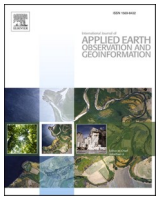
International Journal of Applied Earth Observations and Geoinformation 115 (2022) 103108

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

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



Fusing GEDI with earth observation data for large area aboveground biomass mapping

Yuri Shendryk

Dendra Systems, Sydney, Australia

ARTICLE INFO

Keywords:
GEDI
Biomass
Satellite imagery
Data fusion
Machine learning
Remote sensing
LiDAR

ABSTRACT

An accurate and spatially explicit estimation of biomass is required for sustainable forest management, prevention of biodiversity loss, and carbon accounting for climate change mitigation. This study offers a methodology to generate wall-to-wall aboveground biomass density (AGBD) maps that exclusively relies on open access earth observation (EO) data. Specifically, spaceborne Global Ecosystem Dynamics Investigation (GEDI) LiDAR data were fused with Sentinel-1 synthetic-aperture radar, Sentinel-2 multispectral, elevation, and land cover data to produce biomass maps of Australia and the United States for 2020. The gradient boosting machine learning framework was applied to predict AGBD and its uncertainty at the resolutions of 100 m and 200 m. The performance of models based on (1) Sentinel-2 imagery and land cover and (2) a combination of Sentinel-2 and Sentinel-1 imagery with elevation and land cover data were compared. The most accurate gradient boosting model was identified using a Bayesian hyperparameter optimization with a 5-fold cross-validation. The Sentinel-2 imagery and land cover data analysis resulted in AGBD estimated with the coefficient of determination (R^2) of 0.61 – 0.71, root-mean-square error (*RMSE*) of 59 – 86 Mg/ha, and relative root-mean-square error (*RMSE%*) of 45 – 80%. The accuracy of the models improved with the addition of Sentinel-1 and elevation data: AGBD estimation with R^2 of 0.66 – 0.74, *RMSE* of 55 – 81 Mg/ha, and *RMSE%* of 41 – 77%. It was found that Sentinel-2 and land cover-derived predictors were the most important in estimating annual AGBD. The proposed method also reduced the saturation effect, which is common in high biomass areas when predicting AGBD using satellite imagery. Prediction maps produced in this study could serve as a baseline for current AGB stocks of forested lands equal to 9.8 Pg and 37.1 Pg in Australia and the United States, respectively. Overall, this research highlights methodological opportunities for combining open access EO data to yield more accurate and globally applicable AGB maps through data fusion.

1. Introduction

An accurate and spatially explicit estimation of biomass is critical for understanding terrestrial carbon dynamics and clarifies the role of natural ecosystems in climate change mitigation. Terrestrial ecosystems such as forests, shrublands, and grasslands play a fundamental role in Earth's climate, acting as an overall carbon sink. For example, it was estimated that global forests absorb twice as much carbon as they emit each year from deforestation and other disturbances (Harris et al., 2021; Xu et al., 2021). Therefore, to facilitate the global community's efforts in restoring and protecting forests and other carbon sinks, it is vital to know how much biomass is stored in terrestrial ecosystems and monitor its change over time.

Total plant biomass consists of aboveground biomass (AGB; e.g. trees, shrubs, and grasses) and belowground biomass (e.g. living roots). Generally, only AGB is estimated due to the difficulty of measuring belowground biomass in the field (Lu et al., 2016). There are multiple approaches for estimating AGB, including field and remote sensing-based methods. The most accurate, albeit time-consuming and small scale method to estimate AGB is field-based destructive sampling (Kumar & Mutanga, 2017). Destructive sampling is further commonly used to generate allometric models based on tree measurements such as

Abbreviations: AGBD, Aboveground Biomass Density; CCI, Climate Change Initiative; CNN, Convolutional Neural Network; DEM, Digital Elevation Model; EO, Earth Observation; FIA, Forest Inventory and Analysis; GEDI, Global Ecosystem Dynamics Investigation; LC, Land Cover; LiDAR, Light Detection and Ranging; NDSI, Normalized Difference Spectral Index; RMSE, Root-Mean-Square Error; RSE, Relative Standard Error; SAR, Synthetic-Aperture Radar; SD, Standard Deviation; SLR, Simple Linear Regression; SHAP, SHapley Additive exPlanations.

E-mail addresses: yuri.shendryk@dendra.io, yuri.shendryk@gmail.com.

https://doi.org/10.1016/j.jag.2022.103108
Received 2 August 2022; Received in revised form 23 October 2022; Accepted 12 November 2022
Available online 17 November 2022

diameter at breast height (DBH), height, and volume. In contrast, remote sensing methods informed using allometric models can provide large scale AGB estimates and generally rely on passive (e.g. multispectral) or active (synthetic-aperture radar (SAR) and Light Detection and Ranging (LiDAR)) sensors (Lu et al., 2016). Remote sensing data can be collected from field-based, airborne and spaceborne platforms, with LiDAR sensors commonly providing the most accurate AGB estimates.

Spaceborne remote sensing is currently the most practical tool for spatially explicit monitoring of AGB at continental and global scales (Kumar & Mutanga, 2017; Rodríguez-Veiga et al., 2017). AGB estimation methods using spaceborne remote sensing have substantially evolved in the last five decades, with multispectral (e.g. Landsat, MODIS), SAR (e.g. ALOS PALSAR, ENVISAT ASAR), and LiDAR (e.g. GLAS, ICESat) sensors being most commonly used for AGB estimation. However, the lack of field data and allometric models for calibration and validation, limited coverage as well as low temporal and spatial resolutions hindered their application for tracking progress toward vegetation-based climate change mitigation goals (Kumar & Mutanga, 2017; Rodríguez-Veiga et al., 2017). Nevertheless, monitoring AGB and its change using spaceborne remote sensing has become increasingly feasible over large areas in unprecedented detail (Harris et al., 2021; Santoro et al., 2021). This has been mainly driven by the abundance and open access to high-resolution, multi-source satellite imagery. Specifically, the near-daily availability of satellite multispectral imagery and SAR imagery that overcomes cloud-cover issues enable the mapping of vegetation disturbances down to 3 m resolution (Csillik et al., 2022; Nicolau et al., 2019). This, in turn, creates opportunities for much more timely, transparent, and efficient monitoring of AGB at high resolution for national-level carbon budget inventories and large area comparative studies.

Recent progress in AGB estimation using spaceborne remote sensing data would be impossible without the availability of allometric models and field measurements that they rely on. Field-measured biomass is an integral component of earth observation (EO) derived AGB maps allowing the conversion of remote sensing signals into absolute measurements of biomass (e.g. Megagram (Mg)) and its density (e.g. Mg/ha) that permit time and space comparisons. Unfortunately, manual AGB measurements in the field are time-consuming and require either destructive sampling or measurements of common tree parameters, such as DBH, height, and volume. To alleviate this problem, Global Ecosystem Dynamics Investigation (GEDI) sensor was recently launched and installed at the International Space Station. GEDI collects full-waveform LiDAR data to produce accurate estimates of AGB density (AGBD) within footprints of ~ 25 m in diameter at the near-global scale (Dubayah et al., 2020; Duncanson et al., 2022). Models to produce AGBD estimates from GEDI were developed from discrete-return airborne LiDAR data collocated with field measurements, with the former further used to simulate GEDI waveforms (Kellner et al., 2021). However, GEDI is not a wall-to-wall instrument, and even after a full year of data collection, there are still significant coverage gaps, especially at the equator (Dubayah et al., 2020; Dubayah et al., 2022a).

Generally, the derivation of national and global level wall-to-wall AGBD products relies on spaceborne remote sensing data at a resolution as fine as 30 m informed using field-measured biomass data (Rodríguez-Veiga et al., 2017). For example, there are Climate Change Initiative (CCI) global AGBD maps at 100 m resolution derived from satellite SAR imagery for multiple years (Santoro et al., 2021) as well as multi-year Landsat and PlanetScope time-series derived AGBD products over Canada (30 m) and Peru (100 m), respectively (Csillik et al., 2019; Matasci et al., 2018). There are also multiple wall-to-wall AGBD products in the works that rely specifically on GEDI data. For instance, NASA's Jet Propulsion Laboratory (JPL) 2020 Global Biomass Dataset (Saatchi et al., 2022) and an enhanced biomass product that fuses GEDI, TanDEM-X, and Landsat data (Fatoyinbo et al., 2022)). However, some existing and upcoming AGBD products often rely on proprietary EO data (e.g. TanDEM-X, ALOS-2 PALSAR-2, PlanetScope, etc.), complicating their usability. Furthermore, satellite data-derived AGBD products are prone to saturation in regions of high biomass (Duncanson et al., 2022; Urbazaev et al., 2018) and are difficult to replicate without having access to the privately-owned airborne LiDAR data and field inventory measurements used in their development.

To date, regression analysis, k-nearest neighbors, and random forest were the most popular method for large scale estimation of AGBD using satellite imagery (Csillik et al., 2019; Matasci et al., 2018; Rodríguez-Veiga et al., 2017), while more advanced machine learning approaches (e.g. gradient boosting and convolutional neural networks (CNNs)) are still being underutilized. In this study, gradient boosting was used to estimate AGBD mainly due to its fast-training speed, interpretability, and capability to handle large-scale multi-dimensional and multi-type (e.g. categorical and continuous) data. Gradient boosting uses an ensemble of decision trees for predictions. However, unlike random forest, the decision trees in gradient boosting are generally shallow and added sequentially, with each new tree built to improve the performance of the previous trees (Ke et al., 2017; Microsoft, 2022). As a result, gradient boosting showed good potential in multiple benchmarks (Borisov et al., 2021; Yan et al., 2021) and, while being more interpretable and faster to train, it can outperform CNN models in the case of small training samples of remote sensing data (Su et al., 2021).

In this study, a gradient boosting approach for large area AGBD mapping that solely relies on open access EO data is proposed. The applicability of the proposed approach in different geographic regions is tested and compared against the existing global AGBD product. For this, GEDI data were combined with Sentinel-1 SAR, Sentinel-2 multispectral, elevation, and land cover data to produce wall-to-wall AGBD maps of Australia and the United States for 2020. This data fusion was performed to produce AGBD maps that are more accurate and exhibit less saturation in high biomass areas than what is possible with any single data source alone.

2. Methods

2.1. Study area

The study area spanned Australia and the main contiguous United States. According to the WorldCover map from 2020 (Zanaga et al., 2021), the predominant land covers in Australia were grasslands (54.8%), trees (18.1%), and shrublands (17.0%) in 2020 (Fig. 1A), with the majority of Australia's trees being broadleaf evergreens, such as eucalypts (ABARES, 2018). In contrast, predominant land covers of the contiguous United States included trees (34.8%), grasslands (31.4%), and croplands (15.9%) (Fig. 1B). Forests in the east of the United States mainly consist of broadleaf deciduous trees, except for coniferous forests and plantations in the southern region, while trees in the west of the country are mostly coniferous (Ruefenacht et al., 2008).

2.2. Data

The data consisted of the open access EO datasets available through the Oak Ridge National Laboratory Distributed Active Archive Centre (ORNL DAAC) and Google Earth Engine catalog (Gorelick et al., 2017). These included GEDI Level 4A data (Version 2.1) (Dubayah et al., 2022b) provided by NASA, as well as Sentinel-1 C-band SAR ground range detected imagery, Sentinel-2 Level 2A multispectral imagery, GLO-30 digital elevation model (DEM) (Airbus, 2020) and WorldCover land cover product (Zanaga et al., 2021) provided by the Copernicus Program. All EO datasets were split into tiles of 300 km × 300 km and analyzed using batch processing in Google Earth Engine.

