ISPRS Journal of Photogrammetry and Remote Sensing 95 (2014) 81–92

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

An accurate and computationally efficient algorithm for ground peak identification in large footprint waveform LiDAR data

Wei Zhuang, Giorgos Mountrakis *

Department of Environmental Resources Engineering, State University of New York College of Environmental Science and Forestry, Syracuse, NY 13210, United States

ARTICLE INFO

Article history:
Received 1 January 2014
Received in revised form 2 June 2014
Accepted 4 June 2014
Available online 5 July 2014

Keywords:
Ground identification
Large footprint
Waveform LiDAR
LVIS
Gaussian decomposition
Ground detection

ABSTRACT

Large footprint waveform LiDAR sensors have been widely used for numerous airborne studies. Ground peak identification in a large footprint waveform is a significant bottleneck in exploring full usage of the waveform datasets. In the current study, an accurate and computationally efficient algorithm was developed for ground peak identification, called Filtering and Clustering Algorithm (FICA). The method was evaluated on Land, Vegetation, and Ice Sensor (LVIS) waveform datasets acquired over Central NY. FICA incorporates a set of multi-scale second derivative filters and a *k*-means clustering algorithm in order to avoid detecting false ground peaks. FICA was tested in five different land cover types (deciduous trees, coniferous trees, shrub, grass and developed area) and showed more accurate results when compared to existing algorithms. More specifically, compared with Gaussian decomposition, the RMSE ground peak identification by FICA was 2.82 m (5.29 m for GD) in deciduous plots, 3.25 m (4.57 m for GD) in coniferous plots, 2.63 m (2.83 m for GD) in shrub plots, 0.82 m (0.93 m for GD) in grass plots, and 0.70 m (0.51 m for GD) in plots of developed areas. FICA performance was also relatively consistent under various slope and canopy coverage (CC) conditions. In addition, FICA showed better computational efficiency compared to existing methods. FICA's major computational and accuracy advantage is a result of the adopted multi-scale signal processing procedures that concentrate on local portions of the signal as opposed to the Gaussian decomposition that uses a curve-fitting strategy applied in the entire signal. The FICA algorithm is a good candidate for large-scale implementation on future space-borne waveform LiDAR sensors.

1. Introduction

Large footprint waveform LiDAR sensors, such as Geoscience Laser Altimeter System (GLAS) and Land, Vegetation, and Ice Sensor (LVIS), are capable of capturing the entire waveform of a backscattered signal from ground objects within a footprint (Blair et al., 1999; Zwally et al., 2002). The footprint size of the LiDAR sensors, which partially depends on the flight height, can vary from 1 m to 80 m. For example, LVIS generates a 20–25 m footprint at the flight altitude of 7600–8300 m (Hofton et al., 2008). These sensors have been widely and successfully used in numerous environmental studies (e.g. Lim et al., 2003; Mallet and Bretar, 2009). Ground object heights, for example vegetation and building heights, can be derived from large footprint LiDAR sensors and have been incorporated in biophysical parameter estimation (Anderson et al., 2006; Lefsky et al., 1999), wildlife habitat modeling (Hyde et al., 2005) and urban environment studies (Gong et al., 2011).

Ground object heights are usually estimated through identification of top and ground location in a waveform and calculating the distance in between. For example, building heights are calculated by the distance between the roof top location and the ground peak location in a waveform (Cheng et al., 2011; Gong et al., 2011); forest heights are the distance between the treetop location and the ground peak location in a waveform (Andersen et al., 2005). This method of separate identification of the top and ground locations prevails in height estimation from large footprint waveform LiDAR data because it is robust and does not require a priori knowledge of the study area (Anderson et al., 2011; Duncanson et al., 2010; Popescu et al., 2011; Sun et al., 2011). In previous studies, the top location of a ground object has been successfully associated with the beginning of the returned signal, calculated as the first signal above a noise threshold, as shown in Eq. (1):

* Corresponding author. Address: Department of Environmental Resources Engineering, State University of New York College of Environmental Science and Forestry, 419 Baker Hall, 1 Forestry Dr., Syracuse, NY 13210, United States. Tel.: +1 (315) 470 4824; fax: +1 (315) 470 6958.
E-mail address: gmountrakis@esf.edu (G. Mountrakis).

http://dx.doi.org/10.1016/j.isprsjprs.2014.06.004
0924-2716/

$$\text{Threshold} = \text{Mean noise} + N * \text{standard deviation of noise}, \quad N = 1,\ 2,\ \ldots,\ n \tag{1}$$

where noise was estimated from the raw waveform (Chen, 2010b; Lefsky et al., 1999), N varied in different studies. Compared to top location identification, ground peak identification in the waveform signal has proven more challenging. In relatively open areas such as developed areas, ground peaks are usually strong and thus can be accurately identified (Cheng et al., 2011; Hofton et al., 2006). In vegetated areas, however, ground peak identification usually has larger errors due to either signal overlap between the lower vegetation and the ground (Chen, 2010b) or weak ground peak returns under dense canopies (Chauve et al., 2009).

