remote sensing

MDPI

Article

Estimating Canopy Gap Fraction Using ICESat GLAS within Australian Forest Ecosystems

Craig Mahoney [1,*], Chris Hopkinson [1], Natascha Kljun [2] and Eva van Gorsel [3]

1 Department of Geography, University of Lethbridge, Lethbridge, AB T1K 3M4, Canada; c.hopkinson@uleth.ca
2 Department of Geography, Swansea University, Singleton Park, Swansea SA2 8PP, UK; n.kljun@swansea.ac.uk
3 CSIRO Oceans and Atmosphere, Wilf Crane Crescent, Yarralumla ACT 2600, Australia; evavangorsel@gmail.com
* Correspondence: craig.mahoney@uleth.ca; Tel.: +1-403-332-4043

Academic Editors: Lars T. Waser, Randolph H. Wynne and Prasad S. Thenkabail
Received: 19 September 2016; Accepted: 22 December 2016; Published: 11 January 2017

Abstract: Spaceborne laser altimetry waveform estimates of canopy Gap Fraction (GF) vary with respect to discrete return airborne equivalents due to their greater sensitivity to reflectance differences between canopy and ground surfaces resulting from differences in footprint size, energy thresholding, noise characteristics and sampling geometry. Applying scaling factors to either the ground or canopy portions of waveforms has successfully circumvented this issue, but not at large scales. This study develops a method to scale spaceborne altimeter waveforms by identifying which remotely-sensed vegetation, terrain and environmental attributes are best suited to predicting scaling factors based on an independent measure of importance. The most important attributes were identified as: soil phosphorus and nitrogen contents, vegetation height, MODIS vegetation continuous fields product and terrain slope. Unscaled and scaled estimates of GF are compared to corresponding ALS data for all available data and an optimized subset, where the latter produced most encouraging results (R^2 = 0.89, RMSE = 0.10). This methodology shows potential for successfully refining estimates of GF at large scales and identifies the most suitable attributes for deriving appropriate scaling factors. Large-scale active sensor estimates of GF can establish a baseline from which future monitoring investigations can be initiated via upcoming Earth Observation missions.

Keywords: vegetation; remote sensing; forestry; LiDAR

1. Introduction

Canopy cover information is essential for understanding spatial and temporal variability in vegetation biomass, local meteorological processes and hydrological transfers (i.e., energy in heat variations, gas and water) within vegetated environments [1]. Canopy cover has been monitored via satellite and airborne remote sensing technologies for decades, providing information on vegetation and biomass conditions from local (100's of km^2) to global scales [2–4]. However, the passive nature and/or low resolution of satellite sensors have restricted our ability to simultaneously map canopy structural attributes over large areas and at the tree crown level [5,6].

The use of Airborne Laser Scanning (ALS) for deriving canopy cover indices has been demonstrated at forest stand scales by many [7–13]; however, due to costs and sampling logistics, ALS acquisitions are typically limited in spatial extent. In contrast, the Geoscience Laser Altimeter System (GLAS), which previously operated (2003–2009) onboard the Ice, Cloud and land Elevation Satellite (ICESat), has the potential to retrieve canopy cover indices at near global scales, but has not demonstrated the same success as ALS to date. The large-footprint continuous waveform nature of

GLAS data has been suggested to be sensitive to apparent canopy and ground surface reflectance values; that is, canopy returns reflect stronger or weaker than the ground [14]. Radiative transfer analysis partially corroborates these findings [15], but suggests a greater influence is caused by waveforms being more sensitive to within canopy scattering events, which may lead to less energy irradiating the ground surface [16]. Further compounding the misinterpretation issues, the non-uniform footprint energy distribution effectively irradiates targets with higher intensity at the footprint centre. This systematically alters the footprint backscatter signature, which is embedded in the location and amplitude of peak(s) in the waveform profile, the analysis of which allows the retrieval of Gap Fraction (GF). Given a non-uniform energy distribution, retrieved GF is expected to be different to reality due to artificially enhanced illumination at select locations within the footprint (energy distribution dependent), whereas a uniform energy distribution (similar to that of the Sun) is expected to alleviate this issue.

A practical solution is to scale waveform ground or canopy returns to account for waveform reflectance differences; however, this remains an active area of research. To date, Lefsky et al. [14] demonstrated that scaling the waveform ground return by a factor of two resulted in more highly correlated estimates of Fractional Cover (FC) with respect to ALS data for the relatively small footprint sensor: Scanning Lidar Imager of Canopies by Echo Recovery (SLICER). This approach was adopted by Luo et al. [17] for GLAS with encouraging results. However, by design, the application of such a scaling factor is not tolerant of any deviation attributed to variable environmental characteristics from those for which it was specifically derived, i.e., the deciduous forests of eastern Maryland, USA. Tang et al. [18] scaled Laser Vegetation Imaging Sensor (LVIS) waveform data by field acquired ratios of vegetation and ground reflectance values from spectroradiometers to refine estimates of GF. Whilst this applies a more empirically-based method to estimating canopy cover, the need for extensive field data limits the spatial reach of such analysis. As follow-up studies, Tang et al. [19] and Tang et al. [20] built on previous findings and employed a recursive algorithm to retrieve estimates of Leaf Area Index (LAI) via refined GF estimates from GLAS; however, this method required the use of some locally applicable initial conditions by which scaling factor outputs were governed (i.e. expected minimum and maximum values). Previous studies fail to meet the need for correcting waveform-based canopy cover indices by both empirical and independent means at large scales. Furthermore, with the view of large-scale application, the derivation of a scaling factor based on large-scale measured/estimated predictor attributes is required in order to make waveform scaling viable at large scales; a key focus of the current study.

GLAS data are currently the only near global active (i.e., capable of penetrating the canopy to retrieve within-canopy information) laser data available, and the need to identify which attributes are best suited to spatially scale GLAS estimates of canopy cover is pertinent, as future missions (e.g., NASA's Global Ecosystem Dynamics Investigation Lidar; GEDI) will provide similar data. Establishing robust methodologies for deriving spatially large-scale parameter estimates by current data will not only be applicable to future data, they will allow contemporary products to be established, which will offer a basis from which large-scale monitoring investigations can be conducted.

In the current study, a method for the general application of scaling factors at large scales is developed and tested for sensitivity against predictor attributes. The sensitivity of terrain and vegetation characteristics are investigated as a function of the difference in canopy GF estimates from ALS and GLAS in order to identify which attributes best predict a difference between sources of GF estimation. This will accommodate spatial up-scaling, allowing the use of imputation techniques, such as Random Forest (RF) [21], to map GF continuously at continental scales in future studies. Furthermore, the potential value of restricting GLAS data to an optimal subset, as identified by Mahoney et al. [22], for model training purposes is investigated against training models with all available data. GF estimates are identified as a value-added output in this study, as they provide a basis for estimating the Leaf Area Index (LAI) using the Beer–Lambert law assumptions (not studied here, but the basis of ongoing research).

2. Materials and Methods

2.1. Discrete Return ALS Data

Three Terrestrial Ecosystem Research Network (TERN) supersites and a single independent site in eastern Australia are employed in the current study (Figure 1). Each site is forested and has been surveyed with ALS and GLAS data. Individual ALS surveys were acquired using a Reigl Q560 by Airborne Research Australia (ARA) during the Southern Hemisphere summer and/or fall seasons of 2009, 2012 and 2014.