2.2.1. GEDI data

First, GEDI Level 4A data available within Australia and the United States between January and December 2020 were filtered by rejecting all invalid AGBD measurements not meeting the Level 4A product

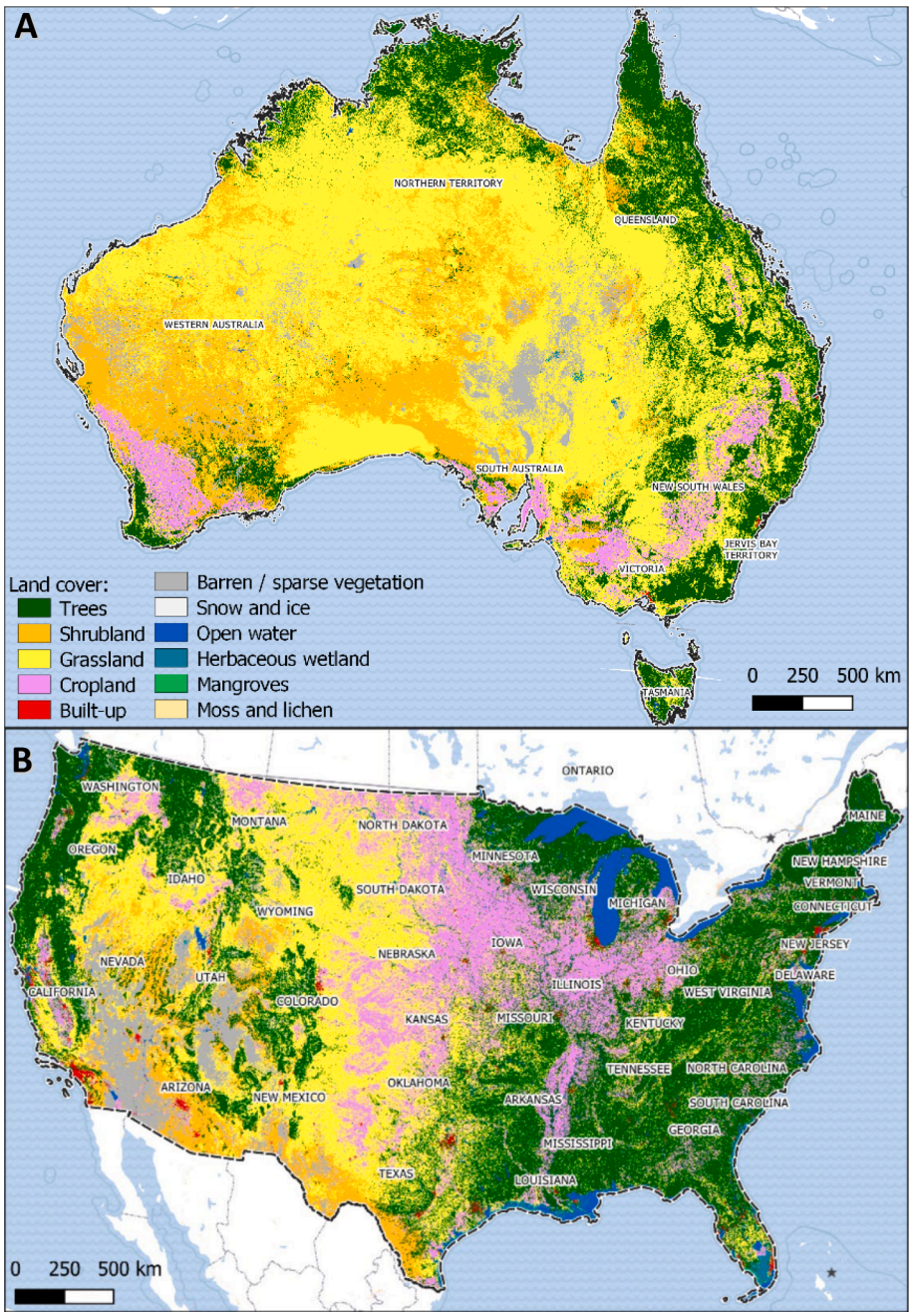


Fig. 1. Land cover of (A) Australia and (B) the main contiguous United States (Zanaga et al., 2021).

quality requirements (*'l4_quality_flag'* = 0) (Dubayah et al., 2022b; Kellner et al., 2021). Then unreliable GEDI measurements with a relative standard error in GEDI-derived AGBD exceeding 50% (*'agbd_se' / 'agbd'* × 100 > 50) were also removed. In addition, GEDI measurements on slopes (derived from GLO-30 DEM) > 30° were removed as they previously showed to adversely affect accurate retrieval of both terrain and tree heights (Liu et al., 2021). Finally, GEDI data within built-up, barren and sparse vegetation, herbaceous wetland, snow and ice, open water, and moss and lichen land cover according to the WorldCover (i.e. map codes of 50, 60, 70, 80, 90, and 100) product were also rejected. Out of all available GEDI measurements within Australia (442,611,946) and the United States (666,739,999), only 22,839,956 (5.2%) and 44,230,815 (6.6%) remained after the data filtering for further processing, respectively (Fig. 2A and 3A). All filtered GEDI AGBD measurements used an optimal algorithm setting group for their derivation (Dubayah et al., 2022b; Kellner et al., 2021). In Australia and the United States this resulted in 85% and 15%, and 53% and 47% of GEDI measurements relying on algorithm setting group 1 and 2, respectively.

Given that GEDI footprint centers had horizontal geolocation accuracy of ± 10 m only (Roy et al., 2021), it was impractical to match GEDI with the EO data directly. Therefore, filtered GEDI measurements were aggregated within 100 m × 100 m and 200 m × 200 m cells, and the average AGBD was calculated for cells containing ≥ 4, ≥ 5, ≥ 6, and ≥ 6, ≥ 7, ≥ 8 GEDI measurements, respectively (Figs. 2 and 3). These thresholds were chosen to evaluate the effect of sample size on predictions while preserving the spatial distribution of GEDI data and maximizing the number of cells with at least two GEDI orbital tracks per cell. Only GEDI cells containing predominantly shrubland, grassland, cropland, tree, and mangrove land covers were preserved for further analysis. In total, up to 175,502 (100 m × 100 m) and 263,699 (200 m × 200 m) GEDI cells across Australia, and up to 489,607 (100 m × 100 m) and 807,775 (200 m × 200 m) GEDI cells across the United States were

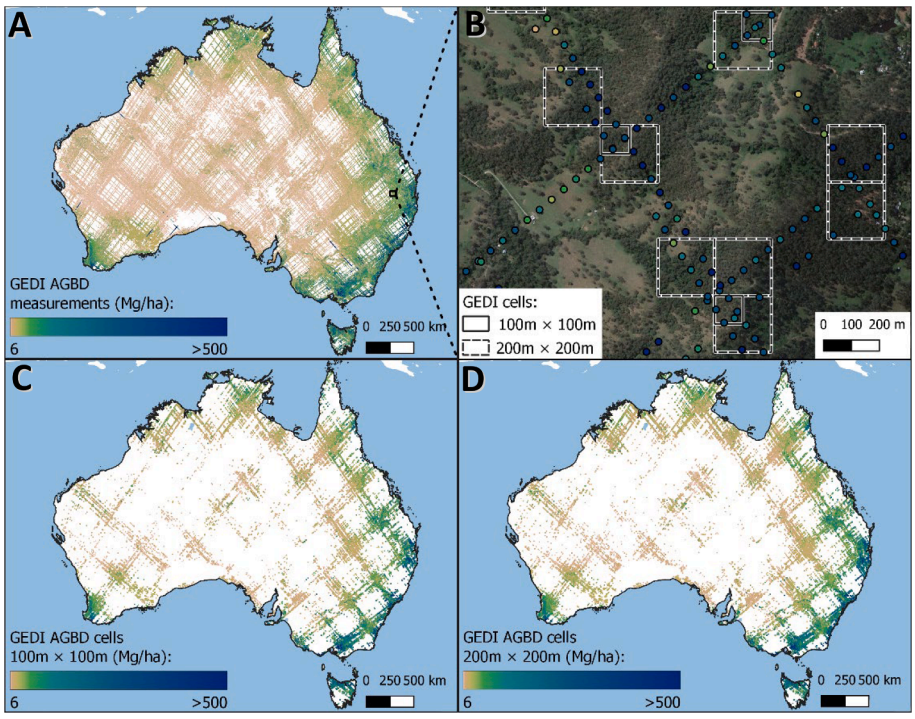


Fig. 2. GEDI data available between January – December 2020 within Australia: (A) after filtering, (B) zoomed-in area showing GEDI measurement aggregation into cells, (C) aggregated into 100 m × 100 m cells containing ≥ 4 GEDI measurements, (D) aggregated into 200 m × 200 m cells containing ≥ 6 GEDI measurements.

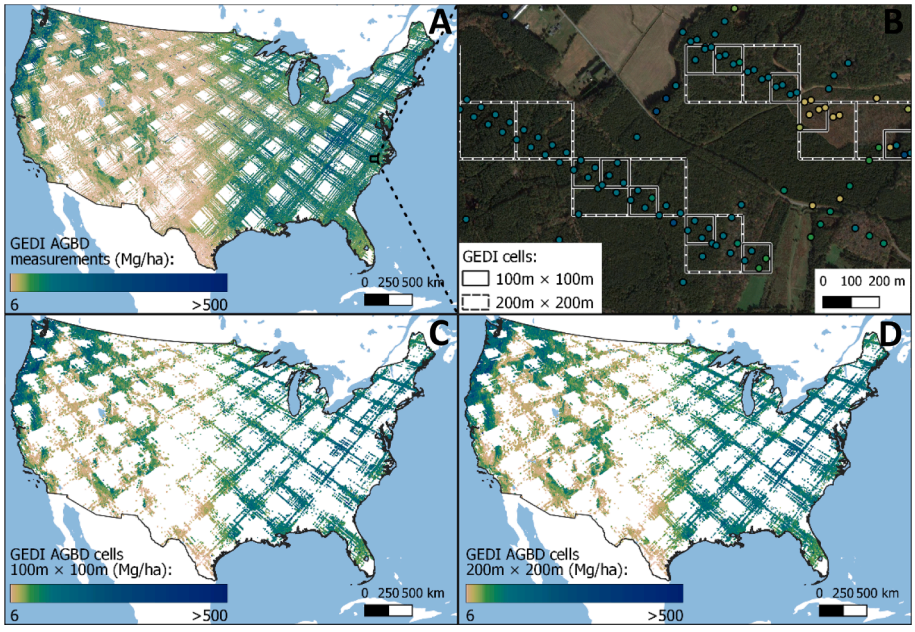


Fig. 3. GEDI data available between January – December 2020 within the United States: (A) after filtering, (B) zoomed-in area showing GEDI measurement aggregation into cells, (C) aggregated into 100 m × 100 m cells containing ≥ 4 GEDI measurements, (D) aggregated into 200 m × 200 m cells containing ≥ 6 GEDI measurements.

used to train machine learning models for estimating AGBD.

2.2.2. Earth observation data

The satellite imagery consisted of C-band SAR imagery collected from Sentinel-1A and 1B satellites and multispectral imagery collected from Sentinel-2A and 2B satellites covering Australia and the United States. Both Sentinel-1 and Sentinel-2 imagery were available in the Google Earth Engine catalog. The dual-polarization (VV-VH) Sentinel-1 ground range detected and logarithmically scaled (dB) SAR imagery were radiometrically calibrated, thermal noise and terrain corrected using Sentinel-1 Toolbox (Veci et al., 2017). Sentinel-2 multispectral imagery was limited to 10 spectral bands at 10 m and 20 m resolutions, which were pre-processed using Sen2Cor to Level 2A surface reflectance (Main-Knorn et al., 2017) (Table 1).

To generate spatially explicit estimates of AGBD across Australia and the United States, both Sentinel-1 and Sentinel-2 imagery were reduced to annual median composites within 300 km × 300 km tiles, and then mosaicked to the full extent of the countries. The median compositing

Table 1
Sentinel-1 and Sentinel-2 imagery used in this study (Note: central wavelength represents an average of Sentinel-2A and 2B imagery).

