

Article

## Canopy Height Estimation in French Guiana with LiDAR ICESat/GLAS Data Using Principal Component Analysis and Random Forest Regressions

Ibrahim Fayad <sup>1,\*</sup>, Nicolas Baghdadi <sup>1</sup>, Jean-Stéphane Bailly <sup>2</sup>, Nicolas Barbier <sup>3</sup>, Valéry Gond <sup>4</sup>, Mahmoud El Hajj <sup>5</sup>, Frédéric Fabre <sup>6</sup> and Bernard Bourgin <sup>7</sup>

<sup>1</sup> IRSTEA, UMR TETIS, 500 rue Jean François Breton, 34093 Montpellier Cedex 5, France; E-Mail: nicolas.baghdadi@teledetection.fr

<sup>2</sup> AgroParisTech, UMR LISAH, 2 place Pierre Viala, 34060 Montpellier, France; E-Mail: bailly@agroparistech.fr

<sup>3</sup> IRD, UMP AMAP, Bd de la Lironde, TA A51/PS2, 34398 Montpellier Cedex 5, France; E-Mail: nicolas.barbier@ird.fr

<sup>4</sup> CIRAD, UPR B&SEF, campus international de Baillarguet, 34398 Montpellier Cedex 5, France; E-Mail: valery.gond@cirad.fr

<sup>5</sup> NOVELTIS, 153 rue du Lac, 31670 Labège, France; E-Mail: mahmoud.elhajj@noveltis.fr

<sup>6</sup> Airbus Defense and Space, 31 rue des Cosmonautes Z.I. du Palays, 31402 Toulouse, France; E-Mail: frederic.fabre@astrium.eads.net

<sup>7</sup> BRGM, 3 avenue Claude Guillemin, 45060 Orléans, France; E-Mail: b.bourgine@brgm.fr

\* Author to whom correspondence should be addressed; E-Mail: ibrahim.fayad@teledetection.fr; Tel.: +33-4-6754-8754.

External Editors: Heiko Balzter and Prasad S. Thenkabail

Received: 10 June 2014; in revised form: 11 November 2014 / Accepted: 17 November 2014 /

Published: 28 November 2014

---

**Abstract:** Estimating forest canopy height from large-footprint satellite LiDAR waveforms is challenging given the complex interaction between LiDAR waveforms, terrain, and vegetation, especially in dense tropical and equatorial forests. In this study, canopy height in French Guiana was estimated using multiple linear regression models and the Random Forest technique (RF). This analysis was either based on LiDAR waveform metrics extracted from the GLAS (Geoscience Laser Altimeter System) spaceborne LiDAR data and terrain information derived from the SRTM (Shuttle Radar Topography Mission) DEM (Digital Elevation Model) or on Principal Component Analysis (PCA) of GLAS waveforms. Results

show that the best statistical model for estimating forest height based on waveform metrics and digital elevation data is a linear regression of waveform extent, trailing edge extent, and terrain index (RMSE of 3.7 m). For the PCA based models, better canopy height estimation results were observed using a regression model that incorporated both the first 13 principal components (PCs) and the waveform extent (RMSE = 3.8 m). Random Forest regressions revealed that the best configuration for canopy height estimation used all the following metrics: waveform extent, leading edge, trailing edge, and terrain index (RMSE = 3.4 m). Waveform extent was the variable that best explained canopy height, with an importance factor almost three times higher than those for the other three metrics (leading edge, trailing edge, and terrain index). Furthermore, the Random Forest regression incorporating the first 13 PCs and the waveform extent had a slightly-improved canopy height estimation in comparison to the linear model, with an RMSE of 3.6 m. In conclusion, multiple linear regressions and RF regressions provided canopy height estimations with similar precision using either LiDAR metrics or PCs. However, a regression model (linear regression or RF) based on the PCA of waveform samples with waveform extent information is an interesting alternative for canopy height estimation as it does not require several metrics that are difficult to derive from GLAS waveforms in dense forests, such as those in French Guiana.

**Keywords:** LiDAR; ICESat/GLAS; canopy height; tropical forest; French Guiana

---

## 1. Introduction

Standing aboveground biomass (AGB) plays a crucial role in the global carbon cycle and is an indispensable factor in environmental and climate modeling, not only for understanding the carbon cycle but also for mitigating the effects of global warming via conservation of carbon sinks. The quantification of aboveground biomass (AGB) and the sequestration of carbon in tropical forests are of major importance, as more than 40% of the global terrestrial carbon stock is contained in these forests (e.g., [1,2]).

Several studies have developed allometric relationships linking the characteristics of a forest to its biomass (e.g., [3–5]). Chave *et al.* [3] developed a pantropical biomass estimation model at the individual tree level that was based on the formula for calculating the mass of a cylinder using stem diameter, height, and wood density (WD). Asner *et al.* [4] proposed a plot aggregate allometry model for tropical areas drawn from the Chave *et al.* [3] model, but they replaced *in situ* canopy height with top-of-canopy height (TCH), as derived from airborne small-footprint LiDAR measurements, and stem diameter with plot-averaged basal area (BA). BA and WD were linked with TCH using linear relationships in the form of  $BA = aTCH$  and  $WD = bTCH + c$ , producing a model for AGB estimation using only TCH. Results showed a RMSE on AGB estimation of 24.7 Mg/ha for the regional models (model coefficients dependent on region) and 26.4 Mg/ha for the generalized model (generalized model coefficients for all regions). Drake *et al.* [5] used a power function to link top-of-canopy height estimated from airborne LiDAR to aboveground biomass ( $AGB = aTCH^b$ ). However, this method is considered plot-aggregate

allometry rather than true allometry, as it reflects the whole-plot properties of forest structures in aggregate and not the properties of each particular tree. This method had an RMSE of 42.2 Mg/ha when tested in five tropical forests with different vegetation types.

The evaluation of canopy height is paramount in AGB estimation, as most allometric relations use canopy heights for the estimation of AGB. Studies have shown that the inclusion of height in biomass allometries significantly improves biomass estimation accuracy [3,6–9]. Chave *et al.* [3] reported that the inclusion of height for stand level estimates of biomass reduced error from 19.5% to 12.8% across all forms of tropical forests and across continents. In addition, Lefsky *et al.* [10], Asner *et al.* [4], and Mitchard *et al.* [11] also found that canopy heights and other LiDAR derived metrics are strongly related to forest biomass. Recently, Asner *et al.* [4] provided allometric relations that allow the estimation of biomass using only LiDAR top-of-canopy heights.

Several studies attempted to estimate canopy heights with either polarimetric SAR interferometry (PolInSAR) [12–15] or SAR tomography [14,16]. Results of PolInSAR, which uses polarimetric separation of scattering phase centers derived from interferometry for canopy height estimation, showed promising results [12,14,17]. However, PolInSAR in forestry is strongly hindered, but not limited to weather changes, atmospheric heterogeneities, and intrinsic phase noise. SAR tomography is an alternative technique for using radar data in canopy height estimation. This technique is an imaging approach, which generates a fully 3D representation of the imaged scene using coherent combination of a greater number of images [14,18,19]. SAR tomography is more robust against various noise sources in comparison to PolInSAR at the expense of the necessity to require many more flight lines. The BIOMASS Earth Explorer mission selected by ESA (European Space Agency) in the framework of its living planet program with a P-band spaceborne SAR satellite will provide strong opportunities for the estimation of both canopy heights and biomass from SAR images. Furthermore, many studies used medium and high resolution optical imagery such those available from MODIS, Landsat, Quickbird, IKONOS and others in order to extrapolate airborne or spaceborne LiDAR derived canopy height estimates (e.g., [20,21]). Lefsky *et al.* [20] results for global canopy height estimation using linear regression with medium 500 m Moderate Resolution Imaging Spectroradiometer (MODIS) data, showed moderately strong relationships for predicting the 90th percentile patch height with a mean RMSE of 5.9 m. Wulder and Seeman [21] used a regression model that relates reflectance in the different spectral bands to airborne large-footprint LiDAR for canopy height estimation.

To this date, canopy height estimation over large areas is best achieved using LiDAR data (either Airborne or Spaceborne). Several studies have estimated canopy height using airborne or spaceborne LiDAR data (e.g., [10,11,22,23]). At regional and global scales, LiDAR data acquired by the Geoscience Laser Altimeter System (GLAS) have been widely used (e.g., [10,20]). Using GLAS data, maximum canopy height within each footprint has been successfully estimated with a precision between 2 and 13 m, depending on forest types and characteristics of the study site (e.g., [10,24–26]). Lefsky *et al.* [10], which applied linear regressions on waveform metrics and ancillary DEM data for canopy height estimation obtained site-specific models with an RMSE between 4.85 and 12.66 m. Hilbert and Schmullius [24] when estimating canopy heights obtained an RMSE of 6.39 m on the canopy height estimation regarding all species and slope classes with a clear negative correlation between accuracy and slope. Lee *et al.* [25] applied a slope correction metric to a GLAS estimation model obtained high correlation between GLAS canopy height estimates and those estimated from a small footprint LiDAR

with an RMSE of 2.2 m. Pang *et al.* [26] estimated the crown-area-weighted mean height with airborne LiDAR measurements using linear regression applied to metrics derived from GLAS waveforms. Their results indicated an RMSE of 3.8 m on the estimation of canopy heights in several coniferous forest sites in western North America.

Lefsky *et al.* [10] linked the maximum canopy height ( $H_{\max}$ ) estimated from GLAS data to AGB using the following linear relationship:  $AGB = \alpha + aH_{\max}^2$ . Boudreau *et al.* [22] linked the GLAS waveform extent (difference between signal start and signal end), the slope ( $\theta$ ) between signal start and the first Gaussian canopy peak and the terrain index (TI) metric derived from the SRTM-DEM to AGB. Saatchi *et al.* [23] and Mitchard *et al.* [11] used Lorey's height (basal-area-weighted canopy height) instead of the maximum height for AGB estimation. In the different studies, it was found that Lorey's height is broadly related to canopy height [20]. However, Asner *et al.* [4] found that Lorey's height does not explain any variations in AGB, basal area, or wood density that cannot be explained by canopy height.

Other studies have relied on optical sensors for AGB estimation. However, optical sensors generally give aggregate spectral signatures (reflectance or vegetation indices) over broad areas with no information on vertical or horizontal forest structure, and these responses have been shown to saturate even at intermediate biomass levels (150–200 Mg/ha) (e.g., [27,28]) and are highly dependent on atmospheric perturbations. Very-high-resolution optical images (metric pixels) have also been used to characterize the spatial (horizontal) characteristics of the canopy through textural indices. This approach provides a much broader range of sensitivity [29] (up to 600 Mg/ha AGB) but is sensitive to variations in scene lighting caused by changes in sun-viewing configurations.

Several studies have also used L-band synthetic aperture radar (SAR), such as PALSAR/ALOS, JERS-1 and SIR-C (radar wavelength of approximately 25 cm) (e.g., [11,30–33]). However, similar to optical sensors, the saturation levels of the radar sensors occur at low to intermediate biomass levels (between 60 and 150 Mg/ha). Le Toan *et al.* [30], Wu *et al.* [32], and Dobson *et al.* [33] reported L-band radar signal saturation of biomass levels at 100 Mg/ha in coniferous forests. However, according to Imhoff *et al.* [34], the saturation levels are closer to 40 Mg/ha because the saturation thresholds occur before the regression maxima. In boreal forests, saturation levels were observed up to 150 Mg/ha (e.g., [11,31]). Luckman *et al.* [35,36] found a saturation point of 60 Mg/ha in the Central Amazon basin. Finally, due to the higher saturation level of SAR P-band, the P-band is able to estimate AGB at higher ranges (e.g., [34,37,38]). Imhoff *et al.* [34] examined AGB levels in broadleaf evergreen forests in Hawaii and coniferous forests in North America and Europe and found saturation levels of 100 Mg/ha for the P-band *versus* 40 Mg/ha for the L-band. Minh *et al.* [37] reported a decrease in sensitivity of approximately 300 Mg/ha for the biomass in dense tropical forests using the P-band. Nizalapur *et al.* [38] found a decrease in sensitivity in a tropical dry deciduous forest of approximately 200 Mg/ha for the P-band and 150 Mg/ha for the L-band.

In a recent study, Zolkos *et al.* [39] compared the performance of the AGB estimation approaches of more than 70 studies using different sensor types (LiDAR, optical, and radar). Their results indicated that LiDAR data (airborne and spaceborne) were significantly better at estimating biomass in comparison to optical and radar sensors. The mean relative standard errors (RSE) of the biomass estimations with different LiDAR data were 39.4 Mg/ha for discrete LiDAR data, 50.2 Mg/ha for full return LiDAR data and 39.6 Mg/ha for GLAS data. However, the RSE was 70 Mg/ha when using optical- or radar-data-based models alone. In spite of these findings, Zolkos *et al.* [39] did not take into account studies using textural

indices extracted from very-high-resolution optical images for the estimation of AGB, which provide a higher sensitivity for AGB of up to 600 Mg/ha (e.g., [29]).

In this study, we used LiDAR data provided by GLAS to estimate canopy heights in French Guiana. Canopy height estimation models based on full waveform data can be divided into two categories: the direct method and statistical models. The direct method enables canopy height estimation in low relief areas using the difference in elevation between signal start and the ground. However, over sloping areas, the direct method overestimates canopy heights because of the additional height introduced by the slope. To remove the effects of the slope, statistical models using GLAS and DEM metrics have been developed. Nevertheless, while the metrics developed in previous studies were very successful in increasing the precision of the canopy height estimation models (e.g., [10,24,26,40]), they presented their own shortcomings. Indeed, in order to use these metrics for better canopy height estimation, the exact position of the top-of-canopy and ground peaks is often required. Over dense vegetated areas such tropical forests, extracting the top-of-canopy and ground peaks is especially difficult using an automated process, as the LiDAR waveform does not often present distinctive peaks [24]. The extraction of these metrics manually is always possible, but becomes time consuming and inefficient when dealing with a large number of GLAS waveforms. Therefore, the aims of the present paper are to test several commonly used canopy height estimation models that utilize metrics derived from GLAS waveforms and SRTM-DEM and to test two techniques, new in the field of forest applied LiDAR: principal component analysis (PCA) and Random Forest. The purpose of using the PCA approach is to eliminate the need for metrics extracted from GLAS in canopy height estimation models, as the extraction of these metrics is error-prone, especially in dense forests, such as those in French Guiana. For the Random Forest regressions, the same metrics derived from GLAS footprints will first be used. Then, the principal components from the PCA of the GLAS waveform will be tested. The results of each model will be validated against canopy height estimates obtained from an airborne LiDAR dataset.

