CANOPY HEIGHTS IN THE BRAZILIAN SAVANNA: INTEGRATING ICESAT-2 DATA WITH LANDSAT AND SENTINEL IMAGERY

Fabiano Rodrigues Pereira [1], Ana Paula Dalla Corte[1,2], Carlos Alberto Silva [2], José Augusto Spiazzi Favarin[1], Thaís Chaves Almeida [1], Carine Klauberg[2], Caio Hamamura[3], Andrew Hudak[4], Lucas Bielak Rezende[1], Inacio Thomaz Bueno[2], Gabriel Máximo da Silva[5]

[1]BIOFIX, Universidade Federal do Paraná, Avenida Prefeito Lothario Meissner, 632, Jardim Botânico, Curitiba, PR, BR 80210-170, fabiano.engflo@gmail.com, anapaulacorte@gmail.com, jasflorestal@yahoo.com.br, florestal.thais@gmail.com, bielak@ufpr.br, [2]Forest Biometrics, Remote Sensing and Artificial Intelligence Lab (SilvaLab). School of Forest, Fisheries and Geomatics Science, University of Florida, PO Box 110410 Gainesville, FL, USA 32611, c.silva@ufl.edu, carine.klaubergs@ufl.edu, ithomazbueno@ufl.edu, [3]Federal Institute of Education, Science and Technology of São Paulo, Avenida Doutor Ênio Pires de Camargo, Capivari 13365-010, SP, Brazil, caiohamamura@gmail.com, [4]Forest Service, U.S. Department of Agriculture, Rocky Mountain Research Station, Forestry Sciences Laboratory, 1221 South Main St., Moscow, ID 83843, United States, andrew.hudak@usda.gov, [5]Instituto Nacional de Pesquisas Espaciais (INPE) Av. dos Astronautas, Brazil, gabriel.maximo@inpe.br

ABSTRACT

This study evaluates the accuracy of ICESat-2 for estimating canopy height in the Brazilian Savanna, a crucial biome facing significant ecological pressures. By implementing a multivariate quality control approach using Cook's distance, we enhanced the reliability of ICESat-2 data and applied Random Forest models to predict canopy heights. Our analysis revealed a strong correlation between ICESat-2 measurements and UAV-derived canopy height data, despite a slight underestimation in height predictions. The models demonstrated satisfactory accuracy, validating the effectiveness of ICESat-2 in large-scale canopy heights modeling. This research represents a significant advancement in utilizing ICESat-2 for canopy height assessment in the Brazilian Savanna, highlighting its potential for informing conservation strategies and carbon market initiatives. Future studies are encouraged to integrate various remote sensing technologies to improve canopy height estimations further.

Key words **—** Brazilian Savanna, photon-counting LiDAR, Cook's distance, ATL08, Remote sensing.

1. INTRODUCTION

The Cerrado, or Brazilian Savanna, is South America's second-largest biome, covering about two million square kilometers (23.3% of Brazil) [1, 2]. Nearly half of its original vegetation has been cleared [3], and the remaining areas face threats from agricultural expansion. Innovative monitoring is essential for understanding biomass distribution and variability, supporting predictive models for vegetation and climate, informing decision-making, carbon markets, and REDD+ programs [4]. Accurate measurement of parameters like canopy height, tied to forest productivity and biodiversity, is essential to understand ecosystems' carbon cycle interactions under climate change. Satellite data are crucial for assessing large-scale vegetation and capturing structural diversity [5, 6, 7]. NASA's ICESat-2, since its 2018 launch, provides high-resolution 3D canopy height data across diverse forests, improving sampling over its predecessor [8, 9, 10, 11]. Originally designed for ice and land elevation, ICESat-2 shows significant potential for canopy height estimation, though its application to model forest canopy height at regional scales, particularly in the Brazilian Savanna, remains limited [12, 13, 14, 15, 16].

ICESat-2 canopy height estimates face uncertainties due to its photon-counting laser system [17], affecting sample quality and representation of actual canopy structure. Stringent quality control is thus essential. This study used Cook's distance, a multivariate technique [18], to enhance ICESat-2 data reliability. With refined data, we tested satellite-derived variables in Random Forest models for canopy height prediction and validated the quality control and predictor variables using a UAV-derived canopy height model (CHM). To our knowledge, this study is the first to apply multivariate quality control to ICESat-2 samples and to develop 30-meter orbital-based canopy height models (CHMs) in the Brazilian Savanna. The proposed quality control analysis and CHM generation workflow utilize freely available satellite data and ICESat-2 observations, making this approach adaptable to other forested regions and beneficial for large-scale aboveground biomass modelling, potentially enhancing fire management strategies in the Brazilian Savanna.