Satellite	Band name	Central wavelength (nm)	Spatial resolution (m)
Sentinel-1	Co-polarized (VV)	5.5×10^7	20
	Cross-polarized (VH)	5.5×10^7	20
Sentinel-2	Blue (B)	492	10
	Green (G)	559	10
	Red (R)	665	10
	Red Edge 1 (RE1)	704	20
	Red Edge 2 (RE2)	740	20
	Red Edge 3 (RE3)	781	20
	Near-infrared 1 (NIR1)	833	10
	Near-infrared 2 (NIR2)	864	20
	Shortwave infrared 1 (SWIR1)	1612	20
	Shortwave infrared 2 (SWIR2)	2194	20

was used to reflect a typical appearance of the landscape in 2020 while preserving original pixel values. For this, all available images of Sentinel-1 and Sentinel-2 imagery between January and December 2020 were used. Sentinel-1 imagery was composited by taking a median of each band of all images in descending orbit for Australia and ascending orbit for the United States at the resolution of 20 m. Sentinel-1 mosaic was also used to calculate *VV*/*VH* ratio, as it has shown to be suitable for monitoring forest phenology (Soudani et al., 2021).

In contrast, Sentinel-2 imagery was first masked from clouds and cloud shadows, where clouds were identified from the Sentinel-2 cloud probability dataset (Sinergise, 2022), while cloud shadows were defined by cloud projection intersection with low-reflectance pixels of near-infrared (*NIR1*) band (Google, 2022). Then, similar to Sentinel-1 imagery, Sentinel-2 imagery was composited by taking a median of all cloud and cloud shadow-masked images at the resolution of 20 m. Sentinel-2 imagery was further used to derive normalized difference spectral indices (*NDSIs*) that previously showed to drive biomass estimation in crops (Shendryk et al., 2021). *NDSIs* were mainly calculated to overcome calibration issues of individual spectral bands associated with atmospheric, illumination, and topographic conditions, thus making them more consistent for a large area analysis. In total, 45 Sentinel-2 derived *NDSIs* were calculated in succession from Blue (*B*) to shortwave infrared (*SWIR2*) spectral bands (Table 1):

$$NDSI_{SB_i,SB_j} = \frac{(SB_i - SB_j)}{(SB_i + SB_j)},$$

where *SB* is the spectral band, and *i* and *j* denote its wavelengths (nm). For clarity, *NDSIs* in this paper are referred to as a combination of band names in Table 1 (e.g. $NDSI_{R,NIR1}$, which is an inverted version of the normalized difference vegetation index (*NDVI*)).

In addition to Sentinel-1 and Sentinel-2 satellite imagery, EO data consisted of digital elevation model (DEM) and land cover data, which were also available in the Google Earth Engine catalog. Global GLO-30 DEM tiles at a native resolution of 30 m were mosaicked and resampled using bilinear interpolation to 20 m to match the resolutions of Sentinel-1 and Sentinel-2 imagery, thus creating a seamless DEM for Australia and the United States. GLO-30 DEM was derived from the WorldDEM product acquired between 2011 and 2015, which was locally infilled using multiple satellite imagery-derived products (Airbus, 2020). DEM was further used to calculate the aspect and slope of the terrain. Similarly, the WorldCover land cover product at a native resolution of 10 m was also resampled to 20 m using mode interpolation to match the rest of the EO datasets. This product contained 11 land cover types (Fig. 1) and was generated by ESA using supervised machine learning of Sentinel-1 and Sentinel-2 imagery collected in 2020 (Zanaga et al., 2021). Finally, all the EO 300 km × 300 km composites were mosaicked and clipped to the extent of Australia and the United States.

2.3. Predictor variables

As GEDI measurements were averaged within 100 m × 100 m and 200 m × 200 m cells, it was intuitive to use these cells for aggregating EO datasets, thus forming a data input for machine learning analysis. It is a common procedure in object-based remote sensing to calculate summary statistics for objects of interest to highlight their spectral, geometrical, and spatial properties (Hossain & Chen, 2019). Therefore, predictor variables in this study were calculated from Sentinel-1, Sentinel-2, and DEM mosaics by extracting summary statistics for each 100 m × 100 m (i.e. 25 pixels at 20 m resolution) and 200 m × 200 m (i.e. 100 pixels at 20 m resolution) GEDI cells. In total, seven statistics were calculated summarizing all pixels within each GEDI cell: average (*avg*), standard deviation (*std*), 2nd percentile (*p2*), 25th percentile (*p25*), 50th percentile (*p50*), 75th percentile (*p75*) and 98th percentile (*p98*). The 2nd and 98th percentiles were used as proxies for minimum and maximum values to minimize the effect of outliers in the data. As opposed to other EO mosaics, the land cover (*LC*) that formed the majority (*maj*) of each 100 m × 100 m and 200 m × 200 m GEDI cell was used as a predictor (i.e. LC^{maj}). For clarity, the predictor variables in this paper are denoted as a combination of the name of the summary statistics and EO data used for their calculation (e.g. $NDSI^{avg}_{R,NIR1}$ corresponds to the average of all pixel values within a GEDI cell of $NDSI_{R,NIR1}$). In total, 358 predictor variables were generated: 21 were extracted from three Sentinel-1 derivatives (i.e. *VV*, *VH*, and *VV/VH*), 315 were extracted from 45 Sentinel-2 derived *NDSIs*, 21 were extracted from three DEM derivatives (i.e. *DEM*, *ASPECT*, *SLOPE*), and final predictor was extracted from the land cover (*LC*).

2.4. AGBD estimation

In this study, Light Gradient Boosting Machine (LightGBM) implementation of the gradient boosting algorithm was used. Multiple LightGBM models were built to calculate potential improvements in model performance when combining satellite imagery with elevation and land cover data and to explore the necessity of using all predictors for estimating AGBD. These included: (1) Sentinel-2 and land cover data (*S2 + LC data*), (2) Sentinel-1, Sentinel-2, DEM, and land cover data (*S1 + S2 + DEM + LC data*), and (3) top 15% of most important Sentinel-1, Sentinel-2, DEM and land cover derived predictors (*top S1 + S2 + DEM + LC data*).

2.4.1. Machine learning

The data consisting of up to 263,699 GEDI cells in Australia and up to 807,775 GEDI cells in the United States were divided into *'train'* (80%) and a hold-out *'test'* (20%) datasets. This was done in a stratified manner by binning AGBD values into four 25th percentile bins. Bayesian optimization (Bergstra et al., 2013) with a 5-fold cross-validation (see Appendix 1 for the list of hyperparameters) was used to select hyperparameters that yielded the most accurate LightGBM model in terms of the lowest quantile loss at the 50th percentile (i.e. mean absolute error). Within the 5-fold cross-validation the *'train'* dataset was further divided into training (80%) and validation (20%) datasets also stratified according to the AGBD values. The main advantage of the Bayesian hyperparameter optimization is that it considers past hyperparameter evaluations when choosing the hyperparameter set for the next assessment, thus minimizing the optimization runtime. Quantile loss was used to estimate the uncertainty of AGBD estimates by calculating a 95% prediction interval (i.e. the difference between quantile loss estimates at the 2.5th percentile and 97.5th percentile). Assuming that the error distribution was Gaussian, the 95% prediction interval was approximately within ± 2 standard deviations (SD), and the 68%

prediction interval was approximately within ± 1 SD of the AGBD estimates. Bayesian hyperparameter optimization was limited to a runtime of 30 minutes, which was sufficient to evaluate from 3 to 1859 model configurations on a 32-core machine depending on the number of predictor variables and AGBD measurements used. According to the average 5-fold cross-validation quantile loss at the 50th percentile, the most accurate model was further evaluated using the *'test'* dataset. To fully leverage the training data, final models for inference (i.e. to generate per-pixel AGBD and its uncertainty maps for Australia and the United States) were trained on all data using optimized hyperparameters (Fig. 4).

The distribution of GEDI-derived AGBD values was extremely right-skewed and ranged between 6.3 and 1970.9 Mg/ha in Australia, and 6.1 and 1978.3 Mg/ha in the United States, which posed an imbalance problem for the LightGBM regressor. Therefore, prior to the machine learning analysis, AGBD values were transformed to the normal distribution using a square root transformation. LightGBM can also have a significant bias in regression (e.g. overestimating low values and underestimating high values); thus, a simple linear regression (SLR) was used to reduce it. First, an SLR was fit with the measured and estimated AGBD values on the *'train'* dataset, and then the estimated values were updated using an SLR model on the *'test'* dataset (Song, 2015).

As mentioned before, the accuracy of the LightGBM models for estimating AGBD was evaluated using a quantile loss at the training stage. In addition, the coefficient of determination (R^2), root-mean-square error (*RMSE*), and relative root-mean-square error (*RMSE%*) were additionally calculated to evaluate final models on the *'test'* dataset. *RMSE* and *RMSE%* were calculated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}, RMSE\% = \frac{RMSE}{\bar{x}} \times 100$$

where x_i is the measured AGBD value, $\hat{x}_i$is the estimated AGBD value, n is the total number of AGBD measurements, and $\bar{x}$ is the mean of the measured AGBD values.

Finally, a procedure for predictor selection was introduced to decrease the computational cost of model training and potentially improve the accuracy of the LightGBM models by reducing the number of input predictors. In this study, SHapley Additive exPlanations (SHAP) method was used to identify the most important predictors. SHAP is a game-theoretic approach that relies on *shapley* values to explain predictions of machine learning models (Lundberg et al., 2020). The importance of predictors is calculated by comparing the model's performance with and without a predictor for every combination of predictors. The results of the SHAP analysis were used to train additional LightGBM models using only 15% (i.e. 54 out of 358) of the most important predictors (i.e. with the highest absolute *shapley* values) corresponding to the *top S1 + S2 + DEM + LC data*.

A total of 36 LightGBM models were trained and optimized for estimating AGBD using predictors extracted from (1) *S2 + LC data* (i.e. Sentinel-2 and land cover data), (2) *S1 + S2 + DEM + LC data* (i.e. Sentinel-1, Sentinel-2, DEM and land cover data) and (3) *top S1 + S2 + DEM + LC data* (i.e. *S1 + S2 + DEM + LC data* reduced to the top 15% most important predictors). This was done at different GEDI cell sizes (i.e. 100 m × 100 m and 200 m × 200 m) containing a different number of GEDI measurements (i.e. ≥ 4, ≥ 5, and ≥ 6 for GEDI cell sizes of 100 m × 100 m and ≥ 6, ≥ 7 and ≥ 8 for GEDI cell sizes of 200 m × 200 m) and for the two countries. In addition, 48 LightGBM models were trained and optimized using the *S1 + S2 + DEM + LC data* for four individual land covers (i.e. trees (forested lands), shrublands, grasslands, and croplands) to evaluate model performance for different vegetation types. This was necessary for understanding the reliability of the proposed methodology in areas for which GEDI L4A data was primarily calibrated (i.e. trees) and areas with limited (i.e. shrublands) and unavailable (i.e. grasslands and croplands) calibration (Duncanson et al., 2022). Finally, models trained on the *top S1 + S2 + DEM + LC data* using GEDI cell size of 100 m × 100 m and containing ≥ 5 GEDI measurements were used for generating per-pixel maps of AGBD and its uncertainty across Australia