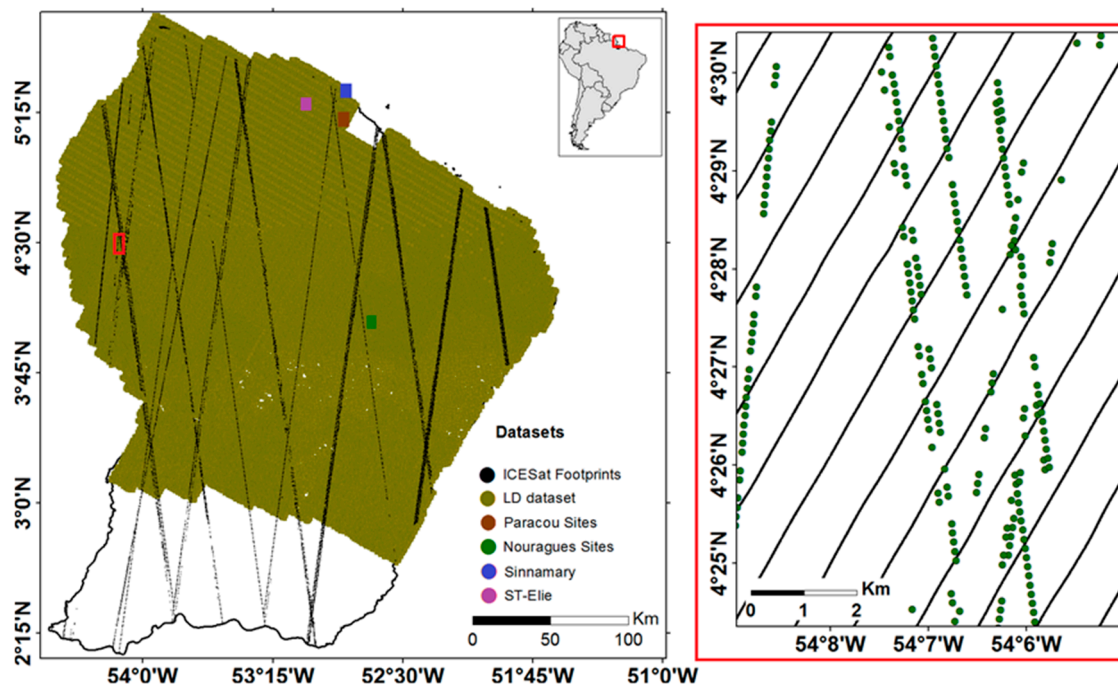
A description of the satellite and airborne LiDAR datasets used in this study is given in Section II, followed in Section III by the presentation of methods for forest height estimation using airborne and satellite LiDAR. The results are shown in in Section IV. Finally, Sections V and VI present the discussion and conclusions respectively.

## **2. Dataset Description**

### *2.1. Study Area*

The study site is French Guiana, which lies in the tropics. It is situated on the northern coast of the South American continent, bordering the Atlantic Ocean, Brazil and Suriname (the central coordinates are 5°15'N and 52°55'W, Figure 1). French Guiana's surface area is 83,534 km<sup>2</sup>, of which 96.75% is forested. The terrain is mostly low-lying; 67.8% of the slopes are lower than 5°, rising occasionally to small hills and mountains. The altitude ranges from 0 to 851 m. The DEM acquired by the Shuttle Radar Topography Mission (SRTM) with a resolution of three arcseconds (90 m) was used.

**Figure 1.** LiDAR datasets acquired for French Guiana (the right image corresponds to the red rectangle in the left image).



## 2.2. Airborne LiDAR Dataset

### 2.2.1. Small-Footprint Low-Density LiDAR Dataset (LD)

A LiDAR dataset was acquired in 1996 during an airborne geophysical survey that covered 4/5 of French Guiana (northern part, Figure 1). Because laser data were acquired for assessing the quality of the survey, and particularly for flight ground clearance, a low sampling frequency was used, and only the first pulse was considered [41]. The data correspond to the elevation of the first obstacle encountered by the laser. The sampling frequency was 10 Hz with a 905-nm wavelength laser and a footprint size of 35 cm (laser beam width of approximately 3 mrad). The laser measurements are therefore considered point data. The database contains laser elevations every 7 m on flight lines spaced 500 m apart and oriented at 30°N, intersected by transverse flight lines spaced 5 km apart and oriented at 120°N. The mean density of this database is approximately 285.2 points/km<sup>2</sup>. Bourguine *et al.* [42] evaluated the quality of this low-density LiDAR dataset (LD), and the accuracy of the terrain elevation was estimated to be approximately  $\pm 2$  m.

### 2.2.2. Small-Footprint High-Density LiDAR Dataset (HD)

LiDAR datasets with high points density (HD) acquired during several airborne surveys in 2004, 2007, 2008 and 2009 by the Altoa Corporation using a helicopter were also used in this study (Table 1). The elevations were recorded using two LiDAR systems: Riegl LMS-Q140i-60 in 2004, 2007 and 2008 and the newer LMS-280i system in 2009. The elevation data were acquired for several small study sites in French Guiana (Figure 1). The mean acquisition density of the HD datasets is 3.5 points/m<sup>2</sup> (between 0.9 and 5.6 points/m<sup>2</sup>). The laser wavelength was 905 nm with a mean footprint size of 45 cm for the

first system and 10 cm for the second, and the precision of the elevation was smaller than 0.1 m [43]. Moreover, the HD, unlike the LD, is a last-return laser elevation measurement, as using the last return increased the percentage of ground returns [43].

**Table 1.** Description of the HD datasets used in this study.

Site	Acquisition Date	Location	Area (km <sup>2</sup> )	Point Density (points/m <sup>2</sup> )
Paracou_2004	2004	5°15.9'N 52°55.9'W	5.35	0.9
Sinnamary	2004	5°24.7'N 52°56'W	6.52	0.9
St-Elie	2007	5°18.2'N 53°3.3'W	4.40	5.3
Nouragues07A	2007	4°5.3'N 52°40.7'W	7.24	3.2
Nouragues07B	2007	4°2.4'N 52°40.6'W	2.42	3.8
Nouragues08A	2008	4°5.1'N 52°41.2'W	1.96	4.5
Nouragues08B	2008	4°3.8'N 52°40.9'W	7.82	3.8
Nouragues08C	2008	4°2.5'N 52°40'W	2.89	4.2
Nouragues08D	2008	4°2.5'N 52°41.0'W	1.08	3.5
Paracou_2009	2009	5°16.1'N 52°55.8'W	12.08	5.6

### 2.3. Spaceborne LiDAR Dataset

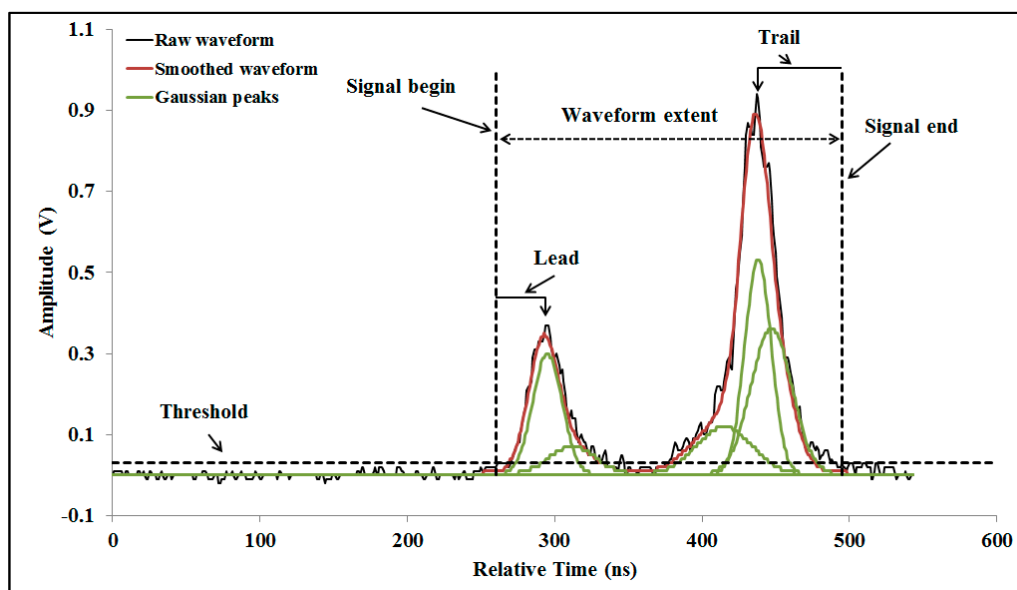
The Geoscience Laser Altimeter System (GLAS) on the Ice, Cloud and land Elevation Satellite (ICESat), which launched in January 2003, used three onboard lasers, L1, L2, and L3, to measure the elevation changes of the polar ice-sheets as well as cloud and aerosol properties. During its operational years, GLAS operated with orbit cycles repeating between every 57 and 197 days for a total of 18 missions. This was due to the unexpectedly short lifetime of the laser system. The GLAS data were acquired in October-November, February-March, and May-June. GLAS acquired full waveform data along profiles with a footprint diameter ranging between 110 and 50 m, spaced every 175 m along the profile. The nominal pointing angle was approximately 0.3° off nadir. Over land and ice surfaces, GLAS measured vertical structures using a 1064-nm laser pulse, and the waveforms were then digitized in 544 or 1000 bins (depending on the mission) with a vertical resolution of 1 ns (15 cm), which corresponds to 81.6 m and 150 m height ranges, respectively. The vertical accuracy of GLAS over flat surfaces has been estimated to be between 0 and 3.2 cm, on average, with a standard deviation lower than 3.3 cm [44].

There are 15 data products (GLA01 to GLA15) available from ICESat GLAS. In this study, only GLA01 (global altimetry data) and GLA14 (global land surface altimetry data) were used. GLA01 contains the raw waveform data, while GLA14 contains information on observation conditions and waveform parameters. The waveforms are decomposed into a maximum of six Gaussian distributions, and the distributions describe the vertical structure of the canopies within the footprints in the GLA14 product. Over flat terrain, it is generally assumed that the last Gaussian peak is the ground return, and the first peak represents reflections from the canopy top [44] (Figure 2).

To exclude unsuitable GLAS data in canopy height estimation, such as data affected by atmospheric conditions and clouds, several filters were applied. (1) Signals with high noise were removed using a signal-to-noise ratio higher than 20. This filter removed 15.6% of the data. (2) GLAS waveforms with delays from either saturation or atmospheric forward scattering were removed. This removed 14.8% of the remaining data. (3) Cloudless waveforms were selected using the cloud detection flag (FRir\_qaFlag = 15).

This filter was used to remove 28.8% of the remaining data. Saturated signals were identified using the GLAS flag (SatNdx > 0). (4) Waveforms with a centroid elevation that was significantly higher or lower than the corresponding SRTM elevation were removed ( $|\text{SRTM-GLAS}| > 100 \text{ m}$ ) [45]. This filter removed 2.5% of the remaining data. From the original database of 101,056 footprints, 46.9% satisfied the filters mentioned above, of which 374 intersected with the LD dataset (where the LD points are at a maximum distance of 50 m from the center of the GLAS's footprint). Finally, the GLAS data, referenced to TOPEX/Poseidon, were converted to WGS84 by subtracting 70 cm from the elevation values.

**Figure 2.** A typical GLAS waveform acquired over a vegetated area on a flat terrain.



### 3. Materials and Methods

#### 3.1. LiDAR Data Processing and Canopy Height Estimation

##### 3.1.1. Processing the LD Dataset

To estimate canopy heights using the airborne low-density LiDAR data, several steps were required. First, the dataset was filtered to remove erroneous elevation measurements. Next, the canopy-top and the ground points were extracted to estimate the canopy heights. The process for canopy height estimation is summarized in the following sections.

##### Data Filtering

Airborne LD LiDAR data showed local-scale fluctuations according to whether the point corresponded to a treetop, a branch at intermediate level, or even a stream or the ground. The analysis of LiDAR data showed important differences due to measurement errors in LiDAR elevations ( $Z$ ) between two neighboring points (a distance of 7 m in the LD dataset). Elevation differences up to 150 m were observed. LiDAR points with a difference in  $Z$  greater than 60 m were discarded. This threshold of 60 m was chosen considering the extreme case in which one laser point represents the top of a tree and its neighboring point reaches the ground, giving an approximate maximum canopy height of 60 m.



Canopy Top Identification

Next, airborne LiDAR data were filtered to select the points that most likely corresponded to canopy tops. This was achieved by selecting the local maximum in a sliding window of  $n$  points ( $n$  being odd numbers). In each window, the local maximum was selected as the point with the maximum amplitude with respect to the line segment joining the boundaries of each window (Figure 3a). The window size was selected so that the variogram of LiDAR elevations ( $Z$ ) no longer displayed an apparent nugget effect. Figure 3b shows that the nugget effect disappears when windows are larger than seven points and that a window of nine points (*i.e.*, 56 m) gives a nearly linear variogram. Windows of a larger size did not improve the results and tended to decrease the number of available points. With a nine-point window, more than a quarter of the filtered LD LiDAR points were conserved (a point every 42 m, on average, along the flight lines), making a total of 3,289,076 top-of-canopy points available over French Guiana (49.21 pts/km<sup>2</sup>).

**Figure 3.** (a) Points selected as top of canopy (local maximum); (b) Variogram of airborne LiDAR elevations from the LD dataset with local maximum points as a function of the size of the filtering window; (c) Canopy height calculation; (d) Ground points selected from a 1000-m window.

