scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Global 30-m annual median vegetation height maps (2000–2022) based on ICESat-2 data and Machine Learning

Maria O. Hunter[1] ✉, Leandro Parente[2], Yu-feng Ho[2], Carmelo Bonannella[2], Laerte Guimarães Ferreira[1], Douglas Morton[3], Davide Consoli[2] & Lindsey Sloat[4]

Accurately measuring vegetation height is essential for understanding ecosystem structure, carbon storage, and biodiversity, yet global height models have overwhelmingly focused on forests, excluding ecosystems with shorter herbaceous vegetation or shrubs. To address this gap in vegetation structure data, we developed the first global estimate of median vegetation height annually from 2000–2022 at 30 m resolution, using ICESat-2 satellite Lidar, Landsat cloud free composites, and other Earth Observation raster data. Thirty two (32) million ICESat-2 20 m segments were used within 10 independent draws to build ensemble Gradient Boosted Tree (GBT) models and estimate 90% prediction intervals. Our model achieves a root mean square error (RMSE) of 2.35 m, R^2 values of 0.515 and a D^2 regression score of 0.62 estimated on the testing set. Comparisons with existing global height products show that our approach increases detail and heterogeneity of height in short vegetation ecosystems. Output maps are publicly available together with reference samples and trained models under CC-BY license.

Background & Summary

Short vegetation or grassy ecosystems, including grasslands, open shrublands, savannas, and tundra, cover roughly 40–50% of the Earth's surface[1,2]. These ecosystems support unique biodiversity, contribute significantly to carbon storage and water cycling, and provide other essential ecosystem services, while also being widely used for livestock grazing. They occur along a continuum of human-mediated transformations, ranging from intensively managed agricultural areas to semi-natural or natural areas[3]. The majority of these areas are classified as natural or semi-natural rather than as cultivated grasslands[4]. The vegetation within these landscapes is diverse, comprising various mixes of grasses, lichens, mosses, shrubs, and sparse trees, resulting in a wide range of vegetation heights and complex structural characteristics[5]. The relative abundance of these vegetation types influences primary productivity, carbon storage in both above- and below-ground biomass, and the availability of habitat for livestock and wildlife[6].

Vegetation height and structure are important for identifying and distinguishing different ecosystem types and functions[7]. Height refers to the vertical growth of plants, while structure describes the spatial arrangement or layering of plants within a landscape. These characteristics influence ecosystem function, biodiversity, and land cover classification. In forests, where canopy closure creates distinct vertical layers, traditional height metrics such as maximum canopy height (H_{max}) and Lorey's height (H_{lor}, a basal area weighted mean canopy height) have been widely used to infer biomass and ecosystem properties.

Several studies have modeled global vegetation metrics by integrating multi-sensor remote sensing data, notably Lidar – a well established active sensing technique for mapping forest structure and aboveground biomass[8–13]. Currently, two spaceborne Lidar missions are active: ICESat-2[14] and GEDI[15]. While GEDI has been used in all recent global models of top of canopy height, its coverage is limited to approximately 52° N to 52° S latitude worldwide[15]). In contrast, ICESat-2 provides unique opportunities for studying vegetation due

[1]Remote Sensing and GIS Laboratory (LAPIG/UFG), Goiânia, Brazil. [2]OpenGeoHub Foundation, Doorwerth, the Netherlands. [3]NASA Goddard Space Flight Center, Greenbelt MD, USA. [4]Land & Carbon Lab, World Resources Institute, Washington DC, USA. ✉e-mail: maria.hunter@ufg.br

to its expanded spatial coverage, footprint size, and sensor type[16]. Given the lack of continuous coverage of in-situ data with vegetation height and structure information (mostly provided by sparse forest inventory initiatives[17–19]), efforts to model these vegetation metrics at large spatial scales frequently rely on satellite Lidar as a source of ground reference. Currently, two approaches have been used to estimate vegetation height at global scale using satellite Lidar:

1. **Lidar data are aggregated at large spatial scales**. Burns *et al.*, 2024 published aggregated metrics of GEDI waveform Lidar at varying spatial resolutions (*i.e.* 1 km, 6 km and 12 km) over the five-year operational period yielding nearly full coverage between 51.6° N and 51.6° S at the 1 km scale, but much lower coverage per year. The lower resolution obscures local variability and aggregates varying land covers in fractured landscapes, but the aggregation of raw data does not incur any modeling effects.
2. **Lidar used as training data in Machine Learning (ML) framework**. In this approach, satellite imagery (*e.g.* Landsat, Setinel-2 and MAXAR) are used as input features to produce wall-to-wall predictions of top of canopy height through ML models[8,10,11]. This can expand coverage beyond the original Lidar data source and highlight local variability, but existing models consistently underestimate portions of the height continuum.

Global products of top of canopy are instrumental for monitoring forest biomass. In these ecosystems, biomass is predominantly stored in trees, and canopy height strongly correlates with carbon stocks, making maximum height-based metrics highly effective[15]. However, this approach does not translate well to non-forest ecosystems, where biomass is distributed more evenly across vegetation layers, and maximum canopy height is often determined by a small number of scattered trees or taller shrubs. In these landscapes, the use of H_{max} or H_{lor} can lead to overestimation of biomass by emphasizing the tallest vegetation rather than the dominant structural component of the ecosystem[20]. Conversely, mean height metrics can be skewed by outliers and may not accurately represent the height distribution of heterogeneous vegetated landscapes.

Recent research has emphasized the need for alternative height metrics that better capture vegetation in open ecosystems[21]. Median height (H_{median}), representing the 50th percentile of height distributions, is a promising metric for characterizing vegetation height across various vegetated landscapes. Unlike maximum height, which is influenced by rare tall individuals, median height provides a more representative measure of the dominant vegetation layer, making it suitable for both forested and open ecosystems. In forests, median height can serve as an alternative to Lorey's height, while in grasslands, shrublands, savannas, and tundra it provides a more accurate representation of the height distribution than traditional canopy height metrics. Considering that short vegetation ecosystems are highly dynamic, with seasonal and interannual variability in vegetation height and changes due to human management, animal grazing, and natural disturbances (*e.g.* fires or flood events), temporally resolved estimates are needed to capture key changes in vegetation structure and conditions[22]. By offering a height metric that is not biased toward trees, a global time series dataset of median vegetation height has the potential to improve land cover classification, biomass estimation, and ecological monitoring, enabling a better representation of ecosystem structure across diverse vegetated landscapes through time.

This study presents a global model of median vegetation height (MVH), integrating multi-sensor remote sensing data to generate improved annual height estimates across all ecosystems. While the primary focus is on improving monitoring in short vegetation ecosystems, the product offers wall-to-wall coverage that includes all terrestrial ecosystems. We provide maps of median height globally at 30-m spatial resolution from 2000–2022, along with upper and lower prediction intervals (*i.e.* 90% of probability – 5th and 95th percentiles), and trend analysis (estimated by per-pixel linear regression and filtered according to the Mann-Kendall test). The exact methodological steps are described in the following sections.

Data and Methods

ICESat-2 data preparation. The ICESat-2 satellite Lidar, in a near polar orbit since September 2018, aims to measure the height of forests and other ecosystems worldwide among other aims. Onboard, the Advanced Topographic Laser Altimeter System (ATLAS) splits a green laser pulse (532 nm) into six beams, divided into three pairs with an energy ratio between strong and weak lasers of approximately 4:1[14]. The small footprint size (11 m) and close along-track sampling distance (0.7 m) provided by the photon-counting sensor provide multiple discrete measurements within each 20 m segment that could not be attained through unmixing of larger footprint Lidar waveforms[16]. The near-polar orbit also covers the full expanse of temperate and polar climatic zones, which are often dominated by short vegetation.

The ICESat-2 ATL08 product produces summary statistics for 100 m segments, limited statistics for 20 m segments, and heights of individual photons together with classification into noise, terrain, canopy and top of canopy (Fig. 1). We downloaded the version 6 of the product for the entire world[23], and based on reported normalized height, we calculated additional metrics for each 20 m segment, including mean height, median height, 95th percentile height, the total number and fraction of photons reflected from vegetation. Information regarding the total number of photons per 20 m segment and the total number of signal photons were also included. Each segment was further defined by: time of ICESat-2 data acquisition, date of acquisition, and whether the selected segment was scanned using the strong or weak beam of the laser. Time of day was divided into night/day depending on the solar altitude angle. Lastly, all 20 m segments were converted to Parquet format resulting in 24.8 billion individual data points (Fig. 2) which are publicly accessible at https://zenodo.org/records/15198654.

Field calibration. To test the application of ICESat-2 satellite Lidar to short vegetation ecosystems we designed an initial field trial to compare ICESat-2 height metrics to field measurements within cultivated grassland (e.g. pasture) systems. The study was designed to include a variety of measured vegetation heights within

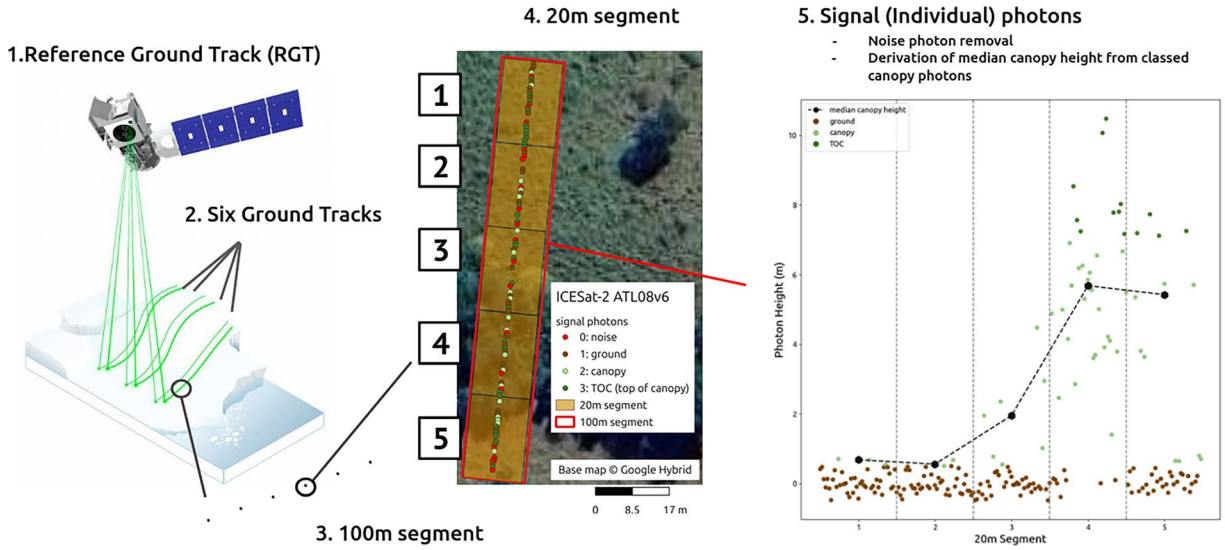

Fig. 1 Schematic of five levels of information available within the ATL08 product of ICESat-2. The Reference Ground Track, Laser Tracks, 100 m segments, 20 m segments and individual photon level information. The location of the centroid of each 20 m segment and information on individual photons including classification and relative height are used to estimate median height of vegetation.

areas consistently defined as cultivated grassland. Field measurements were conducted within the Rio Vermelho watershed in Goiás state in Brazil in October of 2023. Forty (40) sample points were selected from the 2947 ATL08 100m segments collected within the year prior to field work spanning September 2022 through September 2023 (Fig. 2).