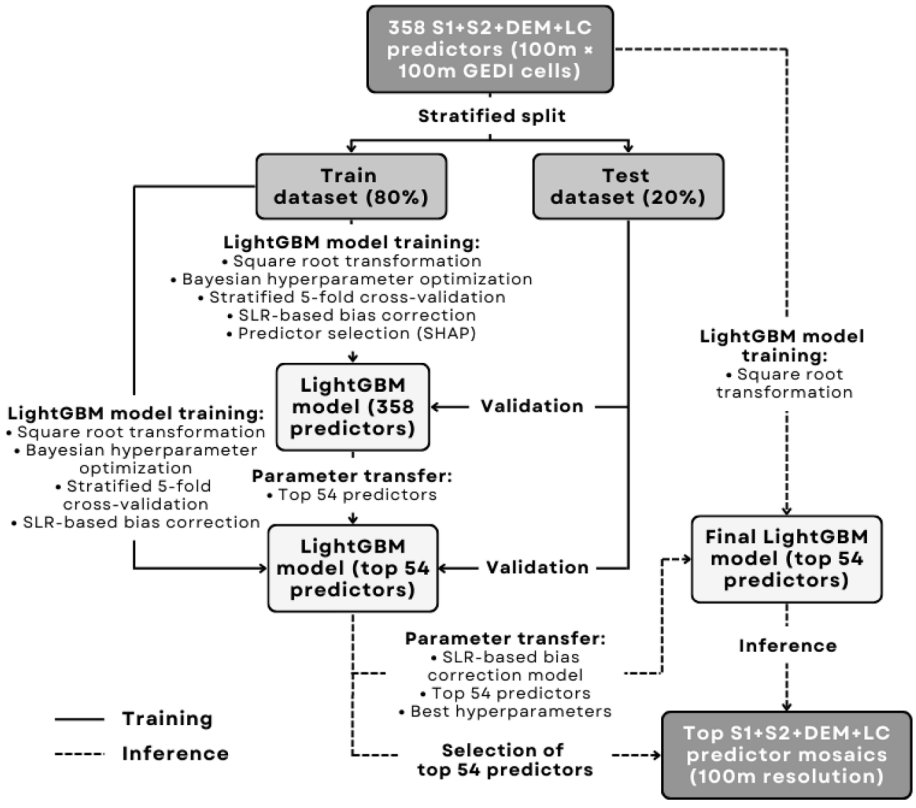


Fig. 4. Example of the machine learning workflow for AGBD estimation at 100 m resolution.

and the United States. The final workflow for generating AGBD and its per-pixel uncertainty can be seen in Fig. 5.

2.5. Independent accuracy assessment

The data from the Forest Inventory and Analysis (FIA) Program of the US Forest Service were used to evaluate the accuracy of the AGBD maps and compare them to the existing products. FIA's national network contains > 500,000 field plots distributed using stratified random sampling across forests (i.e. ≥ 4000 m^2 in size and ≥ 10% canopy cover) of the United States (Gray et al., 2012). Each plot contains four subplots of ~ 7.3 m in radius (i.e. area of ~ 168.1 m^2) where all trees with DBH ≥ 0.13 m were measured (Bechtold & Patterson, 2005). In this study, AGBD for each FIA plot was calculated using a Component Ratio Method in rFIA package (Stanke et al., 2020). To reduce the temporal gap between remote sensing-derived AGBD maps and FIA measurements, FIA data were only limited to plots that were measured after 2010 (i.e. 140,981 plots). Due to geolocation inaccuracies (e.g. some plot locations were randomly displaced up to 1 km from their true locations) of FIA data, the direct evaluation of AGBD maps using FIA plots was impractical. Therefore, FIA plots were further aggregated by calculating the average AGBD within 50 km^2 hexagons that contained at least 4 FIA plots following the minimum number of plots used by FIA to define its strata for estimation (Bechtold & Patterson, 2005). Furthermore, only 50 km^2 hexagons where a samples size of ≥ 4 FIA plots resulted in a relative standard error (RSE) ≤ 25% (assuming simple random sampling) of the mean AGBD estimate produced in this study within each hexagon were selected for further analysis using sgsR package (Goodbody et al., 2022). A relatively fine scale of 50 km^2 for FIA plot aggregation was chosen to highlight potential AGBD saturation issues in high biomass areas that would be otherwise diluted at coarser resolutions, while the RSE ≤ 25% threshold allowed the use of only homogeneous plots for validation. In this way, 4,051 aggregated 50 km^2 hexagon plots that contained 4 – 10 FIA plots were generated to validate AGBD maps (Fig. 6A and 6B).

For comparisons, the most recent available for download global AGBD product developed by CCI for 2018 at 100 m resolution (Fig. 7A) (Santoro et al., 2021) as well as GEDI L4B product for 2019–2021 at 1 km resolution (Dubayah et al., 2022a) were used (Fig. 7B). The CCI product was derived using the BIOMASAR algorithm that inverts a semi-empirical model relating the forest backscatter from SAR C-band Sentinel-1 and L-band ALOS-2 PALSAR-2 to global estimates of AGBD. In contrast, GEDI L4B product was produced by aggregating filtered GEDI L4A data between 18 April 2019 to 4 August 2021 within 1 km × 1 km cells. All three AGBD maps (CCI, L4B, and the one produced in this study) were evaluated using 4,051 FIA-derived plots. To ensure a fair comparison, the AGBD map produced in this study was clipped to the same extent as the CCI product (i.e. all land covers except forests according to the Copernicus Global Land service land cover datasets (Buchhorn et al., 2020) were masked out) and GEDI L4B product (i.e. 1 km × 1 km cells without AGBD estimates due to incomplete spatial coverage as the result of persistent clouds in some areas and the orbital dynamics of the ISS (Dubayah et al., 2022a) were also masked out) within the United States.

While a biomass library containing nearly 11,000 field-measured plots also exists for Australia (TERN, 2016), it was unsuitable for validating AGBD maps in this study due to inconsistencies in its sampling design and as some of its data were used for calibrating both CCI and GEDI-derived AGBD products (Kellner et al., 2021; Santoro et al., 2021).

3. Results

The analysis based on the *S2 + LC data* resulted in AGBD estimated with R^2 of 0.61 – 0.71, *RMSE* of 59 – 86 Mg/ha, and *RMSE%* of 45 – 83%. However, model performance improved with the addition of Sentinel-1 and DEM predictors (i.e. *S1 + S2 + DEM + LC data*): AGBD estimation with R^2 of 0.66 – 0.74, *RMSE* of 55 – 81 Mg/ha, and *RMSE%* of 41 – 77%. Furthermore, model performance deteriorated only marginally when using the *top S1 + S2 + DEM + LC data*: AGBD estimation with R^2 of 0.65 – 0.73, *RMSE* of 56 – 81 Mg/ha, and *RMSE%* of 42 – 79% (Table 2).

After calculating the number of occurrences of the top 15% most important predictors (according to the SHAP approach) across all models, it was found that Sentinel-2 derived *NDSIs* using Green, Red Edge, NIR, and SWIR bands (i.e. $NDSI^{std}_{NIR1,NIR2}$, $NDSI^{std}_{RE3,NIR1}$, $NDSI^{p75}_{SWIR1,SWIR2}$, $NDSI^{p50}_{G,NIR2}$, $NDSI^{p25}_{G,NIR1}$, $NDSI^{p25}_{SWIR1,SWIR2}$, $NDSI^{avg}_{G,NIR2}$, and $NDSI^{avg}_{SWIR1,SWIR2}$), DEM-derived slope ($SLOPE^{std}$, $SLOPE^{p98}$, $SLOPE^{p75}$, $SLOPE^{p50}$, $SLOPE^{avg}$, and $SLOPE^{p25}$), land cover (i.e. LC^{maj}) and Sentinel-1 derived *VV/VH* ratio ($S1^{p25}_{VV/VH}$) were the most important in estimating AGBD.

It is clear from Table 2 that AGBD estimation at fine resolution (i.e. 100 m) came at the cost of accuracy. AGBD prediction models at 200 m resolution generally outperformed the models trained at 100 m resolution with an R^2 improvement of 0.02 – 0.08 and *RMSE* and *RMSE%* reduction of 1 – 22 Mg/ha and 1 – 11%, respectively. Furthermore, there was no substantial performance difference between models trained using different GEDI measurement thresholds (i.e. ≥ 4, ≥ 5, ≥ 6 for 100 m × 100 m GEDI cells (Fig. 8) and ≥ 6, ≥ 7, ≥ 8 for 200 m × 200 m GEDI cells). Overall, this resulted in model performance differences of R^2 of ≤ 0.04, *RMSE* of ≤ 12 Mg/ha, and *RMSE%* of ≤ 6%. It is also important to

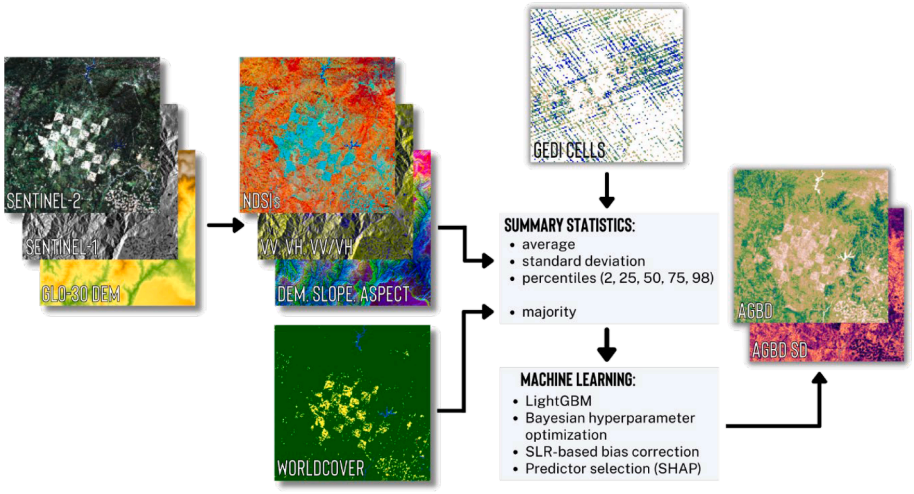


Fig. 5. Workflow for estimating and mapping the AGBD and its uncertainty (i.e. standard deviation (SD)) over Australia and the United States for 2020.

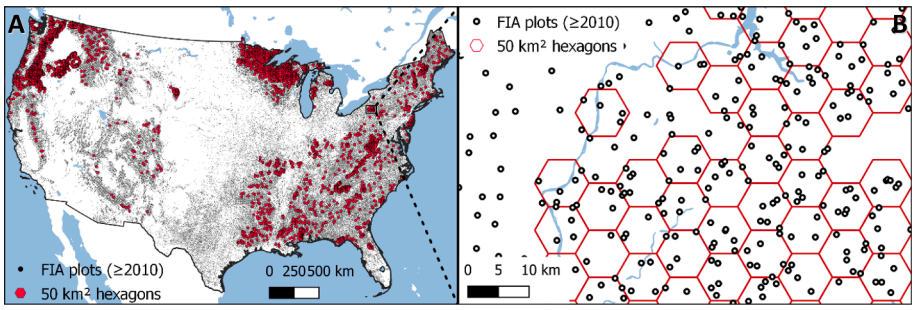


Fig. 6. FIA plot AGBD data: (A) the location of 140,981 FIA plots and 4,051 FIA-derived 50 km^2 hexagon plots across the United States, and (B) zoom-in area showing the aggregation procedure for generating FIA-derived 50 km^2 hexagon plots.

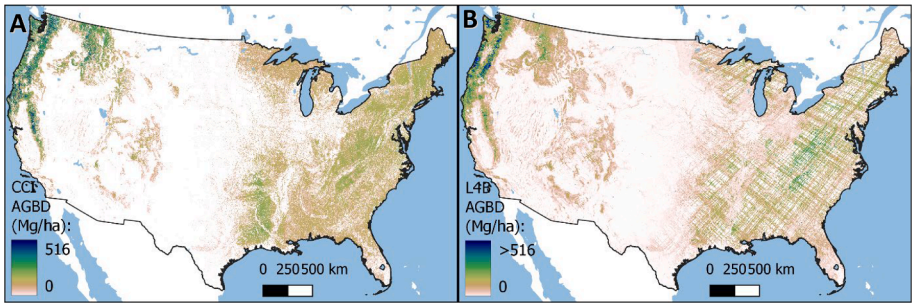


Fig. 7. Third-party AGBD maps used in this study: (A) CCI AGBD product for 2018 at the resolution of 100 m, and (B) GEDI L4B AGBD product for 2019–2021 at the resolution of 1 km.