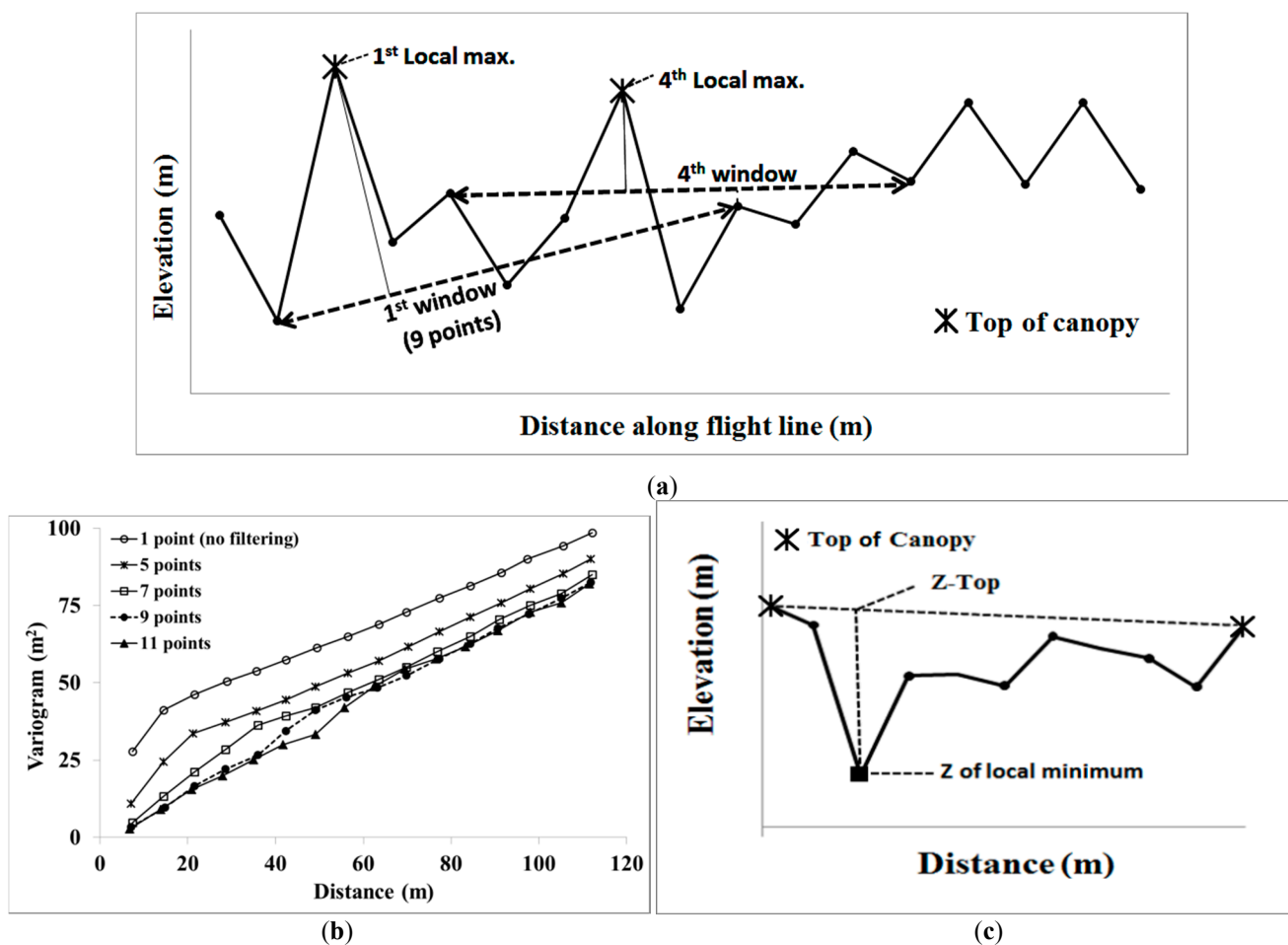
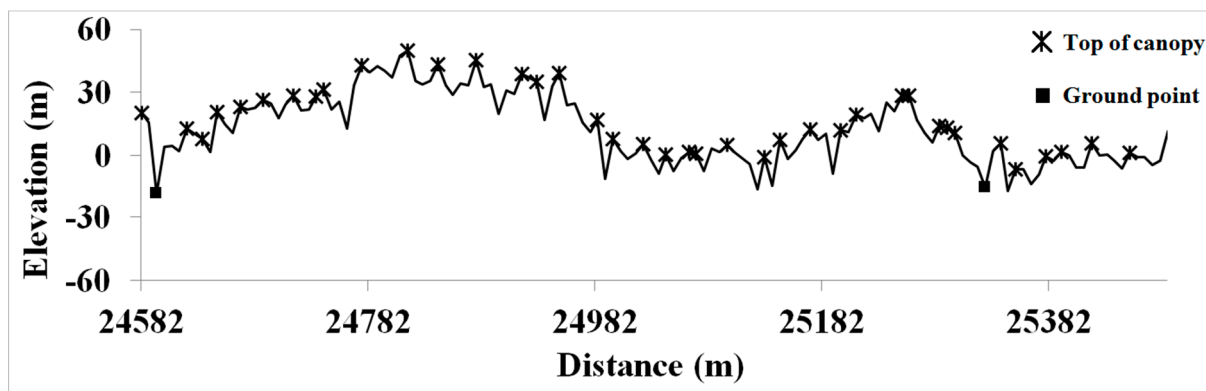


Figure 3. Cont.



(d)

### Identification of Ground Points

Few LiDAR returns reach the ground in tropical forests. Vincent *et al.* [43] estimated that, in last-return mode, only 1% of all laser returns are ground measurements. Bourguine *et al.* [42] estimated the ground returns in the LD dataset to be several hundred meters apart. To select the ground points from the LD dataset, the following procedure was attempted (e.g., [42,46]):

(1) Between two successive points identified as top of canopy, identify the local minimum, *i.e.*, the point that gives the maximum canopy height (Figure 3c). For all points situated between the two top-of-canopy points, the canopy height is calculated as the difference between the elevation of each point ( $Z$ ) and the top-of-canopy elevation ( $Z_{TOP}$ ).  $Z_{TOP}$  is obtained using a linear interpolation between the elevations of two canopy tops.

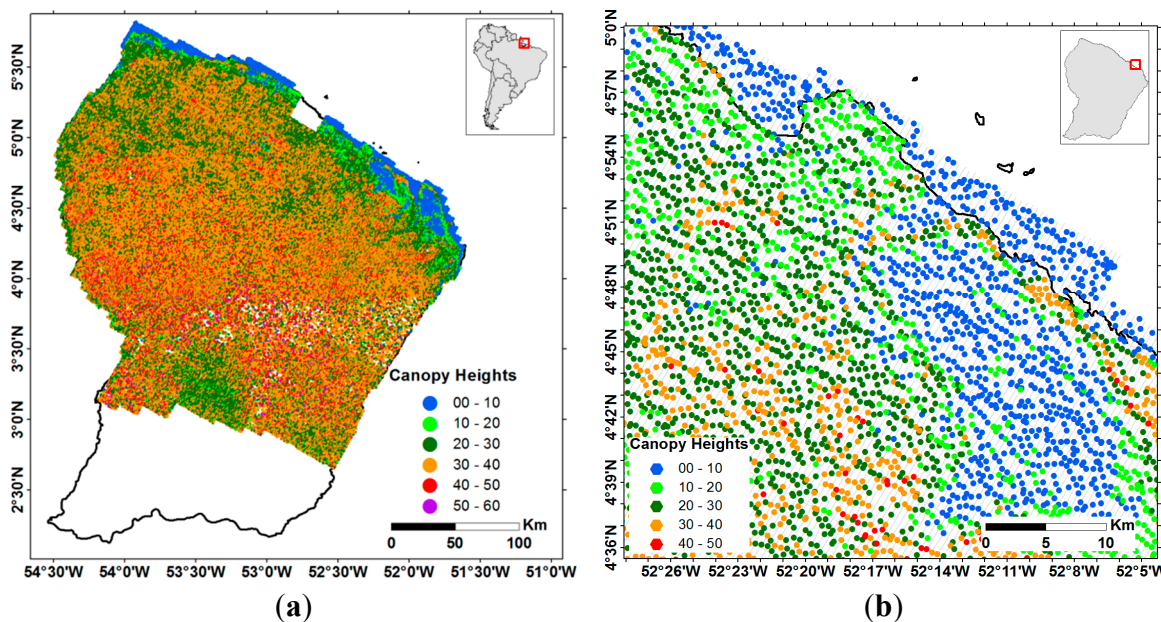
(2) Among the local minimum points selected in the previous step, retain the lowest one inside a non-overlapping moving window (point corresponding to the greatest canopy height) (Figure 3d). With the use of a small window size, the selected ground points are often located above the ground, leading to an underestimation of the canopy height. For too-large windows, too many ground points are eliminated, leading to an excessive smoothing of the estimated canopy height during the subsequent interpolation. Bourguine *et al.* [42] demonstrated that the best window size for this LD dataset is 1000 m. The number of ground points available for French Guiana is 105,438 (1.59 pts/km<sup>2</sup>).

### Canopy Height Estimation

Canopy heights were calculated for the LD dataset using points identified as top of canopy and ground (Figure 3d). The estimation of the canopy height was performed at the level of the 105,438 ground points using linear interpolation between the elevations of the top-of-canopy points (spaced 42 m apart, on average). Canopy height estimation cannot be conducted at the canopy-top level by interpolating the ground points because the distance between ground points (1000 m, on average) is too great to assume a linear trend between the elevations of ground points.

The estimation of canopy height using the LD dataset showed that canopy heights reached a maximum of 69 m with a mean height of approximately 30.4 m. The lower canopy heights (maximum of 20 m) were observed in the coastal marsh areas, situated in the northeastern part of French Guiana (Figure 4).

**Figure 4.** Map of canopy heights calculated from the airborne LiDAR dataset LD for (a) French Guiana and (b) a portion of the coastal marsh. Only 1% of canopy heights were higher than 50 m in all of French Guiana.



### 3.1.2. Processing the HD Dataset

The estimation of canopy height from the airborne high-density (HD) dataset used a similar procedure. However, as the density of points is higher (on average 3.5 pts/m<sup>2</sup>) than that of the LD dataset (on average 285.2 pts/km<sup>2</sup>), several changes were made to account for the difference between the two datasets:

(1) The procedure described in Section 3.1.1 requires flight lines for top-of-canopy and ground point extraction. From the HD dataset, a grid of 1 m × 1 m was created over the study sites. Then, two datasets were created: the first contained the point with the highest elevation in each square of the grid, and the second contained the lowest elevations.

(2) Using the grid of the highest elevations, the procedure developed in Section 3.1.1 for canopy top extraction was applied to extract canopy-top points along the East-West and North-South directions. The window size for the canopy top extraction differed between datasets according to their point density (between 20 and 50 m).

(3) Using the lowest elevations grid, the ground point's extraction procedure detailed in Section 3.1.1 was performed along the horizontal and vertical lines of the grid. However, unlike with the LD dataset, the window sizes used in the selection of ground points were much smaller (between 70 and 120 m, according to the HD dataset). The window sizes of the HD dataset were also determined using an analysis of variograms.

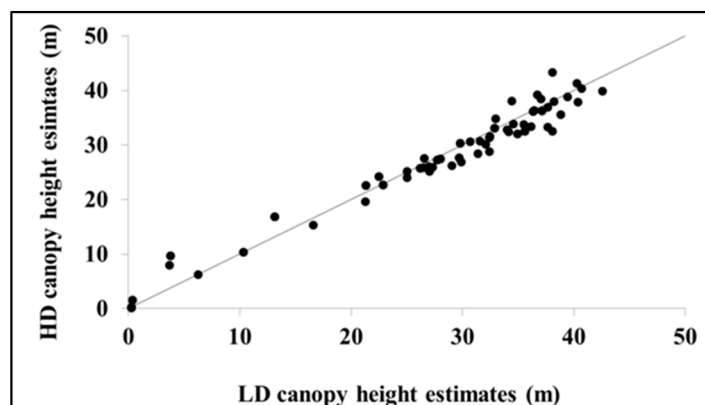
(4) Finally, as the distances between ground points and between canopy-top points were small, the estimation of canopy height was calculated at each canopy-top and ground point. However, unlike the LD dataset, the canopy heights were not estimated using linear interpolation but rather using bilinear interpolation. First, Delaunay triangulations were computed separately for the canopy-top and the ground points. Next, the triangle containing each ground point in the lat/lon plane of the top-of-canopy mesh

was identified, and the ground point was projected on this triangle. Finally, the canopy height was calculated as the difference between the elevation of the projected ground point on the top-of-canopy mesh and the elevation of the actual ground point. A similar procedure was carried out for canopy height estimation at each canopy-top point using the projection of canopy-top points on the Delaunay triangles of the ground points' mesh.

### 3.1.3. Comparison of Canopy Height Estimates from the LD and HD Datasets

The canopy height estimates from the HD dataset are considered near-terrain measurements because of their small footprint size and high density. Unfortunately, the HD dataset does not intersect with the GLAS footprints. To use the LD dataset as reference data for GLAS's canopy height estimation models, the accuracy of the canopy heights of the LD dataset was assessed against the estimates from the HD dataset. For each LD canopy height estimate, the nearest point from the HD dataset, at a maximum distance of 10 m, was chosen. The results of the comparison between canopy heights from the LD and HD datasets showed a mean difference of 0.22 m, an RMSE of 1.57 m, and an  $R^2$  of 93% (Figure 5).

**Figure 5.** Comparison between canopy height estimates from the LD and HD datasets.



## 3.2. GLAS Data Processing

### 3.2.1. GLAS Waveform Metrics Extraction

Several canopy height estimation models from GLAS waveforms have been developed in recent years (e.g., [10,20,40,45,47,48]). They depend on several parameters extracted from waveforms (primarily signal start and end, waveform extent, and leading and trailing edges) and on ancillary data, such as DEMs (slope or terrain index).

Signal start and end are defined as the first and last locations where the waveform intensity exceeds a certain threshold level ( $n \cdot \sigma_b$ ,  $\sigma_b$  is the standard deviation of the background noise) above the mean background noise ( $\mu_b$ ) (Figure 2) [10]. Both  $\mu_b$  and  $\sigma_b$  are found in the GLA14 product. The difference between the signal end and signal start is called the waveform extent. However, there are no consistent optimal thresholds that can be used for every study area. Different thresholds have been used in different studies, including  $3\sigma_b$  [49],  $3.5\sigma_b$  [47],  $4\sigma_b$  [10] and  $4.5\sigma_b$  [48]. The difficulty in identifying the noise threshold could be explained by the difficulty in consistently identifying signal start and signal end.

