

Revised method for forest canopy height estimation from Geoscience Laser Altimeter System waveforms

Michael A. Lefsky^a, Michael Keller^{b,c,d}, Yong Pang^a,
Plinio B. de Camargo^d, Maria O. Hunter^{c,d}

^a Center for Ecological Analysis of Lidar, Department of Natural Resources, Colorado State University, 131 Forestry Building, Fort Collins, CO 80523-1472

lefsky@cnr.colostate.edu

^b USDA Forest Service, International Institute of Tropical Forestry, Jardín Botánico Sur, 1201 Calle Ceiba, San Juan, PR 00926-1119

^c Complex Systems Research Center, University of New Hampshire, Durham NH 03824

^d CENA, Universidade de São Paulo, Piracicaba, São Paulo, Brazil

Abstract. The vertical extent of waveforms collected by the Geoscience Laser Altimeter System (onboard ICESat - the Ice, Cloud, and land Elevation Satellite) increases as a function of terrain slope and footprint size (the area on the ground that is illuminated by the laser). Over sloped terrain, returns from both canopy and ground surfaces can occur at the same elevation. As a result, the height of the waveform (waveform extent) is insufficient to make estimates of tree height on sloped terrain, and algorithms are needed that are capable of retrieving information about terrain slope from the waveform itself. Early work on this problem used a combination of waveform height indices and slope indices from a digital elevation model (DEM). A second generation algorithm was developed using datasets from diverse forests in which forest canopy height has been estimated in the field or by via airborne lidar. Forest types considered in this paper include evergreen needleleaf, deciduous broadleaf and mixed stands in temperate North America, and tropical evergreen broadleaf forests in Brazil. The algorithm described eliminates the need for a DEM, and estimates forest canopy height with an RMSE of 5 m (83% of variance in forest canopy height explained).

Keywords: ICESat, GLAS, forest height, remote sensing.

1 INTRODUCTION

The Geoscience Laser Altimeter System (GLAS), onboard the Ice, Cloud, and land Elevation Satellite (ICESat), is a waveform sampling lidar sensor- it emits short duration (5 ns) laser pulses towards the land surface and records the echo of those pulses as they reflect off the ground surface. When that surface is vegetated, the return echoes, or waveforms, are a function of the vertical distribution of vegetation and ground surfaces within the area illuminated by the laser (the footprint, which has varied between 52 and 90 meters diameter during the GLAS mission with associated changes in eccentricity and the major axis azimuth). For forests on level ground, discrete peaks in the waveform separate the height distribution of reflecting canopy surfaces from that of the underlying ground [1]. In these cases, forest stand height can be calculated as the difference between the elevation of the first returned energy minus the mean elevation of the ground return [2,3].

The vertical extent of each waveform increases as a function of terrain slope and footprint size (the area on the ground that is illuminated by the laser), as modified by the spatial pattern of ground surfaces visible to the laser. Over sloped terrain, returns from both canopy and ground surfaces can occur at the same elevation [1]. As a result, information on the vertical

extent of the waveform extent is insufficient to make estimates of tree height on sloped terrain. Given forest stands of uniform height, highly accurate ancillary topography and information on the spatial pattern of visible ground surfaces, separation of waveforms into ground and canopy components is conceptually and algorithmically simple. However, most forests are non-uniform in height, adequate topographic characterization is rare, and the spatial pattern of visible ground surfaces is usually unknown. Therefore, algorithms are needed that are capable of retrieving information about terrain slope, stand uniformity and the vertical distribution of visible ground surfaces from the waveform itself.

2 METHODS

2.1 Study areas and height data collection

Seven geographically distinct study areas are considered in this work; five study areas have coincident GLAS waveforms and field estimates of height and aboveground biomass, two study areas have coincident GLAS waveforms and estimates of mean height from airborne lidar remote sensing. Of the study areas with field estimates, two were selected to represent coniferous (Oregon, USA) and deciduous (Tennessee, USA) forest types located on high slopes (between 0% and 27% in Oregon and between 1 and 24% in Tennessee). The remaining three study areas with field estimates (in the municipalities of Santarem, Para State; Manaus, Amazonas State; and Canarana, Mato Grosso State; all in Brazil) were selected to represent a range of dry season duration (months with less than 100 mm of rain) within the Amazon basin. Dry season duration was selected as an environmental variable likely to summarize structural trends across the Amazon [4]. The two study areas with coincident discrete return lidar (collected from airborne platforms) are located in mixed coniferous forest in the Tahoe National Forest (California, USA) and eastern deciduous forest in the Bartlett Experimental Forest (New Hampshire, USA). Terrain slope in the Tahoe National Forest ranges from 0 to 60% and for Bartlett Experimental forest, from 3 to 30%)

2.1.1 Study areas with coincident field data

The Oregon sites are in the Willamette National Forest and are predominately associated with temperate coniferous forests of Douglas-fir and western hemlock (*Psuedotsuga menziesii*, *Tsuga heterophylla*) 45 km south of the H.J. Andrews Experimental Forest (44° 00' N, 122° 40' W). The Tennessee sites are in Great Smoky Mountains National Park (35° 34' N, 83° 43' W), and are associated with both northern hardwoods and mixes of hardwoods and pines. Dominant species include oaks (*Quercus* spp.) and other hardwood species, as well as Virginia and white pine (*Pinus virginiana*, *P. strobus*). These sites are more fully described in [5].

There are three study areas in the Amazon Basin. The Reserva Cuieiras near Manaus (2° 35' S, 60° 6' W) experiences a 2-3 month dry season and is composed mainly of old growth forest. In this forest there is a notable difference in structure and biomass between forests on clay plateaus and those in sandy valleys. In general, the plateau forests are taller and have more abundant large trees than the forests in the valleys. The Tapajos National Forest near Santarem (3° 01' S, 54° 57' W) experiences a 4-5 month dry season, and includes plots in old growth forests [6], and secondary forests of a range of ages and biomass densities. The Tanguro site (12° 51' S, 52° 24' W) experiences 5-6 months of dry season and has a more open canopy than the other two Amazon sites. These three sites cover most of the range of dry season length in the Amazon region, but exclude areas that experience less than 2 months of dry season.

For study areas where field plots are located coincident with ICESat footprints we measured forest canopy properties. In the two U.S. sites, we stratified potential plots by

waveform extent and an SRTM-derived slope index, and then randomly selected plots in each class in order to obtain a representative sample. In the Amazon the difficulty of reaching plots selected at random was too great, and we selected plots on the basis of their proximity (up to 2.5 km) to existing roads and open fields.

We modified plot layout and sampling procedures in Tennessee and Oregon from those of the local Forestry Inventory and Analysis programs of the United States Department of Agriculture Forest Service [7]. For trees with diameter at breast height (DBH) greater than 12.7 cm, we recorded DBH, species, and height on four 7.32 m radius subplots (three subplots located 36.6 meters from a central subplot at azimuths of 360°, 120°, and 240°). To ensure sampling of uncommon large trees at the Oregon study area we also tallied all trees with DBH greater than 61 cm on four annular plots 35.9 m in diameter and centered on each of the subplots, and tallied all trees with DBH greater than 81 cm in a single 112.9 m diameter plot centered on the central subplot.