Table 2
AGBD estimation accuracy based on the 'test' dataset (i.e. 20% of the total GEDI cells).

Country	Resolution (m)	GEDI		S1 + S2 + DEM + LC data (358 predictors)			Top S1 + S2 + DEM + LC data (54 predictors)			S2 + LC data (316 predictors)		
		measurements	'test' cells	R^2	RMSE (Mg/ ha)	RMSE %	R^2	RMSE (Mg /ha)	RMSE %	R^2	RMSE (Mg /ha)	RMSE %
United States	100	≥ 4	97,887	0.68	69.5	50	0.66	71.3	52	0.63	75.3	55
United States	100	≥ 5	20,274	0.7	68.7	48	0.68	71.1	50	0.64	74.8	52
United States	100	≥ 6	5,054	0.66	80.6	52	0.65	81.0	53	0.61	86.3	56
United States	200	≥ 6	161,465	0.72	59.3	43	0.71	60.2	44	0.68	64.2	47
United States	200	≥ 7	93,531	0.73	59.0	42	0.72	60.2	43	0.69	64.1	46
United States	200	≥ 8	50,482	0.74	58.6	41	0.72	60.3	42	0.69	63.9	45
Australia	100	≥ 4	35,098	0.71	58.0	77	0.69	59.8	79	0.66	62.4	83
Australia	100	≥ 5	6,762	0.69	61.2	77	0.68	62.3	78	0.65	65.6	82
Australia	100	≥ 6	1,215	0.71	59.7	71	0.72	58.7	70	0.69	61.3	73
Australia	200	≥ 6	52,732	0.74	54.9	69	0.73	56.4	71	0.71	58.5	74
Australia	200	≥ 7	29,978	0.74	55.6	70	0.72	57.0	71	0.7	59.1	74
Australia	200	≥ 8	15,450	0.73	57.1	70	0.72	58.8	72	0.69	61.2	75

note that bias correction using SLR improved the accuracy of the models in terms of R^2 by ≤ 0.03, *RMSE* by ≤ 1.5 Mg/ha, and *RMSE%* by ≤ 2%. However, it could be seen in Fig. 8 that even after bias correction using SLR, there was still a visible overestimation of low AGBD values and underestimation of high values, while prediction errors increased with AGBD estimates (i.e. errors were heteroscedastic and proportional to the estimate).

According to the LightGBM modeling performed for the individual land cover types (Table 3), the models for AGBD estimation of trees trained on the *S1 + S2 + DEM + LC data* performed similarly to the models for AGBD estimation of all land cover types (R^2 of 0.66 – 0.77, *RMSE* of 53 – 75 Mg/ha and *RMSE%* of 38 – 66%). However, the accuracy dropped significantly for models estimating the AGBD of shrublands (R^2 0.32 – 0.52, *RMSE* of 6 – 19 Mg/ha and *RMSE%* of 39 – 65%), grasslands (R^2 0.39 – 0.57, *RMSE* of 12 – 25 Mg/ha and *RMSE%* of 46 – 62%) and croplands (R^2 of 0.14 – 0.68, *RMSE* of 20 – 38 Mg/ha and *RMSE%* of 43 – 85%) (Table 3).

The prediction maps of AGBD at 100 m resolution for Australia and the United States in 2020 are presented in Fig. 9. The maps were generated using models trained on the *top S1 + S2 + DEM + LC data* extracted within GEDI cells with ≥ 5 measurements at 100 m resolution (Fig. 8A2 and 8B2). The uncertainty maps of AGBD were also produced, and their estimates were generally proportional to the magnitude of the AGBD estimate.

According to Fig. 9B, in the United States, the highest levels of AGBD occurred in the coniferous forests of the northwest, while the lowest values occurred in the country's interior, in areas dominated by desert and semi-arid climates. In Australia, the highest AGBD was in the eucalypt forests of the southeast, while the lowest values, similar to the United States, occurred in the central parts of the country, also dominated by desert and semi-arid climates (Fig. 9A). In Australia, AGBD estimates ranged up to 884.2 Mg/ha with a mean of 24.6 Mg/ha, while AGBD uncertainty (i.e. SD) ranged up to 484.8 Mg/ha with a mean of 7.7 Mg/ha. Similarly, in the United States AGBD estimates ranged up to

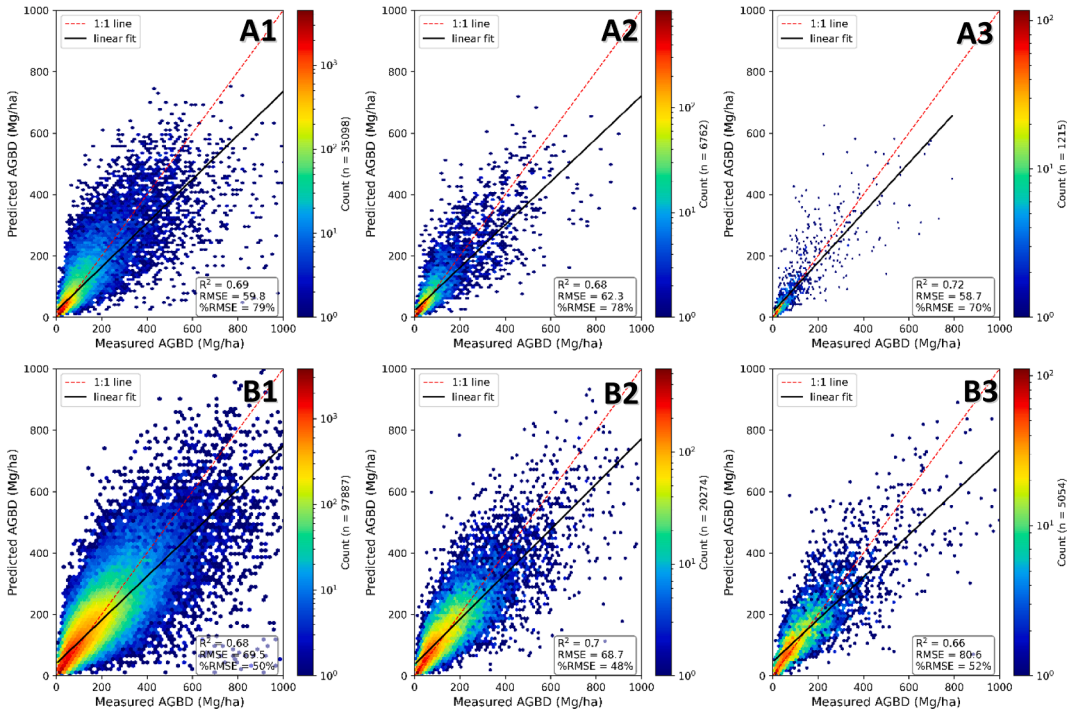


Fig. 8. The ability of models based on the top S1 + S2 + DEM + LC data to estimate AGBD at 100 m resolution on the 'test' dataset (i.e. 20% of the total GEDI cells) in (A) Australia and (B) the United States using (1) ≥ 4, (2) ≥ 5, and (3) ≥ 6 GEDI measurements per GEDI cell.

Table 3
AGBD estimation accuracy based on the 'test' dataset (i.e. 20% of the total GEDI cells) of individual land cover types.

Country	Resolution (m)	GEDI measurements	Trees				Shrublands			
			'test' GEDI cells	R^2	RMSE (Mg/ha)	RMSE %	'test' GEDI cells	R^2	RMSE (Mg/ha)	RMSE %
United States	100	≥4	86,009	0.66	70.8	47	3,917	0.35	19.4	65
United States	100	≥5	18,167	0.67	70.8	46	792	0.34	18.7	63
United States	100	≥6	4,594	0.66	74.8	46	172	0.32	17.9	60
United States	200	≥ 6	143,339	0.72	60.1	40	5,736	0.42	17.9	57
United States	200	≥ 7	84,502	0.72	59.5	39	3,039	0.39	17.3	56
United States	200	≥ 8	46,184	0.73	58.7	38	1,520	0.36	17.4	57
Australia	100	≥ 4	26,821	0.71	60.3	66	3,075	0.45	7.2	47
Australia	100	≥ 5	5,248	0.73	59.6	62	605	0.4	9.9	63
Australia	100	≥ 6	929	0.68	62.8	62	124	0.32	8.9	55
Australia	200	≥ 6	40,486	0.76	55.5	58	4,231	0.52	6.4	41
Australia	200	≥ 7	23,791	0.77	54.0	57	2,228	0.51	7.0	43
Australia	200	≥ 8	12,618	0.77	52.7	57	1,058	0.45	6.2	39
Country	**Resolution (m)**	**GEDI measurements**	**Grasslands**				**Croplands**			
			'test' GEDI cells	R^2	RMSE (Mg/ha)	RMSE %	'test' GEDI cells	R^2	RMSE (Mg/ha)	RMSE %
United States	100	≥ 4	7,662	0.42	23.0	62	165	0.59	25.1	52
United States	100	≥ 5	1,273	0.42	23.7	60	20	0.19	38.4	85
United States	100	≥ 6	281	0.41	24.7	60	3	n/a	n/a	n/a
United States	200	≥ 6	11,900	0.51	21.2	53	325	0.58	28.8	47
United States	200	≥ 7	5,799	0.50	20.7	52	118	0.68	24.9	43
United States	200	≥ 8	2,697	0.51	20.7	51	45	0.4	33.1	63
Australia	100	≥ 4	5,086	0.47	15.1	57	22	0.3	24.2	58
Australia	100	≥ 5	892	0.39	14.9	58	2	n/a	n/a	n/a
Australia	100	≥ 6	159	0.39	14.1	55	0	n/a	n/a	n/a
Australia	200	≥ 6	7,842	0.57	13.0	49	49	0.47	20.0	44
Australia	200	≥ 7	3,881	0.57	12.4	46	18	0.14	30.6	67
Australia	200	≥ 8	1,744	0.52	13.3	49	7	n/a	n/a	n/a

1116.8 Mg/ha with a mean of 62.8 Mg/ha, while AGBD uncertainty ranged up to 360.4 Mg/ha with a mean of 30.1 Mg/ha. Importantly, AGBD uncertainty was on average only 42.2% and 43.0% of the estimated AGBD in Australia and the United States, respectively.