The Gaussian peaks resulting from the decomposition of the GLAS waveform represent canopy features, such as canopy top, canopy trunks, ground or a mix of these elements. The last Gaussian peak does not necessarily represent the ground return. Moreover, there is no general rule to determine the ground peak (e.g., [40,47,49,50]). Duong *et al.* [50] and Sun *et al.* [49] identified the ground as the last peak. Rosette *et al.* [47] and Chen [40] found that the elevation of the stronger of the last two Gaussian peaks has a better correspondence to the ground. In this study, the stronger of the last two Gaussian peaks was selected as the ground return.

The leading edge is defined as the difference between signal start and the first bin that is at half the maximum intensity (Figure 2). The trailing edge corresponds to the difference between signal end and the last bin that is at half the maximum intensity [48] (Figure 2). However, some LiDAR waveforms have a large difference in the intensity between the canopy and the ground peaks. If the ground peak return is significantly lower than the canopy peak, an overestimation of the trailing edge could be observed using Lefsky's metrics. Conversely, with a low intensity return from the canopy peak and a high intensity return from the ground peak, an overestimation of the leading edge could be observed using Lefsky's metrics [10]. Hence, Hilbert and Schullius [24] proposed modified leading edge and trailing edge definitions. The modified leading edge is defined as the elevation difference between signal start and the canopy peak's center, and the modified trailing edge is the difference between signal end and the ground peak's center (Figure 2). These modified metrics better represent the characteristics of the canopy top and the ground surface. This study used the modified leading and trailing edges.

### 3.2.2. Principal Component Analysis of GLAS Waveforms

Principal Component Analysis (PCA) of LiDAR waveforms has been conducted in a handful of studies. Allouis *et al.* [51] used PCA to estimate the water depth in shallow water using airborne LiDAR waveforms. Principal components were then used to perform a regression model between the principal components and water depth. The model relying on PCA for water depth estimation provided the lowest mean error and had the lowest detectable water depth in comparison to other models (mathematical approximation, Heuristic methods, statistical approaches, and convolution methods). However, to convert waveform samples into principal components, further processing of the GLAS waveforms was required. First, the parts of the waveforms useful for canopy height estimation corresponding to the waveform extent were extracted. Next, because not all the waveforms have the same waveform extent, the waveform with the largest extent was identified, and waveforms with shorter waveform extents were padded with the remaining waveform samples after the signal end to give them the same length as the largest waveform extent (same sample count). Note that the first sample of the extracted waveform now corresponds to signal start. In this study, the largest waveform extent had 470 samples. Next, the extracted waveform samples were converted into principal components (PCs), and the number of PCs to be used in the regression model for dominant canopy height estimation (Hmax) was calculated. The number of PCs used in the regression model has a major impact on the performance of the model, as choosing too many PCs will include noise from the sampling fluctuations in the analysis and by choosing too few, relevant information will be lost. A vast literature has developed methods to choose the statistically significant PCs. In this study, the number of statistically significant PCs was determined

using a statistical process based on the study by Karlis *et al.* [52]. The PCs with eigenvalues higher than a certain threshold were selected. The threshold ( $\lambda$ ) was defined as follows:

$$\lambda = 1 + 2 \sqrt{\frac{p-1}{n-1}} \quad (1)$$

where  $p$  is the number of variables (PCs) and  $n$  is the number of observations (waveforms). For our dataset composed of 470 variables and 474 observations, the threshold ( $\lambda$ ) was determined at 2.99. Thus, the first 13 PCs were selected.

### 3.3. Background on GLAS Canopy Height Estimation

#### 3.3.1. Direct Method

The estimation of the canopy height using the direct method is simply the difference between the waveform signal start (canopy top) ( $H_b$ ) and the ground peak ( $H_g$ ):

$$H_{\max} = H_b - H_g \quad (2)$$

The direct method estimates the canopy height with good precision over flat areas. An average difference between GLAS and airborne LiDAR data lower than 3 m was observed in several studies (e.g., [45,53]).

#### 3.3.2. Multiple Regression Models Using GLAS and DEM Metrics

Over sloping areas, both the ground and vegetation peaks are broader and lower in intensity (e.g., [25,26]). The peak identified as the ground peak will no longer represent only the ground but a mix of ground and terrain objects (e.g., [40,54]). In fact, over sloped terrain, waveform extent will increase with the terrain slope and the footprint size [54]. This increase will lead to an earlier detection of the signal start and this will lead to an overestimation of the canopy height [55].

To correct for the effect of terrain slope on the GLAS signal, several studies have developed models to better estimate canopy height. Lefsky *et al.* [48], Pang *et al.* [26], Duncanson *et al.* [55], and Chen [40] developed models based on parameters derived from the waveforms themselves (waveform extent “Wext”, leading edge “Lead” and trailing edge “Trail”). Lefsky *et al.* [10] and Rosette *et al.* [47] developed models based on the waveform extent and terrain index. The terrain index as defined by Lefsky *et al.* [10] is the difference between the maximum and minimum elevations in an  $m \times m$  sampling window applied to a DEM at the GLAS footprint location. The window size depends on the resolution of the DEM. A  $3 \times 3$  window has been deemed best for a 90-m-resolution DEM [10].

The first model was developed by Lefsky *et al.* [10] for the estimation of the tallest canopy within a footprint:

$$H_{\max} = aW_{\text{ext}} - b \cdot \text{TI} \quad (3)$$

This model is based on the waveform extent (Wext) and the terrain index (TI). The incorporation by Lefsky *et al.* [10] of the waveform leading edge extent in Equation (3) resulted in a slight improvement in the canopy height estimation:

$$H_{\max} = aW_{\text{ext}} - b \cdot \text{TI} + c\text{Lead} \quad (4)$$

Pang *et al.* [26] introduced a model to estimate forest canopy height by using metrics derived from the waveforms themselves:

$$H_{\max} = aW_{\text{ext}} - (b(\text{Lead} + \text{Trail}))^c \quad (5)$$

Chen [40] introduced the following model to show how a linear model compares to Equation (5):

$$H_{\max} = aW_{\text{ext}} - b(\text{Lead} + \text{Trail}) \quad (6)$$

Finally, Lefsky *et al.* [20] proposed a modification of the Lefsky *et al.* [48] model to produce a better estimation when the leading and trailing edges are small:

$$H_{\max} = aW_{\text{ext}} - b\text{Lead} - c\text{Trail} \quad (7)$$

In addition, to quantify the contribution of Lead and Trail in the canopy height estimation models, two additional models were analyzed: one that replaces Lead with Trail in Equation (4) and one that removes Lead in Equation (6) (model IDs 7 and 8, respectively, Table 2). Finally, each of the eight models was tested with an added intercept (the bis models, Table 2). The best regression model was selected based on the Akaike information criterion (AIC) [56], the coefficient of determination ( $R^2$ ), and the root mean square error (RMSE). Finally, to assess how the model results will generalize to an independent data set, a 10-fold cross validation was used. Large k-fold values mean less bias towards overestimating the true expected error (as training folds will be closer to the total dataset).

The coefficients used in different models were fitted with least squares regressions using the canopy height estimates from the LD dataset. The corresponding LD canopy height estimate for each GLAS footprint was chosen as the closest point (no farther than 50 m).

**Table 2.** Regression models' fitting statistics calculated with 10-fold cross validation for estimating forest height. R = root mean square error, AIC = Akaike information criterion.

Model	ID	R <sup>2</sup>	RMSE (m)	AIC
$H_{\max} = H_b - H_g$	1	0.50	7.9	3126
$H_{\max} = 0.6527W_{\text{ext}} - 0.0184TI$	2	0.72	4.9	2221
$H_{\max} = 0.5405W_{\text{ext}} - 0.0262TI + 6.427$	2bis	0.73	4.4	2185
$H_{\max} = 0.6682W_{\text{ext}} - 0.0029TI - 0.0261\text{Lead}$	3	0.73	4.7	2223
$H_{\max} = 0.5395W_{\text{ext}} - 0.2557TI - 0.0115\text{Lead} + 6.8876$	3bis	0.73	4.6	2187
$H_{\max} = 0.7555W_{\text{ext}} - \{0.0994(\text{Lead} + \text{Trail})\}^{1.5903}$	4	0.80	3.9	2084
$H_{\max} = 0.6908W_{\text{ext}} - \{0.1315(\text{Lead} + \text{Trail})\}^{1.3109} + 3.3309$	4bis	0.80	3.9	2081
$H_{\max} = 0.7965W_{\text{ext}} - 0.2707(\text{Lead} + \text{Trail})$	5	0.79	3.9	2096
$H_{\max} = 0.6972W_{\text{ext}} - 0.2461(\text{Lead} + \text{Trail}) + 4.1452$	5bis	0.79	3.9	2083
$H_{\max} = 0.6739W_{\text{ext}} - 0.0751\text{Lead} - 0.2959\text{Trail}$	6	0.85	4.0	2064
$H_{\max} = 0.6739W_{\text{ext}} - 0.0751\text{Lead} - 0.2959\text{Trail} + 4.1823$	6bis	0.85	3.9	2056
$H_{\max} = 0.7377W_{\text{ext}} + 0.0235TI - 0.3192\text{Trail}$	7	0.81	3.8	2063
$H_{\max} = 0.6656W_{\text{ext}} - 0.0026TI - 0.28899\text{Trail} + 3.679$	7bis	0.81	3.7	2051
$H_{\max} = 0.7494W_{\text{ext}} - 0.3184\text{Trail}$	8	0.81	3.8	2064
$H_{\max} = 0.6654W_{\text{ext}} - 0.2904\text{Trail} + 3.6344$	8bis	0.81	3.8	2056
$H_{\max} = a_1PC_1 + a_2PC_2 + \dots + a_{13}PC_{13}$	9	0.52	5.9	2373
Most important PCs (PC1, PC2, PC4, PC11) from ID 9	9bis	0.47	6.2	2478
$H_{\max} = W_{\text{ext}} + a_1PC_1 + a_2PC_2 \dots + a_{13}PC_{13}$	10	0.80	3.8	2047

Table 2. Cont.

Model	ID	R <sup>2</sup>	RMSE (m)	AIC
<i>Most important PCs (PC1, PC2, PC4, PC11) from ID 10</i>	10bis	0.79	3.9	2075
$H_{max} = 0.63W_{ext} - 0.10PC_1 + 0.05PC_2 + 0.02PC_3 + 1.3$	11	0.73	4.4	2174
$H_{max} = WC_i + a_1PC_1 + a_2PC_2 \dots + a_{13}PC_{13}$	12	0.78	4.0	2064
<i>Random Forest using: Wext + Lead + Trail + TI</i>	13	0.82	3.4	-
<i>Random Forest using: Wext + Lead + TI</i>	14	0.80	3.6	-
<i>Random Forest using: Wext + Lead</i>	15	0.80	3.6	-
<i>Random Forest using: Wext + TI</i>	16	0.82	3.6	-
<i>Random Forest using: Wext</i>	17	0.73	4.4	-
<i>Random Forest using: First 13 PC</i>	18	0.70	4.7	-
<i>Random Forest using: PC1 + PC2 + PC4 + PC11</i>	18bis	0.69	4.8	-
<i>Random Forest using: Wext and the first 13 PC</i>	19	0.83	3.6	-
<i>Random Forest using: Wext + PC1 + PC2 + PC4 + PC11</i>	19bis	0.82	3.6	-
<i>Random Forest using: WC and the first 13 PC</i>	20	0.81	3.7	-
<i>Random Forest using: WC + PC1 + PC2 + PC4 + PC11</i>	20bis	0.81	3.7	-

### 3.4. Proposed Techniques for Canopy Height Estimation

#### 3.4.1. Multiple Regression Models Using Principal Components

The previous section introduced a number of regression models developed in various studies for the estimation of canopy height. However, these models require several metrics derived from GLAS footprints, such as ground peak, canopy-top peak, leading and trailing edge extents, and metrics derived from ancillary data (SRTM DEM), such as terrain index. Moreover, the extraction of some metrics from GLAS waveforms, such as the location of the ground peak, is error-prone, especially in dense forests, such as those in French Guiana. Processing the GLAS data revealed that a considerable number of waveforms taken only over dense forests had the canopy-top location easily identified. In fact, canopy penetration of the waveform in densely vegetated areas was sometimes insufficient to reach the ground; thus, either the ground peak was unidentifiable or the waveform did in fact reach the ground but the return signal was not strong enough for reliable detection. These difficulties in the detection of the ground peak affect the estimation of the trailing edge extent and, ultimately, the estimation of the canopy height. Therefore, a statistical model for canopy height estimation based only on the waveform samples might be an interesting alternative. In this section, a principal component analysis of GLAS waveforms was conducted. A stepwise linear regression model was built for canopy height estimation using the principal components (PCs). A regression model using PCs takes advantage of model building on orthogonal variables. The regression model using the 13 first PCs for canopy height estimation could be written as follows:

$$H_{max} = a_1PC_1 + a_2PC_2 + \dots + a_{13}PC_{13} \quad (8)$$

where  $PC_i$  are the principal components, and  $a_i$  are the coefficients to be applied to the principal components.

This model based on principal component analysis for canopy height estimation will be compared to the regression models developed in the previous section to quantify the benefits of using waveform data (PCA model) instead of metrics extracted from the waveform.



### 3.4.2. Random Forest Regressions Using GLAS and DEM Metrics

In Section 3.3.2, linear regressions were developed to estimate the canopy height for each GLAS footprint. These regressions linked the canopy height estimated from the LD data to the GLAS and SRTM metrics (waveform extent, leading edge, trailing edge, and terrain index). In this section, the Random Forest (RF) technique was evaluated using the following different configurations:

(1) All the metrics were used to estimate the canopy height (waveform extent “Wext”, leading edge extent “Lead”, trailing edge extent “Trail”, and terrain index “TI”);

(2) The Trail metric was removed because in densely forested areas, such as tropical forests, the LiDAR echo seldom reaches the ground, making the ground peak difficult to identify; thus, the Trail metric is often inaccurate;

(3) To study the effects of Trail and TI on the canopy height estimates, the TI and Trail metrics were removed (only the Wext and Lead were used). This case shows promise in the use of the SRTM DEM in a low relief area;

(4) Only Wext and TI were used to assess the impact of the Lead and Trail metrics on the performance of Random Forest for canopy height estimates;

(5) Only Wext was used. This case evaluated the impact of using Lead, Trail, and TI with Wext for canopy height estimates. The relative importance of the different metrics used in Random Forest for the canopy height estimates was also analyzed. Variable importance is based on two measures. The first is a measure of accuracy obtained by quantifying the mean squared error increase in the model by the removal of a variable. The other importance measure is the Gini index, which quantifies the degree to which a variable produces terminal nodes in the classification forest. Finally, to validate the generalization performance of the Random Forest regressions, the error in the estimation of the canopy height was assessed using a 10-fold cross validation. The performance of the different configurations was assessed by comparing the canopy height estimates from Random Forest regressions and the canopy heights extracted from the LD dataset, which were used as the reference data.