When field data from the Oregon site was analyzed, we determined that field-measured heights were higher than expected for a given DBH (as much as 25% taller for the tallest trees). In particular, the observed frequency of tree heights greater than 65 m was unlikely for forests in this area. While the reason for these overestimates is not certain, we suspect that it is related to below freezing temperatures during field data collection (although the laser rangefinder we used for height estimates is rated to -30° C). To correct for this effect, we applied the methods for statistical imputation described in [8]. Imputation selects a stand-in data value (in this case, tree height) from a database of tree characteristics, using a similarity function which relates the imputed variable to other related variables (in this case, DBH, species and crown class). Two advantages of this approach are that the variability in estimated variables matches those of measured and that multivariate relationships of the data are preserved [9].

For old growth forest in Brazil, we established a main plot (20 x 75 m) along the GLAS transect and two perpendicular side plots (40 x 27.5 m each). In these plots, DBH and maximum height were tallied for all trees with DBH greater than 35 cm. Within the main plot, DBH for all trees with DBH between 10 and 35 cm were recorded on a subplot (10 x 75 m); for a 30% or greater sub-sample of these smaller trees, we recorded maximum height. For secondary forests, we sampled using various densities of randomly-located subplots along a 75 m long transect; where subplot density varied as a function of stem density (19 2 x 2 m plots in recently abandoned agricultural fields, eight or nine 4 x 4 m plots in secondary forests). All stems greater than 10 cm DBH were measured in the sub-plots. Maximum heights of trees in the subplots and of the tallest tree in a 75 m circle centered on the footprint were measured.

2.1.2 Study areas with coincident airborne lidar data

Height indices from airborne lidar remote sensing are highly correlated with field estimates of maximum and mean canopy height [10]. In the context of deriving relationships between GLAS waveforms and forest canopy heights, these height indices have numerous advantages over field estimate of height, including their highly accurate geolocation and high sampling density. A further advantage is the correspondence between the airborne lidar data and the three dimensional geometry of the canopy. Field estimates are based on the total height of individual trees, without reference to crown geometry of individual trees.

Airborne lidar data was collected at Bartlett Experimental Forest in New Hampshire, USA in August of 2005 and at Tahoe National Forest in northern California in September 2005. Old-growth stands at Bartlett forest contain beech (*Fagus grandifolia*), yellow birch (*Betula alleghaniensis*), sugar maple (*Acer saccharum*), and eastern hemlock (*Tsuga canadensis*), while red maple (*Acer rubrum*), paper birch (*B. papyrifera*) and aspen (*Populus tremuloides*)

occupy sites that were once cleared. Dominant species at Tahoe National Forest study area are white fir (*Abies concolor*), red fir (*A. magnifica*) and Douglas fir (*Pseudotsuga menziesii*).

Lidar data was collected using an Optec ALTM 2025/2050 sensor, operated to achieve an average 1 m posting between individual lidar observations. Accuracy reports provided to the investigator by the lidar contractor indicate that the RMS errors of the x,y,z coordinates for each point are within acceptable tolerances (vertical error < 15 cm, horizontal error < 50 cm, as estimated on smooth level surfaces). In cases in which the lidar shots do not reach the actual ground surface (e.g. when dense vegetation is present) the minimum elevation of the lidar shots within a grid cell is not a good estimate of the actual elevation. In these cases, the lowest elevation recorded most often reflects a dense layer of vegetation that the lidar cannot penetrate. Using the first and last return lidar data at each study site, a 2 meter resolution “bare earth” digital elevation model was developed using the Tiffs program (Toolbox for Lidar Data Filtering and Forest Studies [11]). Tiffs uses a morphological approach to develop a “bare earth surface” and achieved the best overall performance when tested with the benchmark dataset provided by ISPRS (International Society of Photogrammetry and Remote Sensing, [11]). Lidar points above the bare earth DEM are considered to be vegetation canopy returns.

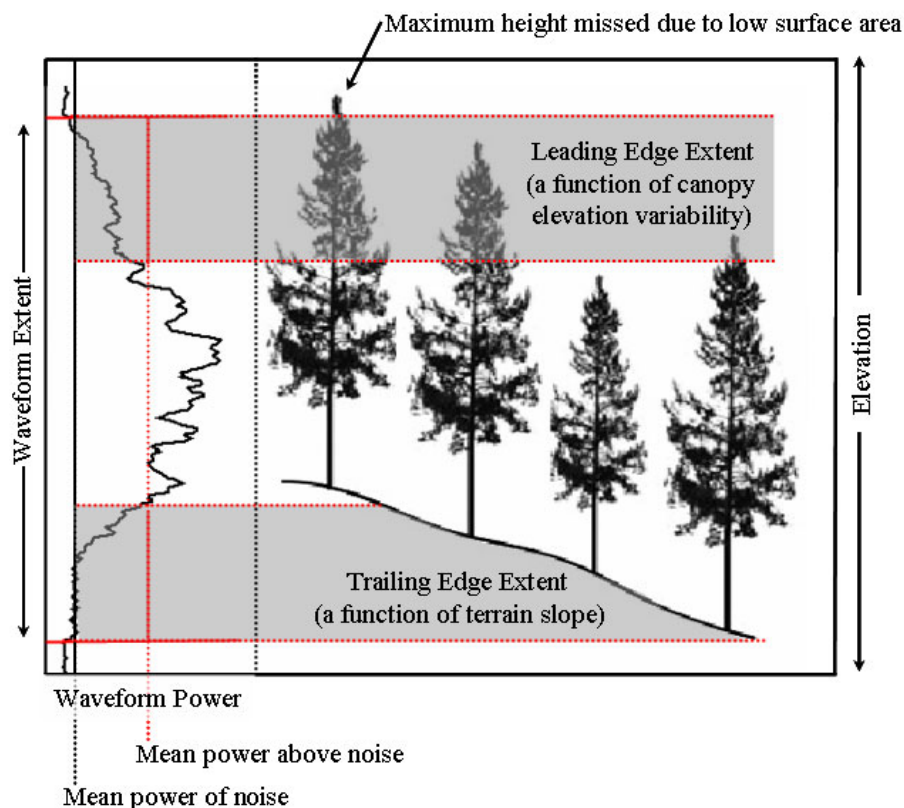


Fig. 1. Definition of total waveform, leading and trailing edge extents, and their relationship to forest canopy structure.

2.2 GLAS data- geographic positions and waveform processing

The ICESat data we used here were from cloud-free profiles acquired between October 2003 and June 2006 [12]. Data for the Amazon study area were collected during observation periods 3a, 3b and 3c; for the Oregon study area during 2a and 2c, for the Tennessee study area during 2a, 2c and 3d; and for the Bartlett and Tahoe study areas during 3d. We excluded data not suitable for determining waveform extent due to upper signal truncation, saturation or very low signal-to-noise ratio [1]. We obtained geolocated footprint locations from the GLA06 Global Elevation Data Product or, where available, from the GLA14 Global Land Surface Altimetry Data Product. See [1] for details of geolocation.

Waveform extent is defined as the vertical distance between the first and last elevations at which the waveform energy exceeds a threshold level (Fig. 1). In this work, the threshold was determined using ICESat data product estimates of the mean and standard deviation of background noise (ICESat product variables D_4NSBGMEAN and D_4NSBGSDDEV). The threshold was set to the mean background noise plus 4.5 times the standard deviation, where 4.5 is a constant derived from experience and supported by analyses comparing waveform extent from GLAS and simulated from coincident airborne lidar. At the leading edge of the waveform, the “signal start” threshold crossing indicates the elevation of the uppermost foliage and/or branches that were detected, and the trailing edge threshold crossing indicates the elevation of the lowest illuminated surface, or the “signal end” [1]. Where sufficient laser energy is reflected from the ground, “signal end” crossing represents the lowest detected ground surface.