A secondary goal was to assess filters including season, time of day, and laser strength. The selection of sample points was based on equal divisions by season of ICESat-2 overpass (wet versus dry season) and vegetation structure. The laser strength and time of day were also noted for each location. For each location sampled, pasture and shrub heights were measured and tree heights were estimated along four 90 m north-south parallel transects with 20 m separation. Field measurements were aligned with ATL08 20 m segments to identify which field measurements were closest to the ICESat-2 field of view. Field vegetation measurements greater than 10 cm in height within each segment were included to calculate the mean, median, 95th percentile and maximum height of vegetation. All metrics were compared using the RMSE and the concordance correlation coefficient[24]. This sample showed the lowest RMSE and highest CCC for median height (3.3 m and 0.19) compared to other metrics tested. The field data are publicly available with additional details on sample design and alignment with ICESat-2 segments in a Zenodo repository[25].

ICESat-2 filtering and sampling design. After our data preparation, about 24.8 billion 20 m ICESat-2 segments collected between January 2019 and October 2022 were available. To improve the overall quality of our training process, these segments were filtered according to all following criteria, established according to our field calibration and existing literature:

- Total number of photons between 3 and 60,
- Number of signal photons less than or equal to 35[26],
- Number of vegetation photons greater than or equal to 3, and
- photons acquired only at night time and by the strong beam.

Aiming to derive a comprehensive set of training data in both space and time, we randomly selected five (5) 20-segments per month and per 1 × 1 degree tile. This procedure was implemented worldwide and repeated ten (10) times, producing ten independent sample sets with a total of 32 million training points (Fig. 2)[27].

Further filtering was conducted using the GLAD global land cover[28] product. This product was overlaid with ICESat-2 segments and height quantiles of ICEsat-2 segments were calculated for each land cover class. Segments with height outside the 1st - 99th percentile height were removed, a total of 655,336 samples. Segments identified within "Terra Firme, True Desert", approximately 13% of samples, were assigned a median height of 0.001 m, the minimum height detectable in byte format.

Landsat cloud-free aggregates. The primary Earth Observation data input for our median vegetation height model was global-scale historical Landsat cloud-free aggregates at 30 m spatial resolution[29]. Spanning from 1997 to 2022, this dataset further processes the GLAD Landsat Analysis Ready Data (ARD) version 2[30] for (i) removing cloud and cloud shadow pixels, (ii) aggregating all 16-day images into bi-monthly temporal composites, and (iii) filling the missing values/data gaps using a time-series reconstruction approach (*i.e.* Seasonally

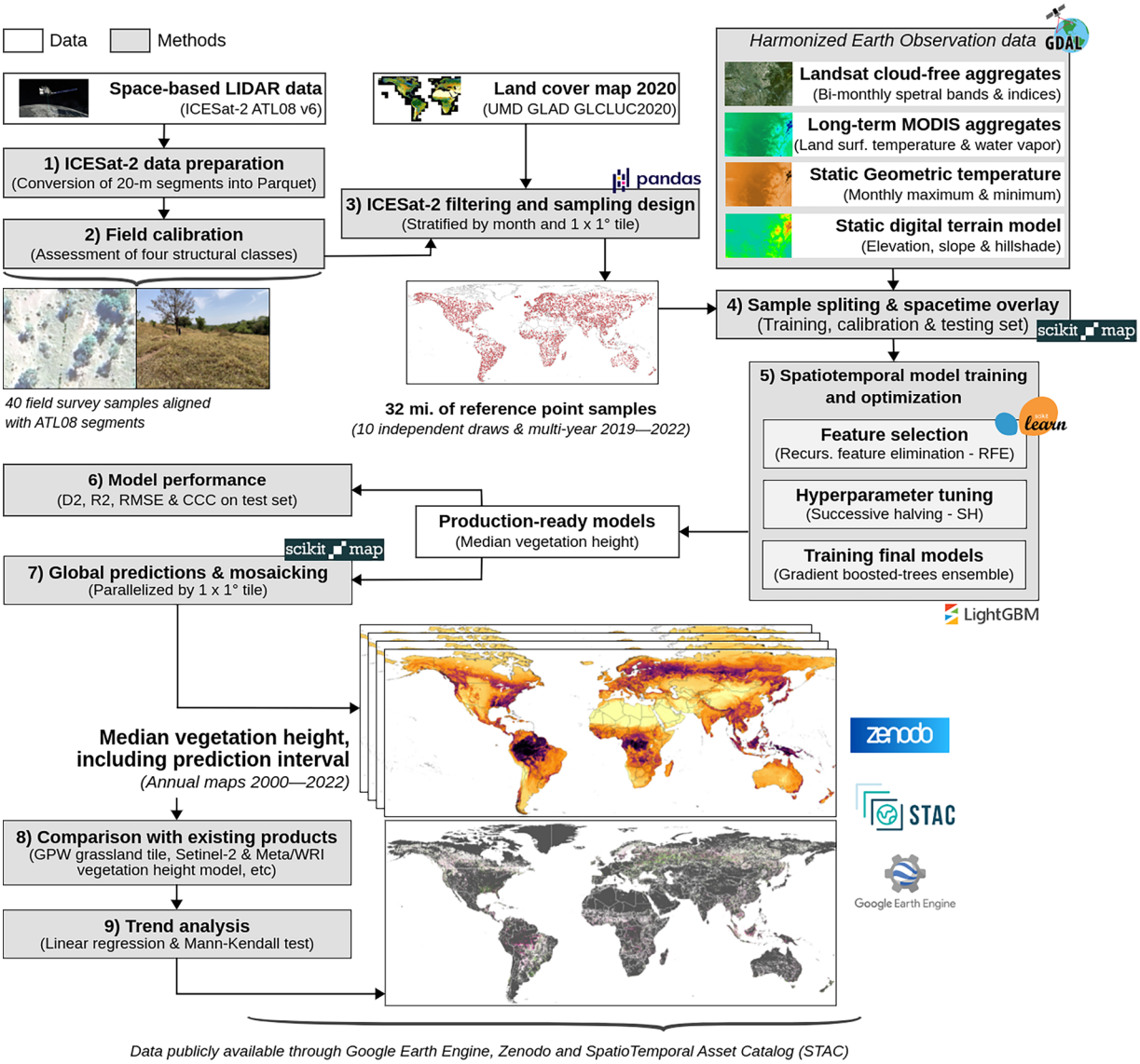

Fig. 2 Processing workflow implemented for producing the median vegetation height maps, including Earth Observation data input, reference samples, preprocessing steps and spacetime machine learning modeling.

Weighted Average). Landsat cloud-free aggregates provide complete, consistent and cross-calibrated images from Landsat 5 through Landsat 9 per year for all Landsat spectral bands (*i.e.* blue, green, red, Near-infrared — NIR, Short-wave infrared 1 — SWIR1, Short-wave infrared 2 — SWIR2, and thermal). Additionally, we derived several key Landsat-based vegetation and water indices, including Bare Soil Index (BSI)[31], Enhanced Vegetation Index (EVI)[32], the Modified Normalized Burn Ratio (NBR2), also called Normalized Difference Tillage Index (NDTI)[33], the Normalized Difference Vegetation Index (NDVI)[34], the Normalized Difference Water Index (NDWI)[35] and the near-infrared reflectance of vegetation (NIRv)[36]. Each index was calculated using distinct linear combinations of reflectance bands, providing information about vegetation health, moisture levels, burn severity, and overall ecological conditions. In addition to spectral indices, the Fraction of Absorbed Photosynthetically Active Radiation (FAPAR) was derived considering the correlation with NDVI[37]. All Landsat-derived index formulas utilized in our modeling are summarized in Table 1. The inclusion of Landsat spectral bands and indices in our modeling framework was motivated by the strong correlation of this data with vegetation structure[28], photosynthesis[36] and soil characteristics[38].

Long-term MODIS aggregates. In addition to Landsat, we used coarser spatial resolution layers that have previously shown strong correlation with vegetation height[39], specifically MODIS Land Surface Temperature and Emissivity (LST&E) (MOD11A2[40]) and Land Aerosol Optical Depth products (MCD19A2 -[41]). These products, available at 1 km spatial resolution, were processed to produce monthly long-term aggregates for daytime and nighttime surface temperatures[42] and column water vapor above the ground[43]. The aggregates were computed by estimating the median value of all monthly composite from 2000 to 2022, and resampled to 30-m by cubic spline[44], resulting in 36 input MODIS features (*i.e.* 12 layers per input variable). The inclusion of long-term

Landsat-derived Index	Abbreviation	Formula
Bare Soil Index	BSI	$\frac{(SWIR1 + RED) - (NIR + BLUE)}{(SWIR1 + RED) + (NIR + BLUE)}$
Enhanced Vegetation Index	EVI	$2.5 \times \frac{NIR - RED}{NIR + 6 \times RED - 7.5 \times BLUE + 1}$
Fraction of Absorbed Photosynthetically Active Radiation	FAPAR	$\frac{(NDVI - 0.03) \times (0.95 - 0.001)}{0.96 - 0.03} + 0.001$
Normalized Difference Tillage Index	NDTI	$\frac{SWIR1 - SWIR2}{SWIR1 + SWIR2}$
Normalized Difference Vegetation Index	NDVI	$\frac{NIR - RED}{NIR + RED}$
Normalized Difference Water Index	NDWI	$\frac{NIR - SWIR1}{NIR + SWIR1}$
Near-infrared reflectance of vegetation	NIRv	$(\frac{NIR - RED}{NIR + RED} - 0.8) \times NIR$

Table 1. List of Landsat-derived indices used by our model.

MODIS aggregates in our modeling framework was motivated by the strong correlation of temperature and water fluxes with global distribution of biomes[45].

Static raster datasets. The elevation data and terrain derivatives used in the modeling were obtained from the Ensemble Digital Terrain Model (EDTM) of the world at 30 m resolution[46], which integrated multiple sources, including ALOS AW3D[47], GLO-30[48], MERIT DEM[49], and various national DTMs. EDTM was used to compute slope, hillshade, positive and negative openness, and geomorphons at 30 and 60 m spatial resolution, using Equi7 spatial reference system[50] and the software GrassGIS[51]. In total nine (9) input features were used to provide terrain characteristics to our model framework, variables that show strong correlation with the distribution of plant functional traits[52].