Several studies have shown that, for many applications, the Random Forest technique is extremely powerful in estimating biophysical parameters (e.g., [57–60]). Random Forest can be used as a classifier or a regression algorithm consisting of an ensemble of regression or classification trees [61]. Each tree is grown with a randomized set of explanatory variables. For the regressions by Random Forest, the estimates are produced by averaging the results produced by each tree.

### 3.4.3. Random Forest Regressions Using Principal Components

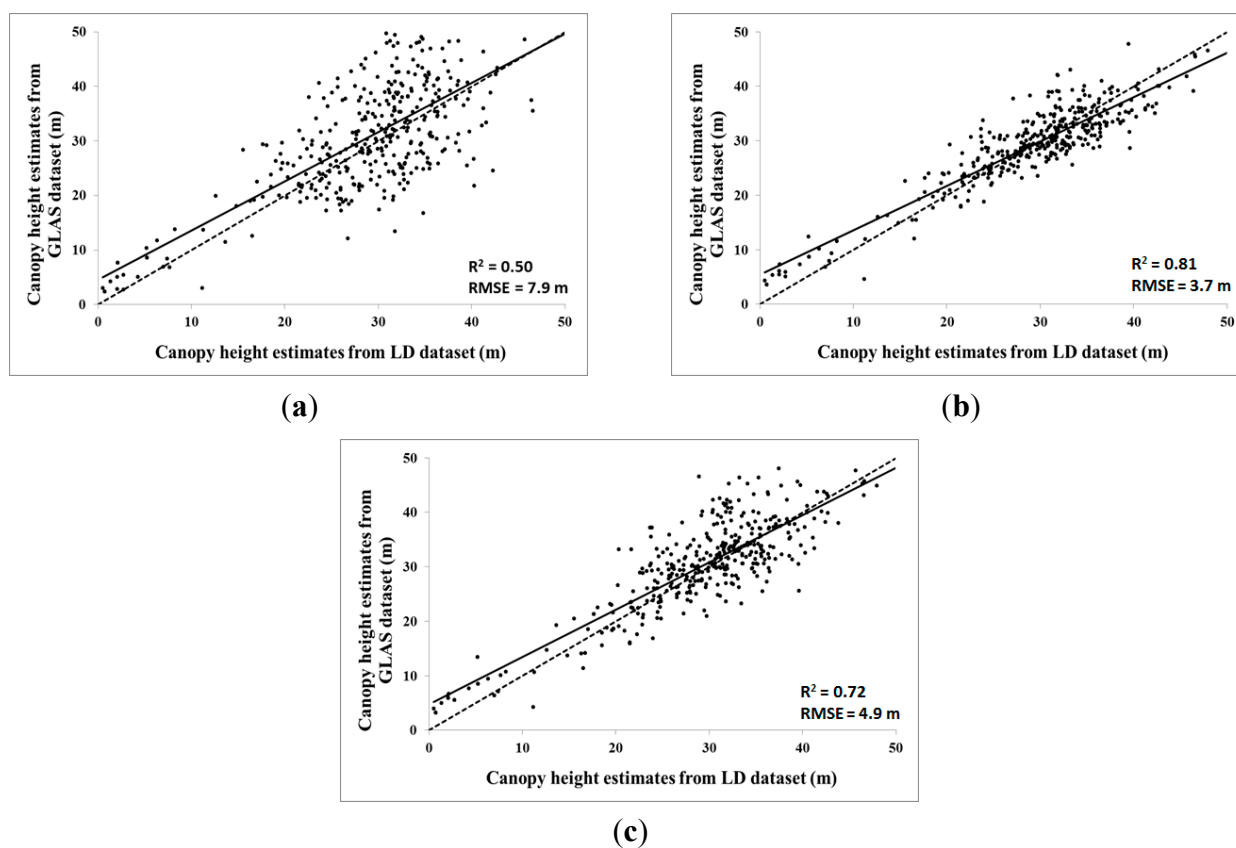
Similar to the previous section, the first 13 principal components were used in the Random Forest regression to link the canopy heights estimated from LD data to these PCs. This model based on principal component analysis and Random Forest regressions was compared to other the models performed in this study.

## 4. Results

### 4.1. Direct Method

The comparison between the canopy height estimates from GLAS waveforms using the direct method and the canopy height estimates from the LD dataset showed a high RMSE of 7.9 m for the estimation of the GLAS canopy height and a low  $R^2$  of 0.50 (Figure 6a). This result can be explained by the fact that most of the footprints were in an area with a slope between  $5^\circ$  and  $10^\circ$ .

**Figure 6.** Canopy height estimates from GLAS data in comparison to estimated canopy heights from the LD dataset: (a) using the direct method (model ID 1, Table 2), (b) using the model with Wext, TI and Trail (model ID 7bis, Table 2), and (c) using the model with Wext and TI (model ID 2, Table 2).



### 4.2. Multiple Regression Models

#### 4.2.1. Using GLAS and DEM Metrics

The results of the regression models with 10-fold cross validation showed that the regression models using the trailing edge extent (model IDs 4 to 8, Table 2) provided slightly better estimations of canopy height. For these models, AIC ranged between 2051 and 2096, RMSE ranged between 3.7 and 4.0 m, and  $R^2$  between 0.79 and 0.81. The best results in estimating forest height were obtained with model ID 7bis (Table 2, Figure 6b). The contribution of the leading edge extent appeared to be weak in comparison to the trailing edge extent when estimating the maximum canopy height. Indeed, model IDs 7 and 7bis,

which used Trail, had better results than model IDs 3 and 3bis (Table 2), which used Lead. Moreover, the use of information calculated from a DEM (terrain index) alone in the regression models had the lowest estimation accuracy for the canopy height (model ID 2, Table 2, Figure 6c) (RMSE of 4.9 m and  $R^2$  of 0.72).

#### 4.2.2. Using Principal Components

The results of the PCA model for canopy height estimation showed an estimation accuracy with an  $R^2$  of 0.52 and an RMSE of 5.9 m (Figure 7a). To reduce the number of PCs involved in the PCA model, stepwise regression was used to extract the most important principal components. The resulting model, which used 6 principal components containing 76.3% of the waveforms' inertia, showed an  $R^2$  of 0.47 and an RMSE of 6.2 m.

**Figure 7.** Comparison between canopy height estimates using the PCA regression models and those estimated from low-density airborne LiDAR data (LD) (a) using the first 13 PCs, (b) using the first 13 PCs with the waveform extent, and (c) using the first three PCs with the waveform extent.

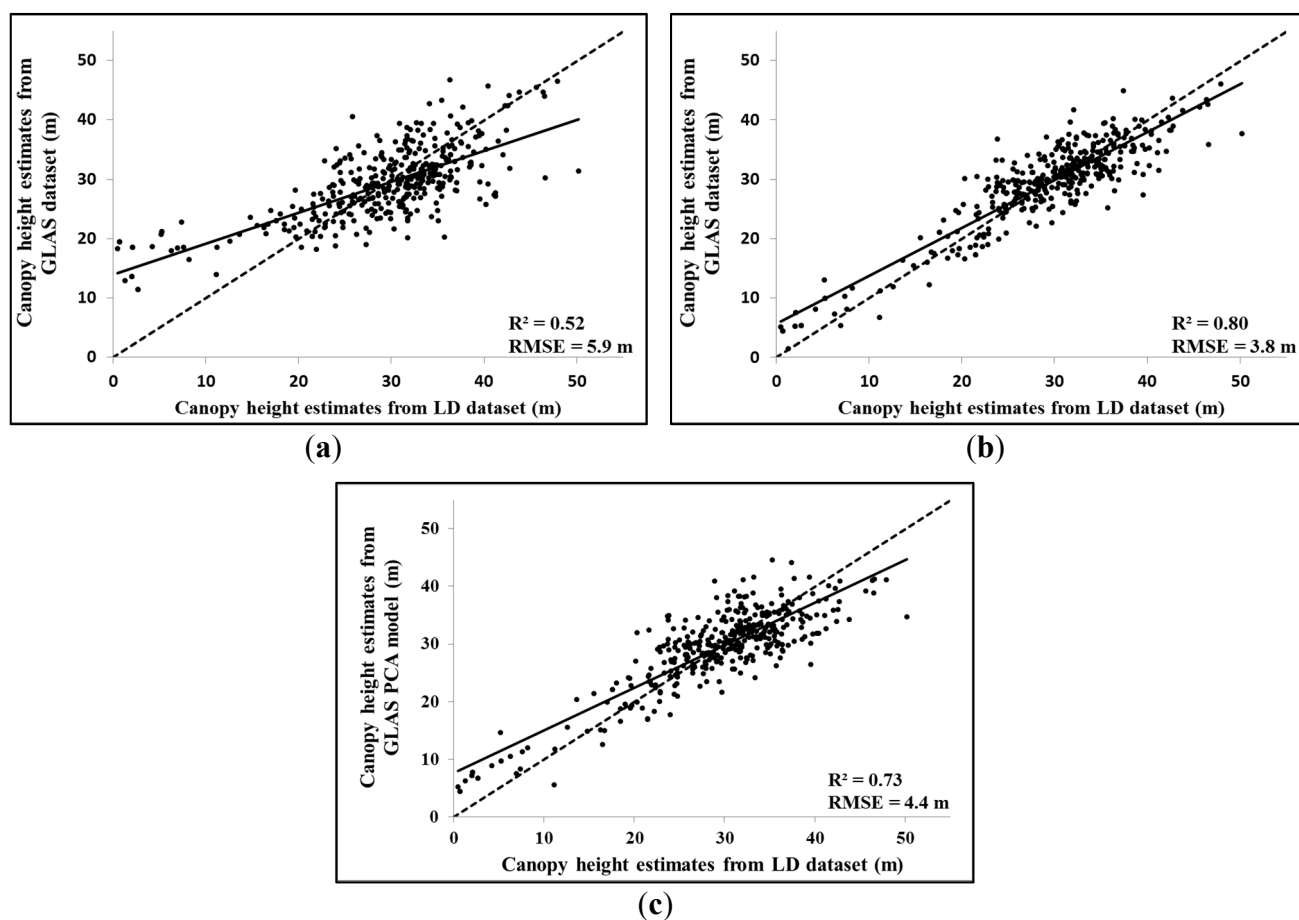


Figure 7a shows that the PCA model appeared to overestimate canopy heights for canopies with heights lower than 20 m. To improve height estimation of these canopies, a regression model incorporating both the first 13 principal components and the waveform extent was performed:

$$H_{max} = W_{ext} + a_1PC_1 + a_2PC_2 + \dots + a_{13}PC_{13} \quad (9)$$

The new PCA regression model for canopy height estimation accounting for the waveform extent showed better canopy height estimation results in comparison to the PCA model without information on the waveform extent, with an RMSE of 3.8 m and an  $R^2$  of 0.80 (Figure 7b). Using only the seven most important components from the stepwise regression, the  $R^2$  decreased to 0.79 and the RMSE increased to 3.9 m. Furthermore, using only the first three principal components with the waveform extent, the  $R^2$  decreased to 0.73, and the RMSE increased to 4.4 m.

Next, the waveform extent was replaced by a waveform extent factor class (WC): (1)  $WC_1$  for waveform extents lower than 20 m, (2)  $WC_2$  for waveform extents between 20 and 40 m, and (3)  $WC_3$  for waveform extents higher than 40 m. The resulting regression model using all the principal components and the WC has the following form:

$$H_{max} = WC_i + a_1PC_1 + \dots + a_{13}PC_{13} \quad (10)$$

where  $WC_i$  is the intercept to be applied to the model depending on the waveform extent ( $i = 1, 2, \text{ or } 3$ ). The values of  $WC_i$  are 7.78, 25.83 and 32.01 for  $WC_1$ ,  $WC_2$ , and  $WC_3$ , respectively.

The new PCA regression model for canopy height estimation with information on the waveform extent showed slightly less canopy height estimation accuracy in comparison to model ID 9 (Table 2), with an RMSE of 4.0 m and an  $R^2$  of 0.78 (Figure 7c).

Like previously, stepwise regression was used to extract the most important PCs. The resulting model using using PCs and containing 76.3% of the waveforms' inertia showed slightly lower performance in comparison to the PCA model that used all the PCs and the WC factor, with an RMSE of 4.2 m and an  $R^2$  of 0.76. Figure 8 shows the canopy height estimates from the LD and GLAS datasets. Good agreement was observed between the two canopy height maps.

**Figure 8.** (a) Map of canopy heights estimated from the LD dataset; (b) Map of canopy heights estimated from the GLAS dataset using the PCA model; (c) Overlapping of the two maps over a small area of French Guiana.