Removal of the effects of terrain slope and canopy height variability relies on two indices of waveform structure. The trailing edge extent (Fig. 1) is most closely related to terrain slope; the leading edge extent (Fig. 1) is related to canopy height variability and must be applied to estimate mean tree height rather than maximum tree height. The trailing edge extent is calculated from the waveform as the height difference between the lowest elevation at which the signal strength of the waveform is half of the maximum signal above the background noise value, and the elevation of the signal end. Similarly, the leading edge extent is determined as the height difference between the elevation of the signal start and the first elevation at which the waveform is half of the maximum signal above the background noise value.

2.3 Statistical Analysis

Stepwise multiple regressions were used to find a parsimonious relationship between waveform indices and tree heights. In previous work [5] we chose to estimate maximum tree height, which does not require any leading edge correction factor and can be easily compared among field and lidar datasets. However, when the upper canopy surface height is variable, it is possible that only a single tree will have the maximum height, and may not return enough energy to be detectable in the waveform; this is especially true in forests dominated by trees with conical crowns (e.g. the Tahoe and Oregon study areas). In these cases, the waveform extent may be less than the height of the tallest tree, a situation we observed in this dataset. In this paper’s datasets, waveform extent is always greater than mean height of the dominant and co-dominant stems. In addition, maximum height (which relies on a single observation) is subject to large sampling error in the field, in contrast to mean tree height which is the average of a number of observations. In this work, we estimate the mean height of the forest plots.

Whereas in [5] we estimated maximum tree height directly, in these analyses the dependant variable in the statistical analysis is the “Height Correction Factor” – the difference between waveform extent and the observed mean tree height. Experience indicated that regressions performed with tree height as the dependant variable were difficult to apply outside of the study areas where they were parameterized because the distribution of forest

tree heights in the model dataset influenced model parameters. Furthermore, rather than keeping the model weight for waveform extent near 1.0 (which would preserve the physical meaning of the resulting model), the parameter for waveform extent would co-vary with other model parameters. In contrast, regressions to estimate the height correction factor have proven to be more easily interpretable in physical terms, and the height correction factor can be directly subtracted from waveform extent to estimate mean height.

While the height correction factor is the dependent variable in this analysis, the height correction is directly dependant on mean tree height (the mean height of the dominant and co-dominant trees or the mean height of lidar returns). In comparison to maximum height, mean height is more sensitive to differences to field methods. For instance, maximum height is not subject to the definition or application of tree crown class (i.e. dominant, co-dominant) which may vary between sites. To use mean height as a dependant variable we must provide, in our methods, for tests that evaluate the potential differences in the canopy structure, crown definition, and sampling method at each site that might lead to differences in the relationship between field estimates of mean height, the leading and trailing edge extents, and waveform extent. We used a two-step process to test for differences in the mean tree height associated with the same waveform extent and other factors. In the first step, we used study area as a contrast variable in our regressions. Contrast variables were coded in a hierarchical manner, in which the differences in the mean residuals of the estimate are considered and combinations of levels are formed that best separate the means of the responses. Each group is further subdivided into its best separated subgroups until there is one less contrasts than the number of levels [13]. Significant coefficients for these variables indicate that the contrast between the groups of sites is significant, with the coefficient indicating the magnitude of the difference. In the resulting equations, contrast variables are denoted using the following syntax:

(Coefficient * [Class A & Class B vs. Class C & Class D])

The contrast variable (within the brackets) evaluates to “1” for all plots in Classes A & B and 0 for Classes C & D. This value (similar to the binary value used in dummy variable regression) can then be multiplied by the coefficient.

While contrast variables test for mean differences in the estimates there remains the question of whether the slopes between the independent variables and observed mean tree heights are different from each other. One method to test for these differences would be to take the product of each independent variable with the contrast variables. We decided against this procedure, as it would have increased the number of independent variables and therefore the probability of over-fitting the statistical model. Instead, we post-hoc regressed estimated and observed values for each site independently and tested for differences between their slopes and intercepts, and those expected for an identity relationship (i.e. slope of 1 and intercept of 0). To test for differences from the identity relationship, we used reduced major axis regressions [14] to derive the coefficients of equations relating estimated and observed mean tree height for each study area, as well as associated regression statistics. The advantage of the reduced major axis is that it avoids the assumption that there is no variance in the independent variables, which is used in ordinary least squares regression. Bootstrap analysis, using random subsets of data selected with replacement from, and with the same size, as the initial dataset was then used to generate distributions of the slope and intercept of the resulting regression equations. If the resulting 5% and 95% confidence intervals of the slope and intercept parameters included the values expected from an identity relationship, then no significant difference was detected. Details of these methods can be found in [15]

3 RESULTS

3.1 Study areas

Minimum, mean and maximum heights (Table 1) for the five study areas reveal limitations with the Tennessee (and to a lesser extent, Oregon and Bartlett) dataset. The minimum mean heights for these datasets are much larger (12 - 36% of these sites maximum values) than the minimum values observed in the Amazon and Tahoe datasets (<7%) where we sample almost all the range of mean height. This is especially relevant in interpreting some results (e.g. the percent of variance explained) from the Tennessee study area.

3.2 Statistical analysis

Stepwise multiple regression was performed in two steps. In the first step, multiple transforms of the leading and trailing edge extents were used to model the correction factor required to reconcile waveform extent and mean canopy height. The results of the first step are presented here for trailing and leading edge variables separately (although they were both included in the same regression)

Trailing Edge Correction Factor =

$$3.4 * \sqrt{\text{Trail}} + 0.92 * \text{Trail} - 88.5 * \frac{\text{Trail}}{\text{Extent}} + 2049.5 + \frac{\text{Trail}}{\text{Extent}^2} - 14171.4 * \frac{\text{Trail}}{\text{Extent}^3}, \quad (1)$$

where Trail is the Trailing Edge Extent, and Extent is the waveform extent.

$$\text{Leading Edge Correction Factor} = 0.72 * \text{Lead} - 21.8 \frac{\text{Lead}}{\text{Extent}}, \quad (2)$$

where Lead is the Leading Edge Extent

In the second step of analysis, the multiple transformations of the trailing edge indices are summarized as a trailing edge correction factor; a similar leading edge correction factor is defined as well, which results in the following equation:

$$\text{Correction Factor} = 8.96 + (1.52 * \text{Leading Edge Extent Factor}) + (1.14 * \text{Trailing Extent Factor}) - (4.02 * [\text{Amazon \& Bartlett vs. Oregon, Tahoe, \& Tennessee}]) + (0.81 * [\text{Oregon vs. Tahoe \& Tennessee}]) + (2.06 * [\text{Tahoe vs. Tennessee}]), \quad (3)$$

The resulting equation explains 93% of the variability in the required correction factor, with an RMSE of 5 m (Fig. 2). Mean tree height is simply calculated as Mean Tree Height = Waveform Extent - Correction Factor

The last three terms in the correction factor equation (3) are the hierarchical contrasts that account for differences in the correction factor at each of the sites. The sum of the multiple contrasts results in the offsets for each site (Table 2). The two conifer sites have offsets to the correction factors that indicate shorter mean heights than expected (a smaller correction factor leads to a larger mean height), while the two fully broadleaf sites offsets that indicate taller mean heights; the Tennessee site, with mixed composition, lies between these two.

