for estimates based on GEDI than the forest inventory, representing a significant gain in precision. Our research demonstrates that GEDI can be used by national forest inventories in countries that seek

reliable estimates of AGBD, and that local calibration using existing field plots may be more appropriate in some applications than using GEDI global models, especially in regions where those models are sparsely calibrated.

1. Introduction

Quantifying carbon stored in forest biomass is essential for understanding terrestrial carbon fluxes and for achieving the objectives set under the United Nations Framework Convention on Climate Change (UNFCCC). National forest inventories (NFIs) are critical for monitoring aboveground biomass density (AGBD) since they generally deploy probability-based sample design strategies that enable straightforward inference about biomass quantities over large areas (hereafter referred to as 'design-based' estimation) (Brown *et al* 1989, Schreuder *et al* 1993, Gregoire 1998). However, the global distribution of field plots remains limited, and many forests are not properly represented due to inaccessibility or financial constraints (de Freitas *et al* 2009, McRoberts *et al* 2010, 2013b).

Light detection and ranging (lidar) can complement field surveys for large-scale assessment of AGBD due to its high correlation with aboveground biomass (Lu 2006, Frolking *et al* 2009, Shugart *et al* 2010, Zolkos *et al* 2013, McRoberts *et al* 2013a, Ene *et al* 2018, Magnussen *et al* 2018), but airborne data acquisition is expensive and wall-to-wall coverage has not been seen as practical (McRoberts *et al* 2010, Wulder *et al* 2012, Lu *et al* 2016). The Global Ecosystem Dynamics Investigation (GEDI; Dubayah *et al* 2020a) mission was launched in 2018 and measures full-waveform lidar data across most of the world's tropical and temperate forests, offering an unprecedented opportunity to expand the use of lidar in forest inventories.

Ståhl *et al* (2016) surveyed methods for using technologies such as lidar in the context of statistically rigorous large-area forest assessments. If a probability sample of field plots exists, lidar can be used to increase precision under a model-assisted approach (e.g. Andersen *et al* 2009, Gregoire *et al* 2011). In cases where sufficient field data for calibrating a robust model using remote sensing data are available but do not constitute a probability sample, model-based approaches can be used to estimate mean biomass (e.g. McRoberts 2010, Chen *et al* 2016, McRoberts *et al* 2018). In the event that model-based methods are used, but ancillary remote sensing data cover only a sample of the area of interest, properties of that sample can be combined with properties of the model to estimate uncertainty of the estimated mean biomass (Ståhl *et al* 2011). This 'hybrid' approach was adopted for inference by the GEDI mission (Patterson *et al* 2019), which observes only a sample of Earth's land surface, and has been applied to GEDI data at multiple scales from 1 km grid cells to entire countries (Dubayah *et al* 2022a).

To our knowledge, all uses of GEDI data with hybrid inference have utilized GEDI's footprint-level models as described in Duncanson *et al* (2022) and Kellner *et al* (2023). Specifically, field biomass measurements from around the world are modeled using simulated GEDI waveforms produced from spatially and temporally coincident airborne lidar (Hancock *et al* 2019). The direct match of field and lidar data reduces dilution of precision that can occur with plots that are offset from, or a different size of, the remotely sensed measurement (Rejou-Mechain *et al* 2014, Labriere *et al* 2018). However, while GEDI takes steps to assess model transferability within a region, this approach limits calibration to the relatively small subset of global plots overflown with airborne lidar.

Vastly more field measurements could be used, and GEDI footprint-level models could be calibrated much more locally, if existing inventory plots could be matched opportunistically to on-orbit GEDI data. Almost all countries collect forest data of some kind, although in many cases available plot data do not comprise a probability sample (MacDicken 2015). As GEDI's continued operation gradually reduces the distance between any forest plot and its nearest GEDI shot, it becomes more relevant to ask if local model calibration can improve the precision and accuracy of biomass estimates made in the context of hybrid inference.

GEDI data are freely available, processed into analysis-ready formats, cover most of the world's tropical and temperate forests, and are readily accessible on platforms such as Google Earth Engine (Gorelick *et al* 2017, Healey *et al* 2020, Dubayah *et al* 2020a). If estimation based on locally calibrated GEDI data can provide comparable advantages to those achieved with airborne lidar, it could potentially benefit national greenhouse gas (GHG) inventories and other applications requiring estimates of AGBD. Therefore, our motivating research questions are:

1. Can on-orbit GEDI data, despite some spatial offset from available forest inventory plots, be used to create useful aboveground biomass models?
2. How does a biomass model constructed with NFI data compare to GEDI's standard (Level 4A) global footprint-level biomass predictions in an area where the Level 4A had little calibration data?

3. How do GEDI-based estimates of AGBD compare to estimates based on field data alone?

We explore these questions in Paraguay using the same hybrid inference framework employed by the GEDI mission. We develop a Paraguay-specific biomass model that uses on-orbit GEDI data and is calibrated with NFI data. Results are compared to the estimates obtained from the design-based framework used by Paraguay's NFI. Our research represents a step towards lidar-based estimation of forest biomass that is accessible to all countries with existing field surveys and within GEDI's spatial coverage.

2. Methodology

2.1. Study area

The study area is the South American country of Paraguay. The Chaco region in western Paraguay contains South America's second largest forest and is a global deforestation hotspot (Grau and Aide 2008). The country has the sixth highest deforestation rate in the world and the highest rate as a proportion of forest area in South America (FAO 2020). Domestically, the land use sector is the main contributor of GHG emissions (MADES-DNCC/PNUD-FMAM 2022). Therefore, it is not surprising that mitigation actions in the forest and land use sector are highlighted in the National Determined Contribution (NDC) submitted to the UNFCCC (DNCC/MADES 2021). The NFI, which was started by the National Forestry Institute (INFONA) in 2014, is the main source of information on GHG fluxes from the land use sector included in all reporting to the UNFCCC, namely the NDC, GHG inventory, and submissions under the Reduced Emissions from Deforestation and Forest Degradation (REDD+) framework. For the NFI, the country was divided into five forest strata: humid forests of the eastern region ('Humid'), subhumid Cerrado forests ('Cerrado'), subhumid flooded forests of the Paraguay River ('Subhumid'), Chaco dry forests ('Chaco'), and palm forests ('Palm') (Cueva 2015). Like most countries, Paraguay estimates AGBD directly from the NFI sample of field survey plots using a design-based inference framework. In accordance with the NFI, we define AGBD as the total mass of live organic matter in trunks, branches, leaves, and stumps per unit area (Cueva 2015).