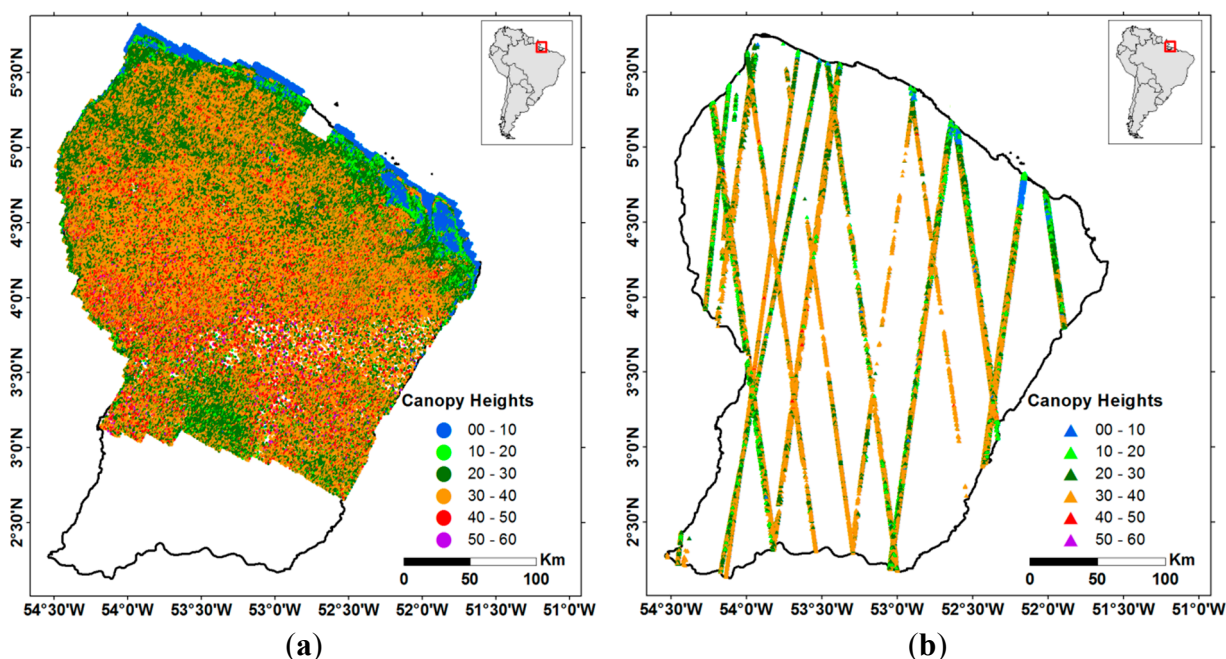
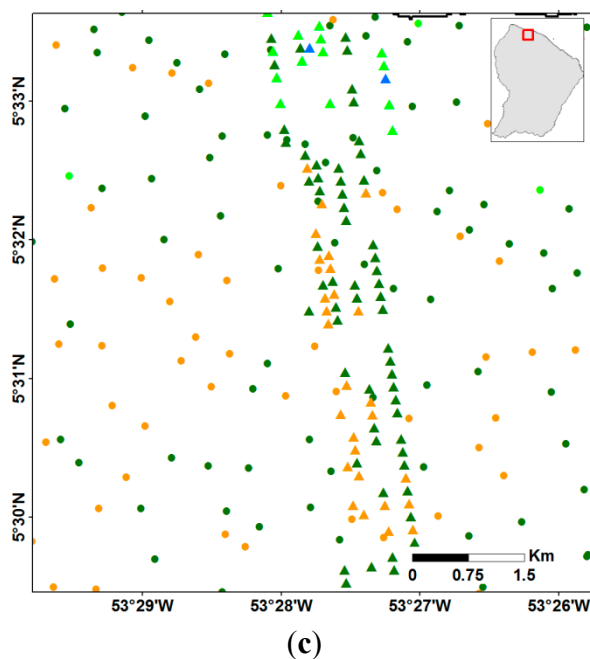


Figure 8. Cont.

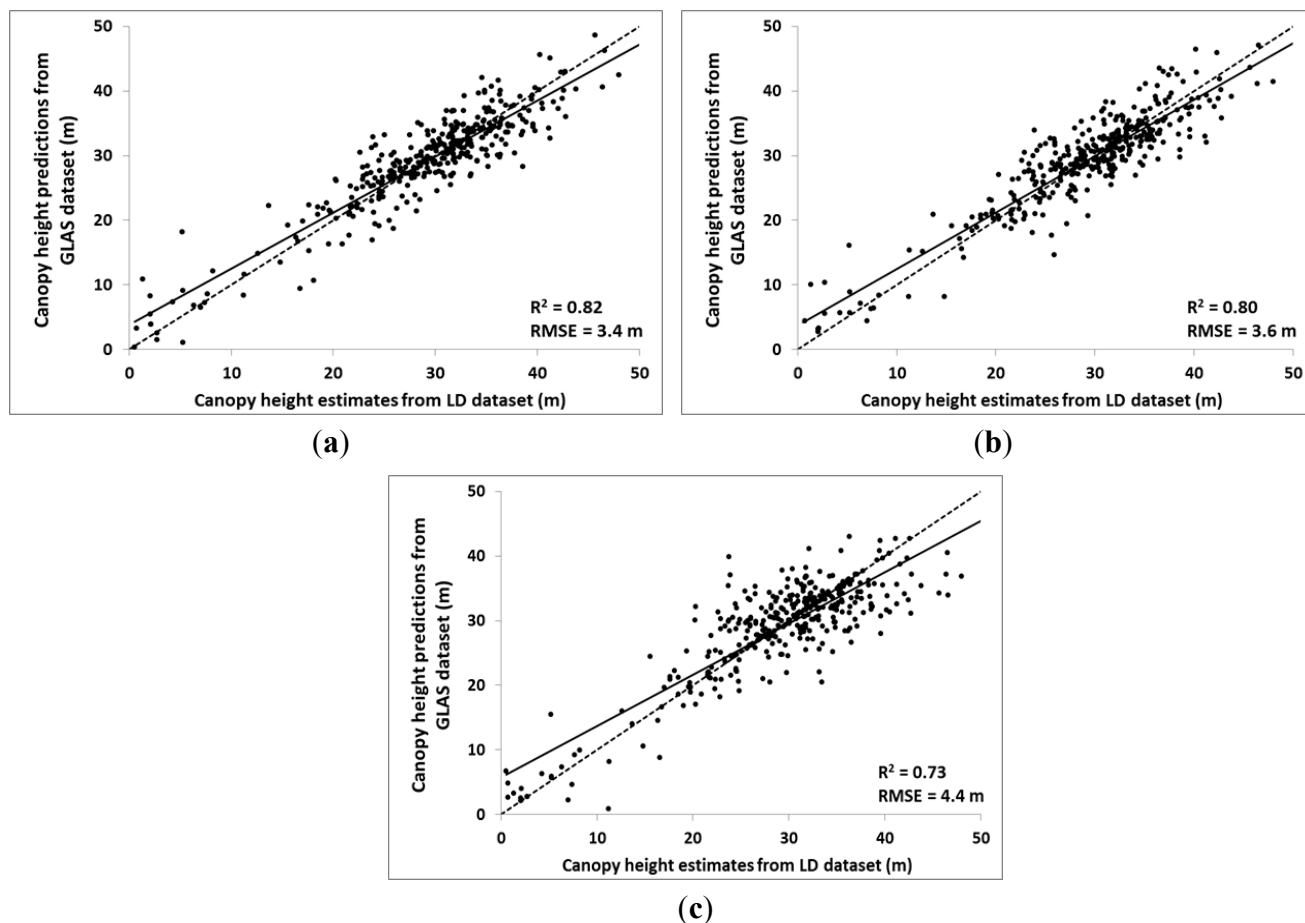


### 4.3. Random Forest Regressions

#### 4.3.1. Using GLAS and DEM Metrics

To analyze the precision of the canopy height estimation using Random Forest, several configurations were tested, and the results reveal that the best configuration for canopy height estimation is the one that uses all the metrics: waveform extent, leading edge, trailing edge, and terrain index (model ID 13, Table 2). The difference between the GLAS canopy height estimates and those estimated from the LD (reference data) in the first configuration had an RMSE of 3.4 m and a coefficient of determination  $R^2$  of 0.82. Moreover, the variable importance test of the metrics showed that the GLAS canopy height is best explained using Wext, with an importance factor almost three times higher than those for the other three metrics; meanwhile, the other metrics (Trail, Lead, and TI) had almost the same importance. Other configurations using Wext, Lead, and TI; Wext and Lead; and Wext and TI (model IDs 10, 11 and 12, respectively, Table 2) showed a slightly lower precision in the canopy height estimation (RMSE) (approximately 3.6 m). The estimation of the GLAS canopy height using only Wext had an RMSE of 4.4 m with an  $R^2$  of 0.73. These results show that, in a low relief area, the use of other metrics in addition to the waveform extent only slightly improved the precision of the estimation of canopy height regardless of which metric was used. The use of one metric (among Trail, Lead and TI) in addition to Wext improved the estimation of canopy heights by approximately 1 m. Moreover, the use of more than one of these metrics in addition to Wext did not improve the estimation of canopy heights. Figure 9 shows examples of the comparison between the GLAS canopy height estimates and the reference canopy heights estimated from the LD dataset.

**Figure 9.** Comparison of estimated canopy heights using Random Forest regressions and estimated canopy heights from the LD dataset for three metrics configurations: (a) Wext + Lead + Trail + TI; (b) Wext + Lead + TI; and (c) Wext.



#### 4.3.2. Using Principal Components

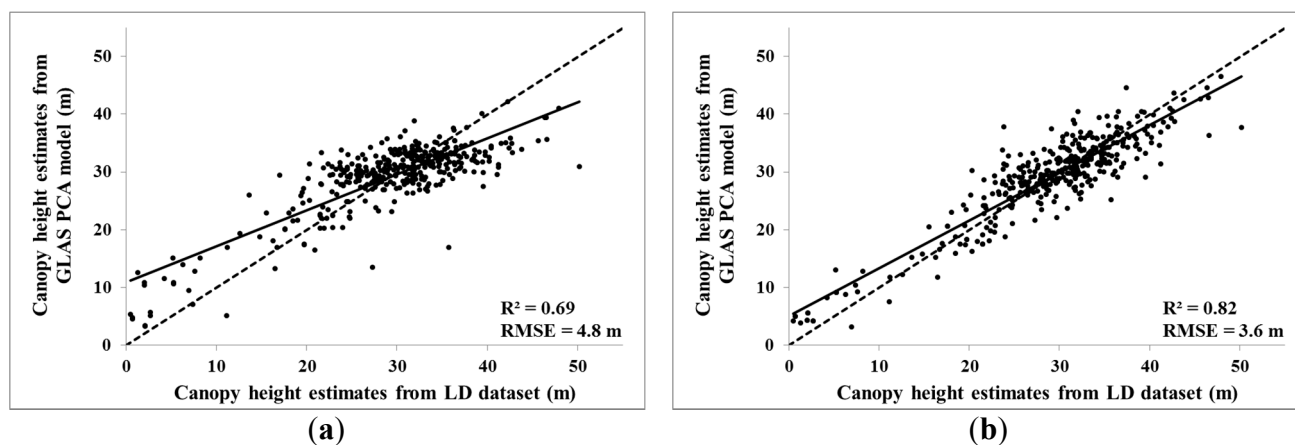
In this section, canopy height estimations with Random Forest regressions using PCs were performed with different configurations. Using the first 13 PCs in the Random Forest regression resulted in better canopy height estimation precision (RMSE = 4.7 m,  $R^2 = 0.7$ ) in comparison to the linear regression model that used the first 13 PCs in Section 4.2.2 (RMSE = 5.9 m,  $R^2 = 0.52$ ). The variable importance test showed that GLAS canopy height is best explained using *PC1*, *PC2*, *PC4*, and *PC11* (variance 62.38%). Using only these four PCs in the Random Forest model had a similar result (RMSE = 4.8 m,  $R^2 = 0.69$ ). Next, the incorporation of the waveform extent in addition to the first 13 principal components greatly improved the precision of the canopy height estimation (RMSE = 3.6 m,  $R^2 = 0.83$ ) in comparison to the RF regressions without Wext. In addition, this result is slightly better to the one obtained using a linear multiple regression with the first 13 PCs and Wext (RMSE = 3.8 m). Using the most important variables (Wext, *PC1*, *PC2*, *PC4*, and *PC11*) in the RF regression yielded similar results, with an RMSE of 3.6 m and an  $R^2$  of 0.82. Finally, replacing the waveform extent by the waveform extent factor class (WC) in addition to the first 13 PCs in the Random Forest regression for canopy height estimation showed similar results (RMSE = 3.7 m,  $R^2 = 0.81$ ). Similar findings were noted when retaining only the most important variables (WC, *PC1*, *PC2*, *PC4*, and *PC11*), with an RMSE of 3.7 m

and an  $R^2$  of 0.81. Figure 10 shows examples of the comparison between the GLAS canopy heights using PCs and the Random Forest technique and the reference canopy heights estimated from the LD dataset.

#### 4.4. Model Performance in Different Forest Conditions

In previous sections, different models were applied on GLAS footprints over French Guiana in order to estimate forest canopy heights. Most models performed well, with an estimation precision lower than 4.5 m on the estimation of canopy heights. In this section, the two best models (model 7bis and 19, Table 2) were tested for different slopes and forest types, in order to analyze how the models would adapt in different forest conditions.

**Figure 10.** Comparison between canopy height estimates using the most important PCs in Random Forest regression models and those estimated from Low Density airborne LiDAR data (LD) using (a) the most important PCs ( $PC1$ ,  $PC2$ ,  $PC4$ , and  $PC11$ ) and (b) the most important PCs with the waveform extent.



In our study site, the distribution of the slopes shows that 80% are lower than five degrees, 17% between five and 10 degrees and 3% higher than 10 degrees. Based on these results, GLAS footprints were divided into two slope categories: GLAS footprints that fall on slopes lower than five degrees and GLAS footprints that fall on slopes higher than five degrees. Because the slopes are relatively weak in French Guiana, model validation for high slopes was not possible. Model validation over these two slope categories showed that the RMSE on the estimation of canopy heights slightly increased from 3.3 to 4.0 m and from 3.5 to 4.8 m for PCA and the linear regression model, respectively (models 7bis and 19, Table 2). However, the PCA model is slightly better at correcting the effects of the slopes in comparison to the linear regression model with a 0.7 m increase in the RMSE vs. 1.3 m for the linear model.

Forest landscape classes in French Guiana were defined in a previous study carried out by Gond *et al.* [62]. Gond *et al.* [62] interpreted 33 remotely sensed landscape types (LTs) using VEGETATION/SPOT images. Five of the 33 classes occupied 78% of the forests in the area. The method utilized in their study used a multivariate analysis of remote sensing data, field observations and environmental data. The defined LTs are as follows:

- LT1 represents dense, closed-canopy forest with small crowns of the same canopy height and small gaps mixed with regular canopies with well-developed crowns of almost the same canopy height without large gaps interlaced with flooded savannas (10%).
- LT2 is a closed canopy forest dominated by well-developed crowns of almost the same canopy height without large gaps.
- LT3 is an irregular- and disrupted-canopy forest where the trees have very different heights and different crown diameters with large gaps mixed with closed-canopy forest dominated by well-developed crowns at almost the same elevation without large gaps. LT3 is also interlaced with liana forests.
- LT4 is similar to LT3 with more liana forest and non-forest land covers.
- LT5 is an open forest associated with wetlands and bamboo thickets. However, no GLAS footprints available over this LT.

Model application over the different LTs showed that the RMSE on the estimation of canopy is consistent across the four LTs (LT1 to LT4). The RMSE ranged between 2.8 and 3.6 m for the PCA model (model 19, Table 2), and between 3.5 and 3.9 m for the linear regression model (model 7bis, Table 2).