2. MATERIAL AND METHODS

2.1. Study area and UAV-LiDAR Data

The research took place in the Canto do Obrieni Privately Protected Area (PPA) managed by the Instituto Araguaia. This 240ha site of Brazilian Savanna encompasses pristine climax Cerrado forests with trees reaching up to 25m in height, unique native grasslands, and seasonally flooded forests [13]. Situated in the state of Tocantins, this area is located in Brazil's Northern region (Figure 1).

The UAV-LiDAR data were collected in August 2024 using the DJI Zenmuse L1 sensor mounted on the Matrice 300 RTK UAV platform. The data acquired from the Zenmuse L1 laser scanner were processed with DJI Terra

software and subsequently exported in LAS format within the UTM coordinate system (EPSG 32722).

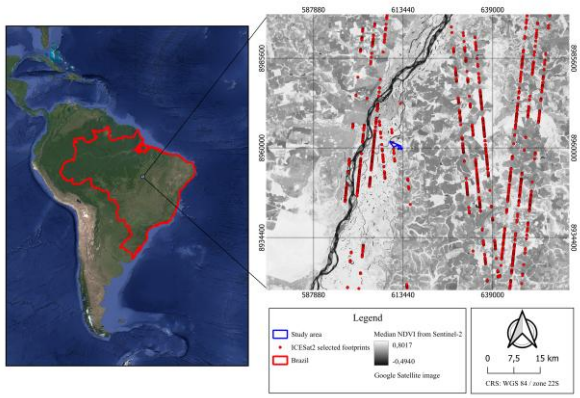


Figure 1. Location map of the study area and the coverage of selected ICESat-2 footprints surrounding it, used for training the canopy height models.

LiDAR data processing used LAStools functions [19]. High-density 3D point clouds were merged with *lasmerge*, clipped to the study area with *lasclip*, and cleaned of noise with *lasnoise*. Ground returns were classified using *lasground_new*, and a 2-meter resolution digital terrain model (DTM) was generated with *las2dem*. This DTM normalized the LiDAR point clouds (*lasheight*) and produced a 2-meter resolution canopy height model (CHM), which validated canopy height estimates from modeling approaches.

2.2. ICESat-2 Data

In this study, the canopy heights from the ICESat-2 vegetation and terrain elevation product (ATL08) were used as the response variable in our models, combined with satellite-derived explanatory variables [20]. The ATLAS instrument onboard ICESat-2 is a photon-counting system that utilizes six beams—three pairs of strong and weak beams—emitting green wavelength lasers (532 nm) with a nominal footprint of about 13 meters [8, 10]. The algorithm for the Land and Vegetation Height data product (ATL08) extracts terrain and canopy height information from the geolocated photon point cloud (ATL03), processed in 100-meter segments along the satellite's ground track [21]. Additionally, it computes various relative height metrics for each segment, providing data on canopy heights relative to the terrain, height variability, and quality flags [20].

In this study, we trained the models using canopy height data within a 50 km buffer of the study area, covering the year leading up to the UAV collection date (Oct/23 to Oct/24). This approach ensured a robust dataset with a substantial volume of data, facilitating the selection of high-quality observations from ICESat-2, particularly since the forests have not experienced significant canopy height changes in a year. Although the ATL08 height products provide various height metrics, we focused on the standard metric output from the ATL08 product, specifically the 98th percentile (P98).

2.3. Satellite- derived explanatory variables

Predictor variables were derived from spectral indices obtained from HLSL30 [21] and Sentinel-2 MSI Level 2A images. Median composites of HLSL30 from the Harmonized Landsat Sentinel-2 (HLS) project and Sentinel-2 MSI images were created for the study area and period using the 'rgee' and 'ICESat2VegR' packages [22], integrating R with Google Earth Engine (GEE). Images with <10% cloud cover were selected, and a quality mask was applied to remove pixels affected by clouds and shadows. Further details are available at [23]. From the HLSL30 images, surface reflectance was calculated for six spectral bands (BLUE, GREEN, RED, NIR, SWIR1, and SWIR2), along with two vegetation indices (NDVI and EVI). For the Sentinel-2 images, ten spectral bands (BLUE, GREEN, RED, RED_EDGE_1 to RED_EDGE_4, NIR, SWIR1, and SWIR2), as well as the indices NDVI and EVI, including variations of NDVI utilizing the RED_EDGE bands were extracted. All these variables served as predictors for ICESat-2 canopy height.