The completion of Paraguay's first NFI was an undeniable achievement given the size and diversity of the country's forests. However, the challenges faced by INFONA reflect similar initiatives in tropical and sub-tropical countries with early-stage NFIs (McRoberts *et al* 2010, McRoberts *et al* 2013b): non-response (the omission of measurements at sampled locations) and plot relocation (modifying plot locations in a manner inconsistent with the original sample design) were frequent in inaccessible forests, uncertainties were high for some estimates of AGBD, and the measurements are now outdated and need to be updated.

2.2. Hybrid estimation

To compare to the design-based approach employed by Paraguay, we applied GEDI's hybrid estimator to the forests in the strata mentioned above and at the national scale. While the estimation approach described by Patterson *et al* (2019) is computationally tractable for relatively small areas such as 1 km grid cells, tracking covariances involving millions of GEDI shots at the scale of a country may be infeasible for some applications. Furthermore, GEDI data may be spatially uneven at the country scale, potentially compromising the assumption of uniform sampling underlying GEDI's hybrid estimator (Patterson *et al* 2019). An additional step can reduce the computational burden and account for spatially uneven GEDI data through aggregation of estimates for constituent grid cells to larger spatial domains. Dubayah *et al* (2022a) described an implementation of GEDI's hybrid estimator applicable for country scale estimation based on individual and relatively local (we use 6 km tiles) estimates, accounting for covariances among tile-level estimates related to both GEDI's sampling design and use of the same model across tiles. For a comprehensive description of the theory and equations required for performing the aggregation process utilized in this study, refer to Dubayah *et al* (2022a, suppl. pp 3–7). We applied the same process here, except with a Paraguay-specific footprint-level biomass model instead of GEDI's continental default models.

We masked GEDI shots within each 6 km grid cell using an official forest area map supplied by INFONA. We further stratified the mask by the forest type classes described above and estimated AGBD and their standard errors (SEs) for each stratum. Relative standard errors (rSEs) were calculated as the ratio of estimated SE to the mean and served as a comparable measure of uncertainty. Notably, Dubayah *et al* (2022a) did not mask out non-forest areas, instead estimating AGBD and SEs for all land (including non-forest). We compared the GEDI-based estimates to field-based estimates using Paraguay's current NFI database and the direct, design-based estimators from the REDD+ reference level submission (Cueva 2015, p 6).

Application of the design-based and hybrid estimators both rely upon a sample of Paraguay's forests, but those samples differed. Whereas the design-based estimator used 286 randomly allocated field plots, the hybrid estimator used approximately 12 million GEDI shots systematically aligned along thousands of cluster samples (ground tracks). It is of interest to

determine if there are systematic differences in the subset of the population represented by these samples since such differences might influence both the estimated mean and its variance. We compared the cumulative distribution functions (CDFs) of the remotely sensed parameters at the locations of 286 NFI plots with the larger distribution of those parameters across the population (the forests of the entire country or within a forest stratum). Parameters included: GEDI relative height (rh) values at the 99th percentile (rh99; approximately equivalent to top height measured by GEDI); percent tree cover (TC) values as defined by the Global Forest Cover Change (GFCC) dataset from Sexton *et al* (2013), and biomass as defined by the European Space Agency's Climate Change Initiative (CCI) dataset from Santoro and Cartus (2021). For GEDI top height, plot locations were matched to the nearest GEDI shot using the criteria described in the next section, omitting plots without a suitable match. The CCI biomass and TC datasets cover all areas defined as forest by INFONA's map, allowing direct matching of plots to remotely sensed variables. We compared the distributions of NFI GEDI top height, TC, and CCI biomass values to the distribution of the remotely sensed observations falling in Paraguay official forest mask. A two-sided Kolmogorov–Smirnov (KS) test was used to test the hypothesis that the empirically observed distributions emanated from the same underlying distribution (Massey Jr 1951).

2.3. Model building dataset

Hybrid inference requires a prediction model that relates field measurements to auxiliary data. To facilitate model creation, we developed a dataset with approximately co-located field measurements and GEDI data. Field plots were associated with the nearest GEDI shot that met the following criteria: labeled as high quality according to the L2A quality flag (a combination of multiple waveform quality metrics), less than 200 m from the center of the field plot, at the same elevation (± 2 m), undisturbed between the time of the field measurement and GEDI shot according to the Hansen Global Forest Change v1.9 dataset (Hansen *et al* 2013) and the Continuous Degradation Detection algorithm (Bullock *et al* 2020), and within the same forest patch as defined by the polygons of connected forest pixels in Paraguay's official forest mask (figure 1). The calibration dataset was created on Google Earth Engine using the GEDI dataset described in Healey *et al* (2020).

The definition of forest used by Paraguay for REDD+ reporting includes disturbed land that is in the process of naturally regenerating (PNC ONU-REDD+ 2016; section 2.3). While the area of forests with no biomass is large due to the high rate of disturbance, the NFI does not include plots in recently disturbed forests. Omitting model calibration at the low end of the biomass spectrum could potentially compromise the assumption underlying model-based inference that the model is correctly specified for the population of interest (Gregoire 1998, Ståhl *et al* 2016). To increase the representation of low-biomass forests, we manually identified 30 GEDI shots in recently disturbed forests showing no aboveground biomass in high-resolution imagery. We determined 10 to be an approximately representative number given Paraguay's high rates of deforestation in recent years (MADES/PNUD/FMAM 2018, FAO 2020), and therefore randomly selected 10 among the 30 to be include in the model building dataset.

2.4. Biomass model selection

Currently, only parametric models may be supported under the hybrid inference paradigm used by GEDI (Patterson *et al* 2019). We therefore used ordinary least squares (OLS) regression in a similar manner to the GEDI L4A models (Duncanson *et al* 2022). Using the R and Python programming languages, we tested one-, two-, three-, and four-parameter AGBD models using our model building dataset and GEDI L2A canopy height and L2B canopy cover metrics (Dubayah *et al* 2020b, 2020c). A 100 m offset was added to all height metrics to preclude the possibility of negative rh values (Duncanson *et al* 2022). All combinations of variables were tested using log, square-root, and non-transformed data for both the response (AGBD) and predictor (GEDI) variables. We employed a ratio method for bias correction during the back-transformation of predictions, as described in Snowdon (1991).