Total AGB stocks of vegetated areas (i.e. forests, mangroves, shrublands, grasslands, and croplands) in Australia were equal to 17.7 Pg (a mean of 24.6 Mg/ha), while in the United States they were equal to 44.1 Pg (a mean of 62.8 Mg/ha). However, total AGB stocks of only forested lands (i.e. trees and mangroves) in Australia were substantially lower, equal to 9.8 Pg (a mean of 70.6 Mg/ha), while in the United States

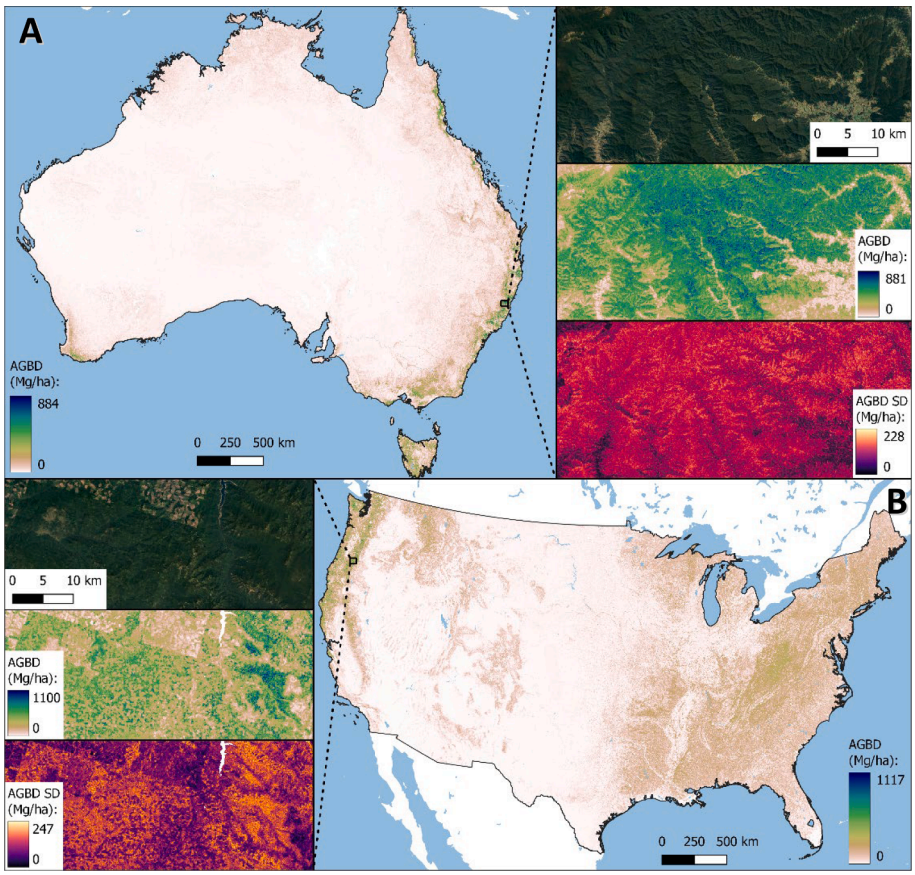


Fig. 9. Prediction maps of AGBD at 100 m resolution across (A) Australia and (B) the United States that were derived using the top S1 + S2 + DEM + LC data in 2020. Zoom-in areas show a satellite image, AGBD, and ± 1 SD of AGBD.

they amounted to 37.1 Pg (a mean of 132.1 Mg/ha). These total AGB estimates were then compared with those from the CCI and GEDI L4B products. According to the CCI AGBD map representing forested lands only, Australia's total AGB stocks were equal to 9.2 Pg (a mean of 80.0 Mg/ha), while in the United States they were equal to 32.7 Pg (a mean of 102.0 Mg/ha). In contrast, the calculation of total AGB stocks using GEDI L4B product was challenging due to the orbital dynamics of the ISS leading to 9.1% and 17.1% of 1 km × 1 km AGBD cells missing from this dataset in Australia and the United States, respectively. However, based on the mean estimates of AGBD of forested lands and their area GEDI L4B total AGB stocks were previously estimated to be 12.1 Pg (a mean of 15.7 Mg/ha) and 42.7 Pg (a mean of 54.7 Mg/ha) in Australia and the United States, respectively (Dubayah et al., 2022a).

The validation of the AGBD map for 2020 against the one from CCI for 2018 and GEDI L4B for 2019 – 2021 using 4,051 FIA-derived 50 km^2 hexagon plots in the United States is presented in Fig. 10. The AGBD map produced in this study outperformed the CCI product in terms of all accuracy metrics (i.e. R^2 increase of 0.05, *RMSE* decrease of 4 Mg/ha, and *RMSE%* decrease of 2%) (Fig. 10A and 10B). Both maps saturated in regions of high AGBD. However, this was much more prominent in the CCI map, where saturation started in areas with an AGBD of only > 300 Mg/ha. In contrast, the AGBD map produced in this study performed on

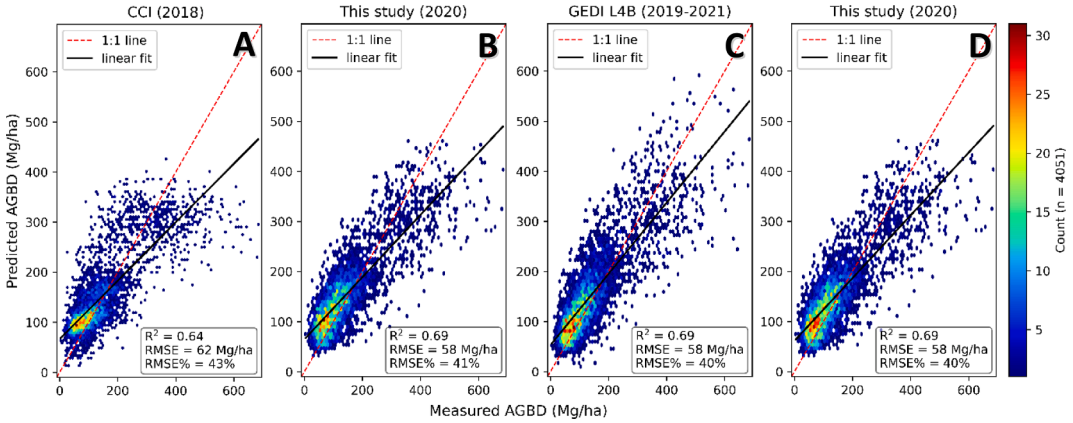


Fig. 10. The validation of AGBD maps using 4,051 FIA-derived 50 km^2 hexagon plots in the United States: (A) CCI AGBD map for 2018 against (B) AGBD map generated in this study for 2020 masked using forested land extent from the CCI product, and (C) GEDI L4B AGBD map for 2019–2021 against (D) AGBD map generated in this study for 2020 masked using available data extent from the GEDI L4B product.

par with the GEDI L4B product for 2019 – 2021 (i.e. R^2 of 0.69, *RMSE* of 58 Mg/ha, and *RMSE%* of 40%) (Fig. 10C and 10D).

4. Discussion

The methodology developed in this study could inform forest management decisions and facilitate carbon budget accounting, as accurate AGBD estimation using the proposed modeling framework is feasible at multiple spatial and temporal scales. The models estimated AGBD with R^2 of 0.68 – 0.74, *RMSE* of 55 – 81 Mg/ha, and *RMSE%* of 42 – 77% in Australia and the United States in 2020. The proposed modeling framework could be used for generating annual wall-to-wall maps of AGBD in other regions (and potentially globally) at a resolution as fine as 100 m. While multiple AGBD estimation studies developed separate models for different plant functional (Duncanson et al., 2022) or land cover types (Saatchi et al., 2022), this study demonstrated that such information could be directly incorporated as a predictor, thus minimizing model training time (by creating a single global model), while avoiding the loss of accuracy due to each model having a smaller number of samples to be trained from.

Generally, AGBD estimation models at 200 m resolution outperformed the models trained at 100 m resolution. This could be attributed to geolocation uncertainty in GEDI measurements and the edge effects of large tree crowns located at the boundaries of GEDI cells (Duncanson et al., 2022). However, further reduction of the resolution to, for example, 50 m is expected to deteriorate the estimation accuracy of AGBD as each GEDI footprint center had horizontal geolocation accuracy of ± 10 only (Roy et al., 2021). This would also result in the reduction of valid GEDI measurements per cell, which, surprisingly, did not lead to a substantial accuracy loss in this study (Tables 2 and 3). In contrast, a further increase of the resolution to, for example, 500 m will most likely improve the estimation accuracy of AGBD, as the effects of geolocation uncertainty and edge effects will be largely diluted at such a scale.

The use of GEDI measurements aggregated within pre-defined cells potentially introduced uncertainties due to the landscape heterogeneity inside the cells. However, the alternative of matching GEDI measurements with the EO data directly would likely lead to even higher uncertainties. Similarly, the calculation of median composites of Sentinel-1 and Sentinel-2 imagery could be considered overly simplistic, resulting in the loss of information necessary to improve the accuracy of AGBD estimation. Therefore, future studies should explore extracting temporal predictors from Sentinel-1 and Sentinel-2 imagery, for example, monthly or seasonal medians, or fitting double logistic or harmonic regressions to EO data (Di Tommaso et al., 2021). As opposed to the median, the mean compositing can result in pixels containing values that were not part of the original imagery, while maximum and minimum compositing reflect the non-typical appearance of the landscape, and in the case of Sentinel-2 imagery can result in pixels contaminated by clouds and cloud shadows due to their high and low reflectance, respectively. However, instead of generating median composites from EO data, a better approach would be calculating geomedians, which maintain the relationship between spectral bands (Roberts et al., 2017). LightGBM models could also be equipped with the ability to learn temporal or geographical priors, by feeding them with EO data timestamps and geographical coordinates (in a suitable cyclic encoding) as additional predictors (Lang et al., 2022). All of this should improve the performance of models, especially in areas prone to disturbances in short periods of time (e.g. croplands and grasslands). Moreover, the addition of proprietary satellite L-band SAR imagery (e.g. ALOS-2 PALSAR-2 or SAOCOM 1A and 1B) that is known to saturate less than C-band SAR imagery (i.e. Sentinel-1) when estimating biomass could improve the performance of the models even further (Imhoff, 1995). Finally, an alternative approach using sparse supervision of CNNs should be able to produce more highly resolved AGBD estimates (Lang et al., 2022). However, this will likely come at the expense of accuracy and substantial loss in model training and inference speeds.

The models for estimating the AGBD of trees (i.e. forested lands) outperformed the ones for estimating AGBD in shrublands, grasslands, and croplands (Table 3) (Note: the accuracy metrics calculated for AGBD estimation in croplands should be interpreted with caution due to a relatively small *'test'* dataset of $\leq$ 325 GEDI cells). This was expected as GEDI data was mainly calibrated in forested areas in Oceania and North America (2,619 training samples) with limited calibration in grasslands and woodlands (83 training samples) and no calibration in croplands (Kellner et al., 2021). Grasslands and croplands are also prone to disturbances in short periods of time (e.g. due to harvesting, bushfires, etc.), which was not captured in the annual median composites of Sentinel-1 and Sentinel-2 imagery. Furthermore, GEDI filtering based on RSE in GEDI-derived AGBD exceeding 50% in this study resulted in the removal of all AGBD measurements $<$ 6 Mg/ha, which were mainly associated with shrublands, grasslands, and croplands. This further highlights the unreliability of GEDI measurements across these land covers, and potentially led to AGBD overestimation in areas of low biomass (<6 Mg/ha) in this study. Interestingly, Grassland, Shrub and Woodland (GSW) model of GEDI Level 4A product showed the highest accuracy among all plant functional types (R^2 of 0.86 and *RMSE%* of 55.4%) (Duncanson et al., 2022). However, EO data were not able to exploit the high accuracy of the GSW model, with shrubland and grassland models achieving R^2 of only 0.32 – 0.57 and *RMSE%* of 39 – 65% in this study (Table 3). This could be partly related to the unrepresentativeness of the GSW model as well as potential noise in the EO data as, for example, bare ground and water pixels were not masked out when calculating summary statistics within 100 m × 100 m and 200 m × 200 m GEDI cells.

In this study, AGBD estimates were mainly driven by predictors extracted from Sentinel-2 derived *NDSIs* using Green, Red Edge, NIR, and SWIR bands, elevation-derived slope, land cover, and Sentinel-1 derived *VV/VH* ratio. Similar *NDSIs* have previously been shown to be important in estimating crop yields (Shendryk et al., 2021), while terrain slope was also reported to be correlated with AGBD (Ferry et al., 2010). However, the importance of the DEM-derived slope could be partly an artifact of the GEDI algorithm leading to height estimation errors over slopes in the Level 2A product, leading to subsequent overestimation of AGBD in such areas. In this study GEDI measurements on slopes $>$ 30° were removed as they previously showed to adversely affect accurate retrieval of both terrain and tree heights (Liu et al., 2021). However, AGBD maps produced in this study were still positively correlated with areas of moderate slopes (>15° and $<$ 30°). Therefore, more conservative filtering of GEDI measurements based on the slope steepness (e.g. $>$ 15°) could further improve the models' performance. It is also important to note that SHAP is only one way of identifying the predictor importance of machine learning models (albeit one of the most stable ones). Different methods (e.g. permutation or gain methods) or the use of SHAP after removing correlated predictor variables could result in the most important predictors being rather different (Hooker et al., 2021).