These estimates of mean tree height explain 83% of variance with the same RMSE of 5 m (Fig. 3). Results of the bootstrap analysis of the estimated vs. observed relationships for each

Table 1. Mean height of Dominant and Co-dominant stems for sites with coincident field plots or mean height of lidar returns for sites with coincident airborne lidar

Study Area	n	Mean Height (m)		
		Mean	Minimum	Maximum
Amazon	52	17.2	1.4	31.6
Bartlett	20	16.8	6.6	19.9
Oregon	20	24.2	6.0	52.1
Tennessee	23	19.7	12.3	34.2
Tahoe	83	13.0	1.9	28.1

Table 2: Sum of site contrasts for each site

Study Area	Correction Factor Offset	Mean Correction Factor without Site Contrast
Amazon	-4.02	16.3
Bartlett	-4.02	17.3
Oregon	4.83	23.5
Tahoe	5.27	27.2
Tennessee	1.15	34.9

Table 3. Tests for differences in slope and intercept of the predicted vs. observed mean forest height taken for each site separately

	n	r ²	RMSE	B0 _{iter}	B1 _{iter}	B0 _{p95}	B0 _{p05}	B1 _{p95}	B1 _{p05}	B0 _{se}	B1 _{se}
Amazon	52	83	4.3	-0.02	0.97	2.11	-2.26	1.11	0.85	0.0	1.0
Bartlett	20	82	2.3	1.77	0.89	8.80	-2.82	1.18	0.49	0.1	2.9
Oregon	20	90	5.9	-0.50	1.08	3.92	-5.71	1.32	0.89	0.1	2.4
Tahoe	83	75	5.5	0.78	0.95	2.19	-0.72	1.09	0.83	0.0	0.7
Tenn.	23	40	4.8	0.66	1.01	6.68	-7.23	1.41	0.70	0.1	3.4
All	198	83	4.9	0.25	0.97	1.44	-0.97	1.05	0.90	0.0	0.6

site (Table 3) indicate that no individual sites have intercepts or slopes that are significantly different from those expected with an identity relationship. With the exception of the Tennessee sites, correlation coefficients vary between 75% and 90% of variance; the low value of variance (40%) explained at the Tennessee site is most likely related to the narrow range of observed values of mean height at that site. Values of the RMSE for mean height generally range from 4.35 to 5.91 with Tennessee falling in the middle of the distribution, suggesting that our interpretation of that site's R^2 is correct. Bartlett is the exception, with the lower RMSE of 2.31

3.3 Analysis of Waveform Indices

The three waveform height indices used in this work are waveform extent, leading edge extent and trailing edge extent of the waveform. The primary challenge in using GLAS data for estimation of forest tree height is illustrated in Fig. 4; waveform extent only explains 22% of variance in mean tree height; an estimate of mean tree height regressed on waveform extent would have an RMSE of 7.4 m. Due to the heteroscedastic residuals, these statistics underestimate the magnitude of these errors; we observed numerous stands that have a total extent of approximately 50 m, but have mean heights of between 10 and 35 m.

The indices used in this work are novel, and therefore require some analysis and explanation. Plotting the trailing edge correction factor as a function of the trailing edge extent (Fig. 5) indicates that the combination of multiple transforms of this variable results in a simple graphical relationship. The identity line plotted in Fig. 5 indicates that the trailing edge correction factor exceeds the trailing edge extent where the trailing edge extent is greater than zero and less than 10 m. Above this point, the ratio between trailing edge correction and extent approximates $\frac{1}{2}$.

Interpretation of the trailing edge extents requires that we have detailed descriptions of the terrain slope and height variability of areas sampled for each waveform. This kind of detailed data is only available for the waveforms that have coincident airborne lidar data; we have used the data from the Bartlett State Forest and Tahoe National Forest in these analyses. Initially we expected that the trailing edge was directly related to slope, but as illustrated in Fig. 6, terrain slope appears to approximate an upper limit to the trailing edge extent, but is not correlated to it. Similarly, our initial interpretation of the leading edge extent was that it was related to the physical extent of the canopy, which is a function of both the variability of tree height and, when the upper canopy surface follows the topography, terrain slope as well. Examination of the leading edge extent as a function of terrain slope indicates that slope constrains the leading edge in a similar manner.

A plot of the leading edge correction as a function of the leading edge extent (Fig. 7) indicates a more complicated relationship than observed for the trailing edge extent. For a given leading edge extent there are a range of correction factors that do not appear to be variance around a single relationship (as for the trailing edge correction factor). Instead, there are a range of correction factors for a leading edge extent that can vary as much as 7.5 m at the low values of leading edge extent; that range decreases as values of leading edge extent increase. The variability at the low end of this range indicates the effect of the lead/extent term, which corrects for observed biases in the calculation of mean height in small stands (as the influence of given leading edge extent decreases with extent). Examination of regressions between the lead edge extent and the leading edge correction for coniferous and deciduous forests indicate that, for a given leading edge extent, the leading edge correction of coniferous forests is larger than deciduous forests, possibly due to the deeper penetration of the laser pulse into these canopies.

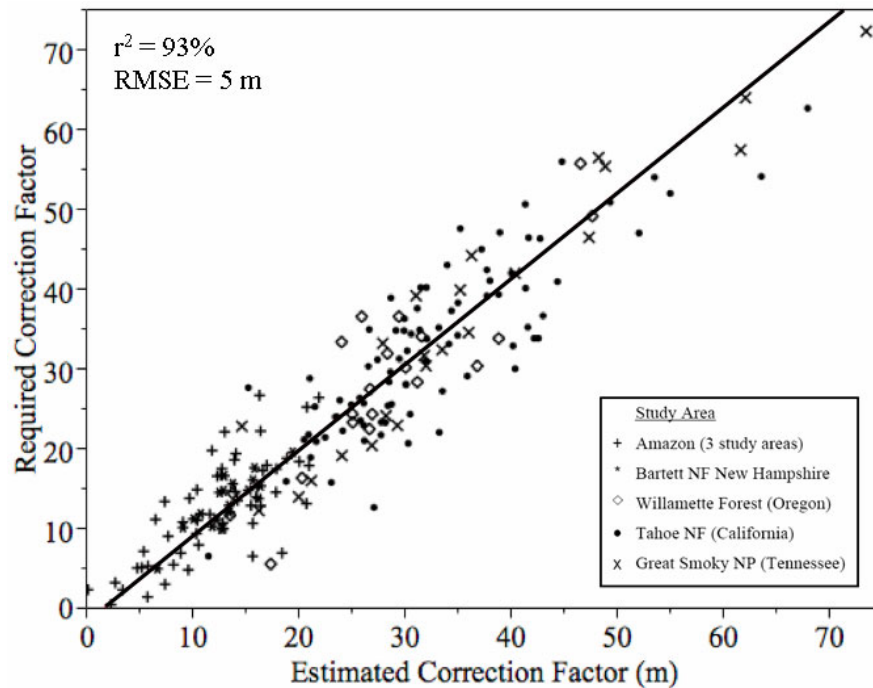


Fig. 2. Estimate of the Required Correction Factor, the difference between the GLAS waveform extent and the mean forest height. The correction factor represents the added waveform extent due to terrain slope and canopy surface variability.