We removed poorly parameterized models after fitting all potential parameter combinations in the calibration dataset. Following the L4A model selection procedure (Kellner *et al* 2023), we discarded parameter combinations with a Pearson's correlation coefficient greater than 0.9 to reduce the effects of multicollinearity (Duncanson *et al* 2022). To further ensure independence in the predictor variables we constrained the L2A rh variables to have a minimum difference of 20 in rh values (meaning, for example, rh50 and rh69 could not be in the same model). Finally, we removed models with statistically non-significant variable coefficients ($p > 0.05$).

We selected a model from a pool of 20 candidate models that exhibited the smallest root-mean-squared error (RMSE) and cross-validated RMSE (CV-RMSE). RMSE was used as a simple and comparable measure of prediction error. CV-RMSE is defined as the average RMSE of predicted biomass for ten subsets withheld from model creation (ten-fold cross validation) and is a measure of transferability

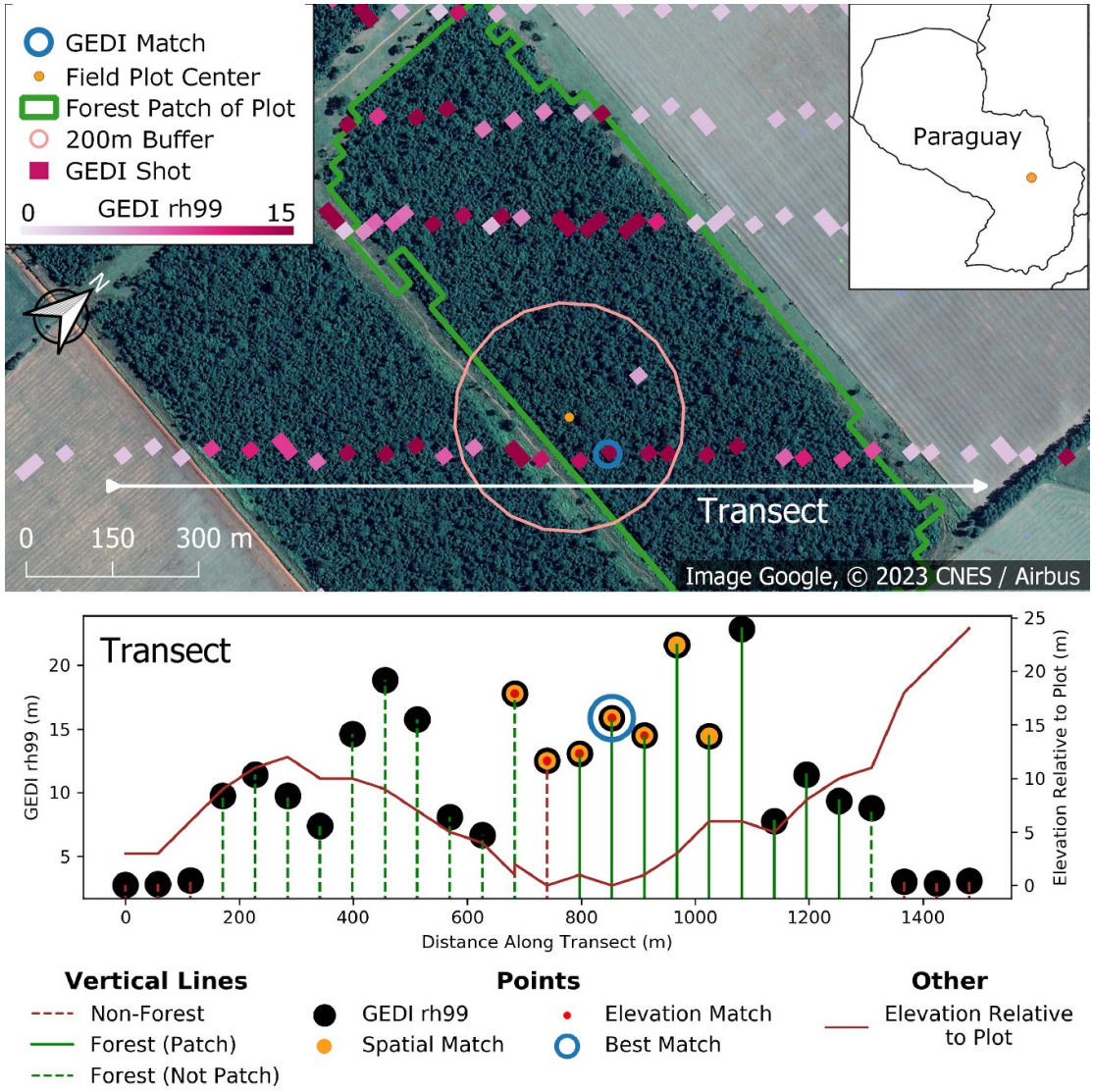

Figure 1. An example of the protocol for matching field plots with GEDI shots. Top: a field plot is matched with the closest GEDI observation that is within a 200 m spatial buffer, in the same forest patch, at the same elevation (±2 m), and in an undisturbed forest according to ancillary forest disturbance datasets. Bottom: the vertical profile of GEDI rh99 for a transect (or subset) of a GEDI ground track near the field plot. Note that the best GEDI match (blue circle) is the closest of three observations that meet all the criteria for linking with field plots.

to 'unseen' data outside of the calibration dataset. The final model was selected manually with consideration for model diagnostic plots, model complexity (e.g. the number and interpretation of coefficients), and alternative metrics such as the Akaike information criterion (AIC) and adjusted R^2.

3. Results

Through our model selection process, we evaluated over 21 000 models for predicting AGBD with the joint use of Paraguay's NFI and on-orbit GEDI lidar data. The models were calibrated with 199 out of 286 field plots from Paraguay's NFI in addition to 10 supplemental plots located in disturbed forests with zero biomass (figure 3). A total of 87 plots were excluded from the model building dataset as they did not meet the GEDI matching criteria described above. The selected model is defined as:

$$\sqrt{\text{AGBD}_i} = \beta_0 + \beta_1 \text{rh5}_i + \beta_2 \sqrt{\text{rh45}_i} + \beta_3 \sqrt{\text{rh91}_i} + \epsilon_i \; \epsilon_i \sim N\left(0, \sigma^2\right)$$

where β_0, β_1, β_2 and β_3 are model parameters to be estimated, and ϵ_i is a model random error. The model parameter estimates were estimated using the OLS estimation method (table 1).