Robson Creek (17.106°S, 145.622°E) is located approximately 25 km southwest of Cairns. The forest exhibits moderate relief and ranges from complex mesophyll vine forest to simple/complex notophyll vine forest with increasing elevation to the north. ALS data were acquired during September 2012 with a mean point density of 5–6 points per square meter (ppm^2). Watts Creek (37.692°S, 145.684°E), located 70 km east of Melbourne, is an old-regrowth open forest dominated by Eucalypts over complex terrain. ALS data were acquired with a mean point density of 5–6 ppm^2. Warra (43.106°S, 146.657°E) approximately 60 km southwest of Hobart, Tasmania, is a cool, temperate wet forest biome over moderately complex terrain. The site consists of moorland, temperate rainforest, riparian and montane conifer forest and shrubs. ALS data were acquired during May 2014 with a mean point density of up to 25 ppm^2. Tumbarumba (35.686°S, 148.234°E) is approximately 120 km southwest of Canberra. It is a Eucalypt-dominated, moderately open forest with complex terrain [23,24]; ALS data were acquired during November 2009 at a mean point density of 5 ppm^2. This site is not to be confused with the TERN site at the Tumbarumba research station, the site employed here is located approximately 7 km east of the research station, but the vegetation and terrain characteristics observed at the research station are still applicable. This location was strategically selected as a study site due to overlap with numerous GLAS footprints.

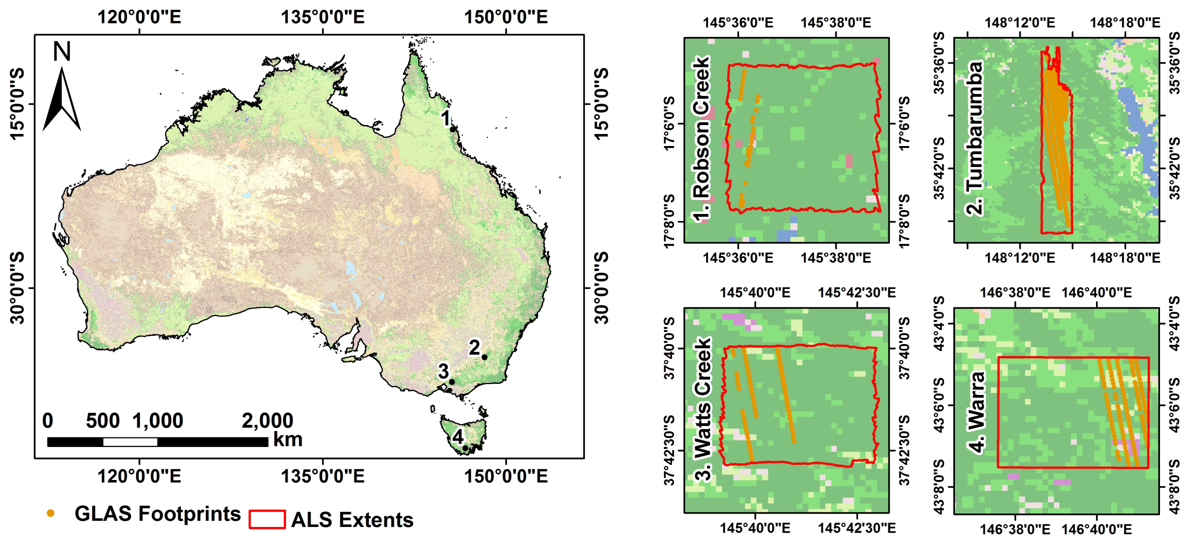

Figure 1. Location of study sites. ALS extents are shown in red, and GLAS footprint centres are indicated by an orange point.

2.2. GLAS Data

The ICESat/GLAS land data (GLA14 data product), Release 33 [25,26], are employed in the current study. GLAS data are continuous waveforms; the returned echo pulse forms an energy profile, which is a function of surfaces encountered during transit, such as vegetation and/or ground surfaces [27–29]. The returned energy profile is fitted with up to six Gaussians as described by Duong et al. [30], the sum of which defines the 'model alternate fit' return pulse.

For this study, only GLAS data acquired within forested regions are employed, where Australia's forest regions are delineated by the tree classes of the National Dynamic Land Cover Dataset (DLCD): closed, open, sparse and scattered; technical details of the DLCD can be found in Lymburner et al. [31]. To further maintain GLAS data integrity, footprints that were saturated upon detection and/or experienced cloud cover whilst in transit are excluded from analysis. Additionally, footprints that exhibit an absolute difference between GLAS and Shuttle Radar Topography Mission (SRTM) elevation estimates >8 m are also removed [32]. Post filtering, a total of 457 GLAS footprints are available for analysis; this reduces to 175 after filtering by the 'optimal' criteria, i.e., high energy (>28 mJ) Laser 3 acquisitions during summertime (cf. Mahoney et al. [22]).

The employed GLAS data (post filtered and optimal) exhibit a temporal range from February 2004–October 2008, where Southern Hemisphere summertime is defined from October to the subsequent March. Little effect on the outcome of this analysis is expected from the temporal disparity between GLAS and ALS data as GLAS waveform profile portions (reflections from canopy and ground) are refined with respect to ALS data regardless of acquisition time. Temporal disparity will likely play a greater role when attempting to model GF from refined GLAS data for specific time periods (with respect to vegetation phenological state); i.e., ALS data will need to be available for that specific period in order to allow GLAS model development.

2.3. ALS Gap Fraction

A number of methods exist for deriving GF from ALS point cloud data. A simple method that uses return 'intensity' is selected here. This selection was made on the basis that intensity has been shown to be more directly related to GF than return count [1,13]; furthermore, intensity ratios are more analogous than return ratios to the energy distribution encountered with GLAS data. GF is defined as the ratio of the sum of intensities from all returns (first, intermediate and last) that exist between the canopy top and 2 m above the ground ($I_{Elevation>2m}$) and the sum of intensities from all returns between the canopy top and the ground itself (I_{All}; Equation (1); Hopkinson et al. [1]). Estimates of GF are calculated directly from the portions of the ALS point cloud data that are within the boundaries of unique GLAS footprints. This method allows a direct comparison of ALS estimates of GF to GLAS equivalents.

$$\mathrm{GF} = 1 - \frac{\sum \mathrm{I}_{\mathrm{Elevation}>2\mathrm{m}}}{\sum \mathrm{I}_{\mathrm{All}}} \tag{1}$$

Further fortification of this methodological selection is supported by the high correlation demonstrated ($R^2 = 0.92$) between this ALS product and field measured GF via Digital Hemispherical Photographs (DHPs) in mature and regenerating mixed wood plots [13]. This method was up scaled and applied to 7 Canadian boreal sites where similar high correlation was also demonstrated ($R^2 = 0.77$) [1].

2.4. GLAS Gap Fraction

The derivation of GF from waveform data (such as GLAS) is analogous to that of the point cloud intensity method described in Equation (1) with minor modification. Where a ratio of summed intensities is noted, waveform GF is calculated from the ratio of summed energy from particular portions of the profile, such as those reflected from the ground and/or vegetation surfaces only. In particular, GF is calculated as one minus the ratio of the sum of the returned energy between the canopy top and 2 m above the ground, and the sum of energy between the canopy top and the ground (Figure 2a). The waveform equivalent of Equation (1) is given by Equation (2), where energy (ϵ) replaces return intensity (I).

$$\mathrm{GF} = 1 - \frac{\sum \epsilon_{\mathrm{Elevation}>2\mathrm{m}}}{\sum \epsilon_{\mathrm{All}}} \tag{2}$$

Whilst simplistic in principle, the large footprint nature of GLAS often presents challenges in locating the true ground elevation. This becomes particularly challenging when attempting analyses over complex terrain [33]. In an attempt to mitigate the influence of complex terrain, we employ the method of Rosette et al. [34] to locate the 'ground' in returned GLAS waveforms. This method makes use of the (up to 6) fitted Gaussians that form the model alternate return pulse, where the ground is considered to correspond to the centre of the Gaussian that exhibits the greatest amplitude of the two that are least elevated within the waveform (Figure 2b).