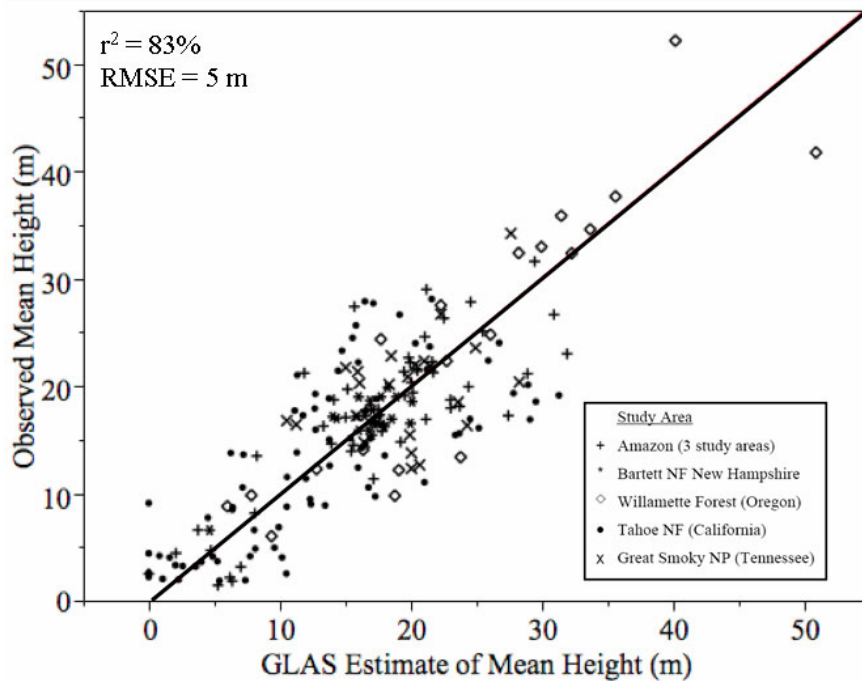


Fig. 3. GLAS estimate of mean forest height (waveform extent – correction factor) vs. observed mean forest height

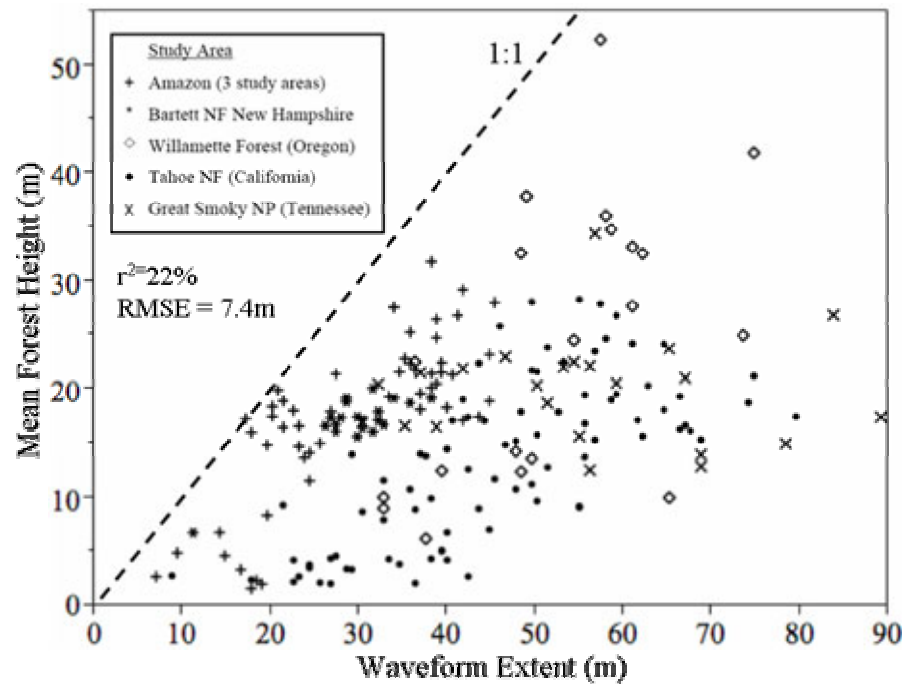


Fig. 4. Relationship between waveform extent and mean forest height, indicating the poor correspondence.

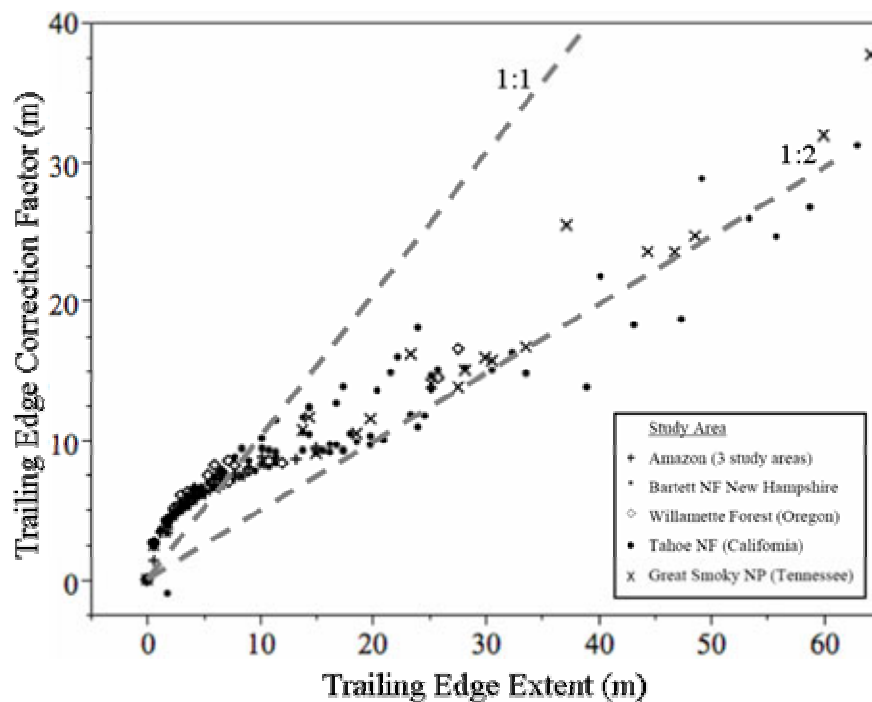


Fig. 5. Relationship between trailing edge extent (a direct waveform index) and the trailing edge correction factor (that part of the total correction factor related to the trailing edge).

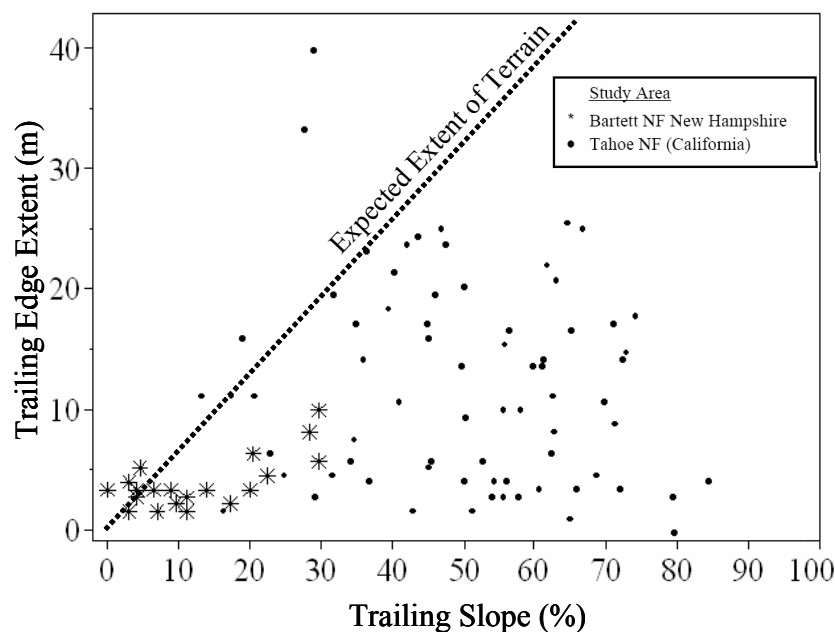


Fig. 6. Relationship between observed trailing edge extent and the extent we would expect to observe if the entire ground surface was represented in the waveform.