The model RMSE is 33.56 Mg ha^{-1}, CV-RMSE is 24.38 Mg ha^{-1}, R^2 is 0.52, adjusted-R^2 is 0.51, cross-validated R^2 is 0.59, and AIC is 894 (refer to figure S1 for regression diagnostic plots). The GEDI mission's continentally calibrated L4A models systematically underestimated AGBD relative to the field data (RMSE = 71.23, figure 2(A)).

Table 1. Ordinary least squares estimation results for the selected AGBD-GEDI model. The predictor variables include relative height (rh) values for the 5th, 45th, and 91st percentiles, defined as the heights at which 5%, 45%, and 91% of the waveform energy is returned for each observation, respectively.

	Parameter	Estimate	Standard error	*t*-value	*p*-value
Intercept	β_0	−64.32	12.67	−5.07	<0.01
rh5	β_1	−0.82	0.152	−5.38	<0.01
rh45[a]	β_2	6.45	2.495	2.59	0.01
rh91[a]	β_3	8.40	1.45	5.79	<0.01

[a] Square-root transformation applied prior to estimating model parameters.

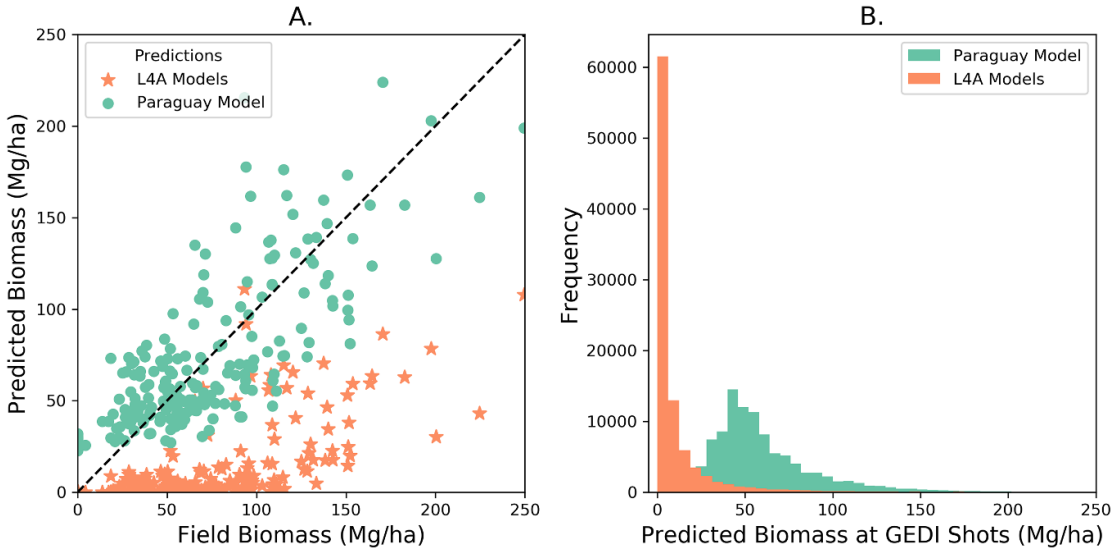

Figure 2. (A). Observed versus predicted aboveground biomass for the locally calibrated GEDI model and L4A. The 'Paraguay model' was fitted from NFI data, with model selection performed using ten-fold cross-validation. (B). A histogram of predicted biomass at GEDI shots for the entire country as calculated by the local GEDI model and L4A.

Consequently, the footprint-level L4A biomass predictions were, on average, under 25 Mg ha^{-1}. By contrast, the footprint-level predictions from the Paraguay model were approximately normally distributed and centered on 50 Mg ha^{-1} (figure 2(B)).

We used GEDI with hybrid inference to estimate mean AGBD at the scale of 6 km tiles (figure 3(B)) and aggregated those estimates at the strata and national level (figure 3(A)). For comparison, we used Paraguay's NFI data, collected under a simple random sample design, to estimate mean AGBD for the same spatial domains. The GEDI estimate of mean AGBD in Paraguay's forests was 65.55 Mg ha^{-1} and the stratum-level estimates ranged from 52.34 Mg ha^{-1} in Chaco to 103.88 Mg ha^{-1} in Humid (table 2). The NFI estimate of mean AGBD for the country was 73.13 Mg ha^{-1}; the 95% confidence interval included the value estimated from GEDI. However, the stratum-level estimates were notably different, although they were statistically equivalent for Chaco, Humid and Cerrado. The NFI estimates correspond to the years 2014 and 2015, while the GEDI estimates are for 2019–2021.

The GEDI estimates of AGBD were more precise than the NFI estimates. For example, the rSE of the NFI estimate was 7.32% for Cerrado compared to GEDI's estimate of 4.20%. The rSEs of the GEDI and NFI estimates ranged from 3.50% to 4.67% and 5.45% to 8.14%, respectively. At the national scale, the SE of the GEDI estimate (2.35 Mg ha^{-1}) was nearly half that of the NFI estimate (4.03 Mg ha^{-1}).

We evaluated the representativeness of the samples used for design-based inference (NFI plot locations) and hybrid inference (GEDI shots) using a KS test applied to the CDF of each sample. Figure 4 shows the CDFs of values of GEDI top height, TC, and biomass (Sexton *et al* 2013, Dubayah *et al* 2020b, Santoro and Cartus 2021) at the locations of the NFI plots in relation to the populations defined by the remote sensing datasets covering the country's forests. Note that for the Humid and Cerrado strata, the NFI sample includes a higher proportion of tall, dense, and high biomass forests than the larger remotely sensed forest population according to the GEDI, GFCC, and CCI datasets, respectively (figure 4 and table 3). In these strata, NFI estimates of AGBD were substantially higher than the GEDI-based estimates (table 2).