2.4. ICESat-2 Data Extraction and quality control

ICESat-2 data for the study area were extracted using the ICESat2VegR package [24] in R [25]. The ATLAS_dataFinder function identified relevant ICESat-2 ATLAS data granules, and ATL03 and ATL08 data were downloaded from the LP DAAC Data Pool using ATLAS_dataDownload. Segment attributes from ATL03 and ATL08 were extracted and clipped to a 50 km buffer around the study area, yielding 40,657 footprints.

Next, regions classified as Forest or Savanna Formation (IDs 3 and 4) were identified using 30 m land cover and land use maps from MapBiomas Collection 9 [26]. The raster data was converted to vector format to outline forest polygons, with only those over 10 hectares considered. Footprints within these polygons were then selected for further analysis. ICESat-2 data have an approximate 5 m geolocation error [10] and apply a 100 m aggregation in the canopy height algorithm, complicating ALS comparisons. To address this, samples were selected from areas of high spatial homogeneity, defined by low STD of Sentinel-derived NDVI in a 90 m × 90 m (3 × 3 pixel) window. Pixels with high STD were excluded using an interquartile range (IQR) threshold: STD* = Q3 + 1.5 × IQR. Data from weak beams, daytime acquisitions, and canopy heights over 40 m were also excluded to enhance reliability [7, 10].

To address interactions affecting ICESat-2 data quality, we used a multivariate approach—Cook's distance (D) [27]—following the quality control method in [13]. Cook's distance, commonly used in regression, is effective for identifying influential data points [28]. High D values indicate strong impacts on canopy height regression. Using the IQR rule, a threshold D* was set to automate low-quality sample detection and can be adjusted for stricter control based on research needs.

2.5. Canopy height modeling and assessment

The random forest (RF) model was applied to extrapolate ICESat-2 canopy height using spectral metrics from satellite

images for wall-to-wall canopy height estimates. Satellite metrics were retrieved at a 30-m scale for each ICESat-2 footprint using seg_gee_ancillary_dt_extract from ICESat2VegR. Three models were tested: Model 1 combined Landsat 8/9 OLI and Sentinel-2, Model 2 used only Landsat 8/9 OLI, and Model 3 used only Sentinel-2. RF modeling, conducted with the 'randomForest' package in R [29], used optimal parameters (ntree = 200, mtry = one-third of input variables).

After the extrapolation of height canopy for the entire study site, the predicted height canopy values within the ALS flight coverage were compared to the ALS HP75 metric (the one with the highest correlation with ICESat-2 height canopy P98 for the area) to validate the modelling accuracy. The UAV-derived canopy heights were spatially matched with models predictions by randomly creating 150 polygons (30 x 30 m) in the UAV-covered area.

The analysis included the Pearson correlation coefficient (r), absolute and relative mean difference (MD and MD%), as well as absolute and relative root mean square difference (RMSD and RMSD%). Additionally, comparison plots were used to assess the predicted ICESat-2 canopy height compared to those obtained from the UAV-CHM.

3. RESULTS

After quality control, 1,306 footprints (~3%) were utilized to train the models. The canopy heights associated with these footprints ranged from 2.54 to 24.32 m, with an average height of 13.76 m and a standard deviation of 4.23 m. During the training phase, the ICESat-2 canopy heights consistently underestimated the actual canopy height across all tested models, with MD and RMSD ranging from 0.01 to 0.04 m and 2.52 to 2.67 m, respectively (Figure 2). The model that combined variables from both satellites exhibited the highest accuracy, with r and RMSD% values of 0.8 and 18.33%, respectively.

Independent validation confirmed that all models achieved satisfactory accuracy in canopy height predictions, with MD, r, and RMSD values ranging from -0.05 to 0.5 m, 0.74 to 0.79, and 2.99 to 3.14 m, respectively. Notably, only the model using Landsat images as predictors demonstrated a slight overestimation of canopy height.

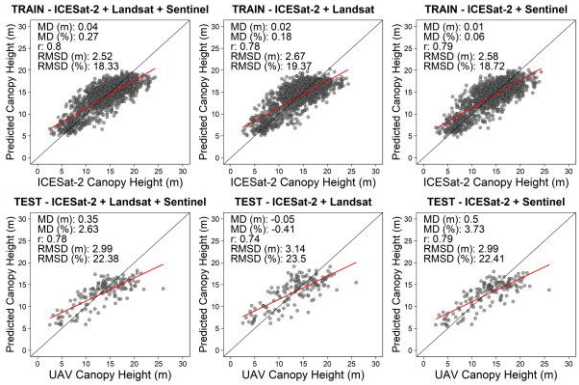
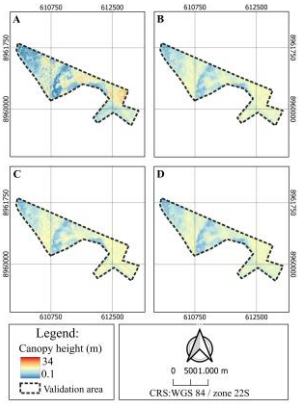


Figure 2. Agreement between predicted ICESat-2 canopy height by different models and UAV-derived canopy height.