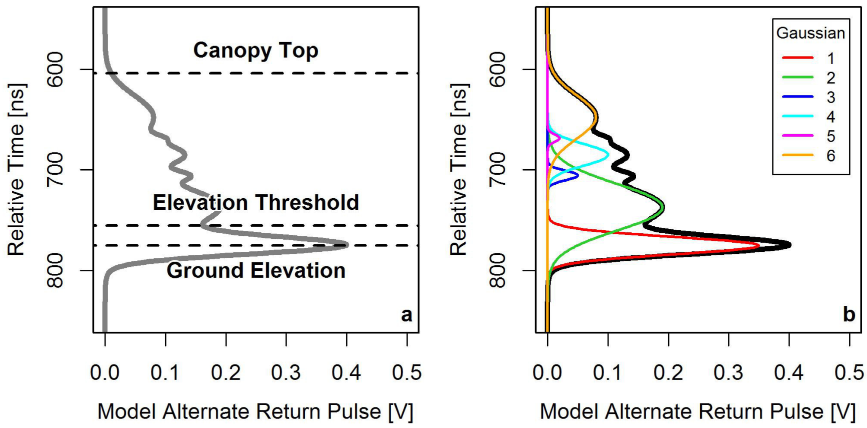

Figure 2. Example waveform indicating (**a**) the locations of the canopy top, ground elevation, and an example height threshold above the ground; and (**b**) the same waveform with fitted Gaussians.

2.5. Gap Fraction Scaling

The literature suggests that a scaling factor applied to either the ground or vegetation portions of waveforms will yield more coherent results with respect to ALS equivalents [35,36]. GF estimates from waveforms can be broken down into energy contributions from canopy and ground components and be represented with the necessary scaling factor (f) by Equation (3).

$$\mathrm{GF_{Waveform}} = \frac{\mathrm{Ground}}{\mathrm{Ground} + (\mathrm{f} \cdot \mathrm{Canopy})} \tag{3}$$

As GF estimates from ALS data have been demonstrated to yield high correspondence with field acquired data, these estimates are assumed to be correct. By this assumption, and the fact that ALS estimates of GF ($\mathrm{GF_{ALS}}$) can be represented by Equation (3) (i.e., $\mathrm{GF_{Waveform}} = \mathrm{GF_{ALS}}$ when f is 1), Equation (3) can be arranged to make f the subject (Equation (4)). This offers a means of scaling waveform estimates of GF to ALS equivalents by physical observations, i.e., GLAS measured canopy and ground components.

$$\mathrm{f} = \frac{\mathrm{Ground} \cdot (1 - \mathrm{GF_{ALS}})}{\mathrm{GF_{ALS}} \cdot \mathrm{Canopy}} \tag{4}$$

Scaling factor values are predicted (from a minimum of 0.25 in 0.25 intervals) for each GLAS footprint based on predictor attributes via a novel approach that utilizes RF regression algorithms; however, the knowledge of which attributes will produce the best predictive results is not a priori known. As a result, all predictor attributes are initially tested, and predictions are refined based on the importance assigned to each attribute by the RF algorithm. RF scores the importance of each variable by the increase in the mean of the error of a tree in a forest when the observed values of this variable are permutated in the Out-Of-Bag (OOB) samples; see Genuer et al. [37]. As the importance measures assigned by RF are subject to the data selected to train the RF model, the order of variable importance may change if model training is repeated. In order to form a more robust estimation of which attributes scaling factors are most sensitive to, an independent analysis of attribute importance is performed also (methodology described below).

2.6. Predictor Attributes

As a finite coincidence is observed between ALS and GLAS data, it is not possible to map scaling factors and refined estimates of GF to individual GLAS footprints beyond ALS extents without the use of imputation algorithms. Such algorithms require the use of predictor attributes to predict a response; the more suitable the predictor data, the better the expected response (with respect to overall accuracy). Pre-empting the use of the scaling factor methodology at large scales, the sensitivity of the scaling factor is investigated with respect to 13 environmental attributes for possible use in imputation (Table 1). It is important to note that each predictor attribute was resampled to the smallest common resolution (250 m in this case) among the available predictors. Additionally, the variability in the predicted scaling factors is gauged against two GLAS instrument attributes: laser campaign and footprint eccentricity, to identify if GLAS energy losses were consistent during its operational lifetime.

Table 1. Summary of predictor attributes used in testing the sensitivity of GLAS GF scaling factors.

Data	Units	Resolution	Description	Reference
Aspect	°	30 m	Derived from Shuttle Radar Topography Mission (SRTM) elevation product. Aspect is the direction of the maximum rate of change in the z-value from each cell in a raster surface.	[38]
Slope	°	30 m	Derived from SRTM elevation product. Slope is the rate of maximum change in z-value from each cell.	[38]
Elevation	m	30 m	National SRTM Digital Elevation Model (DEM), Version 1.0.	[39]
Valley Bottom Flatness (VBF)	-	30 m	Derived from SRTM DEM. VBF is a topographic index that identifies areas of deposited material at a range of scales based on the observations that valley bottoms are low and flat relative to their surroundings and that large valley bottoms are flatter than smaller ones.	[40]
Vegetation Continuous Fields (VCF)	%	250 m	Derived from Moderate Resolution Imaging Spectroradiometer (MODIS) MOD44B product. The VCF collection contains proportional estimates for vegetative cover types: woody vegetation, herbaceous vegetation and bare ground.	[41]
Vegetation classification	-	100 m	The National Vegetation Information System (NVIS) is a comprehensive data system that provides information on the extent and distribution of vegetation types in Australian landscapes based on extensive field data acquisition.	[42]
Vegetation height	m	250 m	Derived product based on the integration of GLAS data with other Australian inventory products.	[43]
Land cover classification	-	250 m	The National Dynamic Land Cover Dataset of Australia is the first nationally-consistent and thematically-comprehensive land cover reference for Australia based on MODIS data.	[31]
Soil type	-	250 m	Based on extensive field acquisitions, the Atlas of Australian Soils was compiled in the 1960s to provide a consistent national description of Australia's soils.	[44]
Soil depth	m	90 m	Derived from extensive field acquisitions and spectroscopic measurements, soil depth profile (A and B horizons)	[45]
Soil nitrogen	%	90 m	Derived from extensive field acquisitions and spectroscopic measurements, a mass fraction of total nitrogen in the soil by weight.	[45]
Soil phosphorus	%	90 m	Derived from extensive field acquisitions and spectroscopic measurements, a mass fraction of total phosphorus in the soil by weight.	[45]
Soil pH	-	90 m	Derived from extensive field acquisitions and spectroscopic measurements, a pH of 1:5 soil/0.01M calcium chloride ($CaCl_2$) extract.	[45]

In order to reliably indicate if the GLAS scaling factor is sensitive to any of the 13 tested attributes, each attribute was split into intervals, where each interval exhibits a variable sample size. The largest common denominator of samples was selected without replacement from each interval, thus forcing each interval to be of the same sample size when analysed. Additionally, to avoid small interval sample sizes and under sampling bias effects, any intervals with <10 samples were excluded from analysis. To avoid any further bias by randomly sampling from each interval only once, the previously described process was repeated 100 times per attribute. This yields a more comprehensive and robust attribute data distribution across each interval.