Even after data filtering based on the Level 4A product quality flag (Kellner et al., 2021), as well as relative standard errors of AGBD, there were still a lot of erroneous GEDI measurements (e.g. high AGBD estimates in areas of moderate slopes (>15° and $\leq$ 30°) as well as in barren and sparse vegetation areas, etc.). It was previously reported that GEDI Level 4A (Version 2.1) orbit granules were affected by the presence of low cloud/fog (Dubayah et al., 2022a), which was not addressed in this study. Some of these erroneous measurements were removed based on land cover filtering prior to averaging AGBD within GEDI cells. However, it was impossible to identify erroneous GEDI measurements in full based on the information provided in GEDI Level 4A (Version 2.1) data, which inadvertently reduced the accuracy of models in this study. With further revisions and calibration of the GEDI Level 4A product as well as additional filtering using outlier detection algorithms, improved accuracy of the AGBD estimation using the proposed machine learning

method is expected. Further filtering of GEDI measurements to only include high-power beams (e.g. *'BEAM0101'*, *'BEAM0110'*, *'BEAM1000'*, and *'BEAM1011'*) collected at night (i.e. *'solar_elevation'* $<$ 0) to ensure penetration of canopies of $>$ 95% canopy cover and avoid the negative impact of background solar illumination on GEDI waveforms could also improve model performance (Dubayah et al., 2021).

The proposed methodology could be potentially used to assess annual or seasonal AGBD changes. However, this should be done with extreme caution as per-pixel uncertainty of AGBD estimates could be substantially higher than seasonal or annual change in biomass due to forest growth or disturbance (e.g. logging or bushfires). For example, in Australia, the average growth rate of forests ranges between 0.5 and 12.9 Mg/ha/year (Australian Government, 2021), while the AGBD uncertainty (i.e. AGBD SD) estimated in this study across Australia ranged up to 484.8 Mg/ha/year with a mean of 7.7 Mg/ha/year (i.e. 42.2% of the estimated AGBD). Therefore, it is advised to use the per-pixel uncertainty of AGBD maps to understand the reliability of the estimates. Model training for monitoring seasonal AGBD changes is also complicated by the fact that GEDI-derived AGBD within cells could vary on average by 30 – 38 Mg/ha from season to season within the same year (based on the estimates of intersecting GEDI 100 m $\times$ 100 m cells containing $\geq$ 3 GEDI measurements in 2020). Furthermore, in this study, the errors of GEDI measurements were not propagated to per-pixel AGBD uncertainty. If errors from GEDI measurements (due to modeling errors, allometries, geolocation, etc.) were to be considered, then the per-pixel uncertainty of AGBD would be either much larger due to error accumulation or lower due to error cancellation. It was reported that GEDI Level 4A product has an uncertainty (i.e. *RMSE%*) of 28.7 – 79.0% depending on the geographic region and plant functional type (Duncanson et al., 2022).

The main advantages of the proposed method for AGBD estimation could be deduced when compared to previous AGBD products (i.e. CCI for 2018 and GEDI L4B for 2019 – 2021). First, the proposed method generated a product with overall higher accuracy and a visibly lower level of saturation in areas with high AGBD (i.e. $>$ 300 Mg/ha) as compared to the CCI product (Fig. 10A and 10B). The proposed method potentially mitigated the saturation in high biomass areas by exploiting a large variety (up to 358) of EO-derived predictors. Second, the proposed method performed on par with the GEDI L4B product (Fig. 10C and 10D), while offering AGBD maps at a much higher resolution (100 m vs 1 km). While it is clear that the GEDI L4B product exhibited limited saturation in high biomass areas, the method proposed in this study could effectively fill its current observation gaps and increase its spatial resolution while maintaining accuracy. In addition, while the validation using FIA-derived 50 km^2 hexagons was applicable to forested lands only, it was off the AGBD mapping resolution of 100 m and 200 m in this study. Finer resolution validation sets for forested lands as well as croplands, grasslands, and shrublands in both Australia and the United States are needed to fully understand the accuracy of the proposed method. Third, the proposed method solely relies on open access EO data that could be readily accessed in the Google Earth Engine catalog. While, for example, the CCI product for 2018 relies on freely distributed annual ALOS-2 PALSAR-2 mosaics, its application at different temporal or spatial resolutions would require a substantial investment to purchase additional satellite imagery. Fourth, the method proposed in this study offers a straightforward way to estimate AGBD uncertainty using a quantile loss, and it could be easily modified to derive AGBD at different spatial and temporal resolutions. Finally, comparisons of total AGB stocks of forested lands showed the method proposed in this study estimated more AGB for both countries as compared to the CCI product (9.8 Pg vs 9.2 Pg in Australia, and 37.1 Pg vs 32.7 Pg in the United States). However, the predictions of this study were lower than that of the GEDI L4B product, which estimated total AGB stocks at 12.1 Pg and 42.7 Pg in Australia and the United States, respectively. These differences are likely related, in part, to models' accuracy, different degrees of saturation in high biomass areas, and disagreement in the extent of forested lands between all three products.

5. Conclusions

In this study, a machine learning approach for fusing open access GEDI, Sentinel-1, Sentinel-2, elevation, and land cover data for large area AGBD mapping is proposed. Models performed well with R^2 of 0.66 – 0.74, *RMSE* of 55 – 81 Mg/ha, and *RMSE%* of 41 – 77%. The AGBD estimation was mainly driven by Sentinel-2, and land cover-derived predictors, while their fusion with Sentinel-1 and elevation-derived predictors boosted the performance of the models (R^2 increase of 0.03 – 0.05, *RMSE* decrease of 4 – 6 Mg/ha and *RMSE%* decrease of 4 – 6%). Unsurprisingly, the AGBD prediction at fine resolution (i.e. 100 m) came at the cost of accuracy, with the most reliable predictions in forested areas. Improving AGBD estimation in shrublands, grasslands, and croplands will potentially require additional GEDI calibration in those land cover types. The proposed method solely relies on open access EO data and provides less saturation in regions of high biomass as compared to previous AGBD products. Prediction maps produced in this study could serve as a baseline for current AGB stocks of forested lands equal to 9.8 Pg and 37.1 Pg in Australia and the United States, respectively. This study also highlights methodological opportunities for combining GEDI and EO data towards annual and seasonal AGBD mapping at the global scale through data fusion. Overall, using the proposed methodology could potentially overcome the common challenges of existing EO-based methods: 1) signal saturation at high AGBD levels, 2) reliance on proprietary datasets and 3) temporal limitation to annual estimates.

Declaration of Competing Interest

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

Data availability

Data will be made available on request.

Acknowledgments

This work was funded by Dendra Systems. I would like to thank Dr. Susan Graham and Jonathan Miller from Dendra Systems for providing constructive feedback and suggestions on the draft of the manuscript. This work would also not be possible without the open data policies of NASA's GEDI mission and ESA's Copernicus Program, which I greatly appreciate.

Appendix A

In this study, hyperparameters of LightGBM in regression mode were optimized using Bayesian hyperparameter optimization. Refer to LightGBM (Ke et al., 2017) and (Microsoft, 2022) for hyperparameter description (*Note: *alpha* for quantile loss was set to 0.5 (i.e. 50th percentile, which corresponds to the mean absolute error. Once the best hyperparameters were identified, alpha was changed to 0.025 (i.e. 2.5th percentile) and 0.975 (i.e. 97.5th percentile) to estimate AGBD uncertainty).

Hyperparameter	Range
n_estimators	*1000*
early_stopping_rounds	*50*
loss (*objective*)	*"quantile"*
*alpha**	*0.025, 0.5, 0.975*
learning_rate	*0.1*
max_depth	*2 – 15 (step: 1)*
num_leaves	*10 – 1000 (step: 5)*
boosting_type	*"gbdt"*
subsample	*0.5 – 1.0 (step: 0.1)*
colsample_bytree	*0.5 – 1.0 (step: 0.1)*
min_child_samples	*20 – 1000 (step: 5)*

References

ABARES, 2018. Australia's state of the forests report 2018. https://www.agriculture.gov.au/sites/default/files/abares/forestsaustralia/documents/sofr_2018/web accessible pdfs/SOFR_2018_web.pdf.

Airbus, 2020. Copernicus DEM Product Handbook (v3.0). November, 1–38.

Australian Government, 2021. National Inventory Report 201 9 (The Australian Government Submission to the United Nations Framework Convention on Climate Change Australian National Greenhouse Accounts) (Vol. 2, Issue April).

Bechtold, W.A., Patterson, P.L., 2005. The Enhanced Forest Inventory and Analysis Program — National Sampling Design and Estimation Procedures. USDA General Technical Report, SRS-80, 85.

Bergstra, J., Yamins, D., Learning, D. C. B. T.-P. of the 30th I. C. on M. (2013). Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures (S. Dasgupta & D. McAllester (eds.); Vol. 28, Issue 1, pp. 115–123). PMLR. http://proceedings.mlr.press/v28/bergstra13.pdf.

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., Kasneci, G., 2021. Deep neural networks and tabular data: A survey. February, 1–20. http://arxiv.org/abs/2110.01889.

Buchhorn, M., Smets, B., Bertels, L., Lesiv, M., Tsendbazar, N., Masiliunas, D., Linlin, L., Herold, M., Fritz, S., 2020. Copernicus Global Land Service: Land Cover 100m: Collection 3: epoch 2015: Globe (Version V3.0.1). Remote Sensing 12 (6), 1044.

Csillik, O., Kumar, P., Mascaro, J., O'Shea, T., Asner, G.P., 2019. Monitoring tropical forest carbon stocks and emissions using Planet satellite data. Sci. Rep. 9 (1), 1–12. https://doi.org/10.1038/s41598-019-54386-6.

Csillik, O., Reiche, J., De Sy, V., Araza, A., Herold, M., 2022. Rapid remote monitoring reveals spatial and temporal hotspots of carbon loss in Africa's rainforests. Commun. Earth Environ. 3 (1), 1–8. https://doi.org/10.1038/s43247-022-00383-z.

Di Tommaso, S., Wang, S., Lobell, D.B., 2021. Combining GEDI and Sentinel-2 for wall-to-wall mapping of tall and short crops. Environ. Res. Lett. 16 (12), 125002.

Dubayah, R., Armston, J., Healey, S.P., Bruening, J.M., Patterson, P.L., Kellner, J.R., Duncanson, L., Saarela, S., Ståhl, G., Yang, Z., Tang, H., Blair, J.B., Fatoyinbo, L., Goetz, S., Hancock, S., Hansen, M., Hofton, M., Hurtt, G., Luthcke, S., 2022a. GEDI launches a new era of biomass inference from space. Environ. Res. Lett. 17 (9), 095001.

Dubayah, R., Armston, J., Kellner, J.R., Duncanson, L., Healey, S.P., Patterson, P.L., Hancock, S., Tang, H., Bruening, J., Hofton, M., Blair, J.B., Luthcke, S.B., 2022b. GEDI L4A Footprint Level Aboveground Biomass Density, Version 2.1. ORNL DAAC, Oak Ridge, Tennessee, USA.

Dubayah, R., Blair, J. B., Beck, J., Wirt, B., Armston, J., Hofton, M., Luthcke, S., Tang, H., 2021. GLOBAL Ecosystem Dynamics Investigation (GEDI) Level 2 User Guide For SDPS PGEVersion 3 (P003) of GEDI L2A Data and SDPS PGEVersion 3 (P003) of GEDI L2B Data. 3, 1–25.