Finally, we incorporated geometric temperature transformations, estimated as functions of latitude, day of the year, and elevation, following the methodology of Kilibarda[53]. These transformations generate monthly estimates of minimum and maximum temperatures, adding 24 new input features. By incorporating Earth's geometric and temporal dynamics, these variables enable the model to differentiate between locations with similar temperature patterns but distinct latitudinal positions or seasonal aspects. This distinction is especially useful for capturing regional temperature variations.

Sample splitting and spacetime overlay. To prepare the samples for our modeling approach, we (i) ran a spacetime overlay with all harmonized EO data and (ii) split them into training, calibration and testing sets. During the spacetime overlay, the pixel values of the Landsat aggregates were associated with each sample by matching the location (*i.e.* geographical coordinates) and the ICESat-2 acquisition year with 84 Landsat features in a specific year (*i.e.* seven reflectance bands and seven spectral indices for six bi-monthly periods). For long-term MODIS aggregates (*i.e.* 36 spatial layers) and static layers (*i.e.* 24 geometric temperature and nine DTM layers), the overlay considered only the sample locations, resulting in a total of 153 input features for the modeling.

Overlaid samples were split first into 90% for training/calibration, and 10% for testing, using as group strata a regular grid of 1 × 1 degree tiles. Globally distributed, the regular grid was used for ensuring that all samples inside a specific tile were assigned to either the training/calibration or testing set. Within the tiles with model training/calibration samples, we then selected 10% of samples (*e.g.* 2.7 million samples) for model optimization (*i.e.* Feature selection and hyper-parameter tuning). We kept the testing set completely isolated from model training and optimization, which allowed us to conduct an independent and unbiased technical validation of our models on a global scale (Fig. 3).

Spatiotemporal model training and optimization. Our modeling approach considered an ensemble of ten Gradient Boosted Tree (GBT) models, optimized and trained on about 29 million ICESat-2 samples (*i.e.* training and calibration set) using `LightGBM`[54]. To increase the diversity among the models, we implemented a bootstrap strategy[55] and randomly selected a different set of initial hyper-parameters for each model. The final predicted value was estimated by averaging all individual model predictions. The ten predicted values were also used to derive a prediction interval of 90% probability (*i.e.* 5th and 95th percentiles) of median vegetation height.

Optimization ran for each model separately using the calibration set (10% of bootstrapped samples), first by selecting the most important features by Recursive Feature Elimination — RFE[56], and later searching/tuning the model hyper-parameters by Successive Halving — SH[57] to maximize model performance. RFE considered a standard Random Forest model with 100 trees, measurement of quality of split by Poisson criterion and default values for other hyper-parameters (fitted using `scikit-learn`[58]). As most median vegetation heights are concentrated between 1 to 5 meters at global scale, the ML models need to address the very skewed distribution of the target variable, which requires the usage of Poisson criterion in the training[59]. For each RFE iteration, the 7 least important features were removed according to gini importance, resulting in 40 features as the final selection (*i.e.* about 26% of the total number of features).

The selected features were then used to run SH using `scikit-learn`[58], for iteratively assessing different combinations of hyper-parameter candidates bounded by a customized search space. This assessment used an early stopping strategy (on half of the calibration set) and five-fold spatial blocking cross-validation (based on 1 × 1 degree tile) to minimize the risk of over-fitting[60]. The implemented SH started with about 113,000 samples,

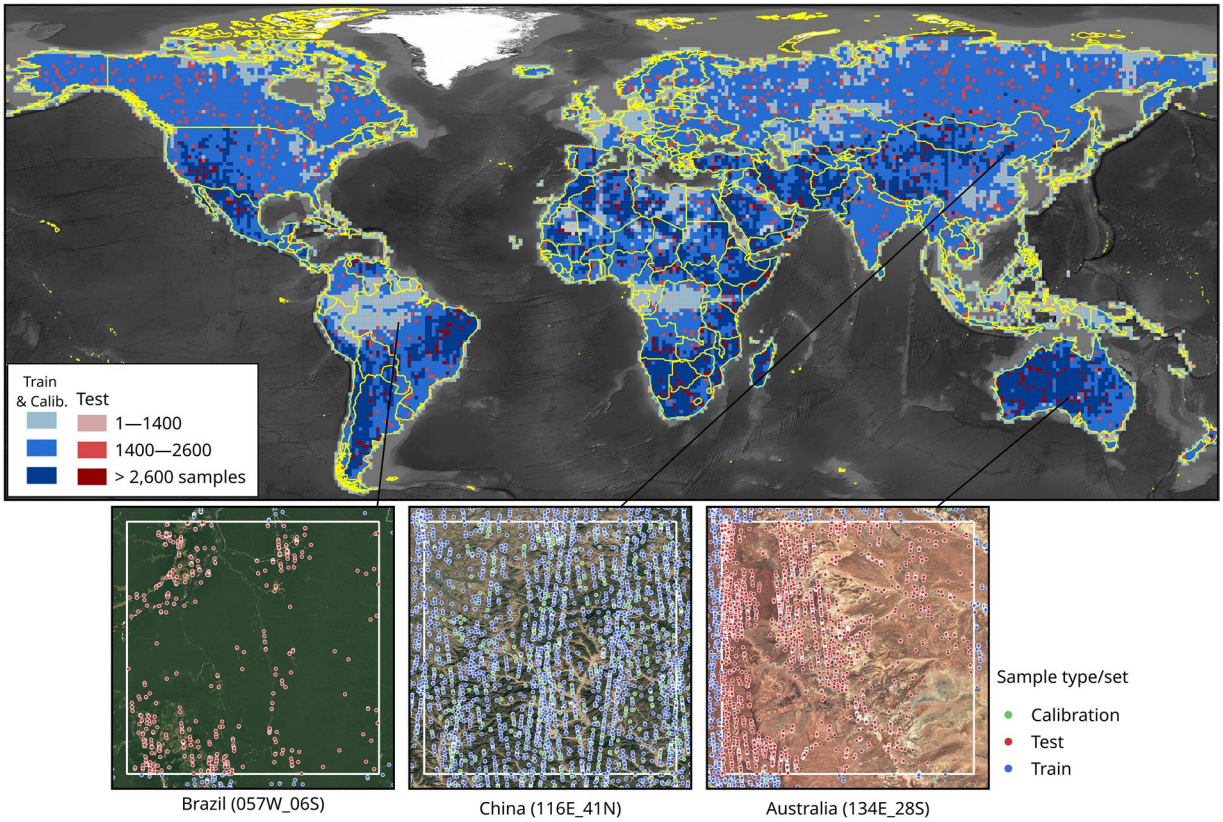

Fig. 3 Spatial distribution of 32 million ICESat-2 reference samples used for modeling global median vegetation height. Reference samples were split into 90% for model training/calibration (within the same group strata – blue tiles), and 10% for testing model performance (completely isolated group strata – red tiles).

selecting the best candidates (*i.e.* dropping half of the less accurate candidates) and doubling the number of samples per iteration until reaching the full size of the calibration set (*i.e.* about 2.7 million samples). We used the D^2 regression score metric (see equation (1)—[61]) to select the most accurate candidates, which is suitable for evaluating skewed distributions (*e.g.* Poisson distribution). Once the last iteration was done, the hyper-parameters with high D^2 were selected for each ML model. The optimization found ten sets of hyper-parameters, which were used to train the ten final models with 90% of the total number of samples (*i.e.* train and calibration set combined).

Model performance. For estimating the model performance of median vegetation height modeling, the testing set (*i.e.* 10% of samples, completely isolated tiles/group strata) was used to calculate the D^2 regression score (see equation (1)[61]), Root Mean Square Error (RMSE), R^2 and Concordance Correlation Coefficient (CCC) in transformed space (see equation (3)[62]).

$$D^2 = 1 - \frac{D_{\text{model}}}{D_{\text{null}}} \tag{1}$$

$$D(y_i, \hat{y}_i) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i \log\frac{y_i}{\hat{y}_i} - (y_i - \hat{y}_i)\right] \tag{2}$$

$$CCC = \frac{2 \cdot Cov\,(y_{ln1}, \hat{\hat{y}}_{ln1})}{Var(y_{ln1}) + Var(\hat{\hat{y}}_{ln1}) + (\overline{y_{ln1}} - \overline{\hat{y}_{ln1}})^2} \tag{3}$$

where:

D_{model} = Mean Poisson deviance of the fitted model
D_{null} = Mean Poisson deviance of the null model, which predicts the mean $\bar{y}$ of the observed values.
$Cov(y, \hat{y})$ = Covariance between the observed and predicted values
$\text{Var}(y)$ = Variance of the observed values

Dataset	License	Version	GEE Asset ID
ICESat-2 ATL08[23,27]	CC-0	v006	NA
GEDI 98th Percentile Height[9]	CC-0	v1	LARSE/GEDI/GRIDDEDVEG_002/V1/1KM/gediv002_rh-98-a0_vf_20200101_20201231
GEDI 50th Percentile Height[9]	CC-0	v1	LARSE/GEDI/GRIDDEDVEG_002/V1/1KM/gediv002_rh-50-a0_vf_20200101_20201231
ETH Global Sentinel-2 10 m Canopy Height[72]	CC-BY-4.0	v1	users/nlang/ETH_GlobalCanopyHeight_2020_10m_v1
High Resolution Canopy Height Map by WRI and Meta[8]	CC-BY-4.0	v1	projects/meta-forest-monitoring-okw37/assets/CanopyHeight

Table 2. Vegetation and Canopy Height Datasets.

$Var(\hat{y})$ = Variance of the predicted values
y_i = Observed value
$\hat{y}_i$ = Predicted value
$y_{\ln 1}$ = Natural logarithm of 1 + observed value
$\hat{y}_{ln1}$ = Natural logarithm of 1 + predicted value
n = Number of samples

Global predictions and mosaicking. Global predictions were produced per 1 × 1 degree tile and on a yearly basis from 2000 to 2022, resulting in annual values of median vegetation height at 30 m spatial resolution. Although the ML modeling focused on open ecosystems, we extend predictions to all vegetated global land areas, excluding only pixels mapped as stable water according to UMD GLAD GLCLUC[28], an effort that enables a broader analysis of spatiotemporal predictions. The ten GBT models were compiled to a native C binary using lleaves[63], reducing the prediction time by a factor of 5. Per pixel prediction intervals (5th and 95th percentiles) were estimated through the individual predicted values retrieved from the final ensemble GBT models.

The entire processing workflow was executed on a High-Performance Computing (HPC) system, with tasks distributed across processing nodes using SLURM[64] and Docker containers[65]. Generating the final predictions required approximately 90,300 CPU hours and 2.9 terabytes of RAM. The predicted tiles were then mosaicked into Cloud-Optimized GeoTIFF (COG) files and made publicly accessible via Google Earth Engine and SpatioTemporal Asset Catalog (STAC).