The mean scaling factor required per subsequent intervals (per attribute) is analysed for linearity via a Harvey–Collier (HC) test for linearity [46] and fitted with a linear regression model if the mean of the recursive residuals did not significantly (α = 0.95) differ from zero. Rejection of the hypothesis of no significant difference of the recursive residuals from zero results in the execution of local polynomial regression fitting [47]. The Squared Sum of fitted model Residuals (SSR) is assessed to yield an independent measure of variable importance. Throughout the regression analyses, mean values of f are not regressed against attribute interval values, but rather against a scalar that ranges from 1 to the number of valid (>10 samples) intervals. This allows for analytical consistency between numerical and categorical attributes, where the latter would otherwise exhibit some unordered category corresponding numerical code that may alter the magnitude of the calculated SSR, thus potentially changing the importance of certain attributes. This independent measure of variable importance is favoured over that built into the RF algorithm as RF importance measures are subject to the RF training data, which varies with model training repetition; hence, the order of variable importance may also vary. A general workflow of obtaining SSR information for ranking predictor importance independently is illustrated in Figure 3; note that this workflow assumes calculations of ALS and GLAS GF.

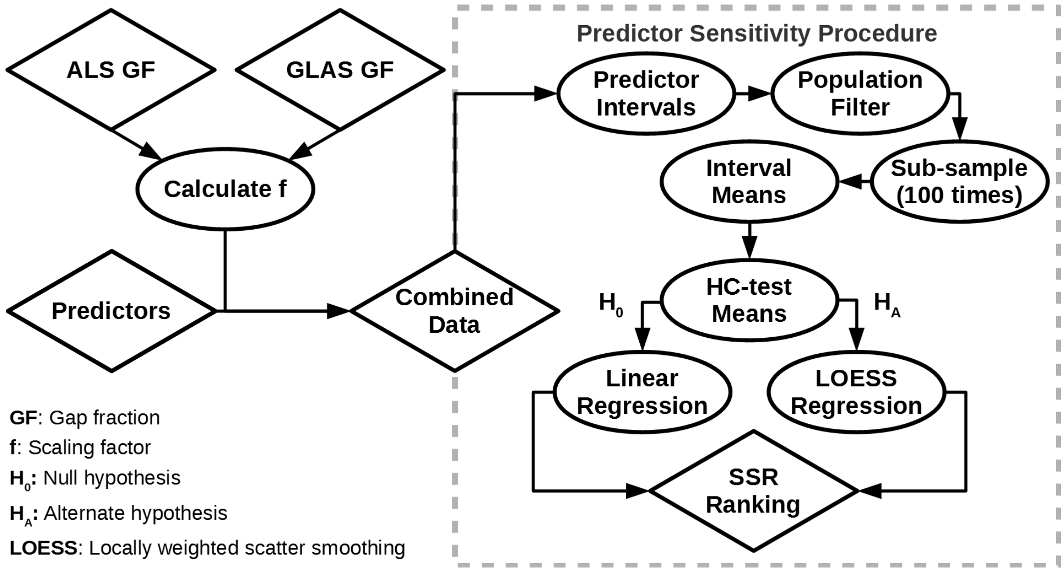

Figure 3. Methodological workflow for obtaining Squared Sum of fitted model Residuals (SSR) for independent predictor ranking from ALS and GLAS GF.

2.7. Gap Fraction Comparisons

ALS estimates of GF are directly compared with spatially coincident, unscaled, GLAS waveform equivalents for all data, defined as 'controls'. This yields a baseline against which comparative improvements or losses, via the application of scaling factors, can be assessed. Scaled estimates of GLAS GF are based on Equation (4), where f is predicted for each footprint via RF imputation where the RF technique was selected, as it is well documented and robust; however, the existence of other techniques for potential use should be acknowledged. Two sets of scaling factor predictions are made via RF: the first trained using all available GLAS data (coincident with ALS data), the second using GLAS data that are high energy (>28 mJ) Laser 3 footprints acquired during summertime only (as per Mahoney et al. [22]). This restriction is imposed to investigate if such filters yield optimal results for GLAS GF estimates and are not only applicable to GLAS estimates of canopy height.

The RF algorithm is a decision tree ensemble technique employed in regression form here, where a total of 75% of a parent dataset of corresponding response (scaling factor) and predictor information is randomly sampled, defined as the 'training' set. The training set is used to grow a 'tree' of nodes, where a number of predictors are sampled and split via a threshold at each node to maximise information gain or Kullback–Leibler divergence [48]. Node creation is repeated until information gain becomes negligible, thus producing a 'leaf'. The training data selection and tree growth process is repeated 500 times, forming a 'forest' and trained RF model. In order to retrieve predictions from the trained RF model, each line of predictors (of the same format as the trained model) is processed through the previously created 500 trees. The pathway taken to a single leaf per tree is determined by the value of predictors sampled at each node and their values relative to the previously trained node thresholds. The mean of the leaf values obtained from each tree is taken as the final predicted output of the model.

The direct comparisons of unscaled and scaled estimates of GLAS GF with ALS equivalents are quantified via a set of model summary statistics: root mean squared error (RMSE), the coefficient of determination (R^2), fraction of predictions within a factor of 2 of observations (F_2), fractional bias (F_B) and mean prediction bias ($\bar{P}_B$). The difference in model summary statistics is noted between ALS and unscaled GLAS estimates of GF for all available GLAS data and the optimal subset. This comparison was also performed for two forms of scaled GLAS estimates of GF. The first set of refined GLAS GF estimates was modified based on predictions of f made from an RF model trained via all available GLAS data, where the second predictions of f were made via an RF model trained using the optimal GLAS subset only. The percentage difference of each summary statistic is calculated between unscaled and scaled GLAS GF estimates for all and the optimal datasets. Assessing the greatest accuracy gains between all and optimal datasets helps to identify if the restrictions of the latter are worth imposing for predicting large-scale estimates of GF across Australia.

It is important to note that predicted scaling factors made for all data are not expected to exhibit any 'tuning' bias even though the same data was used in RF model training. In the training process, RF grows numerous 'trees' (in this case, 500), each from a random selection of 75% of all available data. During the prediction process, each 'tree' provides a modelled response (500 in total) given the same predictors. The mean of all outputs from all 'trees' is taken as the final modelled response, thus negating any possible data 'tuning' bias.

3. Results

3.1. Consistency Assessment

The sensitivity of the derived GLAS scaling factors is tested with respect to footprint eccentricity and laser campaign (Figure 4). The maximum percent differences of mean and standard deviation between each interval as a function of footprint eccentricity are 2.34% and 1.46%, respectively; where equivalent values as a function of laser campaign are 2.74% and 2.29%, respectively. Mean scaling factor values vary less with eccentricity than the laser campaign, a suspected consequence of the differing number of intervals between each attribute. However, neither attribute alludes to any decay in GLAS's consistency of estimating GF (inferred by no notable differences of scaling factor) as a function of its operational lifetime.