The KS test statistics were consistently larger for the NFI than GEDI when compared to the GFCC TC and CCI Biomass datasets (table 3). A small test statistic (e.g. <0.1) indicates that the sample

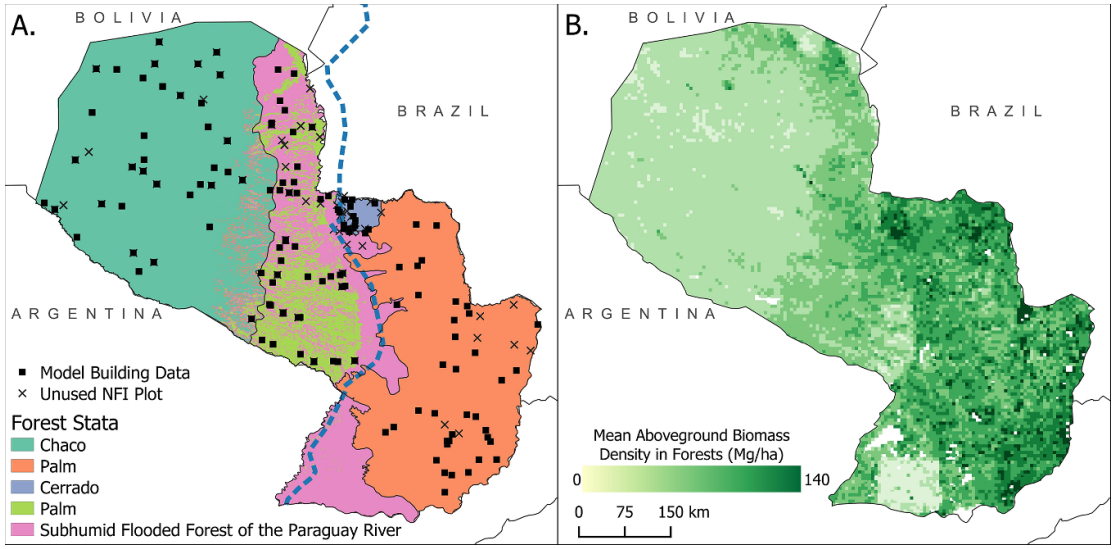

Figure 3. (A). The model building dataset and excluded NFI plots overlayed on the forest strata used by Paraguay for greenhouse gas reporting. (B). Predicted aboveground biomass density using a GEDI model that was calibrated with co-located NFI plots and GEDI shots and averaged within 6 km grids for hybrid estimation.

Table 2. Estimates of mean aboveground biomass density, their associated standard errors (SEs), and relative standard errors (rSE) generated using only national forest inventory (NFI) data and design-based inference and hybrid inference (GEDI).

	NFI			GEDI		
	Mean (Mg ha^{-1})	SE (Mg Ha^{-1})	rSE (%)	Mean (Mg ha^{-1})	SE (Mg ha^{-1})	rSE (%)
Chaco	59.26	3.23	5.45	52.34	2.33	4.44
Humid	111.49	8.11	7.27	103.88	4.26	4.10
Cerrado	109.16	7.99	7.32	95.06	4.00	4.20
Subhumid	108.78	8.79	8.08	81.11	2.84	3.50
Palm	34.29	2.79	8.14	57.01	2.66	4.67
Country	73.19	4.03	5.51	65.55	2.35	3.59

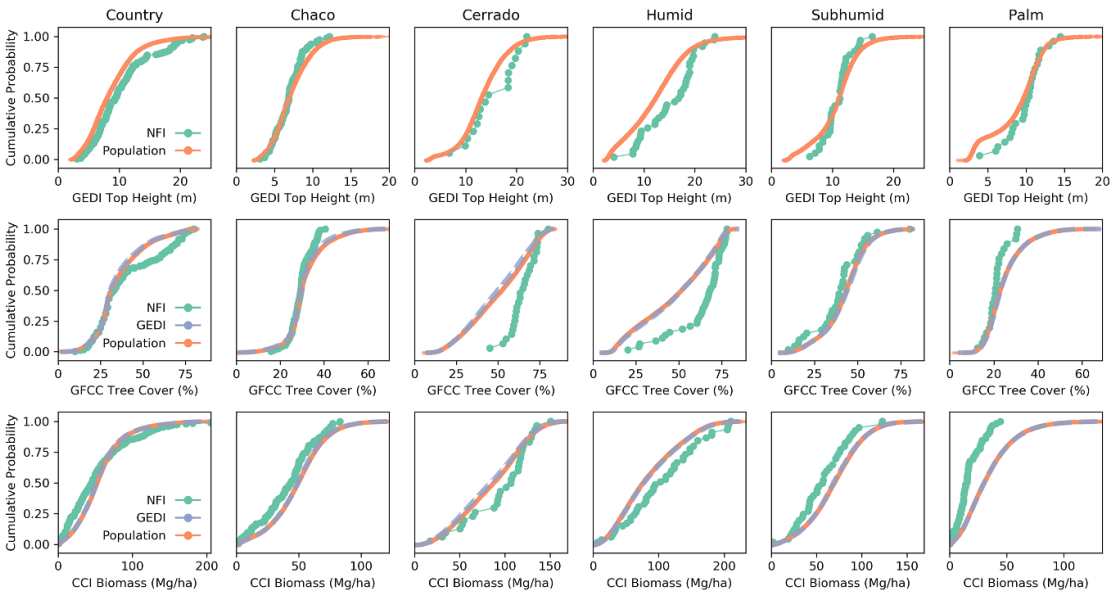

Figure 4. The cumulative distribution functions of the national forest inventory (NFI) plot locations and population for the three independent datasets: top height as approximated by GEDI rh99, GFCC tree cover, and CCI biomass. The NFI distribution was defined by the model building dataset for rh99 and the full NFI dataset (i.e. not subsampled) for GFCC tree cover and CCI biomass.

Table 3. The Kolmogorov–Smirnov (KS) test statistics for summarizing agreement between the underlying distribution of a sample to the population. Results are shown at the strata and national scale.

	Population				
	GEDI top height	GFCC tree cover		CCI biomass	
		Sample			
	NFI	NFI	GEDI	NFI	GEDI
Chaco	0.135	0.151*	0.021**	0.16**	0.011**
Humid	0.313**	0.422**	0.014**	0.202*	0.012**
Cerrado	0.345*	0.478**	0.035**	0.231	0.03**
Subhumid	0.165	0.218*	0.021**	0.209	0.015**
Palm	0.184	0.279*	0.022**	0.409**	0.01**
Country	0.155**	0.166**	0.012**	0.15**	0.009**

Note 1: The asterisks represent the confidence level in which we can reject the null hypothesis that the sample was drawn from the same distribution as the population (*95%, **99%).

Note 2: The *p*-value of GEDI when applied as a sample will almost always be close to zero due to the large sample size.