Comparison with existing products. The comparison with existing vegetation height products considered a sampling design previously used to acquire land cover reference samples for mapping grasslands worldwide, based on feature space coverage sampling and composed of 7,005 regular tiles (*i.e.* 1 × 1 km Global Pasture Watch (GPW) tiles)[4,66]. We retrieved all ICESat-2 ATL08 data for these tiles, resulting in 3,576 globally distributed unique locations with reference median vegetation height values (including 103,549 ICESat-2 20 m segments), which were excluded from our ML modeling. These tiles were then aggregated by grasslands (60,078 segments in 1675 tiles), forested ecosystems (18,004 segments in 1304 tiles), and all land cover types to further explore the representation of our model at global and continental scales. Additionally, the independent satellite Lidar data collected by the GEDI instrument and aggregated to 1 km resolution[9] was compared with our predicted height values. From all GEDI height metrics (*i.e.* based on the full energy waveform captured by the sensor), we selected the median and top of canopy (98th percentile height) for direct comparison with our modeling results. Finally, we compared the predicted height values with two products modeling top of tree canopy height, the ETH Global Sentinel-2 10 m Canopy Height (Sentinel Canopy Height)[67] and High Resolution Canopy Height Map by WRI and Meta (WRI/Meta Height Model)[8], which have 10 m and 1 m spatial resolution, respectively. Additional details of the product version and Google Earth Engine (GEE) asset IDs used in this analysis are available in Table 2.

Trend analysis. Aiming to identify hotspots of vegetation dynamics, related for example to deforestation, restoration and shrub encroachment, we ran a per-pixel trend analysis on all 23 years (2000–2022) of median vegetation height predictions. For each pixel, a linear regression was trained (between time and vegetation height) and a Mann-Kendall test was performed to detect monotonic trends on median vegetation height[68]. The trend represents the average median height change in meters per year. All significant trend values (*i.e.* $P < 0.05$) were retained and mosaicked into Cloud-Optimized GeoTIFF (COG) files (Fig. 4). The final trend map is publicly accessible via Google Earth Engine and SpatioTemporal Asset Catalog (STAC).

Data Records

The global maps of median vegetation height (see Fig. 4) described in this paper are available from 2000–2022 in COG (Cloud Optimized GeoTIFF) format under the Creative Commons license CC-BY, archived in Zenodo (https://doi.org/10.5281/zenodo.15198654[27]), and publicly accessible in OpenLandMap SpatioTemporal Asset Catalog (STAC - https://stac.openlandmap.org/gpw_gsvh-30m/collection.json) and Google Earth Engine (GEE – https://developers.google.com/earth-engine/datasets/publisher/global-pasture-watch). Using the COG format enables users to seamlessly load the maps in various GIS solutions (*e.g.* Quantum GIS, MapServer, GeoServer, etc.) and programming environments (*e.g.* JupyterLab, RStudio, Google Colab, etc.), without the need to download the entire raster file locally.

A total of 70 global mosaics (*i.e.* one trend map and 23 years of predicted median vegetation height, including upper and lower boundary prediction intervals – 5th and 95th percentiles) are available in the WGS84

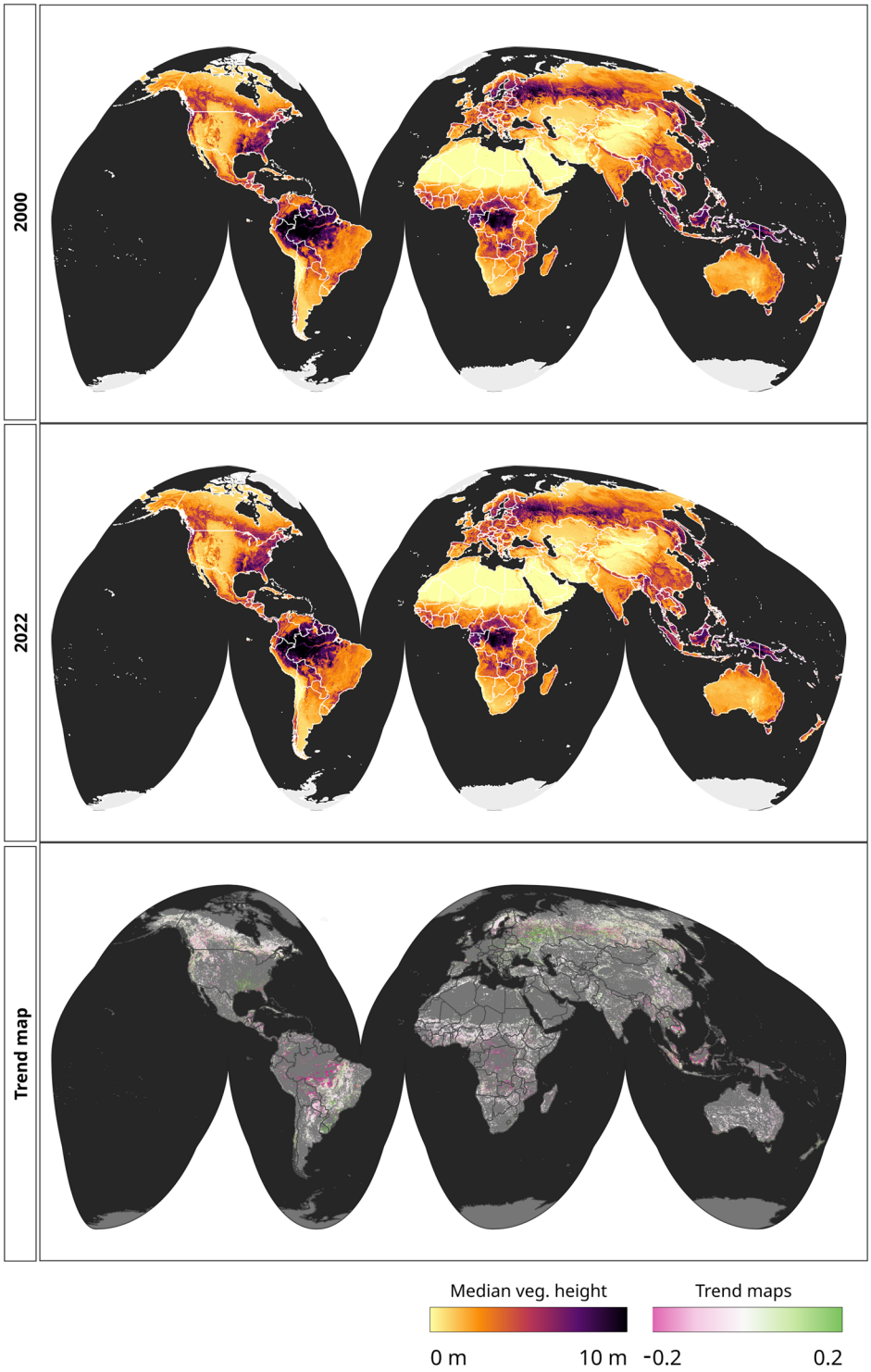

Fig. 4 Global short vegetation height maps from 2000–2022, including per-pixel trend maps, with median canopy height information.

Coordinate Systems (*i.e.* EPSG:4326), pixel size equal to 0.00025 (STAC and GEE) and 0.00075 degrees (Zenodo). The total size of the dataset is about 10 terabytes. All raster files are in `16-bit integer` format with pixel values ranging from 0–20 meters and value `-3200` representing no-data (*i.e.* ocean waters, lakes and water bodies), following a naming convention that organizes the most important data properties in ten fields:

1. Project name : Global Pasture Watch (`gpw`)
2. Product name : *Short vegetation height* (`short.veg.height`)
3. Procedure combination : Ensemble Gradient Boosting Trees (`egbt`), trend/intercept derived per-pixel via

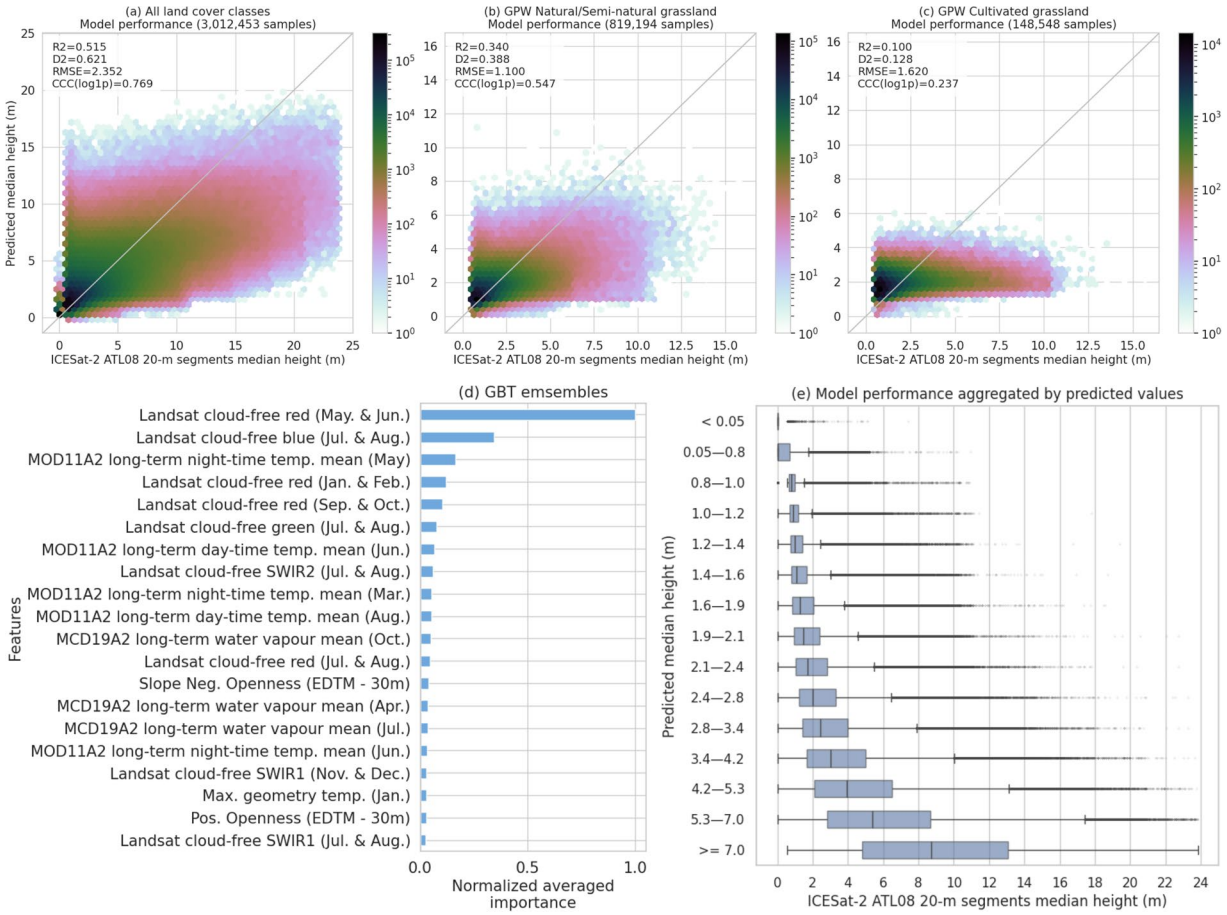

Fig. 5 Model performance and feature importance of median height modeling. Model performance on ICESat testing set samples is presented for all land cover types (**a**), natural/semi-natural grasslands (**b**), and cultivated grasslands (**c**). The normalized averaged feature importance, derived from 10 Gradient Boosting Tree (GBT) models trained to predict median vegetation height, is displayed in plot (**d**). Plot (**e**) shows aggregated predicted values with an approximately equal number of samples per height interval (*i.e.* 200,000 samples).

linear regression (`trend`), Mann-Kendall test (`mk`).