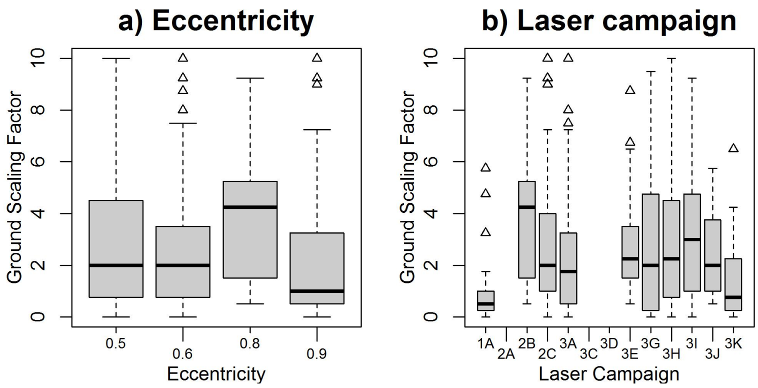

Figure 4. Boxplots illustrating how GLAS scaling factors vary as a function of (**a**) footprint eccentricity and (**b**) laser campaign (Lc). Boxes indicate the median (black line), first (Q1) and third (Q3) quartiles (lower and upper box edges, respectively), up to 1.5-times the Interquartile Range (IQR) beyond the box (whiskers) and outliers greater than 1.5-times the IQR beyond the box (triangles). Note that whiskers extend to a point, which is no more than 1.5-times the IQR from the box; if no outliers exist beyond this range, the whiskers are truncated and do not always appear symmetrical as a result.

3.2. Scaling Factor Sensitivity

The sensitivity of GLAS scaling factors is tested against 13 predictor attributes to test their respective predictive capabilities in imputation methods. An independent estimate of the importance of each attribute is noted by the model fitted sum of squared residuals (Table 2). Only one of the attributes rejected the HC null hypothesis of linearity and, hence, was fitted with a localized non-linear regression model; all other attributes were fitted with linear models. Note that a model performance rating could not be calculated for valley bottom flatness or land cover, as both only exhibited two intervals, which would have resulted in biased model fit.

Table 2. Independent predictor attribute importance as determined by a measure of the Sum of Squared Residuals (SSR). Note, residuals are expressed as a percentage form the true value. The fit type 'L' represents a linear fit, whereas 'NL' represents a non-linear fit, as determined from the Harvey-Collier (HC) test.

Attribute	SSR	Fit Type
Phosphorus	0.01	L
Height	0.01	L
Nitrogen	0.06	L
VCF	0.15	L
Slope	0.16	NL
Aspect	0.20	L
Elevation	0.22	L
pH	0.44	L
NVIS	0.70	L
Soil	1.53	L
Depth	2.00	L
VBF	-	L
Land cover	-	L

Results suggest that the best attributes for predicting GLAS scaling factors are those that exhibit an SSR up to the arbitrary threshold of 0.16: phosphorus, height, nitrogen, VCF and slope. This threshold was selected as the SSR values increase dramatically beyond this point, but are relatively close prior to this value. Scaled GF results were predicted via RF based on these attributes alone and were compared against RF predictions made using all available attributes; no significant difference in outputs was noted between the predicted means informed by a two-sample *t*-test (α = 0.95).

3.3. Gap Fraction Comparisons

ALS estimates of GF are compared by unscaled and scaled GLAS equivalents for all available GLAS data in Figure 5a and for the optimal GLAS subset in Figure 5b. Scaling factor values for individual GLAS footprints are predicted from an RF model trained using the five most important attributes noted above. The associated model summary statistics for each GF comparison are given in Table 3.

Results suggest that unscaled GLAS estimates of GF relate poorly to ALS equivalents for all available and optimal GLAS data, where scaled estimates exhibit a far greater correspondence. However, it is noted where GF is expected to be large (ALS GF > 0.8), scaled estimates tend to overestimate (due to predictions of $f < 1$ in Equation (3)); in this region, unscaled estimates appear closer to corresponding ALS results. The improvement in correspondence between GLAS and ALS GF is greatest when GLAS waveforms are scaled by the RF model trained with optimal GLAS data. Whilst improvements are noted between unscaled and scaled GLAS data when waveforms are scaled by the RF model trained using all available data, the optimal data predictions offer greater improvement for all model summary statistics. GF data appear in clusters centred at approximately 0.1 and 0.8 (ALS) due to the relatively small coincidence between ALS and GLAS data and very little to none of which were acquired for forest regions that exhibit GF between 0.3 and 0.7.

Table 3. Summary statistics (and percent change) for comparisons of ALS and GLAS estimates of GF for unscaled and scaled results per RF predicted scaling factors made across all available GLAS footprints and an optimized subset of GLAS footprints. Note: N is the population size, R^2 the coefficient of determination, RMSE the root mean squared error, F_2 the fraction of predictions within a factor of 2 of observations, F_B the fractional bias and $\bar{P}_B$ the mean prediction bias.

GLAS GF	Dataset	N	R^2	RMSE	F_2	F_B	$\bar{P}_B$
Unscaled	All Data	309	0.77	0.18	0.32	0.20	−0.09
Scaled			0.88	0.11	0.59	−0.01	0.01
% difference			14	−39	84	−105	111
Unscaled	Optimized Data	102	0.83	0.16	0.34	0.21	−0.10
Scaled			0.89	0.09	0.67	−0.01	0.01
% difference			7	−43	97	−105	90

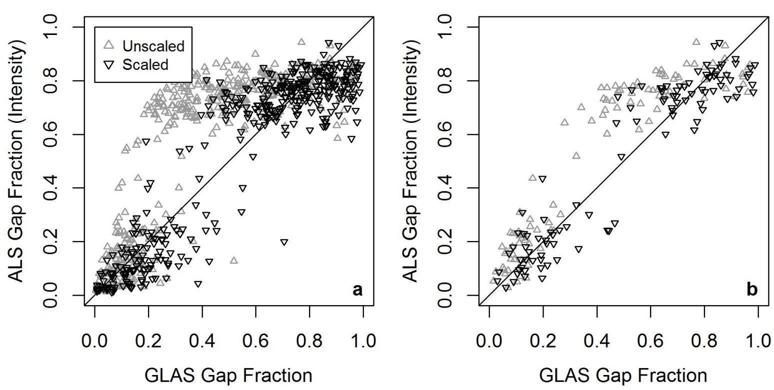

Figure 5. Comparison of ALS and GLAS unscaled (grey) and scaled (black) estimates of GF where scaling factors were predicted via an RF model trained with the best suited predictor attributes using (**a**) all available training data and (**b**) an optimized subset.

4. Discussion

GLAS waveforms require scaling in order to retrieve accurate estimates of GF. This is a legacy of the non-uniform energy distribution with which the sensor illuminated its targets and suggested differences in reflectance values between vegetation and ground surfaces [14,35]. In the novel approach to predicting scaling factors for unique GLAS waveforms based on other remote sensing data products, soil phosphorus and nitrogen contents were identified as two of the most important attributes for scaling factor predictions, suggesting that soil mineral contents are key in influencing the density of vegetation canopies. Furthermore, it is possible that soil texture and colour may vary as a function of these variables also, thereby influencing soil reflectance properties. Vegetation height ranked second most important overall; this is somewhat expected in a forest environment (i.e. considering trees only), although caution is expressed, as this relationship may degrade with the inclusion of additional vegetation, such as shrubs, which can be short, but exhibit high cover densities. The Vegetation Cover Index (VCF) ranked fourth most important, which is a trivial connection between vegetation cover indices. Finally, slope is ranked as fifth most important, indicating that canopy density becomes limited with the vegetation's ability to anchor itself, although this is expected to be somewhat species dependent.