Note 3: The abbreviations in the table represent the National Forest Inventory (NFI), Global Ecosystem Dynamics Investigation (GEDI), Global Forest Cover Change (GFCC; Sexton *et al* 2013), and the European Space Agency's Climate Change Initiative (CCI; Santoro and Cartus 2001).

and population share the same underlying distribution. The KS test also revealed large divergence between the NFI and GEDI, although the results were not statistically significant ($p > 0.05$) for some strata.

3.1. Discussion

GEDI is the first spaceborne lidar that acquires data in most of the world's temperate and tropical forests at a spatial resolution and sampling density that is compatible with forest inventories. GEDI's baseline estimation strategy relies upon pre-calibration of footprint-level biomass models with a global set of paired field-airborne lidar data (Duncanson *et al* 2022). However, GEDI was designed to enable broader forest monitoring applications (Dubayah *et al* 2020a), and its public availability through NASA's active archive system (Dubayah *et al* 2020b) and Google Earth Engine (Gorelick *et al* 2017, Healey *et al* 2020) facilitate operational uses of GEDI in tropical and sub-tropical countries such as Paraguay.

A primary concern with using on-orbit GEDI data with an existing NFI is the potential degradation of biomass model fit due to spatial and temporal mismatches between field measurements and GEDI shots. In our judgment, the model employing GEDI height metrics exhibited an adequate fit to Paraguay's NFI data (figure 2(A)). Further, the model we developed using on-orbit GEDI data exhibited lower SEs compared to using NFI data alone (table 2). Nevertheless, it is important to acknowledge that model-based frameworks utilizing GEDI may underestimate AGBD if growth induces systematic discrepancies in forest structure during the time difference between data collected by the NFI and GEDI.

There were differences both at the national level and at the level of some ecozone strata in the estimated mean AGBD (table 1). Insight into these differences may be gained from comparison of the CDF of all remotely sensed measurements retrieved over a national forest mask and the CDF of those same measurements retrieved at (or near) NFI plot locations. In the Cerrado and Humid strata, where estimates of mean AGBD were larger for the NFI estimates, KS tests (table 2) suggested that the samples drawn by the NFI corresponded to taller, denser, and higher-biomass forests on the basis of remotely sensed data than one might expect from a random sample of those data. We observed the opposite pattern in the Palm, Chaco, and Subhumid strata (figure 4). While the remotely sensed data merely correlates with real biomass, this result suggests that sampling differences may play a role in deviations of GEDI and NFI estimates.

It was earlier stated that Paraguay's NFI may relocate plots deemed inaccessible, although there was no formalized method for doing so. Practical issues preventing full execution of a designed sample are common in forest inventories, and can in some cases compromise ensuant inferences (estimates of forest attributes and their uncertainties) (Westfall *et al* 2022). As with any probability sample, it is also possible that a given (relatively small) set of observations may not accurately represent the larger population (Schreuder *et al* 1993). In either case, model-based methods such as hybrid estimation make fewer assumptions about the empirical distribution of the data used in the analysis; we need only to assume that the model applies to the population of interest (Gregoire 1998). This highlights a key advantage of the method we have demonstrated; calibration data in a country need not come from a probability sample. In fact, it may be

reasonable in some cases to apply models developed in a country with extensive plot resources to similar forests occurring in a country with no formal forest inventory.

We used the Paraguay-specific biomass model for hybrid inference rather than the regional GEDI L4A models (Dubayah *et al* 2022b, Duncanson *et al* 2022). A lack of paired-field data in the South American deciduous broadleaf tree (DBT) plant functional type means that there is currently no South American L4A DBT model. This greatly impacts Paraguay, where DBT is the predominant type. Instead, the Evergreen Broadleaf Tree model, which was mostly calibrated in Amazonian forests, is used for all of Paraguay. In the future, the inclusion of more field plots into the global GEDI calibration dataset should enable a South American DBT model and improve the L4A predictions in Paraguay. Validation at Paraguay's field plots showed that locally calibrated models yielded better prediction of AGBD than the current L4A model in Paraguay (figure 2). This benefit may be lower, particularly in relation to the lack of fit introduced by pairing spatially offset data sources, in regions where L4A calibration is more locally representative.

It is likely that the aggregation process recommended by Dubayah *et al* (2022a) yielded different estimates than using the country (or strata) as a single estimation unit. It is known that differential cloud occurrence and unplanned resonance in the International Space Station's orbit (especially in 2021) has caused the probability of inclusion in GEDI's sample to vary strongly over space. Humid ecosystems with frequent cloud cover will often have higher AGBD than arid regions; since GEDI is sensitive to cloud cover, there will also be less observations in cloudy regions. Therefore, the mean AGBD (i.e. without aggregation) may be skewed towards regions with lower cloud cover and biomass. Regions where the frequency of GEDI observations are spatially uneven can result in a sample that does not represent an area of interest and, when that sample is used as a basis of inference, an AGBD estimator that is biased (Dubayah *et al* 2022a). Given GEDI's beam configuration (8 beams per overpass, spaced approximately 600 m apart), we determined that 6 km was a domain size for which we could reasonably assume the kind of uniform sampling that is presumed by Patterson *et al*'s (2019) hybrid estimator.

Lidar-based approaches to estimating carbon stocks have rarely been adopted by national forest monitoring programs (FAO 2020, Melo *et al* 2023). The preference for design-based inference through direct application of NFI data is understandable: NFIs are locally managed, do not require an airborne campaign or biomass model, require minimal technical capacity, and have a long history of use in many countries. Nevertheless, NFIs must confront logistical and continuity challenges that can impinge on comprehensive and statistically rigorous forest resource assessments. We have demonstrated here that integration of NFI and GEDI data through hybrid inference can reduce uncertainties. Furthermore, hybrid inference using GEDI minimized the potential impact of inventory data that may not accurately reflect the population (figure 4).