#### 4.5. Error on the Estimation of Biomass

The objective of this section is to analyze the impact of the canopy height estimation precision on the Above ground Carbon Density (ACD) and Above Ground Biomass (AGB) estimation precision. Asner *et al.* [63] proposed a general plot aggregate allometry in order to estimate the above ground carbon density (ACD):

$$ACD = aH^\alpha \cdot BA^\beta \cdot W_D^\gamma \quad (11)$$

where H is the LiDAR derived top-of-canopy height, BA the basal area and  $W_D$  the wood density. Moreover, Asner *et al.* [4] showed that the basal area (BA) and the wood density ( $W_D$ ) were dependent on the LiDAR derived top-of-canopy height for all the studied tropical forests (Hawaii, Madagascar, Peru, Panama, and Colombia). Hence, according to their study the previous allometric relation could be written as:

$$ACD = aH^\alpha (b \cdot H)^\beta (c + d \cdot H)^\gamma \quad (12)$$

The relationship between the precision on the estimation of canopy heights and the precision on the estimation of ACD and AGB can be written as:

$$\frac{\Delta AGB}{AGB} = \frac{\Delta ACD}{ACD} = \left[ (\alpha + \beta) + \left( \frac{d \cdot \gamma}{d + \frac{c}{H}} \right) \right] \frac{\Delta H}{H} \quad (13)$$

where  $\Delta AGB/AGB$  is the relative precision on the estimation of above ground biomass,  $\Delta ACD/ACD$  is the relative error on the estimation of the above ground carbon density. The coefficients  $\alpha$ ,  $\beta$ ,  $\gamma$  were estimated by Asner *et al.* [4] using 754 field plots across five tropical countries (Hawaii, Madagascar, Peru, Panama, and Colombia) and many vegetation types. The coefficients c and d were also estimated by Asner *et al.* [4] for different regional forests. In our analysis of the canopy height estimation precision impact on the AGB and ACD estimation precision, the chosen coefficients c and d were those estimated from the moist Colombian forest [4]. These coefficients were chosen due to the fact that the French



Guiana's forest is a moist tropical forest and close in location to the Colombian forest. Finally, accuracy on the estimation of canopy heights of 3.6 m will lead to a relative error on the estimation of the ACD and AGB of about 14.1% (for a mean canopy height of 30 m). The United Nations Program on Reducing Emissions from Deforestation and forest Degradation (REDD) recommends biomass errors within 20 Mg/ha or 20% of field estimates for evaluating forest carbon stocks, but should not exceed errors of 50 Mg/ha for a global biomass map at a resolution of 1 ha [64,65]. Finally, in the case of high relief, where the precision on the estimation of canopy height exceeds 5 m, the precision on the estimation of biomass will be at best 20%.

## 5. Discussion

Our findings regarding the strong correlation between the waveform extent and the *in situ* canopy heights are in accordance with the studies of Lefsky *et al.* [10], Hilbert and Schmillius [24], and Baghdadi *et al.* [45]. They found that this metric is one of the most important metrics used in canopy height estimation models. However, waveform extent is not the sole metric used for canopy height estimation, as it can be affected by external sources, such as terrain relief. Thus, in order to obtain more precise canopy height estimation results, additional metrics are required. Previous studies developed metrics, such as the trail, lead, and the terrain index (TI), in order to increase the canopy height estimation precision. The TI index was first developed by Lefsky *et al.* [24] and the lead and trail were first introduced in Lefsky *et al.* [48]. These metrics were later used in many other studies like Hilbert *et al.* [24], Pang *et al.* [26], Chen *et al.* [40] and Baghdadi *et al.* [45]. These metrics, which were mainly used for the correction of the slope, proved to be very useful, as they increased significantly the precision on the estimation of canopy height models [10,24,40]. Moreover, the waveform extent and the trail metrics proved to be very successful in estimating canopy heights even in low relief areas like our study site. Indeed, the linear regression models that used the waveform extent and the trail metric showed a decrease in RMSE of at least 3.9 m in comparison to the direct method (for example, RMSE reaches 3.7 m in using the linear model with Wext, Trail and TI in comparison to RMSE of 7.9 m for the direct method). In contrast, the contribution of the lead in the canopy height estimation models seemed to be weak in this study. Similar findings were noted in the study of Baghdadi *et al.* [45], which also estimated canopy heights over flat terrain.

Our results also demonstrated that canopy height estimation using random forest regressions is better in comparison to the linear models, even when using the same metrics. Indeed, the random forest model, which uses only the waveform extent and the terrain index (TI), showed a 1.3 m decrease in RMSE in comparison to the linear model, which uses the same metrics. This is probably due to the fact that the relation between the GLAS metrics and canopy heights is not strictly linear.

The metric based estimation methods applied in this study include some potential error sources. These error sources are related to the precision of the extracted GLAS metrics especially metrics extracted using vegetation or ground information, such as the lead and trail. Indeed, over dense vegetated areas, the precision on the localization of the ground peak decreases significantly, and this will lead to lower precisions on the estimation of the trail metric and ultimately on the canopy height estimation. To solve this issue, another technique used in this study for canopy height estimation was the principal component analysis (PCA) of the waveform. This technique does not require metrics to be extracted from the GLAS

waveform in order to estimate canopy heights, as it works using the principal components of the raw LiDAR waveforms. The results of the PCA based models for canopy height estimation showed promising results when estimating canopy heights using either linear regressions or random forest regressions, with an RMSE of 5.9 and 4.7 m for the linear regressions and RF models, respectively. In addition, adding the waveform extent metric to these models showed slightly better estimation results in comparison to the metric based methods, with an RMSE ranging between 3.8 to 4.4 m for the linear regression models and around 3.6 m for the random forest models.

Other sources of error on the estimation of canopy heights are terrain slopes. Indeed over sloping areas, canopy height estimation precision decreases with the increase of the slope [25,40,45]. In our study area of low relief, an increase in RMSE of 0.7 m for the best PCA model and 1.3 m on for the best metric based model were noted in 5° to 10° slope areas in comparison to flat areas (0 to 5° slopes). However, over higher slopes (>10°), the error on the estimation of canopy heights is expected to be higher. In this study, the SRTM 90 m DEM was used, which is the only available DEM over large areas. The future availability of finer DEMs, such as the SRTM 30 m or the TanDEM-X 12 m, might improve the estimation of canopy heights.

Results showed that the canopy height estimation error using ICESat/GLAS (RMSE about 3.6 m in this study) leads to a relative error on the estimation of aboveground biomass of about 14%. This relative error will increase to about 20% for canopy height estimation precision of 5 m or higher. Thus, the United Nations Program on Reducing Emissions from Deforestation and forest Degradation (REDD) recommendations may not be satisfied over forested areas with steep slopes because the canopy height estimation precision will be higher than those estimated in this study.

## 6. Conclusion

In this study, the performance of the most frequently used linear regression models for canopy height estimation, which use metrics extracted from GLAS waveforms, was first evaluated. Then, models based on two seldom-used techniques for canopy height estimation from GLAS waveforms were introduced. The first included regression models using the principal component analysis (PCA) of GLAS waveforms. The second was based on the Random Forest technique. The Random Forest technique first used the metrics derived from the GLAS waveforms and then used PCs. The evaluation of these different models was performed with a large database consisting of GLAS data and canopy heights estimated from small-footprint airborne LiDAR measurements.

Within the GLAS footprints, which fell mostly on flat and sometimes moderately sloping terrain (slope < 15°), the direct method based on the difference between the ground peak and signal start showed an accuracy precision of 7.9 m (RMSE). The linear regression models that used a combination of waveform extent (Wext), modified trailing and leading edge extents (Trail and Lead) [24] and terrain index (TI) showed better accuracies for canopy height estimation in comparison to the direct method, with an RMSE between 3.7 and 4.9 m. In addition, the results reveal that the most relevant metrics in the estimation of forest heights are Wext and Trail. The linear regression model based on Wext and Trail estimated canopy height with an RMSE of 3.8 m. However, this model requires the Trail metric, which is difficult to extract with good accuracy in densely vegetated forests, such as those in French Guiana, affecting canopy height estimation due to the large contribution of the Trail metric to the linear

regression models. The contribution of Lead and TI calculated from the SRTM DEM appears to be very weak.

The regression model using the first 13 PCs and incorporating the waveform extent provided canopy height estimates with an RMSE of 3.8 m. The PCA regression models appear to be the best among the tested models, as they do not use difficult-to-extract metrics, such as the Trail metric.

The PCA model only requires the determination for each GLAS waveform, of which class contains Wext (Wext lower than 20 m, Wext between 20 m and 40 m, or Wext higher than 40 m). Thus, even if the estimation of Wext depends on the signal start and signal end metrics, which are sometimes difficult to calculate with certainty, the error in the estimation of Wext does not affect the estimation of canopy height because the Wext classes are defined in large intervals (20 m).

The Random Forest model using all metrics (Wext, Trail, Lead, and TI) had an RMSE of 3.4 m. Using only one of the Trail, Lead or TI metrics in addition to Wext slightly increased the RMSE to 3.6 m. Using only Wext, which has a relative importance factor almost three times higher than those for the other metrics, produced canopy height estimates with a precision of 4.4 m. Finally, using the PCs in the Random Forest regressions showed similar canopy height estimation results in comparison to using the PCs in the linear regression models, with an RMSE of 3.7 m when using the waveform extent and the four most important PCs.

The results of this study showed that using solely the raw GLAS waveforms (the section between signal start and signal end) and a rough estimate of the waveform extent, it is possible to estimate canopy heights with accuracies similar to or slightly better than those of the commonly used linear regression models. In addition, the precise estimation of some metrics that are difficult to calculate in densely vegetated forests is no longer required. Moreover, the use of Random Forest regressions using either GLAS and DEM metrics or the PCs extracted from the GLAS waveforms did not appear to increase estimation precision. In conclusion, the models introduced in this study provide good canopy height estimates in the dense tropical forest of French Guiana. Moreover, these canopy height estimates can be used for AGB estimation.

## Acknowledgments

The authors wish to thank the French Space Study Center (CNES, DAR 2014 TOSCA) for supporting this research. The authors acknowledge the National Snow and Ice Data Center (NSDIC) for the distribution of the ICESat/GLAS data. The authors also acknowledge the French Geological Survey (BRGM) and, in particular, José Perrin for providing the low-density LiDAR dataset. The authors wish to thank Lilian Blanc (Cirad) and Grégoire Vincent (IRD) for providing the high-density LiDAR dataset. We extend our thanks for Noveltis and EADS/Astrium for their financial support.

## Author Contributions

Fayad I., Baghdadi N. and Bailly J.S. had the original idea for the study, with all co-authors carried out the design. Fayad performed out the experiments, and the results were interpreted by all co-authors. Fayad I. and Baghdadi N. drafted the manuscript which was then improved and revised by all authors. All authors read and approved the final manuscript.

## Conflicts of Interest

The authors declare no conflict of interest.

## References

1. Pan, Y.; Birdsey, R.A.; Fang, J.; Houghton, R.; Kauppi, P.E.; Kurz, W.A.; Phillips, O.L.; Shvidenko, A.; Lewis, S.L.; Canadell, J.G.; *et al.* A large and persistent carbon sink in the world's forests. *Science* **2011**, *333*, 988–993.
2. Beer, C.; Reichstein, M.; Tomelleri, E.; Ciais, P.; Jung, M.; Carvalhais, N.; Rödenbeck, C.; Arain, M.A.; Baldocchi, D.; Bonan, G.B.; *et al.* Terrestrial gross carbon dioxide uptake: Global distribution and covariation with climate. *Science* **2010**, *329*, 834–838.
3. Chave, J.; Andalo, C.; Brown, S.; Cairns, M.A.; Chambers, J.Q.; Eamus, D.; Fölster, H.; Fromard, F.; Higuchi, N.; Kira, T.; *et al.* Tree allometry and improved estimation of carbon stocks and balance in tropical forests. *Oecologia* **2005**, *145*, 87–99.
4. Asner, G.P.; Mascaró, J. Mapping tropical forest carbon: Calibrating plot estimates to a simple LiDAR metric. *Remote Sens. Environ.* **2014**, *140*, 614–624.
5. Drake, J.B.; Knox, R.G.; Dubayah, R.O.; Clark, D.B.; Condit, R.; Blair, J.B.; Hofton, M. Above-ground biomass estimation in closed canopy neotropical forests using lidar remote sensing: Factors affecting the generality of relationships. *Global Ecol. Biogeogr.* **2003**, *12*, 147–159.
6. Feldpaush, T.R.; Lloyd, J.; Lewis, S.L.; Brienen, R.J.W.; Gloor, M.; Mendoza, A.M.; Lopez-Gonzalez, G.; Banin, L.; Salim, K.A.; Affum-Baffoe, K.; *et al.* Tree height integrated into pantropical forest biomass estimates. *Biogeosciences* **2012**, *9*, 3381–3403.
7. Lima, A.J.N.; Suwa, R.; de Mello Ribeiro, G.H.P.; Kajimoto, T.; dos Santos, J.; da Silva, R.P.; de Souza, C.A.S.; de Barros, P.C.; Noguchi, H.; Ishizuka, M.; *et al.* Allometric models for estimating above- and below-ground biomass in Amazonian forests at São Gabriel da Cachoeira in the upper Rio Negro, Brazil. *Forest Ecol. Manag.* **2012**, *277*, 163–172.
8. Maia Araújo, T.; Higuchi, N.; de Carvalho Júnior, J.A. Comparison of formulae for biomass content determination in a tropical rain forest site in the state of Pará, Brazil. *Forest Ecol. Manag.* **1999**, *117*, 43–52.
9. Vieira, S.; de Camarago, P.B.; Selhorst, D.; da Silva, R.; Hutyrá, L.; Chambers, J.Q.; Brown, I.F.; Higuchi, N.; dos Santos, J.; Wofsy, S.C.; *et al.* Forest structure and carbon dynamics in Amazonian tropical rain forests. *Oecologia* **2004**, *140*, 468–479.
10. Lefsky, M.A.; Harding, D.J.; Keller, M.; Cohen, W.B.; Carabajal, C.C.; Del Bom Espirito-Santo, F.; Hunter, M.O.; de Oliveira, R., Jr. Estimates of forest canopy height and aboveground biomass using ICESat. *Geophys. Res. Lett.* **2005**, *32*, doi:10.1029/2005GL023971.
11. Mitchard, E.T.A.; Saatchi, S.S.; White, L.J.T.; Abernethy, K.A.; Jeffery, K.J.; Lewis, S.L.; Collins, M.; Lefsky, M.A.; Leal, M.E.; Woodhouse, I.H.; *et al.* Mapping tropical forest biomass with radar and spaceborne LiDAR in Lopé National Park, Gabon: Overcoming problems of high biomass and persistent cloud. *Biogeosciences* **2012**, *9*, 179–191.