4. Variable type : mean predicted value (`m`), 5th (`p.05`) and 95th percentiles (`p.95`)
5. Spatial resolution : `30m`
6. Begin of time reference : date of first Landsat composite used by the modeling (`20220101`)
7. End of time reference : date of last Landsat composite used by the modeling (`20221231`)
8. Spatial extent : global (`go`)
9. Coordinate system : World Geodetic System 1984, used in GPS (`epsg.4326`)
10. Version : `v1`

Technical Validation

Model performance and feature importance. Model performance was evaluated using the testing set (Fig. 3), resulting in an RMSE of 2.352 m, and R^2, D^2 and CCC in transformed space of 0.515, 0.621 and 0.769, respectively (Fig. 5a). Ten Gradient Boosted Tree models were trained independently and merged to form the final ensemble model. The Landsat cloud-free red band for the months of May and June was the most important feature across all individual models. All other variables show changes in the order of their importance between individual models, however, Landsat blue, red, green and SWIR-2 bands together with MOD11A2 long-term day and night time land surface temperature are the most important features for modeling median vegetation height. The overall feature importance is reported in Fig. 5d. The model presented better performance on natural/semi-natural grassland compared to ICESat-2 samples acquired over cultivated grassland (Fig. 5b,c), which might be explained by a stronger correlation of our input features with less intensively managed vegetation. Comparing predicted height to measured height in 15 approximately equal sized sample bins, mean values are within the expected range at predicted heights below 2 m and below the expected range at heights from approximately 2–5 m but within the inter–quartile range (Fig. 5e). Predicted height above 7.0 m represents less than 7% of model results.

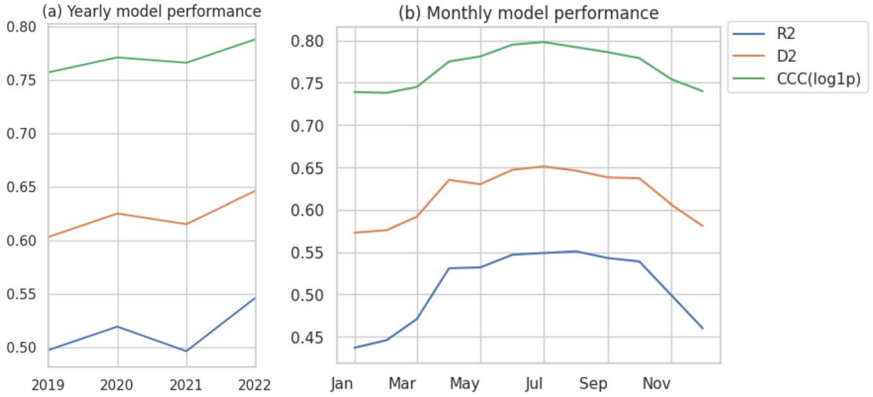

Fig. 6 Yearly (**a**) and monthly (**b**) model performance conducted using 3,012,453 ICESat samples of testing set.

Given the distribution of ICESat-2 samples over time, we also grouped the testing set by year and by month to temporally assess the model performance (Fig. 6). In this regard, the model showed slightly poorer performance in 2019 and better in 2022, which might be related to differences in data availability and quality at the beginning of the ICESat-2 mission, however all accuracy metrics (R^2, D^2, *CCC*) were within 0.05 (Fig. 6a). Greater variability in model performance was observed in the monthly model performance. The largest variability is in R^2 values, with a minimum reported R^2 of 0.43 in January and a maximum of 0.55 in June through August. The monthly variation in CCC was lower, but showed a similar pattern, ranging from 0.74 in December-February to 0.80 in July (Fig. 6b). This may be due to seasonal variations in phenology or variations in data availability due to seasonal snow or cloud cover reducing the number of available training samples in the northern hemisphere winter.

Model comparison with existing products. Regarding the analysis based on the regular GPW tiles (*i.e.* 1 × 1 km), at the global scale, we find RMSE of 2.55 m, and CCC of 0.61 when all land covers were considered and RMSE of 2.17 m and CCC of 0.20 considering areas identified as grasslands. The distribution of tiles across continents is unequal with the largest number of tiles within the African continent (1369) and the smallest within Oceania (106). However, these two continents have the best model performance: In Oceania the RMSE is 1.95 m and the CCC is 0.76 and in Africa the RMSE is 2.08 m and the CCC is 0.69. We observe a higher RMSE in Asia (3.78 m) compared to all other continents (less than 2.46 m). CCC are also lower in Asia, 0.49, as opposed to 0.57 –0.76 in other regions.

The distribution of height within these tiles follows a long-tailed negative exponential distribution. Half of the measurements have a median height of 1.80 m or less with 90% of measurements between 0.69 and 9.91 m. The distribution of modeled values for the same locations range from 0.76 to 7.19 m with a median value of 2.12 m. It is important to note that the reference vegetation height refers to ICESat-2 data, which has a minimum vegetation height of 50 cm, whereas the model predicts areas with zero height. This is reflected in the minimum values predicted for individual tiles: 0.51 m for ICESat-2 and 0.004 m for our model. Heights decrease when considering only grassland areas, with 95% of values below 4.5 m for ICESat-2 and below 3.25 m for the model. Median values are 1.8 m and 2.12 m for ICESat-2 and our model, respectively. Considering forest areas with canopy height greater than 10 m following Potapov *et al.* (2022), the inter-quartile range of tile median height measured by ICESat-2 ranges from 2.3–8.3 m with a mean of 4.95 m. Model values show an inter-quartile range of median height per tile of 3.66–6.88 m, and a mean of 5.3 m (Fig. 7).

Considering the gridded product of spaceborne GEDI Lidar, it is important to note that GEDI metrics are based on the full waveform of returned energy, and not only vegetation. As such, in areas with short vegetation, the median height of GEDI is strongly concentrated at 0 m and below. Approximately 65% of tiles have an estimated median height less than 0 m and 95% of tiles have median height less than 3.5 m. In forest areas, approximately 16% of tiles maintain an estimated median height of less than 0 m. The 98th percentile of waveform height is used to estimate top of canopy height and varies between 2.4 and 40.3 m within sampled grassland areas and between 2.5 and 57.9 m in forest areas (Fig. 7). Comparing the results of the MVH model[27] to existing global canopy height models within GPW tiles, we find that both the WRI/Meta Canopy Height model and the Sentinel Canopy Height model estimate many pasture regions as having zero height. The Sentinel model estimates 30% of tiles with zero height and half with height of 5 m or greater. The WRI/Meta Canopy Height model estimates 93% of tiles to have median height of 0 m. The dominance of zero height in these models is expected as they are focused on tree canopy height as opposed to vegetation height. Within forest areas, the Sentinel model estimates 85% of tiles with median height at or above 10 m. The WRI/Meta Canopy Height model estimates 30% of tiles with median height over 10 m and 16% of tiles with median height of 0 m.

Lang *et al.* (2023) reports an overestimation of canopy height for vegetation with less than 20 m height, but minimal bias in temperate and tropical grasslands when compared to GEDI validation data. This is consistent with our analysis: within grassland areas we find a mean height underestimation of 0.2 m comparing the Sentinel map with aggregate GEDI data. When we consider all land covers, the Sentinel model overestimates GEDI 98th percentile height by 1.5 m. However, it is important to note that cross-referencing high resolution imagery and

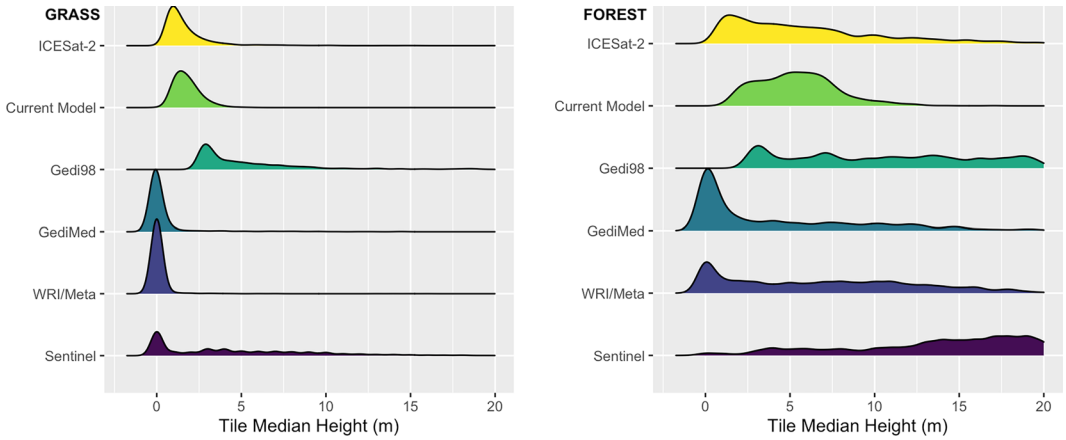

Fig. 7 Distribution of median height per tile in (GRASS) grassland areas as defined by Parente *et al.* (2024) and (FOREST) forest areas with canopy height over 10 m (following GLAD GLCLUC[28]) for ICESat-2 segments and model heights for areas coincident with segments of the Median Vegetation Height Model and other models and data sets.

height information shows that the effective spatial resolution of the Sentinel product is less than 10 m. This is attributed to the footprint size of GEDI (25 m) and the geolocation uncertainty (15-20 m), and given these limitations, the effective resolution of the Sentinel map is similar to, or lower than, the map produced here[67]. Considering the high resolution canopy height map of WRI/Meta, the canopy delineation approach is not well suited to short vegetation ecosystems as it was not trained to delineate non-tree canopies such as those of more densely spaced herbs and shrubs. To better demonstrate the differences between existing products and our predictions, we developed a map explorer available at https://gpw-lapig.projects.earthengine.app/view/hct.