Consistent with the training phase, the models showed minimal differences in accuracy, as reflected in the canopy height prediction maps (Figure 3). Furthermore, the maps demonstrate that all three models effectively predicted canopy height, closely aligning with the spatial distribution depicted in the CHM map used for validation.

Figure 3. Canopy height maps from CHM and those derived from the canopy height models. A: CHM; B: ICESat-2 + Landsat 8/9 OLI + Sentinel-2; C: ICESat-2 + Landsat 8/9 OLI; D: ICESat-2 + Sentinel-2.

4. DISCUSSION

In this study, we assessed the accuracy of ICESat-2's vegetation canopy height product in a Brazilian Savanna area located in northern Brazil. Our findings indicated a strong correlation between ICESat-2 canopy height and the P75 metric of UAV-CHM, with a slight underestimation consistent with previous studies in other forested regions [15, 21].

Our study found that ICESat-2 canopy height exhibited the strongest correlation with the UAV height metric P75, contrasting with prior research in forested regions, where canopy height typically aligns most closely with the 90th to 95th percentile relative height [13, 15, 20]. This finding suggests that selecting a specific ALS height metric for comparison with ICESat-2 canopy height may not need to be standardized, as height distribution patterns can vary across different areas. However, the consistently high correlations between various ALS and ICESat-2 heights metrics reinforce the high detection accuracy of the ICESat-2 canopy height product, establishing it as a reliable foundational dataset for subsequent spatially continuous mapping.

Moreover, the estimation errors (RMSD = 2.99 m to 3.14 m) obtained in this study are comparable to those in previous studies, such as [15] (RMSE = 2.64 to 2.93 m), [21] (RMSE = 0.85 to 3.25 m), and [18] (RMSE = 4.49 to 5.83 m). These satisfactory results in ICESat-2 canopy height mapping also highlight ICESat-2's potential to advance applications in the Brazilian Savanna.

In summary, we generated wall-to-wall canopy height maps using Random Forest (RF) models to link ICESat-2 canopy height data with spectral imagery. Results may vary with different modeling approaches or additional data sources and improved canopy height estimates have been noted by integrating LiDAR, RADAR, and multispectral imagery with machine learning for forest mapping [13]. Thus, future studies using diverse machine learning techniques or NASA-ISRO NISAR data could enhance the spatial resolution and accuracy of canopy height maps in the Brazilian Savanna

when combined with ICESat-2 data.

5. CONCLUSION

The present study underscores the potential of ICESat-2 as a valuable tool for canopy height estimation in the Brazilian Savanna, as we obtained reliable ICESat-2 data using a multivariate quality control technique and effectively applied Random Forest models to predict canopy heights. Our results indicate a strong correlation between ICESat-2 canopy heights and UAV-derived data, despite a slight underestimation. The models achieved satisfactory accuracy, demonstrating that ICESat-2 can effectively inform large-scale canopy height and ecological assessments in the region. Future research could benefit from integrating diverse remote sensing technologies to further enhance canopy height estimation in this biome.

6. ACKNOWLEDGMENT

This study was funded by NASA's Carbon Monitoring System (CMS). Fabiano Rodrigues Pereira is a doctoral fellow supported by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Funding Code (88887.956692/2024-00). This work was partially conducted by the Biofix Laboratory, Federal University of Paraná (UFPR). Araguaia Institute for providing the area for the study. Ana Paula Dalla Corte – Funding Code (305422/2021-9; 402350/2021-9).