4. Conclusion

Our research shows that statistical estimation of AGBD using on-orbit GEDI data and incidental field data is possible and offers numerous advantageous than using an NFI alone. Based on our findings, we offer the following suggestions for integrating GEDI with an NFI:

1. It is possible to establish repeatable, objective rules for matching on-orbit GEDI waveforms to NFI biomass measurements. In this study, paired GEDI-field data collected in this manner identified a GEDI-biomass relationship that supported relatively more precise hybrid estimates of mean aboveground biomass;
2. At least in cases such as Paraguay, where L4A model-building data do not currently reflect local conditions (Duncanson *et al* 2022), opportunistic parametrization of biomass models with on-orbit data and local field plots may be preferable;
3. Hybrid inference using GEDI data can be used for precise estimation of AGBD, even though it does not rely on a designed sample of field plots. This approach has the advantage of leveraging millions of remotely sensed observations across an area the size of Paraguay;
4. For large areas applications of hybrid inference, aggregation of estimates for smaller areas (6×6 km tiles in our case) can reduce computation and help account for spatial variability in the frequency of GEDI observations;
5. By comparing the cumulative distribution of GEDI data at field plots with the distribution for the population, we can identify NFI samples that may not accurately represent the estimation domain.

GEDI is the first space-based mission capable of supplying the biomass metadata that McRoberts *et al* (2022) identified as requisite for model-based inference (Dubayah *et al* 2022a). In GEDI's implementation of hybrid inference, this means that the covariance matrices for the L4A model parameters are publicly available. Our work marks, to our knowledge, the first use of on-orbit GEDI data with an NFI to create more locally representative L4A-like models (and corresponding parameter covariance matrices). Working directly with local inventory data and specialists can also build country ownership and take advantage of local expertise.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgments

This research was funded by NASA Grant 80HQTR21T0013 (GEDI Science Team) and NASA Interagency Agreement NNL22OB15A. We thank Google Earth Engine for establishing the GEDI data asset used in our analysis. We also thank the CEOS Biomass Harmonization Subgroup and the USGS SilvaCarbon program for facilitating our institutional collaborations. We are especially grateful for the INFONA's national forest inventory for providing the field data used in our analysis. The findings and conclusions in this publication are those of the authors and should not be construed to represent any official USDA or U.S. Government determination or policy.

Conflict of interest

The author declares no competing interests.

ORCID iDs

Eric L Bullock https://orcid.org/0000-0003-3279-6771
Sean P Healey https://orcid.org/0000-0003-3498-4266
Joana B Melo https://orcid.org/0000-0002-7147-3281

References

Andersen H-E, Barrett T, Winterberger K, Strunk J and Temesgen H 2009 Estimating forest biomass on the western lowlands of the Kenai Peninsula of Alaska using airborne lidar and field plot data in a model-assisted sampling design *IUFRO Division 4 – Extending Forest Inventory and Monitoring over space and time* (*Quebec City, Canada, 19–22 May 2009*) pp 19–22

Brown S, Gillespie A J R and Lugo A E 1989 Biomass estimation methods for tropical forests with applications to forest inventory data *For. Sci.* **35** 881–902

Bullock E L, Woodcock C E and Olofsson P 2020 Monitoring tropical forest degradation using spectral unmixing and Landsat time series analysis *Remote Sens. Environ.* **238** 110968

Chen Q, McRoberts R E, Wang C and Radtke P J 2016 Forest aboveground biomass mapping and estimation across multiple spatial scales using model-based inference *Remote Sens. Environ.* **184** 350–60

Cueva K 2015 *Metodologia de Procesamiento Y Análisis de Datos del Inventario Forestal Nacional (IFN)* (San Lorenzo) (available at: https://infona.gov.pyhttps://infona.gov.py/sistema-nacional-de-monitoreo-forestal-del-paraguay/)

de Freitas J V *et al* 2009 The new Brazilian national forest inventory *Proc. 8th Annual Forest Inventory and Analysis Symp. (Monterey, CA, 16–19 October 2006) Gen. Tech. Report WO-79* vol 79, ed R E McRoberts, G A Reams, P C Van Deusen and W H McWilliams (Washington, DC: US Departme)

DNCC/MADES 2021 *Actualización de la NDC de la República del Paraguay al 2030* (Asunción) (available at: https://unfccc.int/sites/default/files/NDC/2022-06/ACTUALIZACI%C3%93N%20DE%20LA%20NDC%20DEL%20PARAGUAY_Versi%C3%B3n%20Final.pdf)

Dubayah R *et al* 2020a The Global Ecosystem Dynamics Investigation: high-resolution laser ranging of the Earth's forests and topography *Sci. Remote Sens.* **1** 100002

Dubayah R *et al* 2022a GEDI launches a new era of biomass inference from space *Environ. Res. Lett.* **17** 095001

Dubayah R *et al* 2022b GEDI L4A footprint level aboveground biomass density, version 2.1 *NASA ORNL Biogeochemical Dynamics DAAC* (https://doi.org/10.3334/ORNLDAAC/2056)

Dubayah R, Hofton M, Blair M J B, Armston J, Tang H and Luthcke S 2020b GEDI L2A elevation and height metrics data global footprint level V002 *NASA EOSDIS Land Processes DAAC* (https://doi.org/10.5067/GEDI/GEDI02_A.002) (Accessed 1 May 2021)

Dubayah R, Tang H, Armston J, Luthcke S, Hofton M and Blair M J B 2020c GEDI L2B canopy cover and vertical profile metrics data global footprint level V002 *NASA EOSDIS Land Processes DAAC* (https://doi.org/10.5067/GEDI/GEDI02_B.002) (Accessed 1 May 2021)

Duncanson L *et al* 2022 Aboveground biomass density models for NASA's Global Ecosystem Dynamics Investigation (GEDI) lidar mission *Remote Sens. Environ.* **270** 112845

Ene L T, Gobakken T, Andersen H-E, Næsset E, Cook B D, Morton D C, Babcock C and Nelson R 2018 Large-area hybrid estimation of aboveground biomass in interior Alaska using airborne laser scanning data *Remote Sens. Environ.* **204** 741–55

FAO 2020 *Global Forest Resources Assessment 2020: Main report* (Food and Agriculture Organization of the United Nations) (https://doi.org/10.4060/ca9825en)

Frolking S, Palace M W, Clark D B, Chambers J Q, Shugart H H and Hurtt G C 2009 Forest disturbance and recovery: a general review in the context of spaceborne remote sensing of impacts on aboveground biomass and canopy structure *J. Geophys. Res. Biogeosci.* **114** G00E02

Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D and Moore R 2017 Google Earth Engine: planetary-scale geospatial analysis for everyone *Remote Sens. Environ.* **202** 18–27

Grau H R and Aide M 2008 Globalization and land-use transitions in Latin America *Ecol. Soc.* **13** 16

Gregoire T G 1998 Design-based and model-based inference in survey sampling: appreciating the difference *Can. J. For. Res.* **28** 1429–47