Usage Notes

Vegetation types and prediction intervals. Our modeling framework was based on Ensemble Gradient Boosting Trees (EGBT) and enabled us to run model optimization on a Poisson loss function, a suitable method for skewed/asymmetric distributions, such as global median vegetation height (Fig. 7). Although computationally intensive and susceptible to overfitting depending on the number of iterations, we mitigated model overfitting by employing early stopping to limit the number of iterations on the training step[69], and estimated prediction intervals using a distribution of predicted values. This approach – despite improving overall model performance – exhibited a tendency toward the mean, resulting in overly narrow and optimistic prediction intervals (Figs. 8 and 9). In the spatial domain, our height predictions vary with latitude, reaching a maximum of 7.5 m on average at the equator. Smaller scale variability in median vegetation height within planted and natural ecosystems is shown for regions in Spain and Mozambique (Fig. 8a,b). We also highlight height changes in transitions from open grasslands to woodland areas for the Brazilian Cerrado and from bare soil to shrubland in southern Australia in Fig. 8c,d.

Considering the vegetation types separated by UMD GLAD GLCLUC[28], median heights estimated by ICESat-2 range from 1.1 m in wetland areas with sparse vegetation, to 9.5 m in wetlands with stable tree cover estimated between 20–25 m height (Table 3). Heights estimated by the MVH model for these ecosystems are 1.0 m and 8.6 m, respectively. Ranked heights presented consistency across the analyses, with the exception of the shortest classes. Specifically, wetland saltpan areas are estimated by ICESat-2 to have a median height of 1.37 m, taller than wetland areas with sparse vegetation cover. In contrast, the MVH model estimates a shorter height of 0.86 m in saltpan areas. In this case, ICESat-2 may overestimate height given the limitations of the instrument in measuring the shortest and sparsest vegetation. It is expected that both ICESat-2 median height and MVH model height are lower than the top of canopy heights referenced in the GLAD vegetation type as these refer to the top of canopy height.

The variability in median height as estimated by the MVH model is lower than within the ICESat-2 training data. The 2nd percentile of median heights varies less in the training data than in the MVH model, ranging from 0.56 to 0.83 m for ICESat-2 and from 0.04 to 3.09 m for the MVH model. The reduced variability is due to the upper estimates of median height; Comparing the 98th percentile of the distribution of median height, the MVH model is lower than the ICESat-2 height by 3.7 to 8 m. This effect is most significant in tall forests where the 98th percentile of median height is underestimated by 7 – 7.9 m. Despite this underestimation, it is important to note that relative patterns are generally preserved, so users can still detect where vegetation is taller or more structurally complex compared to surrounding areas even in cases where the absolute values are not matching the landscape on the ground. Therefore, due to known underestimation of canopy height in the upper extremes, care should be taken when interpreting absolute values in regions with very tall forests. Relative patterns and trends are still informative.

The annual time series maps presented in this study allow for tracking changes in height and reveal patterns of structural change over time. The median height of vegetation is affected by the presence and density of all vegetation layers, including herbaceous grasses, woody shrubs, and trees. Areas with positive trends in median height may reflect woody encroachment, pasture abandonment, secondary regrowth or a variety of other land

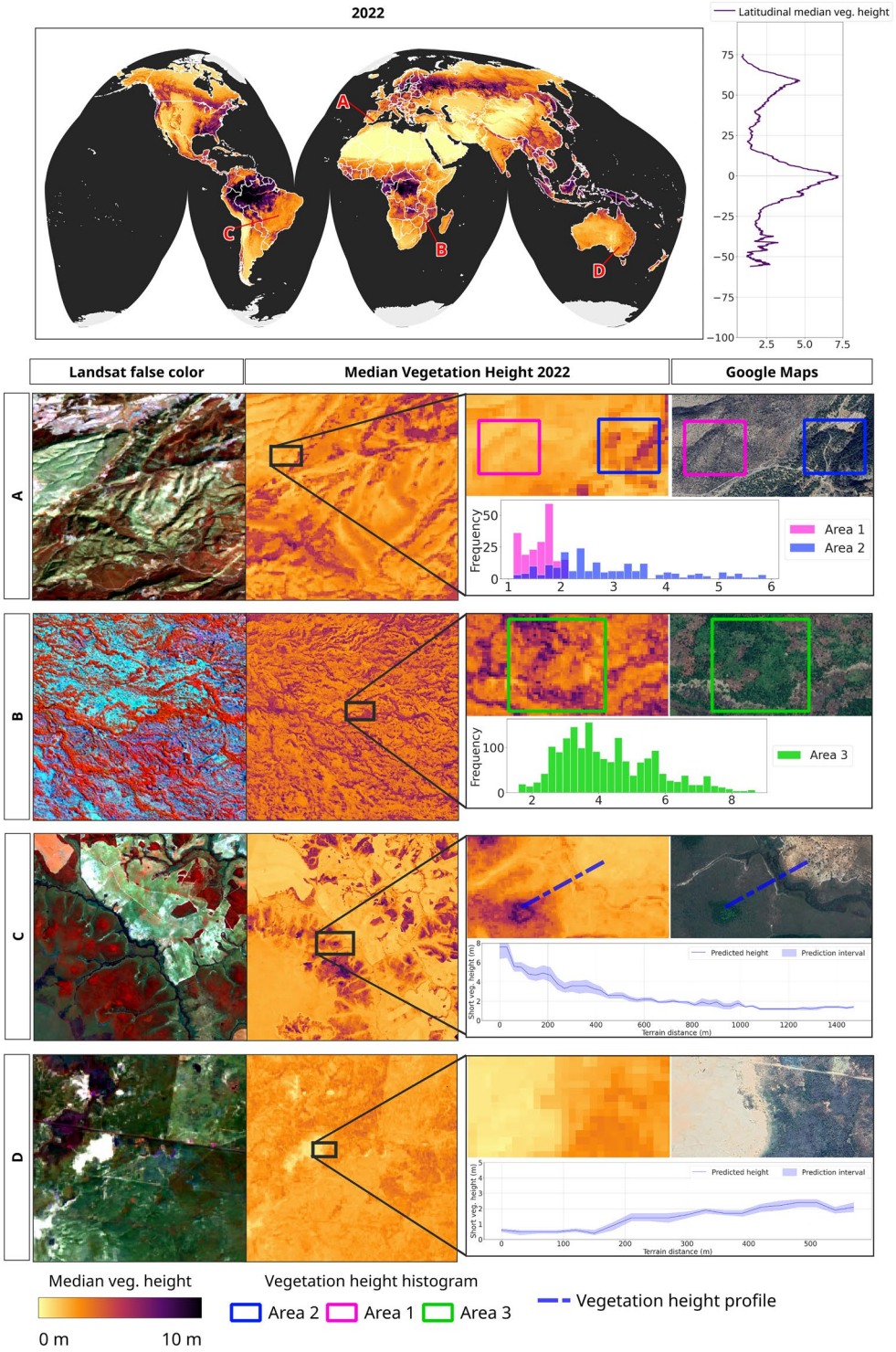

Fig. 8 Global short vegetation height map for 2022 highlighting areas in (**A**) Moratalla, Spain; (**B**)Inhaminga, Mozambique; (**C**) Emas National Park (Cerrado), Brazil; (**D**) Ivanhoe, Australia.

cover or land use changes. Conversely, negative trends may indicate pasture management, degradation due to increased bare soil, and thinning or disturbance within wooded areas. At the global scale, large regions with significant trends in height are apparent, and positive trends are observed in Russia and Uruguay whereas negative trends can be seen in Northern Brazil and Bolivia, scattered regions of Africa and throughout Indonesia. At local scales, height changes may correlate with changes in land cover and land use as shown in Fig. 9 in Kazahkstan, Paraguay, Brazil, and Angola, where both positive and negative trends are highlighted with prediction intervals around modeled height. In general, prediction intervals widen with increasing height, ranging less than 0.2 m around a short vegetation height of 1 m and from 6–8.5 m around a short vegetation height of 7 m, as can be seen

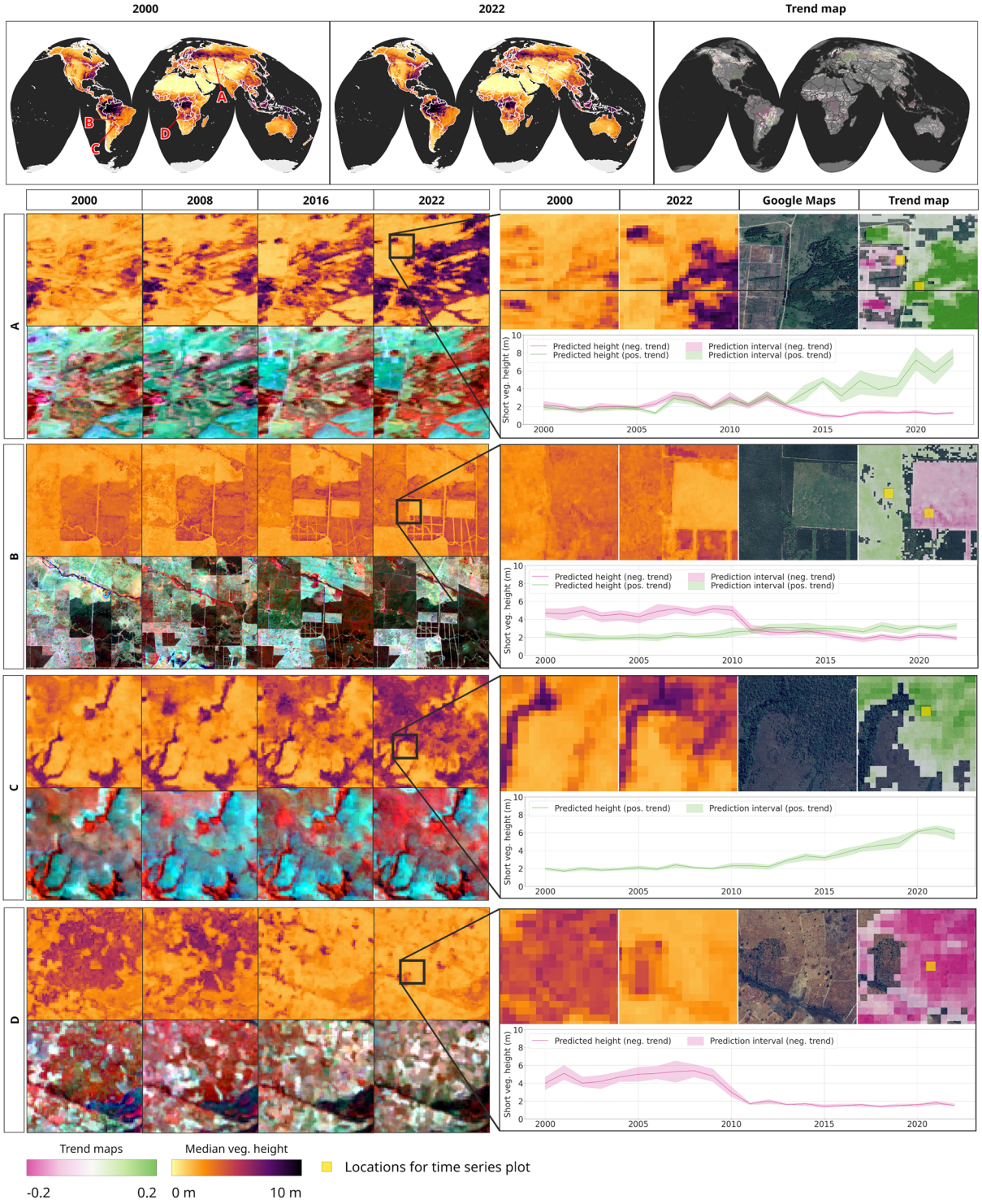

Fig. 9 Global short vegetation height maps from 2000–2022, including per-pixel trend maps, highlighting areas in (**A**) Alekseev, Kazakhstan; (**B**) Estancia Quebrachos, Paraguay; (**C**) Serranópolis, Brazil; (**D**) Kaiuera, Angola.

in these regions. However, it is important to note that the trends presented here are monotonic increases over the full time period. In this sense, the lack of a significant trend does not indicate that changes in height have not occurred over the full time series, but rather that they have not produced a linear trend. Many applications require consideration of a shorter time period or a different start year. Availability of the full time series of modeled MVH[27] allows users to calculate trends tailored to specific use-cases and time scales.