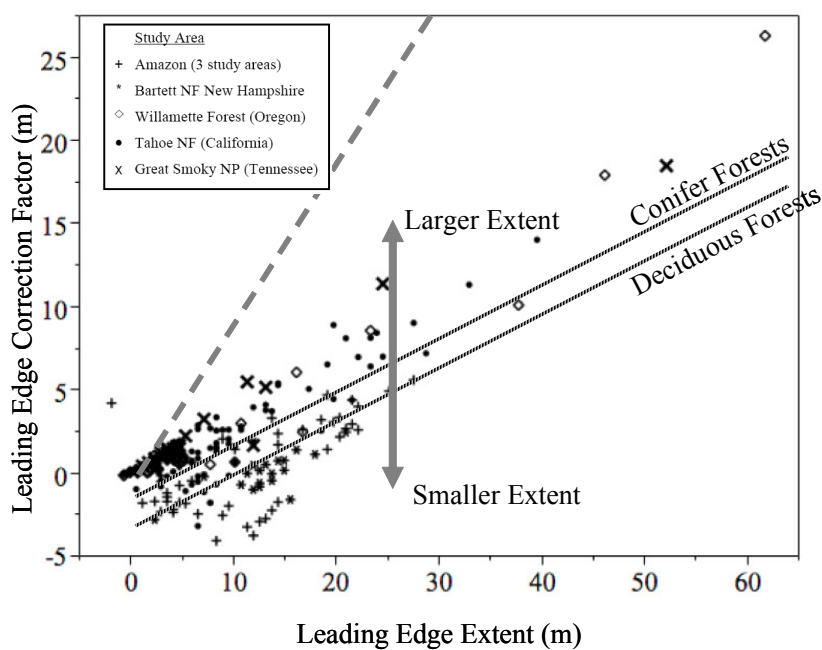


Fig. 7. Relationship between leading edge extent (a direct waveform index) and the leading edge correction factor (that part of the total correction factor related to the leading edge).

4 DISCUSSION

4.1 Accuracy of estimates

The results of this work indicate that we can create estimates of mean height (specifically the mean height of dominant and co-dominant trees or mean height of lidar returns) using GLAS data. Both the percent of variance explained (83%) and distribution of residuals (5 m) are consistent with the requirements of a global dataset, although these estimates are less precise than we would expect from airborne lidar datasets. We anticipate that, at a global extent, this is the likely upper limit of precision for estimates of mean height from GLAS waveforms.

A preliminary analysis of our ability to estimate canopy height (Lefsky and Harding, Unpublished) indicated that, when sites with a range of terrain slope were considered, the expected upper limit for variance explained was 80%. That study compared the mean heights of synthetic waveforms created by accumulating SLICER waveforms (which were collected with a 10 m footprint) from within a 75 m footprint with the mean heights derived from the individual SLICER waveforms [16]. We found that ~55% of the variance in height estimates was explained by waveform extent, 15% of variance was explained by terrain slope and ~10% of variance was explained by the height variability of the upper canopy surface. The remaining 20% of variance was identified as due to the interaction of upper canopy slope and terrain slope. These two slopes will be additive if canopy rises as the terrain declines or, if the canopy follows the terrain, there may be no additional extent in the waveform. We have not been able to identify which of these cases is operative from the waveform itself and therefore this variance is, and will most likely continue to be, unexplained.

4.2 Waveform indices

The removal of the effect of topographic slope on waveform extent is the single greatest challenge for using GLAS to estimate forest tree height and aboveground biomass. A previous approach to the problem [5] was limited in several respects. Using the experience developed during the first round of analyses an entirely new and more satisfactory approach to the problem of forest tree height estimation has been developed.

In this paper, we describe a second generation algorithm for estimating mean canopy height from GLAS waveforms, (the first generation algorithm is described in [5]). This method relies only on indices of waveform pattern, the leading and trailing edge extents. These indices have a direct, and physically meaningful, relationship to the influence of terrain slope and canopy height variability that leads to waveform extent that is greater than the mean height of the canopy. The complexity of the relationships that relate the leading and trailing edge extents to this additional height is a direct consequence of the waveform measurement itself.

Waveforms record the vertical distribution of energy returned from the first intercepted surface; these surfaces mask whatever lies beneath them. Therefore, the extent of the topographic correction is generally less than the total vertical extent of ground surfaces within the waveform's footprint, as much of the ground surface is not recorded in the waveform (Fig. 6). Correction for topography based on internal information in the waveform thus provides a better correction than the true topographic slope. The leading edge is, as expected, a function of both upper canopy variability and terrain slope, making it more difficult to interpret. The difference between coniferous and deciduous crowns indicated in Fig. 7 is likely due to the greater height variability of taller conifer forests.

In addition to the leading edge and trailing edge extents, the correction factors associated with these indices also use transformations of these variables based on their ratio with waveform extent, and in the case of the trailing edge correction, with waveform extent raised to a range of powers. Fig. 5 indicates that terms of the equation used to estimate the trailing

edge correction result in a non-linear curve; the ratio of the correction factor to the trailing edge extent exceeds 1 for smaller stands but converges on a ratio of approximately $\frac{1}{2}$ for stands with larger trailing edges. Given the quality of the estimates of the overall correction factor (Fig. 2) it is reasonable to treat this variable ratio as a realistic effect rather than an artifact of the data analysis.

For stands with the largest trailing edges, it is straightforward to interpret the ratio between the trailing edge extent and the trailing edge correction factor. We would expect that, when the full range of ground surface are sampled by the waveform, that a symmetric ground return would result and that the topographic correction would be $\frac{1}{2}$ of that total vertical range. For stands with low trailing edge extents, the correction factor is larger than 50% of the trailing edge, which suggests that we are not sampling the entire vertical range of ground surfaces. Figure 6 indicates that the trailing correction factor that is required to estimate mean height is often much lower than what we would expect from the local slope. In closed canopy forests (such as those in this study) this relationship is expected; the adaptation of this method to more open forests may require the addition of a canopy cover index as well, to estimate the effect of this better sampling of the vertical range of terrain.

Both the trailing and leading edge correction factors contain terms that are scaled by waveform extent, and powers thereof. These scaled terms have different behaviors in each correction factor. In the trailing edge correction, the summation of the scaled terms cancels out the effect of the extent terms at all but the highest levels of the trail index. However, the variability in the trailing edge correction at trailing extents above 20m is due to these terms. In contrast, the extent scaled term in the leading edge correction has the expected effect- it results in greater variability at lower trail values. Greater extents in the dataset are associated with higher slope and the greater leading edge correction would appear to be a correction for the slope component of the leading edge effect, not the difference in canopy height variability.