Gregoire T G, Ståhl G, Næsset E, Gobakken T, Nelson R and Holm S 2011 Model-assisted estimation of biomass in a LiDAR sample survey in Hedmark County, Norway *Can. J. For. Res.* **41** 83–95

Hancock S, Armston J, Hofton M, Sun X, Tang H, Duncanson L I, Kellner J R and 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 M C *et al* 2013 High-resolution global maps of 21st-century forest cover change *Science* **342** 850–3

Healey S P, Yang Z, Gorelick N and 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** 2840

Kellner J, Armston J and Duncanson L 2023 Algorithm theoretical basis document for GEDI footprint aboveground biomass density *Earth Space Sci.* **10** e2022EA002516

Labriere N *et al* 2018 *In situ* reference datasets from the Tropisar and AfriSAR campaigns in support of upcoming spaceborne

biomass missions *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **11** 1–11
Lu D 2006 The potential and challenge of remote sensing-based biomass estimation *Int. J. Remote Sens.* **27** 1297–328
Lu D, Chen Q, Wang G, Liu L, Li G and Moran E 2016 A survey of remote sensing-based aboveground biomass estimation methods in forest ecosystems *Int. J. Digit. Earth* **9** 63–105
MacDicken K G 2015 Global forest resources assessment 2015: what, why and how? *For. Ecol. Manage.* **352** 3–8
MADES-DNCC/PNUD-FMAM 2022 Informe del Inventario Nacional de Gases de Efecto Invernadero de Paraguay, serie 1990-2017 *Proyecto IBA3* (Asunción) (available at: https://unfccc.int/sites/default/files/resource/2021_IIN_PY%20Versi%C3%B3n%20Final_compressed.pdf)
Magnussen S, Nord-Larsen T and Riis-Nielsen T 2018 Lidar supported estimators of wood volume and aboveground biomass from the Danish national forest inventory (2012–2016) *Remote Sens. Environ.* **211** 146–53
Massey F J Jr 1951 The Kolmogorov-Smirnov test for goodness of fit *J. Am. Stat. Assoc.* **46** 68–78
McRoberts R E 2010 Probability-and model-based approaches to inference for proportion forest using satellite imagery as ancillary data *Remote Sens. Environ.* **114** 1017–25
McRoberts R E, Næsset E and Gobakken T 2013a Inference for lidar-assisted estimation of forest growing stock volume *Remote Sens. Environ.* **128** 268–75
McRoberts R E, Næsset E, Gobakken T, Chirici G, Condés S, Hou Z, Saarela S, Chen Q, Ståhl G and Walters B F 2018 Assessing components of the model-based mean square error estimator for remote sensing assisted forest applications *Can. J. For. Res.* **48** 642–9
McRoberts R E, Næsset E, Saatchi S and Quegan S 2022 Statistically rigorous, model-based inferences from maps *Remote Sens. Environ.* **279** 113028
McRoberts R E, Tomppo E O and Næsset E 2010 Advances and emerging issues in national forest inventories *Scand. J. For. Res.* **25** 368–81
McRoberts R E, Tomppo E O, Vibrans A C and de Freitas J V 2013b Design considerations for tropical forest inventories *J. For. Res.* **33** 189–202
Melo J, Baker T, Nemitz D, Quegan S and Ziv G 2023 Satellite-based global maps are rarely used in forest reference levels submitted to the UNFCCC *Environ. Res. Lett.* **18** 034021
Patterson P L *et al* 2019 Statistical properties of hybrid estimators proposed for GEDI—NASA's global ecosystem dynamics investigation *Environ. Res. Lett.* **14** 065007
PNC ONU-REDD+ 2016 *Nivel de Referencia de las Emisiones Forestales por Deforestación en la República del Paraguay para pago por resultados de REDD+ bajo la CMNUCC* (Asunción) (available at: https://redd.unfccc.int/files/paraguay_2016_frel_submission_modified.pdf)
Rejou-Mechain M *et al* 2014 Local spatial structure of forest biomass and its consequences for remote sensing of carbon stocks *Biogeosciences* **11** 6827–40
MADES/PNUD/FMAM 2018 *Segundo Informe Bienal de Actualización (IBA2). Proyecto IBA2* (Asunción) (available at: https://unfccc.int/sites/default/files/resource/56910784_Paraguay-BUR2-1-Informe%20Bienal%20de%20Actualizacion_PY_Dic%202018_.pdf)
Santoro M and Cartus O 2021 ESA biomass climate change initiative (Biomass_cci): global datasets of forest above-ground biomass for the years 2010, 2017 and 2018, v2 *Centre for Environmental Data Analysis* (https://doi.org/10.5285/84403d09cef3485883158f4df2989b0c)
Schreuder H T, Gregoire T G and Wood G B 1993 *Sampling Methods for Multiresource Forest Inventory* (New York: Wiley)
Sexton J O *et al* 2013 Global, 30-m resolution continuous fields of tree cover: Landsat-based rescaling of MODIS vegetation continuous fields with lidar-based estimates of error *Int. J. Digit. Earth* **6** 427–48
Shugart H H, Saatchi S and Hall F G 2010 Importance of structure and its measurement in quantifying function of forest ecosystems *J. Geophys. Res. Biogeosci.* **115** G00E13
Snowdon P 1991 A ratio estimator for bias correction in logarithmic regressions *Can. J. For. Res.* **21** 720–4
Ståhl G *et al* 2016 Use of models in large-area forest surveys: comparing model-assisted, model-based and hybrid estimation *For. Ecosyst.* **3** 5
Ståhl G, Holm S, Gregoire T G, Gobakken T, Næsset E and Nelson R 2011 Model-based inference for biomass estimation in a LiDAR sample survey in Hedmark County, Norway *Can. J. For. Res.* **41** 96–107
Westfall J A, Schroeder T A, McCollum J M and Patterson P L 2022 A spatial and temporal assessment of nonresponse in the national forest inventory of the US *Environ. Monit. Assess.* **194** 1–14
Wulder M A, White J C, Nelson R F, Næsset E, Ørka H O, Coops N C, Hilker T, Bater C W and Gobakken T 2012 Lidar sampling for large-area forest characterization: a review *Remote Sens. Environ.* **121** 196–209
Zolkos S G, Goetz S J and Dubayah R 2013 A meta-analysis of terrestrial aboveground biomass estimation using lidar remote sensing *Remote Sens. Environ.* **128** 289–98