Current limitations. ICESat-2 produces a global sample, covering all latitudes, and has benefits for measuring short vegetation including a small footprint size and frequent along-track measurements that yield

Land cover class	Testing set proportion (%)	ICESat ATL08 median height (m)	Predicted median height (m)
Cropland, Stable	6.79	1.61 (0.56—7.73)	1.76 (0.92—3.58)
Terra Firma, dense short veg. (79-100% cov.)	27.51	1.98 (0.59—7.31)	2.09 (0.89—4.51)
Terra Firma, semi-arid (11-47% cov.)	5.84	1.23 (0.57—3.76)	1.00 (0.20—2.29)
Terra Firma, semi-arid (47-75% cov.)	11.22	1.50 (0.58—4.76)	1.52 (0.83—2.99)
Terra Firma, stable tree cover 3-5m	3.34	2.95 (0.69—9.71)	3.17 (1.56—5.99)
Terra Firma, stable tree cover 6-10m	8.33	4.00 (0.74—11.68)	4.01 (1.66—7.47)
Terra Firma, stable tree cover 11-15m	6.21	5.16 (0.71—14.99)	4.90 (1.67—9.55)
Terra Firma, stable tree cover 16-20m	6.06	6.05 (0.71—18.16)	5.76 (1.95—11.12)
Terra Firma, stable tree cover 21-25m	4.75	8.82 (0.80—21.79)	8.18 (3.09—14.79)
Wetland, dense short veg. (79-100% cov.)	4.70	1.80 (0.57—7.61)	2.02 (0.85—4.95)
Wetland, salt pan (3-10% cov.)	0.06	1.37 (0.56—6.25)	0.86 (0.04—3.56)
Wetland, sparse vegetation (11-47% cov.)	0.11	1.07 (0.56—3.76)	1.02 (0.27—1.90)
Wetland, sparse vegetation (47-75% cov.)	0.18	1.53 (0.57—5.25)	1.60 (0.69—2.96)
Wetland, stable tree cover 3-5m	0.40	3.13 (0.66—10.38)	3.43 (1.55—6.68)
Wetland, stable tree cover 6-10m	1.05	3.80 (0.68—11.50)	3.94 (1.68—7.72)
Wetland, stable tree cover 11-15m	0.86	5.22 (0.67—15.18)	4.89 (1.80—9.60)
Wetland, stable tree cover 16-20m	0.88	6.21 (0.68—17.78)	5.98 (1.99—11.93)
Wetland, stable tree cover 21-25m	0.38	9.53 (0.83—21.44)	8.61 (3.00—13.46)

Table 3. Averaged median vegetation height, expected (ICESat-2 ATL08) and predicted (EGBT), for multiple land classes according to UMD GLAD product. Height values in parentheses are the 2nd and 98th percentiles, respectively.

multiple return photon measurements over individual land segments. However, cloud cover and smoke prohibit signal penetration, and the 532 nm green laser has lower signal penetration than the 1064 nm NIR laser used in the GEDI sensor. Also, we restricted training data to strong-beam, night-time measurements to minimize signal noise, and included additional filters to maximize the quality of our training data. Despite these efforts, there are many sources of data uncertainty that may not have been fully accounted for. An additional limitation of the ICESat-2 data set is that vegetation is not identified at heights less than 50 cm from the ground surface due to the method of separating terrain from vegetation photons. This results in the overestimation of height in the shortest or most sparse landcovers (as is seen in wetland saltpans). Below we discuss some of the limitations of our model:

- **Difficulty detecting very short vegetation**: All Lidar sensors struggle with very low canopies, especially under partial ground occlusion or when the vegetation cover fraction is sparse. In those scenarios, even height retrievals filtered for the highest quality may underestimate height or classify vegetation returns as ground signals. Our product is no exception, and caution should be used in areas with a dominant vegetation height under half the RMSE or 1.175 m. Future algorithmic refinements, leveraging photon-level data (ATL03 product) or improved noise filtering, may mitigate underestimation of low-stature vegetation.
- **Vegetation heterogeneity**: In open landscapes with scattered shrubs, a single 30-m pixel can contain a mixture of herbaceous and shrub layers. The median height metric may be pulled upward by a small fraction of woody cover. Users aiming to isolate purely herbaceous biomass estimates could thus combine our datasets with fractional cover maps to account for sub-pixel heterogeneity[70]. In future versions, more information on structural complexity leveraging ICESat-2's ability to capture vertical structure may better disentangle these overlapping layers.
- **Sub-annual and seasonal patterns**: Many grassy ecosystems experience substantial seasonal changes in height and structure, but current ICESat-2 data density is insufficient to map sub-annual or seasonal patterns of vegetation height. As Lidar acquisitions grow or emerging satellite missions offer denser sampling, more frequent snapshots may enable models to differentiate growing versus senescent phases. Such finer temporal resolution would improve our ability to monitor intra-annual changes in height and structure and detect disturbances such as grazing pressure, fire regimes, or harvesting events.

Finally, although validated against both field measurements and ICESat-2 itself, broader validation using airborne or drone Lidar would further strengthen model confidence. For example, initiatives like the USGS 3D Elevation Program (3DEP) or certain commercial airborne Lidar mapping companies (e.g. NV5 Geospatial, Woolpert) conduct large-scale, systematic flights over the United States. As these datasets become more comprehensive and openly available, they could serve as an additional benchmark to refine model performance across a broad range of open ecosystems. These annual maps and and trends from 2000–2022 offer a powerful lens for understanding vegetation height and its dynamics, supporting applications in habitat monitoring, carbon accounting, and restoration monitoring. While data density, photon-level retrievals, and airborne

Lidar campaigns continue to expand, these products will only become more robust, helping researchers and decision-makers better account for the carbon and biodiversity values of non-forest ecosystems.

Code availability

All modeling pipelines presented in this paper were implemented in Python, and the source code is publicly available (MIT License) at: https://github.com/wri/global-pasture-watch. For reproducibility purposes, we have archived a snapshot of all reference samples and trained models (https://doi.org/10.5281/zenodo.15194973[71]) in Zenodo.

Received: 17 April 2025; Accepted: 31 July 2025;
Published online: 23 August 2025

References

1. Bardgett, R. D. *et al*. Combatting global grassland degradation. *Nature Reviews Earth & Environment* **2**, 720–735 (2021).
2. O'Mara, F. P. The role of grasslands in food security and climate change. *Annals of botany* **110**, 1263–1270 (2012).
3. Gibson, D. J. *Grasses and grassland ecology* (Oxford University Press, 2009).
4. Parente, L. *et al*. Annual 30-m maps of global grassland class and extent (2000–2022) based on spatiotemporal machine learning. *Scientific data* **11**, 1–22, https://doi.org/10.1038/s41597-024-04139-6 (2024).
5. Sala, O. E. & Paruelo, J. M. Ecosystem services in grasslands. *Nature's services: Societal dependence on natural ecosystems* 237–251 (1997).
6. Grace, J., José, J. S., Meir, P., Miranda, H. S. & Montes, R. A. Productivity and carbon fluxes of tropical savannas. *Journal of Biogeography* **33**, 387–400 (2006).
7. Migliavacca, M. *et al*. The three major axes of terrestrial ecosystem function. *Nature* **598**, 468–472 (2021).
8. Tolan, J. *et al*. Very high resolution canopy height maps from rgb imagery using self-supervised vision transformer and convolutional decoder trained on aerial lidar. *Remote Sensing of Environment* **300**, 113888 (2024).
9. Burns, P., Hakkenberg, C. & Goetz, S. Gridded gedi vegetation structure metrics and biomass density at multiple resolutions. ornl daac, oak ridge, tennessee, usa (2024).
10. Li, W., Niu, Z., Shang, R., Qin, Y. & Chen, H. High-resolution mapping of forest canopy height using machine learning by coupling icesat-2 lidar with sentinel-1, sentinel-2 and landsat-8 data. *International Journal of Applied Earth Observation and Geoinformation* **92**, 102163 (2020).
11. Potapov, P. *et al*. Mapping global forest canopy height through integration of gedi and landsat data. *Remote Sensing of Environment* **253**, 112165 (2021).
12. Lefsky, M. A. *et al*. Estimates of forest canopy height and aboveground biomass using icesat. *Geophysical research letters* **32** (2005).
13. Steutker, D. & Glenn, N. Lidar measurement of sagebrush steppe vegetation heights. *Remote Sensing of Environment* **102**, 135–145 (2006).
14. Magruder, L. A., Brunt, K. M. & Alonzo, M. Early icesat-2 on-orbit geolocation validation using ground-based corner cube retro-reflectors. *Remote Sensing* **12**, 3653 (2020).
15. Duncanson, L. *et al*. Aboveground biomass density models for NASA's Global Ecosystem Dynamics Investigation (GEDI) lidar mission. *Remote Sensing of Environment* **270**, 112845, https://doi.org/10.1016/j.rse.2021.112845 (2022).
16. Neuenschwander, A. *et al*. Ice, cloud, and land elevation satellite (icesat-2) project algorithm theoretical basis document (atbd) for land - vegetation along-track products (atl08), version 6. icesat-2 project. https://doi.org/10.5067/8ANPSL1NN7YS (2022).
17. Matasci, G. *et al*. Three decades of forest structural dynamics over canada's forested ecosystems using landsat time-series and lidar plots. *Remote Sensing of Environment* **216**, 697–714 (2018).
18. Nilsson, M. *et al*. A nationwide forest attribute map of sweden predicted using airborne laser scanning data and field data from the national forest inventory. *Remote Sensing of Environment* **194**, 447–454 (2017).
19. Maltamo, M., Næsset, E. & Vauhkonen, J. Forestry applications of airborne laser scanning. *Concepts and case studies. Manag For Ecosys* **27**, 460 (2014).
20. Scarth, P., Armston, J., Lucas, R. & Bunting, P. A structural classification of australian vegetation using icesat/glas, alos palsar, and landsat sensor data. *Remote Sensing* **147**, 11 (2019).
21. Bazzo, C. O. G., Kamali, B., Hütt, C., Bareth, G. & Gaiser, T. A review of estimation methods for aboveground biomass in grasslands using uav. *Remote Sensing* **15**, 639 (2023).
22. Coverdale, T. C. & Davies, A. B. Unravelling the relationship between plant diversity and vegetation structural complexity: A review and theoretical framework. *Journal of Ecology* **111**, 1378–1395 (2023).
23. Neuenschwander, A. *et al*. Atlas/icesat-2 l3a land and vegetation height, version 6 http://nsidc.org/data/ATL08/versions/6 (2023).
24. Chatterjee, S. A new coefficient of correlation. *Journal of the American Statistical Association* **116**(536), 2009–2022 (2021).
25. Hunter, M., Teles, N. & Silva Costa, J. V. ICESat-2 Validation Survey for Pastures of the Rio Vermelho Watershed https://doi.org/10.5281/zenodo.14860218 (2025).
26. Milenkovic, M. *et al*. Assessing amazon rainforest regrowth with gedi and icesat-2 data. *Science of Remote Sensing* **5**, 1000051 (2022).
27. Parente, L. *et al*. Global pasture watch - annual short vegetation height maps at 30-m spatial resolution (2000-2022) https://doi.org/10.5281/zenodo.15198654 (2025).
28. Potapov, P. *et al*. The global 2000-2020 land cover and land use change dataset derived from the landsat archive: first results. *Frontiers in Remote Sensing* **3**, 856903 (2022).
29. Consoli, D. *et al*. A computational framework for processing time-series of earth observation data based on discrete convolution: global-scale historical landsat cloud-free aggregates at 30 m spatial resolution", keywords = "analysis ready data (ard), discrete convolution, earth observation data, gap-filling, geographic information system (gis), high performance computing (hpc), imputation, landsat, time-series processing, time-series reconstruction. *PeerJ* **12** (2024).
30. Potapov, P. *et al*. Landsat analysis ready data for global land cover and land cover change mapping. *Remote Sensing* **12**, 426 (2020).
31. Roy, P., Sharma, K. & Jain, A. Stratification of density in dry deciduous forest using satellite remote sensing digital data-an approach based on spectral indices. *Journal of biosciences* **21**, 723–734 (1996).
32. Huete, A. *et al*. Overview of the radiometric and biophysical performance of the modis vegetation indices. *Remote Sensing of Environment* **83**, 195–213 (2002).
33. Van Deventer, A., Ward, A., Gowda, P. & Lyon, J. Using thematic mapper data to identify contrasting soil plains and tillage practices. *Photogrammetric engineering and remote sensing* **63**, 87–93 (1997).
34. Tucker, C. J. Red and photographic infrared linear combinations for monitoring vegetation. *Remote sensing of Environment* **8**, 127–150 (1979).
35. Gao, B.-C. NDWI-A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment* **58**, 257–266 (1996).
36. Badgley, G., Field, C. B. & Berry, J. A. Canopy near-infrared reflectance and terrestrial photosynthesis. *Science advances* **3**, e1602244 (2017).