With respect to large-scale derivations of GF via the presented waveform scaling methodology, the use of the five identified most important attributes is expected to be paramount; however, the need for attribute pre-selection will be a function of the selected imputation technique. For example, if RF were to be employed, the a priori identification of the most important attributes becomes somewhat irrelevant as RF is capable of selecting which attributes contribute most to yield the most accurate model. Conversely, when employing other techniques, such as kNN, the selection of an optimal set of predictor attributes is of great importance in the context of overall model accuracy [49]. In such instances, the use of the presented attribute pre-selection analysis is advocated in order to identify attributes that best characterize the prediction/response variable. In the case of RF techniques, attribute pre-selection can still be applied in the pursuit of a parsimonious model, that is, computational processing time is reduced at the cost of marginal losses in overall accuracy. Therefore, by accepting this compromise, models can be applied over large scales with increased efficiency by a priori removing attributes that contribute little to the model; an approach already advocated by McRoberts et al. [49]. However, in the case that a truly parsimonious model is desired, it is suggested that future studies filter predictor attributes that exhibit strong correlation with each other in order to maximise computational efficiency.

The presented attribute pre-selection methodology focussed on the identification of most important attributes for scaling GLAS estimates of GF by analysis of physical relationships only, i.e., neglecting relationships formed by complex interplay in statistical algorithms (such as RF). Such analysis was favoured as knowledge of the relationship between the environment and f is required to understand why f is an inherent requirement when attempting to obtain GF from GLAS (or any waveform system) data. The identification of these attributes will allow a better physical understanding of how such attributes contribute to how f varies as a function of landscape subtleties; this is otherwise unobtainable by sole reliance on statistical ensemble models. The quantification of such physical understandings is not pursued in the current study as the primary objective is to demonstrate a proof-of-concept for refining initial GLAS GF estimates as a function of ALS data. It is acknowledged that the same set of most important attributes could be identified via RF importance measures provided all attributes were employed in model building; however, without additional analysis, a physical understanding of why these were identified would not be known.

Even based on the relatively small ALS-GLAS coincidence noted here, the improved correspondence of GLAS results to ALS equivalents is evident. However, it is noted that GLAS-refined GF in Figure 5a, and Figure 5b are more outlying than unscaled GLAS GF equivalents; this is a somewhat expected result based on the inherent inability of RF to predict extreme values. This is due to RF prediction values being the mean of all ensemble results, which by definition will be less than the maximum and/or (above average) outlying prediction results, which are expected to occur minimally in the ensemble results. Modifications to the RF procedure have been attempted in order to mitigate this issue but at the cost of overall prediction accuracy [50]. Furthermore, as a consequence of little coincidence, few GF data are available between 0.4 and 0.7; hence, any predictions made in this region are expected to exhibit greater uncertainties. Such a gap in results requires additional (coincident) ALS and GLAS data to validate any GF predictions made in this range. Given the increasing availability of ALS data globally, this may be achieved elsewhere across the globe, although any calculations made elsewhere will likely be subject to vastly different environmental characteristics and vegetation species to those noted in this Australia-centric study, which may require a new sensitivity analysis be conducted. Additionally, with increasing ALS data availability, the current methodology for scaling GLAS waveforms is pertinent as initial model training requires scaling factors be determined in areas of data coincidence. With increased ALS-GLAS data coincidence, uncertainties in scaling factor predictions are expected to decrease.

It is important to note that the predictor sensitivity results presented in this study are valid at 250 m spatial resolution only, but are expected to be applicable at large geographies, particularly Australia-wide or even globally. However, it is important to note for the latter that applicability will depend on the availability of similar predictor attributes at a 250 m resolution. The validity of these results is expected to hold for predictions made at coarser resolutions (>250 m) due to the effect of aggregation, which effectively reduces local variability. Conversely, local variability tends to increase with finer resolutions (<250 m); as such, caution is advised before predictions can be made. Such variability cannot be assumed to be adequately described by the predictor sensitivity analysis presented in this study due to the resolution of the available predictor attributes on which the analysis was performed. Prior to making finer resolution predictions, a finer resolution predictor attribute sensitivity analysis is required.

Given the global (spatial) distribution of coincident ALS and GLAS data, caution must be expressed when employing the developed method for deriving scaling factors in extreme environments (i.e., fragmented landscapes, such as northern Canada). The effects of surface fragmentation may neutralize the relationships between scaling factors and predictor attributes as continuous canopy cover within GLAS footprints may be difficult to achieve; this would likely affect the waveform profile and thus any initially retrievable GF, as well as the calculated scaling factor. As a result, it is likely that significantly different landscapes will require a different approach for refining GLAS GF estimates or an independent predictor sensitivity analysis. Furthermore, the effects of high latitudes with respect to GLAS measurements have been briefly documented for canopy heights only [51], but may apply to GLAS derivations of GF, as well. Whilst a latitudinal effect is suspected to influence GLAS GF and waveform scaling factors, the severity of such effects is unknown; a further sensitivity analysis may quantify such effects.

The use of this methodology can also be extended to other waveform instruments, past or future. Of particular interest are the upcoming spaceborne missions of ICESat-2 and the Global Ecosystem Dynamics Investigation (GEDI) LiDAR. The provided GLAS data can provide a baseline estimate of GF at large scales; such future missions may allow change detection and monitoring investigations.

5. Conclusions

A model rooted in ALS data was developed to derive GLAS waveform ground return scaling factors to refine GLAS estimates of GF. Scaling factors were predicted for unique GLAS footprints via an RF model built on the identified most important five (of thirteen) investigated predictor attributes, namely: soil phosphorus and nitrogen contents, vegetation height, MODIS VCF (MOD44B product) and slope. It is important to note that these products may not be best suited to refining GLAS estimates of GF across the globe; however, the presented methodology will enable the identification of the best suitors of available data. Moreover, the presented predictor attribute pre-selection methodology incorporates adaptability in to predicting GLAS GF refinements at large scales. That is, the pre-selection can be ignored if using an imputation (or other) technique that exhibits feature selection, but should be strongly considered for techniques that require an optimal attribute set.

Comparisons of scaled GLAS GF with ALS equivalents (not scaled) corresponded more closely than unscaled counterparts. Two RF models were utilized for predicting scaling factors: one based on all available GLAS data (quality controlled) and an optimal subset, as per Mahoney et al. [22]. Post scaling, the relationship between GLAS and ALS GF was best for predictions trained using the optimal GLAS subset ($R^2 = 0.89$, RMSE = 0.10). This represented the greatest percent change (all model summary statistics considered) in the results with respect to unscaled GF equivalents, suggesting that restricting GLAS data to what is deemed optimal holds value for GF purposes, not only vegetation height.

The sensitivities of GLAS GF scaling factors with respect to ALS equivalents identified in this study offer a first step towards developing a physically-based means of scaling GLAS waveform returns to yield refined estimates of GF. However, results are based on predictor attributes that are unique to Australia that may not be available elsewhere across the globe. The presented methodology allows the identification of which attributes are best suited to predicting GLAS scaling factors wherever predictor data are available. This will allow for the development of refined estimates of GLAS GF at near global scales, the first large-scale estimate of GF from an active sensor. The method can also be applied to other waveform systems, which is of particular use given that upcoming laser altimetry missions, such as the GEDI LiDAR and ICESat-2, are expected to be operational in the near future. The effects of site specific environmental characteristics on scaling factors require investigation in future studies in order to quantify the influence of such characteristics on waveform derivations of canopy density metrics and correction factors at large scales.