12. Santoro, M.; Askne, J.; Dammert, P.B.G. Tree height estimation from multi-temporal ERS SAR interferometric phase tomography. In Proceedings of FRINGE 2003 Workshop, Frascati, Italy, 1–5 December 2003.
13. Praks, J.; Hallikainen, M.; Antropov, O.; Molina, D. Boreal forest tree height estimation from interferometric TanDEM-X images. In Proceedings of 2012 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Munich, Germany, 22–27 July 2012; pp. 1262–1265.
14. Reigber, A.; Moreira, A. First demonstration of airborne SAR tomography using multibaseline L-band data. *IEEE Trans. Geosci. Remote Sens.* **2000**, *38*, 2142–2152.
15. Castel, T.; Beaudoin, A.; Trouche, G. Analysis of SAR interferometry for tree height estimation over hilly forested area. *Agricultura* **2002**, *1*, 15–23.
16. Neumann, M.; Hensley, S.; Laval, M.; Ahmed, R. Forest Structure Characterization Using JPL's UAVSAR Multi-Baseline polarimetric SAR interferometry and tomography. In Proceedings of The 4th Asia-Pacific Conference on Synthetic Aperture Radar, Tsukuba, Japan, 23–27 September 2013.
17. Garestier, F.; Dubois-Fernandez, P.C.; Champion, I. Forest height inversion using high-resolution P-Band Pol-InSAR data. *IEEE Trans. Geosci. Remote Sens.* **2008**, *46*, 3544–3559.
18. Huang, Y.; Ferro-Famil, L.; Reigber, A. Under-foilage object imaging using SAR tomography and polarimetric spectral estimators. *IEEE, Trans. Geosci. Remote Sens.* **2011**, *50*, 2213–2225.
19. Guillaso, S.; Reigber, A. Scatterer characterisation using polarimetric SAR tomography. In Proceedings of 2005 IEEE International Geoscience and Remote Sensing Symposium, Seoul, Korea, 25–29 July 2005; pp. 2685–2688.
20. Lefsky, M.A. A global forest canopy height map from the moderate resolution imaging spectroradiometer and the geoscience laser altimeter system. *Geophys. Res. Lett.* **2010**, *37*, doi:10.1029/2010GL043622.
21. Wulder, M.A.; Seaman, D. Forest inventory height update through the integration of LIDAR data with segmented Landsat imagery. *Can. J. Remote Sens.* **2003**, *29*, 536–543.
22. Boudreau, J.; Nelson, R.F.; Margolis, H.A.; Beaudoin, A.; Guindon, L.; Kimes, D.S. Regional aboveground forest biomass using airborne and spaceborne LiDAR in Québec. *Remote Sens. Environ.* **2008**, *112*, 3876–3890.
23. Saatchi, S.S.; Harris, N.L.; Brown, S.; Lefsky, M.; Mitchard, E.T.A.; Salas, W.; Zutta, B.R.; Buermann, W.; Lewis, S.L.; Hagen, S.; *et al.* Benchmark map of forest carbon stocks in tropical regions across three continents. *Proc. Natl. Acad. Sci.* **2011**, *108*, 9899–9904.
24. Hilbert, C.; Schmullius, C. Influence of surface topography on ICESat/GLAS forest height estimation and waveform shape. *Remote Sens.* **2012**, *4*, 2210–2235.
25. Lee, S.; Ni-Meister, W.; Yang, W.; Chen, Q. Physically based vertical vegetation structure retrieval from ICESat data: Validation using LVIS in White Mountain National Forest, New Hampshire, USA. *Remote Sens. Environ.* **2011**, *115*, 2776–2785.
26. Pang, Y.; Lefsky, M.; Andersen, H.-E.; Miller, M.E.; Sherrill, K. Validation of the ICESat vegetation product using crown-area-weighted mean height derived using crown delineation with discrete return lidar data. *Can. J. Remote Sens.* **2008**, *34*, S471–S484.

27. Ploton, P.; Pélissier, R.; Barbier, N.; Proisy, C.; Ramesh, B.R.; Coutron, P. Canopy texture analysis for large-scale assessments of tropical forest stand structure and biomass. In *Treetops at Risk*; Lowman, M., Devy, S., Ganesh, T., Eds.; Springer New York: New York, NY, USA, 2013; pp. 237–245.
28. Lu, D.; Chen, Q.; Wang, G.; Moran, E.; Batistella, M.; Zhang, M.; Laurin, G.V.; Saah, D. Aboveground forest biomass estimation with Landsat and lidar data and uncertainty analysis of the estimates. *Int. J. For. Res.* **2012**, *2012*, 1–16.
29. Proisy, C.; Mougin, E.; Fromard, F.; Karam, M.A. Interpretation of polarimetric radar signatures of mangrove forests. *Remote Sens. Environ.* **2000**, *71*, 56–66.
30. Le Toan, T.; Quegan, S.; Davidson, M.W.J.; Balzter, H.; Paillou, P.; Papathanassiou, K.; Plummer, S.; Rocca, F.; Saatchi, S.; Shugart, H.; *et al.* The BIOMASS mission: Mapping global forest biomass to better understand the terrestrial carbon cycle. *Remote Sens. Environ.* **2011**, *115*, 2850–2860.
31. Sandberg, G.; Ulander, L.M.H.; Fransson, J.E.S.; Holmgren, J.; Le Toan, T. L- and P-band backscatter intensity for biomass retrieval in hemiboreal forest. *Remote Sens. Environ.* **2011**, *115*, 2874–2886.
32. Wu, S.-T. Potential application of multipolarization SAR for pine-plantation biomass estimation. *IEEE Trans. Geosci. Remote Sens.* **1987**, *GE-25*, 403–409.
33. Dobson, M.C.; Ulaby, F.T.; LeToan, T.; Beaudoin, A.; Kasischke, E.S.; Christensen, N. Dependence of radar backscatter on coniferous forest biomass. *IEEE Trans. Geosci. Remote Sens.* **1992**, *30*, 412–415.
34. Imhoff, M.L. Radar backscatter and biomass saturation: Ramifications for global biomass inventory. *IEEE Trans. Geosci. Remote Sens.* **1995**, *33*, 511–518.
35. Luckman, A.; Baker, J.; Kuplich, T.M.; da Costa Freitas Yanasse, C.; Frery, A.C. A study of the relationship between radar backscatter and regenerating tropical forest biomass for spaceborne SAR instruments. *Remote Sens. Environ.* **1997**, *60*, 1–13.
36. Luckman, A.; Baker, J.; Honzák, M.; Lucas, R. Tropical forest biomass density estimation using JERS-1 SAR: Seasonal variation, confidence limits, and application to image mosaics. *Remote Sens. Environ.* **1998**, *63*, 126–139.
37. Minh, D.H.T.; Le Toan, T.; Rocca, F.; Tebaldini, S.; d’Alessandro, M.M.; Villard, L. Relating P-band synthetic aperture radar tomography to tropical forest biomass. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 967–979.
38. Nizalapur, V.; Sekhar Jha, C.; Madugundu, R. Estimation of above ground biomass in Indian tropical forested area using multifrequency DLRESAR data. *Int. J. Geomat. Geosci.* **2010**, *1*, 167–178.
39. Zolkos, S.G.; Goetz, S.J.; Dubayah, R. A meta-analysis of terrestrial aboveground biomass estimation using lidar remote sensing. *Remote Sens. Environ.* **2013**, *128*, 289–298.
40. Chen, Q. Retrieving vegetation height of forests and woodlands over mountainous areas in the Pacific Coast region using satellite laser altimetry. *Remote Sens. Environ.* **2010**, *114*, 1610–1627.
41. Bourguin, B.; Baghdadi, N. Assessment of C-band SRTM DEM in a dense equatorial forest zone. *Compt. Rendus Geosci.* **2005**, *337*, 1225–1234.

42. Bourguine, B.; Baghdadi, N.; Hosford, S.; Daniels, P. Generation of a ground-level DEM in a dense equatorial forest zone by merging airborne laser data and a top-of-canopy DEM. *Can. J. Remote Sens.* **2004**, *30*, 913–926.
43. Vincent, G.; Caron, F.; Sabatier, D.; Blanc, L. LiDAR shows that higher forests have more slender trees. *Bois. Forêts Tropiques* **2012**, *314*, 51–56.
44. Harding, D.J.; Carabajal, C.C. ICESat waveform measurements of within-footprint topographic relief and vegetation vertical structure. *Geophys. Res. Lett.* **2005**, *32*, doi:10.1029/2005GL023471.
45. Baghdadi, N.; Le Maire, G.; Fayad, I.; Bailly, J.S.; Nouvellon, Y.; Lemos, C.; Hakamada, R. Testing different methods of forest height and aboveground biomass estimations from ICESat/GLAS data in Eucalyptus plantations in Brazil. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2014**, *7*, 290–299.
46. Baghdadi, N.; Cavelier, S.; Chiles, J.-P.; Bourguine, B.; Toutin, T.; King, C.; Daniels, P.; Perrin, J.; Truffert, C. Merging of airborne elevation data and Radarsat data to develop a Digital Elevation Model. *Int. J. Remote Sens.* **2005**, *26*, 139–163.
47. Rosette, J.A.B.; North, P.R.J.; Suarez, J.C. Vegetation height estimates for a mixed temperate forest using satellite laser altimetry. *Int. J. Remote Sens.* **2008**, *29*, 1475–1493.
48. Lefsky, M.A.; Keller, M.; Pang, Y.; De Camargo, P.B.; Hunter, M.O. Revised method for forest canopy height estimation from Geoscience Laser Altimeter System waveforms. *J. Appl. Remote Sens.* **2007**, *1*, doi:10.1117/1.2795724.
49. Sun, G.; Ranson, K.J.; Kimes, D.S.; Blair, J.B.; Kovacs, K. Forest vertical structure from GLAS: An evaluation using LVIS and SRTM data. *Remote Sens. Environ.* **2008**, *112*, 107–117.
50. Duong, H.; Lindenbergh, R.; Pfeifer, N.; Vosselman, G. ICESat full-waveform altimetry compared to airborne LASER scanning altimetry over The Netherlands. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 3365–3378.
51. Allouis, T.; Bailly, J.-S.; Pastol, Y.; Le Roux, C. Comparison of LiDAR waveform processing methods for very shallow water bathymetry using Raman, near-infrared and green signals. *Earth Surf. Process. Landf.* **2010**, *35*, 640–650.
52. Karlis, D.; Saporta, G.; Spinakis, A. A simple rule for the selection of principal components. *Commun. Stat.—Theory Methods* **2003**, *32*, 643–666.
53. Neuenschwander, A.L.; Magruder, L.A.; Tyler, M. Landcover classification of small-footprint, full-waveform lidar data. *J. Appl. Remote Sens.* **2009**, *3*, doi:10.1117/1.3229944.
54. Carabajal, C.C.; Harding, D.J. SRTM C-band and ICESat laser altimetry elevation comparisons as a function of tree cover and relief. *Photogramm. Eng. Remote Sens.* **2006**, *72*, 287–298.
55. Duncanson, L.I.; Niemann, K.O.; Wulder, M.A. Estimating forest canopy height and terrain relief from GLAS waveform metrics. *Remote Sens. Environ.* **2010**, *114*, 138–154.
56. Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*; Parzen, E., Tanabe, K., Kitagawa, G., Eds.; Springer New York: New York, NY, USA, 1998; pp. 199–213.
57. Houghton, R.A.; Butman, D.; Bunn, A.G.; Krankina, O.N.; Schlesinger, P.; Stone, T.A. Mapping Russian forest biomass with data from satellites and forest inventories. *Environ. Res. Lett.* **2007**, *2*, doi:10.1088/1748-9326/2/4/045032.

58. Kankare, V.; Vastaranta, M.; Holopainen, M.; Rätty, M.; Yu, X.; Hyyppä, J.; Hyyppä, H.; Alho, P.; Viitala, R. Retrieval of forest aboveground biomass and stem volume with airborne scanning LiDAR. *Remote Sens.* **2013**, *5*, 2257–2274.
59. Le Maire, G.; Marsden, C.; Nouvellon, Y.; Grinand, C.; Hakamada, R.; Stape, J.-L.; Laclau, J.-P. MODIS NDVI time-series allow the monitoring of Eucalyptus plantation biomass. *Remote Sens. Environ.* **2011**, *115*, 2613–2625.
60. Mutanga, O.; Adam, E.; Cho, M.A. High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *Int. J. Appl. Earth Obs. Geoinf.* **2012**, *18*, 399–406.
61. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
62. Gond, V.; Freycon, V.; Molino, J.-F.; Brunaux, O.; Ingrassia, F.; Joubert, P.; Pekel, J.-F.; Prévost, M.-F.; Thierron, V.; Trombe, P.-J.; *et al.* Broad-scale spatial pattern of forest landscape types in the Guiana Shield. *Int. J. Appl. Earth Obs. Geoinf.* **2011**, *13*, 357–367.
63. Asner, G.P.; Mascaró, J.; Muller-Landau, H.C.; Vieilledent, G.; Vaudry, R.; Rasamoelina, M.; Hall, J.S.; van Breugel, M. A universal airborne LiDAR approach for tropical forest carbon mapping. *Oecologia* **2012**, *168*, 1147–1160.
64. Houghton, R.A.; Hall, F.; Goetz, S.J. Importance of biomass in the global carbon cycle. *J. Geophys. Res.* **2009**, *114*, doi:10.1029/2009JG000935.
65. Hall, F.G.; Bergen, K.; Blair, J.B.; Dubayah, R.; Houghton, R.; Hurtt, G.; Kellndorfer, J.; Lefsky, M.; Ranson, J.; Saatchi, S.; *et al.* Characterizing 3D vegetation structure from space: Mission requirements. *Remote Sens. Environ.* **2011**, *115*, 2753–2775.

© 2014 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/4.0/>).