4.3 Cross-site analysis

In our first generation height correction algorithm, we estimated maximum tree height. In the work, mean tree heights (which are less sensitive to errors in field observations) are combined with estimates of mean return height from airborne lidar remote sensing. In this work, we test for the effect of methodological differences between the various height datasets using a combination of dummy variable regression and post-hoc analysis of site-specific regression slopes.

Using airborne lidar datasets to assist in estimating the topographic correction factor allows the incorporation of hundreds (eventually, thousands) of additional observations from study areas with minimal cost and effort, as existing airborne lidar data collections have been targeted with GLAS. In addition, the resulting observations of mean tree height are extremely consistent and are defined by the geometry of the tree surface, rather than by field procedures that vary among study areas, improving our confidence in the regressions. Through the analysis of these combined (lidar and field) datasets, general relationships have been developed that were not obvious when only field data was considered. The airborne lidar data has a higher number of observations, and more consistent relationships with the GLAS waveforms. As a consequence, relationships that were useful for estimating the correction factor at the field sites, but inconsistently significant from a statistical perspective, became more consistently included in the estimates of the height correction factor.

In our statistical analysis, we allow for the possibility of differences in the relationship between GLAS waveform indices and field estimates of height to account for the issues involved in combining field data estimates of the dominant and co-dominant height and estimates of mean height derived from airborne lidar data. The results of that analysis (Table 2) indicated that the primary site contrasts were not due to the source of mean heights (i.e.

lidar vs. field measurements), but rather between sites with broadleaf (Amazon, Bartlett) and conifer composition (Tahoe, Oregon). Tennessee, which is mixed composition, has a site offset that is roughly half way between those two. These results indicate that, for a given set of waveform indices, conifer sites have mean heights that are approximately 9 meters shorter than broadleaf sites. To put this result in context, the mean correction factor without site contrasts is 23.9 m; the average for each site is given in Table 2. The site contrast doesn't exceed 25% of the average site correction factor, or 14% of average site waveform extent, for any study site, and therefore may be a secondary effect. Nevertheless, this result clearly requires explanation.

Given the contrasting geometry of conifer and broadleaf crowns, the physical structure of canopies might offer an explanation for this effect. Within the estimation equation, there are three possible sources for the discrepancy between conifer and broadleaf canopies; their effect on the extent of the waveform, leading edge and trailing edge. We have already indicated (Fig. 1) that, given the low surface area of the uppermost portion of a conifer crown, we would expect that waveform extent would be underestimated relative to broadleaf crowns. However, correcting for this effect would require adding additional height to conifer stands, the opposite of the observed effect and therefore not a likely mechanism. Given the deeper penetration of energy into conifer canopies, we would expect that, for conifer and broadleaf stands with the same height of dominant and co-dominant stems, we would have greater leading edge extent for the conifer stands (as implied by Fig. 7), resulting in a greater relative leading edge correction factor. This would again *reduce* the apparent height of conifer canopies, and therefore is not likely to be the correct mechanism.

On the basis of canopy structure, the final possibility for explaining the site contrast values for the conifer stands is the extent of the trailing edge. We have seen that the total gap fraction and its spatial coverage can modify the fraction of terrain slope that is represented in the trailing edge (as illustrated in Figs. 6 & 8). We also know that, relative to conifers, broadleaf trees show a greater tendency to expand their crowns to fill any gaps between them. Therefore, it is reasonable to hypothesize that the spatial pattern of gap fraction might lead to greater values of trailing edge extent for conifer crowns. This would result in a greater relative trailing edge correction factor for conifer stands, and we would again expect that we would need to add height to conifer stands, not subtract them.

On the basis of this logic, there is no reason to conclude that we are looking at an effect of canopy structure because any effect we expect would be opposite in sign from those observed. The next possible source of the site contrasts is statistical; perhaps there is some property of the datasets that would lead the least-squares regression to incorporate these site contrasts. In fact, the average trailing edge of the conifer stands is 4.5m larger than that of the broadleaf stands, the average leading edge is 6.1m larger than that of the broadleaf stands, and therefore the total difference between the sum of the leading and trailing edges is 10.6m. At the same time, the total waveform extent of the conifer stands is 19.3m taller than broadleaf stands. These coarse level differences suggest that, given that the regression has the site contrasts available, they are likely to explain variance and therefore will be included in the analysis, perhaps masking other mechanisms. One observation that supports this hypothesis is that the site contrast between the broadleaf and other sites is the second variable to be entered into the stepwise regression. Our conclusion is that, while the addition of the site contrast variables does explain an additional 5% of variance and reduces the RSME by 1m, this is related to the mean stature of the various sites and apparently not due to differences in the way that canopy structure is represented in the waveform. Nevertheless, the need for site contrast variable does suggest that there are effects that may not be adequately addressed by our current method.

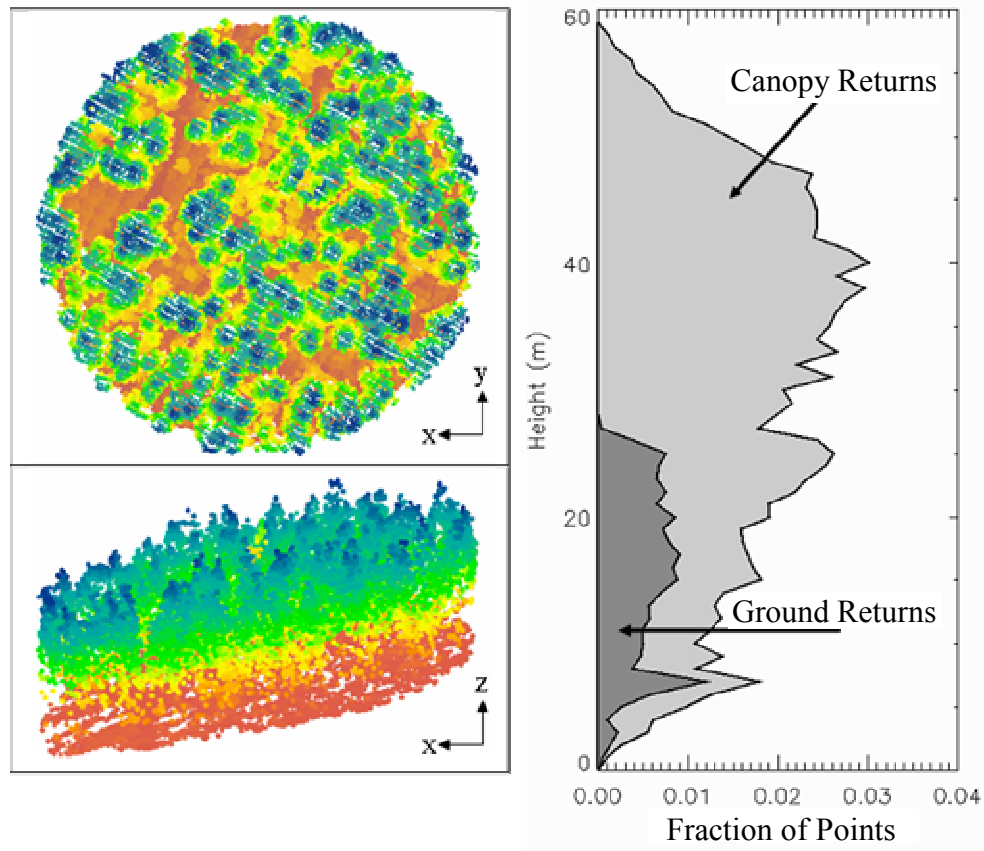


Fig. 8. At left, two views of a 65m diameter subset of discrete return lidar collected at Tahoe National Forest, indicating the uneven spatial distribution of terrain observations (top) and the overall terrain slope (bottom). At right, the vertical distribution of discrete return lidar observations from the canopy and terrain.