7. REFERENCES

[1] J. M. C. Silva, J. M. Bates. Biogeographic patterns and conservation in the South American Cerrado: a tropical savanna hotspot. Bioscience, 52(3):225–234, 2002.
[2] J. Bonanomi, et al. Protecting forests at the expense of native grasslands: Land-use policy encourages open-habitat loss in the Brazilian cerrado biome. Perspectives in Ecology and Conservation, 17(1):26–31, 2019.
[3] C. M. Souza, Jr., et al. Reconstructing three decades of land use and land cover changes in Brazilian biomes with Landsat archive and Earth Engine. Remote Sensing, 12(17):2735, 2020.
[4] M. B. T. Da Costa, et al. Beyond trees: Mapping total aboveground biomass density in the Brazilian savanna using high-density UAV-lidar data. Forest Ecology and Management, 491:119155, 2021.
[5] D. Schimel, et al. Observing terrestrial ecosystems and the carbon cycle from space. Global Change Biology, 21:1762–1776, 2015.
[6] S. Tao, et al. Global patterns and determinants of forest canopy height. Ecology, 97:3265–3270, 2016.
[7] L. Malambo, S. C. Popescu. Assessing the agreement of ICESat-2 terrain and canopy height with airborne lidar over US ecozones. Remote Sensing of Environment, 266:112711, 2021.
[8] L. A. Magruder, et al. Early ICESat-2 on-orbit geolocation validation using ground-based corner cube retro-reflectors. Remote Sensing, 12:3653, 2020.
[9] T. Markus, et al. The Ice, Cloud, and Land Elevation Satellite-2 (ICESat-2): Science requirements, concept, and implementation. Remote Sensing of Environment, 190:260–273, 2017.
[10] A. Neuenschwander, et al. Validation of ICESat-2 terrain and canopy heights in boreal forests. Remote Sensing of Environment, 251:112110, 2020.
[11] T. A. Neumann, et al. The Ice, Cloud, and Land Elevation Satellite – 2 mission: A global geolocated photon product derived from the Advanced Topographic Laser Altimeter System. Remote Sensing of Environment, 233:111325, 2019.
[12] X. Zhu, et al. Consistency analysis of forest height retrievals between GEDI and ICESat-2. Remote Sensing of Environment, 281:113244, 2022.
[13] T. Zhang and D. Liu. Improving ICESat-2-based boreal forest height estimation by a multivariate sample quality control approach. *Methods in Ecology and Evolution*, 14(7): 1623-1638, 2023.
[14] F. Jiang, et al. Mapping the forest canopy height in northern China by synergizing ICESat-2 with Sentinel-2 using a stacking algorithm. *Remote Sensing*, 13(8): Article 8, 2021.
[15] W. Li, et al. 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.
[16] S. Nandy, et al. Mapping forest height and aboveground biomass by integrating ICESat-2, Sentinel-1, and Sentinel-2 data using random forest algorithm in northwest Himalayan foothills of India. *Geophysical Research Letters*, 48(14): e2021GL093799, 2021.
[17] S. C. Popescu, et al. Neuenschwander, R. Sheridan, L. Narine, and K. M. Walsh. Photon counting LiDAR: An adaptive ground and canopy height retrieval algorithm for ICESat-2 data. *Remote Sensing of Environment*, 208: 154–170, 2018.
[18] Instituto Araguaia. Annual Report 2022. Brasil, 2022.
[19] M. Isenburg. "LAStools—Efficient LiDAR Processing Software" (Version 1.8, Licensed). Available online: http://rapidlasso.com/LAStools (accessed on 11 November 2019).
[20] A. L. Neuenschwander, et al. Ice, Cloud, and Land Elevation Satellite 2 (ICESat-2) algorithm theoretical basis document (ATBD) for land-vegetation along-track products (ATL08). National Aeronautics and Space Administration, Washington, DC, USA, 2019.
[21] A. Neuenschwander, K. Pitts. The ATL08 land and vegetation product for the ICESat-2 Mission. Remote Sensing of Environment, 221:247–259, 2019.
[22] C. Aybar et al. rgee: An R package for interacting with Google Earth Engine. *Journal of Open Source Software*, 5(51): 2272, 2020.
[23] M. Claverie et al. Harmonized Landsat-8 Sentinel-2 (HLS) product user's guide. National Aeronautics and Space Administration (NASA), Washington, DC, USA, 2017.
[24] C. A. Silva and C. Hamamura. ICESat2VegR: An R Package for NASA's Ice, Cloud, and Elevation Satellite (ICESat-2) Data Processing and Visualization for Terrestrial Applications. Version 0.0.1, accessed on 13 Jun. 2024. Available at: https://CRAN.R-project.org/package=ICESat2VegR.
[25] R CORE TEAM. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2024. Disponível em: https://www.R-project.org/.
[26] MapBiomas. 2024. Collection of Brazilian Land Cover & Use Map Series (http://mapbiomas.org/).
[27] R. D. Cook. Detection of influential observation in linear regression. *Technometrics*, 19(1): 15–18, 1977.
[28] K. S. Kannan and K. Manoj. Outlier detection in multivariate data. *Applied Mathematical Sciences*, 47(9): 2317–2324, 2015.
[29] L. Breiman. Random forests. *Mach. Learn.*, 45: 5–32, 2001.