Acknowledgments: Hopkinson acknowledges early funding and data support from the Commonwealth Scientific and Industry Research Organisation of the Australian Government, as well as ongoing lab funding through the Campus Alberta Innovates Program. Mahoney acknowledges post-doctoral fellow support through the NSERC (Natural Sciences and Engineering Research Council) funded Amethyst program. MODIS (MOD44B) Vegetation Continuous Fields (VCF) data were obtained from the Global Land Cover Facility (http://www.landcover.org). The National Vegetation Information System (NVIS) data were obtained from the Department of the Environment, Water, Population and Communities (DSEWPaC), Australian Government; NVIS is a collaborative effort of Australia's states and territories. ICESat GLAS data were obtained from the National Snow and Ice Data Center (NSIDC; http://nsidc.org). Airborne Laser Scanning (ALS) and all other predictor data are available through the AusCover (http://www.auscover.org.au); AusCover is the remote sensing data products facility of the Terrestrial Ecosystem Research Network (TERN, http://www.tern.org.au). LiDAR data for Tumbarumba were collected with support from (National Centre for Earth Observation) NCEO EO mission support 2009. Special thanks to Jorg Hacker and Airborne Research Australia (ARA) and Fugro Spatial Solutions for carrying out the airborne campaigns.

Author Contributions: C.M. and C.H. conceived of and designed the experiments. C.M. performed the experiments and analysed the data. N.K. and E.v.G. contributed materials. C.M. wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ALS	Airborne Laser Scanning
ASRIS	Australian Soil Resource Information System
DLCD	Dynamic Land Cover Dataset
FC	Fractional Cover
GEDI	Global Ecosystem Dynamics Investigation
GF	Gap Fraction
GLAS	Geoscience Laser Altimeter System
ICESat	Ice, Cloud and land Elevation Satellite
LAI	Leaf Area Index
LiDAR	Light Detection And Ranging

LVIS	Land, Vegetation, and Ice Sensor
MODIS	Moderate Resolution Imaging Spectroradiometer
NASA	National Aeronautic Space Administration
NVIS	National Vegetation Inventory System
RMSE	Root Mean Squared Error
SLICER	Scanning Lidar Imager of Canopies by Echo Recovery
SRTM	Shuttle Radar Topography Mission
VBF	Valley Bottom Flatness
VCF	Vegetation Continuous Fields

References

1. Hopkinson, C.; Chasmer, L. Testing LiDAR models of fractional cover across multiple forest ecozones. *Remote Sens. Environ.* **2009**, *113*, 275–288.
2. Hall, F.G.; Botkin, D.B.; Strebel, D.E.; Woods, K.D.; Goetz, S.J. Large-scale patterns of forest succession as determined by remote sensing. *Ecology* **1991**, *72*, 628–640.
3. Welles, J.M.; Cohen, S. Canopy structure measurement by gap fraction analysis using commercial instrumentation. *J. Exp. Bot.* **1996**, *47*, 1335–1342.
4. Korhonen, L.; Korpela, I.; Heiskanen, J.; Maltamo, M. Airborne discrete-return LiDAR data in the estimation of vertical canopy cover, angular canopy closure and leaf area index. *Remote Sens. Environ.* **2011**, *115*, 1065–1080.
5. Tian, Y.H.; Wang, Y.J.; Zhang, Y.; Knyazikhin, Y.; Bogaert, J.; Myneni, R.B. Radiative transfer based scaling of LAI retrievals from reflectance data of different resolutions. *Remote Sens. Environ.* **2003**, *84*, 143–159.
6. Fernandes, R.A.; Miller, J.R.; Chen, J.M.; Rubinstein, I.G. Evaluating image-based estimates of leaf area index in boreal conifer stands over a range of scales using high-resolution CASI imagery. *Remote Sens. Environ.* **2004**, *89*, 200–216.
7. Parker, G.G.; Lefsky, M.A.; Harding, D.J. Light transmittance in forest canopies determined using airborne laser altimetry and in-canopy quantum measurements. *Remote Sens. Environ.* **2001**, *76*, 298–309.
8. Lovell, J.L.; Jupp, D.L.; Culvenor, D.S.; Coops, N.C. Using airborne and ground-based ranging LiDAR to measure canopy structure in Australian forests. *Can. J. Remote Sens.* **2003**, *29*, 607–622.
9. Todd, K.W.; Csillag, F.; Atkinson, P.M. Three-dimensional mapping of light transmittance and foliage distribution using lidar. *Can. J. Remote Sens.* **2003**, *29*, 544–555.
10. Barilotti, A.; Turco, S.; Alberti, G. LAI Determination in Forestry Ecosystem by LiDAR Data Analysis. In Proceedings of Workshop 3D Remote Sensing in Forestry Report, Vienna, Austria, 14–15 February 2006.
11. Morsdorf, F.; Kötz, B.; Meier, E.; Itten, K.; Allgöwer, B. Estimation of LAI and fractional cover from small footprint airborne laser scanning data based on gap fraction. *Remote Sens. Environ.* **2006**, *104*, 50–61.
12. Thomas, V.; Treitz, P.; McCaughey, J.H.; Morrison, I. Mapping stand-level forest biophysical variables for a mixedwood boreal forest using lidar: An examination of scanning density. *Can. J. For. Res.* **2006**, *36*, 34–47.
13. Hopkinson, C.; Chasmer, L. Using discrete laser pulse return intensity to model canopy transmittance. *Photogramm. J. Finl.* **2007**, *20*, 16–26.
14. Lefsky, M.A.; Cohen, W.; Acker, S.; Parker, G.G.; Spies, T.; Harding, D. LiDAR remote sensing of the canopy structure and biophysical properties of Douglas-fir western hemlock forests. *Remote Sens. Environ.* **1999**, *70*, 339–361.
15. Yang, W.; Ni-Meister, W.; Lee, S. Assessment of the impacts of surface topography, off-nadir pointing and vegetation structure on vegetation LiDAR waveforms using an extended geometric optical and radiative transfer model. *Remote Sens. Environ.* **2011**, *115*, 2810–2822.
16. North, P.; Rosette, J.; Suárez, J.; Los, S. A Monte Carlo radiative transfer model of satellite waveform LiDAR. *Int. J. Remote Sens.* **2010**, *31*, 1343–1358.
17. Luo, S.; Wang, C.; Li, G.; Xi, X. Retrieving leaf area index using ICESat/GLAS full-waveform data. *Remote Sens. Lett.* **2013**, *4*, 745–753.
18. Tang, H.; Dubayah, R.; Swatantran, A.; Hofton, M.; Sheldon, S.; Clark, D.B.; Blair, B. Retrieval of vertical LAI profiles over tropical rain forests using waveform LiDAR at La Selva, Costa Rica. *Remote Sens. Environ.* **2012**, *124*, 242–250.