37. Robinson, N. P. *et al*. Terrestrial primary production for the conterminous United States derived from Landsat 30 m and MODIS 250 m. *Remote Sensing in Ecology and Conservation* **4**, 264–280 (2018).
38. Tian, X. *et al*. Time series of landsat-based bimonthly and annual spectral indices for continental europe for 2000–2022. *Earth System Science Data* **17**, 741–772 (2025).
39. Lefsky, M. A. A global forest canopy height map from the moderate resolution imaging spectroradiometer and the geoscience laser altimeter system. *Geophysical Research Letters* **37** (2010).
40. Wan, Z., Hook, S. & Hulley, G. MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 1km SIN Grid V061 (2021).
41. Lyapustin, A. & Wang, Y. MODIS/Terra+Aqua Land Aerosol Optical Depth Daily L2G Global 1km SIN Grid V006 (2018).
42. Hengl, T. & Parente, L. Long-term modis lst day-time and night-time temperatures, sd and differences at 1 km based on the 2000–2020 time series https://doi.org/10.5281/zenodo.6458406 (2022).
43. Parente, L., Simoes, R. & Hengl, T. Monthly aggregated water vapor modis mcd19a2 (1 km): Long-term data (2000-2022) (2023).
44. Purinton, B. & Bookhagen, B. Beyond vertical point accuracy: Assessing inter-pixel consistency in 30 m global dems for the arid central andes. *Frontiers in Earth Science* **9**, 758606 (2021).
45. Bonannella, C., Hengl, T., Parente, L. & de Bruin, S. Biomes of the world under climate change scenarios: increasing aridity and higher temperatures lead to significant shifts in natural vegetation. *PeerJ* **11**, e15593 (2023).
46. Ho, Y.-f. *et al*. Global ensemble digital terrain model and land relief parameterization at 30 m resolution: a community-based open data service to support regional and global modeling In preparation (2025).
47. Tadono, T. *et al*. Generation of the 30 m-mesh global digital surface model by alos prism. *The international archives of the photogrammetry, remote sensing and spatial information sciences* **41**, 157–162 (2016).
48. Strobl, P. The new copernicus digital elevation model. *GSICS Quarterly* **14**, 17–18 (2020).
49. Yamazaki, D. *et al*. Merit dem: A new high-accuracy global digital elevation model and its merit to global hydrodynamic modeling. In *AGU fall meeting abstracts*, vol. 2017 (2017).
50. BM, B. *et al*. Tuw-geo/equi7grid: v0.2.4 https://doi.org/10.5281/zenodo.8252376 (2023).
51. Neteler, M., Bowman, M. H., Landa, M. & Metz, M. GRASS GIS: A multi-purpose open source GIS. *Environmental Modelling & Software* **31**, 124–130 (2012).
52. Wang, Q., Wang, C.-J. & Wan, J.-Z. Relationships between topographic variation and plant functional trait distribution across different biomes. *Flora* **293**, 152116 (2022).
53. Kilibarda, M. *et al*. Spatio-temporal interpolation of daily temperatures for global land areas at 1 km resolution. *Journal of Geophysical Research: Atmospheres* **119**, 2294–2313 (2014).
54. So, B. Enhanced gradient boosting for zero-inflated insurance claims and comparative analysis of catboost, xgboost, and lightgbm. *Scandinavian Actuarial Journal* 1–23 (2024).
55. James, G., Witten, D., Hastie, T., Tibshirani, R. & Taylor, J. Tree-based methods. In *An Introduction to Statistical Learning: with Applications in Python*, 331–366 (Springer, 2023).
56. Demarchi, L. *et al*. Recursive feature elimination and random forest classification of natura 2000 grasslands in lowland river valleys of poland based on airborne hyperspectral and lidar data fusion. *Remote Sensing* **12**, 1842 (2020).
57. Jamieson, K. & Talwalkar, A. Non-stochastic best arm identification and hyperparameter optimization. In *Artificial intelligence and statistics*, 240–248 (PMLR, 2016).
58. Shaharum, N. *et al*. Image classification for mapping oil palm distribution via support vector machine using scikit-learn module. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **42**, 133–137 (2018).
59. Mathlouthi, W., Fredette, M. & Larocque, D. Regression trees and forests for non-homogeneous poisson processes. *Statistics & Probability Letters* **96**, 204–211 (2015).
60. Raskutti, G., Wainwright, M. J. & Yu, B. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *The Journal of Machine Learning Research* **15**, 335–366 (2014).
61. O'Brien, C. M.*Statistical learning with sparsity: the lasso and generalizations* (Wiley Periodicals, Inc., 2016).
62. Steichen, T. J. & Cox, N. J. A note on the concordance correlation coefficient. *The Stata Journal* **2**, 183–189 (2002).
63. Boehm, S. lleaves https://github.com/siboehm/lleaves (2024).
64. Yoo, A. B., Jette, M. A. & Grondona, M. Slurm: Simple linux utility for resource management. In *Workshop on job scheduling strategies for parallel processing*, 44–60 (Springer, 2003).
65. Boettiger, C. An introduction to docker for reproducible research. *ACM SIGOPS Operating Systems Review* **49**, 71–79 (2015).
66. Parente, L. *et al*. Global Pasture Watch - Grassland reference samples based on visual interpretation of VHR imagery (2000-2022) https://doi.org/10.5281/zenodo.14035457 (2024).
67. Lang, N., Jetz, W., Schindler, K. & Wegner, J. D. A high-resolution canopy height model of the earth. *Nature Ecology & Evolution* **7**, 1778–1789 (2023).
68. Guo, M., Li, J., He, H., Xu, J. & Jin, Y. Detecting global vegetation changes using mann-kendal (mk) trend test for 1982–2015 time period. *Chinese Geographical Science* **28**, 907–919 (2018).
69. Mohammed, A. & Kora, R. A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University-Computer and Information Sciences* **35**, 757–774 (2023).
70. Guerschman, J. P. *et al*. Assessing the effects of site heterogeneity and soil properties when unmixing photosynthetic vegetation, non-photosynthetic vegetation and bare soil fractions from landsat and modis data. *Remote Sensing of Environment* **161**, 12–26 (2015).
71. Parente, L. *et al*. Global pasture watch - global reference samples and machine learning model for prediction of short vegetation height https://doi.org/10.5281/zenodo.15194973 (2025).
72. Lang, N., Schindler, K. & Wegner, J. D. ETH_GlobalCanopyHeight_10m_2020_version1 https://doi.org/10.3929/ethz-b-000609802 (2022).

Acknowledgements

The research presented here was supported by the Land & Carbon Lab at WRI with funding from the Bezos Earth Fund and by the Open-Earth-Monitor Cyberinfrastructure project, which received funding from the European Union's Horizon Europe research and innovation program under grant agreement No. 101059548. L.F. also acknowledges ongoing support from CAPES (Coordination of Superior Level Staff Improvement) and CNPq (the Brazilian Research Council). The authors are grateful for the contributions of Jairo Matos de Rocha to initial versions of ATL03 and ATL08 databases in support of the Brazil field campaign. The authors would also like to acknowledge the support of Tharles Andrade and Jairo Matos de Rocha in the development of the HCT viewer.

Author contributions

M.H. was the primary author and together with L.P. conceived, designed and coordinated the implementation of the mapping framework. M.H., L.F. and L.P. conceived of the field campaign that was implemented and published by M.H. M.H., L.P.,Y.H. conceived, designed and coordinated the ATL08 data processing and filtering, implemented by Y.H. L.P. and D.C. implemented the EO data pre-processing, model training, predictive modeling

and data publication. M.H. and L.P. performed visual assessment and technical validation of the results. L.P., M.H., Y.H. prepared data visualizations. M.H., L.P., L.S., C.B. contributed to writing the manuscript with input from D.M.. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.O.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.