Dubayah, R., Blair, J.B., Goetz, S., Fatoyinbo, L., Hansen, M., Healey, S., Hofton, M., Hurtt, G., Kellner, J., Luthcke, S., Armston, J., Tang, H., Duncanson, L., Hancock, S., Jantz, P., Marselis, S., Patterson, P.L., Qi, W., Silva, C., 2020. The global ecosystem dynamics investigation: high-resolution laser ranging of the earth's forests and topography. Sci. Remote Sens. 1 (September 2019), 100002 https://doi.org/10.1016/j.srs.2020.100002.

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., Alonso, A., Andersen, H.-E., Aplin, P., Baker, T.R., Barbier, N., Bastin, J.F., Biber, P., Boeckx, P., Bogaert, J., Boschetti, L., Boucher, P.B., Boyd, D.S., Burslem, D.F.R.P., Calvo-Rodriguez, S., Chave, J., Chazdon, R.L., Clark, D.B., Clark, D.A., Cohen, W.B., Coomes, D.A., Corona, P., Cushman, K.C., Cutler, M.E.J., Dalling, J.W., Dalponte, M., Dash, J., de-Miguel, S., Deng, S., Ellis, P.W., Erasmus, B., Fekety, P.A., Fernandez-Landa, A., Ferraz, A., Fischer, R., Fisher, A.G., García-Abril, A., Gobakken, T., Hacker, J.M., Heurich, M., Hill, R.A., Hopkinson, C., Huang, H., Hubbell, S.P., Hudak, A.T., Huth, A., Imbach, B., Jeffery, K.J., Katoh, M., Kearsley, E., Kenfack, D., Kljun, N., Knapp, N., Král, K., Krůček, M., Labrière, N., Lewis, S.L., Longo, M., Lucas, R.M., Main, R., Manzanera, J.A., Martínez, R.V., Mathieu, R., Memiaghe, H., Meyer, V., Mendoza, A.M., Monerris, A., Montesano, P., Morsdorf, F., Næsset, E., Naidoo, L., Nilus, R., O'Brien, M., Orwig, D.A., Papathanassiou, K., Parker, G., Philipson, C., Phillips, O.L., Pisek, J., Poulsen, J.R., Pretzsch, H., Rüdiger, C., Saatchi, S., Sanchez-Azofeifa, A., Sanchez-Lopez, N., Scholes, R., Silva, C.A., Simard, M., Skidmore, A., Stereńczak, K., Tanase, M., Torresan, C., Valbuena, R., Verbeeck, H., Vrska, T., Wessels, K., White, J.C., White, L.J.T., Zahabu, E., Zgraggen, C., 2022. Aboveground biomass density models for NASA's Global Ecosystem Dynamics Investigation (GEDI) lidar mission. Remote Sens. Environ. 270, 112845.

Fatoyinbo, L., Lee, S.-K., Hansen, M., Huang, C., 2022. GEDI Products. https://gedi.umd.edu/data/products/.

Ferry, B., Morneau, F., Bontemps, J.D., Blanc, L., Freycon, V., 2010. Higher treefall rates on slopes and waterlogged soils result in lower stand biomass and productivity in a tropical rain forest. J. Ecol. 98 (1), 106–116. https://doi.org/10.1111/j.1365-2745.2009.01604.x.

Goodbody, T.R., Coops, N.C., Queinnec, M., 2022. Structurally Guided Sampling. R package version 1.3.1. https://github.com/tgoodbody/sgsR.

Google, 2022. Sentinel-2 Cloud Masking with s2cloudless. https://developers.google.com/earth-engine/tutorials/community/sentinel-2-s2cloudless.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031.

Gray, A., Brandeis, T., Shaw, J., McWilliams, W., Miles, P., 2012. Forest inventory and analysis database of the United States of America (FIA). Biodivers. Ecol. 4 (Jovan), 225–231. https://doi.org/10.7809/b-e.00079.

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. Nat. Clim. Change 11 (3), 234–240. https://doi.org/10.1038/s41558-020-00976-6.

Hooker, G., Mentch, L., Zhou, S., 2021. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. Statist. Comput. 31 (6), 1–22. https://doi.org/10.1007/s11222-021-10057-z.

Hossain, M.D., Chen, D., 2019. Segmentation for Object-Based Image Analysis (OBIA): a review of algorithms and challenges from remote sensing perspective. ISPRS J. Photogramm. Remote Sens. 150 (February), 115–134. https://doi.org/10.1016/j.isprsjprs.2019.02.009.

Imhoff, M.L., 1995. Radar backscatter and biomass saturation: ramifications for global biomass inventory. IEEE Trans. Geosci. Remote Sens. 33 (2), 511–518. https://doi.org/10.1109/36.377953.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y., 2017. LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 2017-Decem(Nips), 3147–3155.

Kellner, J.R., Armston, J., Duncanson, L., 2021. Algorithm Theoretical Basis Document (ATBD) for GEDI Level-4A (L4A) Footprint Level Aboveground Biomass Density.

Kumar, L., Mutanga, O., 2017. Remote sensing of above-ground biomass. Remote Sensing 9 (9), 1–8. https://doi.org/10.3390/rs9090935.

Lang, N., Jetz, W., Schindler, K., Wegner, J.D., 2022. A high-resolution canopy height model of the Earth. http://arxiv.org/abs/2204.08322.

Liu, A., Cheng, X., Chen, Z., 2021. Performance evaluation of GEDI and ICESat-2 laser altimeter data for terrain and canopy height retrievals. Remote Sens. Environ. 264, 112571. https://doi.org/10.1016/j.rse.2021.112571.

Lu, D., Chen, Q., Wang, G., Liu, L., Li, G., Moran, E., 2016. A survey of remote sensing-based aboveground biomass estimation methods in forest ecosystems. Int. J. Digital Earth 9 (1), 63–105. https://doi.org/10.1080/17538947.2014.990526.

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2 (1), 56–67. https://doi.org/10.1038/s42256-019-0138-9.

Main-Knorn, M., Pflug, B., Louis, J., Debaecker, V., Müller-Wilm, U., Gascon, F., 2017. Sen2Cor for Sentinel-2. October, 3. https://doi.org/10.1117/12.2278218.

Matasci, G., Hermosilla, T., Wulder, M.A., White, J.C., Coops, N.C., Hobart, G.W., Bolton, D.K., Tompalski, P., Bater, C.W., 2018. Three decades of forest structural dynamics over Canada's forested ecosystems using Landsat time-series and lidar plots. Remote Sens. Environ. 216 (August), 697–714. https://doi.org/10.1016/j.rse.2018.07.024.

Microsoft, 2022. LightGBM documentation. https://lightgbm.readthedocs.io/en/latest/Parameters.html.

Nicolau, A.P., Herndon, K., Flores-Anderson, A., Griffin, R., 2019. A spatial pattern analysis of forest loss in the Madre de Dios region, Peru. Environ. Res. Lett. 14 (12), 124045.

Roberts, D., Mueller, N., McIntyre, A., 2017. High-dimensional pixel composites from earth observation time series. IEEE Trans. Geosci. Remote Sens. 55 (11), 6254–6264. https://doi.org/10.1109/TGRS.2017.2723896.

Rodríguez-Veiga, P., Wheeler, J., Louis, V., Tansey, K., Balzter, H., 2017. Quantifying forest biomass carbon stocks from space. Current Forestry Reports 3 (1), 1–18. https://doi.org/10.1007/s40725-017-0052-5.

Roy, D.P., Kashongwe, H.B., Armston, J., 2021. Science of Remote Sensing The impact of geolocation uncertainty on GEDI tropical forest canopy height estimation and change monitoring. Sci. Remote Sens. 4, 100024 https://doi.org/10.1016/j.srs.2021.100024.

Ruefenacht, B., Finco, M.V., Nelson, M.D., Czaplewski, R., Helmer, E.H., Blackard, J.A., Holden, G.R., Lister, A.J., Salajanu, D., Weyermann, D., Winterberger, K., 2008. Conterminous U.S. and Alaska forest type mapping using forest inventory and analysis data. Photogramm. Eng. Remote Sens. 74 (11), 1379–1388. https://doi.org/10.14358/PERS.74.11.1379.

Saatchi, S., Xu, L., Yang, Y., 2022. JPL 2020 Global Biomass Dataset. https://ceos.org/gst/jpl-biomass.html.

Santoro, M., Kay, H., Lucas, R., Quegan, S., 2021. CCI Biomass product user guide (year 3, version 3.0).

Shendryk, Y., Davy, R., Thorburn, P., 2021. Integrating satellite imagery and environmental data to predict field-level cane and sugar yields in Australia using machine learning. Field Crops Res., 260, 107984. https://doi.org/10.1016/J.FCR.2020.107984.

Sinergise, 2022. Sentinel Hub's cloud detector for Sentinel-2 imagery. https://github.com/sentinel-hub/sentinel2-cloud-detector.

Song, J., 2015. Bias corrections for Random Forest in regression using residual rotation. J. Korean Statist. Soc. 44 (2), 321–326. https://doi.org/10.1016/j.jkss.2015.01.003.

Soudani, K., Delpierre, N., Berveiller, D., Hmimina, G., Vincent, G., Morfin, A., Dufrêne, É., 2021. Potential of C-band Synthetic Aperture Radar Sentinel-1 time-series for the monitoring of phenological cycles in a deciduous forest. Int. J. Appl. Earth Obs. Geoinf. 104, 102505.

Stanke, H., Finley, A.O., Weed, A.S., Walters, B.F., Domke, G.M., 2020. rFIA: An R package for estimation of forest attributes with the US Forest Inventory and Analysis database. Environ. Modell. Software 127, 104664. https://doi.org/10.1016/j.envsoft.2020.104664.

Su, H., Wang, A., Zhang, T., Qin, T., Du, X., Yan, X.H., 2021. Super-resolution of subsurface temperature field from remote sensing observations based on machine learning. Int. J. Appl. Earth Obs. Geoinf. 102, 102440 https://doi.org/10.1016/j.jag.2021.102440.

TERN, 2016. Biomass Plot Library-National collation of tree and shrub inventory data, allometric model predictions of above and below-ground biomass, Australia. Made Available by the AusCover Facility of the Terrestrial Ecosystem Research Network (TERN).

Urbazaev, M., Thiel, C., Cremer, F., Dubayah, R., Migliavacca, M., Reichstein, M., Schmullius, C., 2018. Estimation of forest aboveground biomass and uncertainties by integration of field measurements, airborne LiDAR, and SAR and optical satellite data in Mexico. Carbon Balance Manage. 13 (1) https://doi.org/10.1186/s13021-018-0093-5.

Veci, L., Lu, J., Foumelis, M., Engdahl, M., 2017. ESA's Multi-mission Sentinel-1 Toolbox. EGU General Assembly Conf. Abstr. 19, 19398.

Xu, L., Saatchi, S.S., Yang, Y., Yu, Y., Pongratz, J., Bloom, A.A., Bowman, K., Worden, J., Liu, J., Yin, Y.i., Domke, G., McRoberts, R.E., Woodall, C., Nabuurs, G.-J., de-Miguel, S., Keller, M., Harris, N., Maxwell, S., Schimel, D., 2021. Changes in global terrestrial live biomass over the 21st century. Sci. Adv. 7 (27), eabe9829.

Yan, J., Xu, Y., Cheng, Q., Jiang, S., Wang, Q., Xiao, Y., Ma, C., Yan, J., Wang, X., 2021. LightGBM: accelerated genomically designed crop breeding through ensemble learning. Genome Biol. 22 (1), 1–24. https://doi.org/10.1186/s13059-021-02492-y.

Zanaga, D., Van De Kerchove, R., De Keersmaecker, W., Souverijns, N., Brockmann, C., Quast, R., Wevers, J., Grosu, A., Paccini, A., Vergnaud, S., Cartus, O., Santoro, M., Fritz, S., Georgieva, I., Lesiv, M., Carter, S., Herold, M., Li, L., Tsendbazar, N.-E., … Arino, O. (2021). *ESA WorldCover 10 m 2020 v100*. https://doi.org/10.5281/ZENODO.5571936.