19. Tang, H.; Brolly, M.; Zhao, F.; Strahler, A.H.; Schaaf, C.L.; Ganguly, S.; Zhang, G.; Dubayah, R. Deriving and validating Leaf Area Index (LAI) at multiple spatial scales through lidar remote sensing: A case study in Sierra National Forest, CA. *Remote Sens. Environ.* **2014**, *143*, 131–141.
20. Tang, H.; Dubayah, R.; Brolly, M.; Ganguly, S.; Zhang, G. Large-scale retrieval of leaf area index and vertical foliage profile from the spaceborne waveform lidar (GLAS/ICESat). *Remote Sens. Environ.* **2014**, *154*, 8–18.
21. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
22. Mahoney, C.; Hopkinson, C.; Held, A.; Kljun, N.; van Gorsel, E. ICESat/GLAS Canopy Height Sensitivity Inferred from Airborne LiDAR. *Photogramm. Eng. Remote Sens.* **2016**, *82*, 351–363.
23. Van Gorsel, E.; Leuning, R.; Cleugh, H.A.; Keith, H.; Kirschbaum, M.U.; Suni, T. Application of an alternative method to derive reliable estimates of nighttime respiration from eddy covariance measurements in moderately complex topography. *Agric. For. Meteorol.* **2008**, *148*, 1174–1180.
24. Hopkinson, C.; Lovell, J.; Chasmer, L.; Jupp, D.; Kljun, N.; van Gorsel, E. Integrating terrestrial and airborne LiDAR to calibrate a 3D canopy model of effective leaf area index. *Remote Sens. Environ.* **2013**, *136*, 301– 314.
25. Brenner, A.C.; Zwally, H.J.; Bentley, C.R.; Csathó, B.M.; Harding, D.J.; Hofton, M.A.; Minster, J.B.; Roberts, L.; Saba, J.L.; Thomas, R.H.; et al. *Geoscience Laser Altimeter System (GLAS) Algorithm Theoretical Basis Document 4.1: Derivation of Range and Range Distributions From Laser Pulse Waveform Analysis for Surface Elevations, Roughness, Slope, and Vegetation Heights*; NASA Goddard Space Flight Center: Greenbelt, MD, USA, 2003.
26. Zwally, H. Schutz, H.; Bentley, C.; Bufton, J.; Herring, T.; Minster, J.; Spinhirne, J.; Thomas, R. *GLAS/ICESat L2 Global Land Surface Altimetry Data, Version 33*; National Snow and Ice Data Center: Boulder, CO, USA, 2011.
27. Abshire, J.B.; Sun, X.; Riris, H.; Sirota, J.M.; McGarry, J.F.; Palm, S.; Yi, D.; Liiva, P. Geoscience Laser Altimeter System (GLAS) on the ICESat mission: On-orbit measurement performance. *Geophys. Res. Lett.* **2005**, *32*, L21S02.
28. Harding, D.J.; Carabajal, C.C. ICESat waveform measurements of within-footprint topographic relief and vegetation vertical structure. *Geophys. Res. Lett.* **2005**, *32*, L21S10.
29. Neuenschwander, A.L.; Urban, T.J.; Gutierrez, R.; Schutz, B.E. Characterization of ICESat/GLAS waveforms over terrestrial ecosystems: Implications for vegetation mapping. *J. Geophys. Res. Biogeosci.* **2008**, *113*, G02S03.
30. Duong, V.H.; Lindenbergh, R.; Pfeifer, N.; Vosselman, G. Single and two epoch analysis of ICESat full waveform data over forested areas. *Int. J. Remote Sens.* **2008**, *29*, 1453–1473.
31. Lymburner, L.; Tan, P.; Mueller, N.; Thackway, R.; Lewis, A.; Thankappan, M.; Randall, L.; Islam, A.; Senarath, U. *The National Dynamic Land Cover Dataset*; Commonwealth of Australia (Geoscience Australia): Canberra, Australia, 2011.
32. Los, S.; Rosette, J.; Kljun, N.; North, P.; Chasmer, L.; Suárez, J.; Hopkinson, C.; Hill, R.; van Gorsel, E.; Mahoney, C.; et al. Vegetation height and cover fraction between 60°S and 60°N from ICESat GLAS data. *Geosci. Model Dev.* **2012**, *5*, 413–432.
33. Mahoney, C.; Kljun, N.; Los, S.O.; Chasmer, L.; Hacker, J.M.; Hopkinson, C.; North, P.R.J.; Rosette, J.A.B.; van Gorsel, E. Slope Estimation from ICESat/GLAS. *Remote Sens.* **2014**, *6*, 10051–10069.
34. Rosette, J.A.B.; North, P.R.J.; Suárez, J.C. Vegetation height estimates for a mixed temperate forest using satellite laser altimetry. *Int. J. Remote Sens.* **2008**, *29*, 1475–1493.
35. Means, J.E.; Acker, S.A.; Harding, D.J.; Blair, J.B.; Lefsky, M.A.; Cohen, W.B.; Harmon, M.E.; McKee, W.A. Use of large-footprint scanning airborne lidar to estimate forest stand characteristics in the Western Cascades of Oregon. *Remote Sens. Environ.* **1999**, *67*, 298–308.
36. Lefsky, M.A.; Cohen, W.B.; Parker, G.G.; Harding, D.J. Lidar Remote Sensing for Ecosystem Studies Lidar, an emerging remote sensing technology that directly measures the three-dimensional distribution of plant canopies, can accurately estimate vegetation structural attributes and should be of particular interest to forest, landscape, and global ecologists. *Bioscience* **2002**, *52*, 19–30.
37. Genuer, R.; Poggi, J.M.; Tuleau-Malot, C. Variable selection using random forests. *Pattern Recognit. Lett.* **2010**, *31*, 2225–2236.
38. Burrough, P. Principles of geographical information systems for land resources assessment. *Geocarto Int.* **1986**, *1*, 54.
39. Gallant, J.C.; Dowling, T.I.; Read, A.M.; Wilson, N.; Tickle, P.; Inskeep, C. *1 Second SRTM Derived Digital Elevation Models User Guide*; Commonwealth of Australia (Geoscience Australia): Canberra, Australia, 2011.
40. Gallant, J.C.; Dowling, T.I. A multiresolution index of valley bottom flatness for mapping depositional areas. *Water Resour. Res.* **2003**, *39*, 1347.

41. DiMiceli, C.M.; Carroll, M.L.; Sohlberg, R.A.; Huang, C.; Hansen, M.C.; Townshend, J.R.G. *Annual Global Automated MODIS Vegetation Continuous Fields (MOD44B)*; University of Maryland: College Park, MD, USA, 2011.
42. ESCAVI. *Australian Vegetation Attribute Manual: National Vegetation Information System*; Technical Report, Executive Steering Committee for Australian Vegetation Information (ESCAVI); Report No. 6.0; Department of Environment and Heritage: Canberra, Australia, 2003.
43. Mahoney, C.; Hopkinson, C.; Held, A.; Simard, M. Continental-Scale Canopy Height Modeling by Integrating National, Spaceborne, and Airborne LiDAR Data. *Can. J. Remote Sens.* **2016**, *42*, 574–590.
44. Australian Soil Resources Information System (ASRIS). *Australian Soil Resource Information System*; CSIRO: Canberra, Australia, 2014.
45. Rossel, V.R.A.; Chen, C.; Grundy, M.J.; Searle, R.; Clifford, D.; Campbell, P.H. The Australian three-dimensional soil grid: Australia's contribution to the GlobalSoilMap project. *Soil Res.* **2015**, *53*, 845–864.
46. Harvey, A.C.; Collier, P. Testing for functional misspecification in regression analysis. *J. Econom.* **1977**, *6*, 103–119.
47. Cleveland, W.S.; Grosse, E.; Shyu, M.J. *Statistical Models in S*; Chapter Local Regression Models; Chapman and Hall: New York, NY, USA, 1992; pp. 309–376.
48. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86.
49. McRoberts, R.E.; Næsset, E.; Gobakken, T. Optimizing the k-Nearest Neighbour technique for estimating forest aboveground biomass using airborne laser scanning data. *Remote Sens. Environ.* **2015**, *163*, 13–22.
50. Simard, M.; Pinto, N.; Fisher, J.B.; Baccini, A. Mapping forest canopy height globally with spaceborne lidar. *J. Geophys. Res. Biogeosci.* **2011**, *116*, G04021.
51. Nelson, R. Model effects on GLAS-based regional estimates of forest biomass and carbon. *Int. J. Remote Sens.* **2010**, *31*, 1359–1372.