5 CONCLUSION

The proof of concept presented in [5] was difficult to apply outside of the study areas where it was demonstrated because it sought to estimate the variable maximum height directly. As a consequence, the distribution of forest tree heights in the model dataset influenced the final results. By moving to the analysis of the correction factor- the height required to reconcile the observed mean height and the waveform extent- the resulting relationships are independent of forest height. The most important contribution to the correction factor is the mathematically complex but graphically simple relationship between the trailing edge extent and topographic slope, resulting in a trailing edge correction that (when considered in a univariate context) has a 1:1 relationship with the required topographic correction factor. The relationship estimating canopy height is now simpler and comprehensible, and the estimation of forest height is no longer dependent on the use of indices derived from SRTM (as in [5]). SRTM is only available at low resolution (90m) for much of the world and its estimate of topography is also dependant on tree height [17]. The revision of these methods has also facilitated the ongoing development of theoretical approaches to the problem of topographic correction.

We fully expect that this second generation approach to height correction will not be the last. Specifically, we acknowledge that there are cases that do not conform to the assumptions we have used in the paper. Short stands on steep slopes remain a problem, although this may be a limitation of the data rather than the height correction algorithm. A more tractable problem involves errors that arise when forests on moderate slopes have very low or very high cover. When a low power ground return occurs on moderate slopes, the ground return power is distributed over the range of the slope and the maximum power of the ground return is further reduced. When the trailing edge extent for that waveform is defined the threshold (half the maximum power over the background noise level) will exceed the power of the ground return and will identify the last elevation of the canopy return instead. On low slopes, this is rare effect, as the ground return does not become spread out by topography to the same degree. On high slopes, the trailing edge of the canopy return is inseparable from the ground return, and the resulting trailing edge extent is appropriate. Similarly, when *canopy* cover is low, the power of the canopy return on moderate slopes is further reduced and the leading edge index will identify the first elevation of the ground return.

These effects have not been observed in our field or lidar datasets. We have only seen them in datasets of waveforms simulated from modeled stand structure (Yong et al. In Prep). Nevertheless, the simulated waveforms are realistic enough that we take these results seriously. We are actively working to identify real-world examples of the effects, and to incorporate them into the height correction algorithm.

References

- [1] D. J. Harding and C. C. Carabajal, "ICESat waveform measurements of within-footprint topographic relief and vegetation vertical structure," *Geophys. Res. Lett.* **32**, L21S10 (2005) [doi:10.1029/2005GL023471].
- [2] M.A. Lefsky, W.B. Cohen, S.A. Acker, T.A. Spies, G.G. Parker, and D. Harding, "Lidar remote sensing of biophysical properties and canopy structure of forest of Douglas-fir and western hemlock," *Rem. Sens. Environ.* **70**, 339-361(1999) [doi:10.1016/S0034-4257(99)00052-8].
- [3] D. J. Harding, M. A. Lefsky, G. G. Parker, and B. Blair. "Lidar altimeter measurements of canopy structure: methods and validation for closed-canopy broadleaf forest," *Rem. Sens. Environ.* **76** (3), 283-297 (2001) [doi:10.1016/S0034-4257(00)00210-8].
- [4] M. Keller, A. Alencar, G. P. Asner, B. Braswell, M. Bustamante, E. Davidson, T. Feldpausch, E. Fernandes, M. Goulden, P. Kabat, B. Kruijt, F. Luizao, S. Miller, D. Markewitz, A. D. Nobre, C. A. Nobre, N. Priante Filho, H. Rocha, P. Silva Dias, C. von Randow, and G. L. Vourlitis, "Ecological research in the large scale biosphere atmosphere experiment in Amazonia (LBA): early results ecological applications," *Ecol. Appl.* **14**, S3-S16 (2004) [doi: 10.1890/03-6003].
- [5] M. A. Lefsky, D. J. Harding, M. Keller, W. B. Cohen, C. C. Carabajal, F. Del Bom Espirito-Santo, M. O. Hunter, and R. de Oliveira Jr., "Estimates of forest canopy height and aboveground biomass using ICESat," *Geophys. Res. Lett.* **32**, L22S02, (2005) [doi:10.1029/2005GL023971].
- [6] M. Keller, M. W. Palace, and G. C. Hurtt, "Biomass estimation in the Tapajos National Forest, Brazil: examination of sampling and allometric uncertainties," *For. Ecol. Mgmt.* **154**, 371-382 (2001) [doi:10.1016/S0378-1127(01)00509-6].
- [7] *U.S.D.A. Forest inventory and analysis national core field guide*, U.S.D.A. Forest Service, Washington Office, Internal report (2004).
- [8] M. A. Lefsky, A. T. Hudak, W. B. Cohen, and S. A. Acker, "Patterns of covariance between forest stand and canopy structure in the Pacific Northwest," *Rem. Sens. Environ.* **95**(4), 517-531 (2005b)[doi:10.1016/j.rse.2005.01.004].

- [9] M. A. Moeur and R. S. Stage, "Most similar neighbor: an improved sampling inference procedure for natural resource planning," *Forest Sci.* **41**(2), 337- 359 (1995).
- [10] M. A. Lefsky, W. B. Cohen, D. J. Harding, G. G. Parker, S. A. Acker, and S. T. Gower, "Lidar remote sensing of aboveground biomass in three biomes," *Global. Ecol. Biogeogr.* **11**(5), 393-400 (2002)[doi:10.1046/j.1466-822x.2002.00303.x].
- [11] Q. Chen, P. Gong, D. D. Baldocchi, and G. Xie, "Filtering airborne laser scanning data with morphological methods," *Photogramm. Eng. Rem. Sens.* **73**(2), 175-185 (2007).
- [12] B. E. Schutz, H. J. Zwally, C. A. Shuman, D. Hancock, and J. P. DiMarzio, "Overview of the ICESat mission," *Geophys. Res. Lett.* **32**, L21S01 (2005) [doi:10.1029/2005GL024009].
- [13] *JMP, Version 6*, SAS Institute Inc., Cary, NC, (2005)
- [14] S. Van Huffel, Ed., *Recent advances in total least squares techniques and errors-in-variables modeling*, Philadelphia Society for Industrial and Applied Mathematics (1997).
- [15] M. A. Lefsky, A. T. Hudak, W. B. Cohen, and S. A. Acker, "Geographic variability in lidar predictions of forest stand structure in the Pacific Northwest," *Rem. Sens. Environ.* **95**(4), 532-548 (2005) [doi:10.1016/j.rse.2005.01.010].
- [16] J. N. Blair, D. B. Coyle, J. L. Bufton, and D. J. Harding, "Optimization of an airborne laser altimeter for remote sensing of vegetation and tree canopies," in *Surface and Atmospheric Remote Sensing: Technologies, Data Analysis and Interpretation, Proc. Int. Geosci. Rem. Sens. Symp.* **2**, 939-941 (1994) [doi: 10.1109/IGARSS.1994.399307].
- [17] C. Carabajal and D. J. Harding, "ICESat validation of SRTM C-band digital elevation models," *Geophys. Res. Lett.* **32**, L22S01 (2005) [doi:10.1029/2005GL023957].