A popular Ground Peak Identification Algorithm (GPIA) is based on Gaussian Decomposition (GD). GD is a curve fitting algorithm that uses the summation of numerous Gaussian functions to fit a waveform in order to satisfy a certain statistical criterion, such as an intensity threshold of Root Mean Square Error (RMSE) (Hofton et al., 2000; Wagner et al., 2006). The ground is usually located at a waveform peak identified by the centroid of the last Gaussian function or the centroid of Gaussian function with larger amplitude in the last several Gaussian functions from Gaussian decomposition (Popescu et al., 2011; Sun et al., 2008). Harding and Carabajal (2005) suggested that in low topographic relief areas the centroid of the last Gaussian function can be used to estimate the ground location. Chen (2010a) compared different Gaussian functions in GD used to identify ground peak in mountainous areas and found that using the centroid of the Gaussian function with the strongest amplitude from the lowest two functions provided the best estimate for ground location in the GLAS waveform.

Three parameters (i.e. center location, amplitude and width of a Gaussian function) are usually estimated in GD using an optimization algorithm, such as the Levenberg–Marquardt (LM) algorithm (Hofton et al., 2000) or maximum likelihood estimation using the Expectation Maximization (EM) algorithm (Soderman et al., 2005). Although optimization algorithms can be used in GD for fast parameter estimation, the iterative procedures embedded in the optimization process can be time-consuming, particularly for large-scale studies with high data volume. Furthermore, the parameter settings of the optimization algorithms such as total iteration number and ending RMSE, may significantly impact algorithmic performance and efficiency. As a result several more efficient algorithms, such as Zero-Crossing (ZC), constant fraction and local maximum, ZC is developed the transition point of the second derivatives of a waveform and have been applied to peak detection with small footprint waveform data (Wagner et al., 2004). Even though the zero-crossing algorithm had high computational efficiency, it was easily affected by small peaks caused by the ringing effect (a signal oscillation from the sensor) or background noise in a waveform (Chauve et al., 2009) and resulted in false peak detection. A comparison has been made among the three algorithms to detect ranges between a small waveform LiDAR sensor and ground objects, and has proven the algorithms to be sensitive to noise (Wagner et al., 2004). Therefore, the algorithms may not be suitable for large footprint waveform data where significant background noise may be present.

Another issue associated with GD-based algorithms using the centroid of a detected Gaussian function is the fact that this centroid may not accurately represent the ground peak due to weak ground signal or overlapping signal caused by slope and low-lying vegetation (Chauve et al., 2009; Dubayah et al., 2010). GD tends to put more effort on waveform peaks with larger intensity, because a better fit for the large peaks results in smaller RMSE. For example, in a forest with dense canopy (e.g. secondary even-age deciduous forest) with steep terrain, the waveform signal from canopy and ground can easily overlap because of the time distribution of the emitted pulse. In many cases GD does not assign a separate Gaussian function for the ground peak in a waveform because the added Gaussian function could increase the RMSE of the overall fit. Therefore, the ground peak is misidentified resulting in an incorrect lower tree height.

Other GPIAs based on statistical models have been proposed for height estimation without direct ground peak identification. These methods build regression models from the waveform extent (e.g. leading edge and trailing edge) to avoid ground peak identification errors caused by slope effects and overlapping signal with lower vegetation (Lefsky et al., 2005, 2007; Pang et al., 2007). However, as the leading and trailing edge vary for different waveforms and forest stands, a large number of sample plots are needed for a regional scale study area with different terrain situations and forest structures to build the models (Chen, 2010b), which may not be feasible for a large-scale mapping task.

This study proposes an alternative GPIA method called FIltering and Clustering Algorithm (FICA) to detect ground peaks in large footprint waveforms. Our goal is to create an accurate, computationally efficient algorithm that exhibits low accuracy variability at different landscape conditions (e.g. land cover type, slope and canopy coverage (CC)). FICA performance in ground peak identification is investigated and compared with two existing GPIAs (Zero-crossing and Gaussian decomposition) in terms of accuracy and computational efficiency.

2. Methods

2.1. Proposed FICA method

The procedural steps of the algorithm are shown in Fig. 1. The ground detection process is organized in three major steps: preprocessing, peak detection, and post-processing.

2.1.1. Pre-processing

The waveform was initially smoothed by a Gaussian filter in order to decrease the effect of background noise (Eq. (2)). In previous studies using GLAS waveform datasets (Chen, 2010b; Sun et al., 2008), the filter width (σ) and filter maximum length (2L), both in terms of time, were set as $2\sigma_t$ and $6\sigma_t$, respectively, where σ_t was the width of the Gaussian function simulating the transmitted waveform pulse:

$$F(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{t^2}{2 * \sigma^2}\right), -L < t < L,\ t \text{ is the time bin} \tag{2}$$

In our approach we tested multiple smoothing options to identify optimal conditions. Although a Gaussian filter with a larger width could remove more noise, it sacrifices details in a signal (Gonzalez and Wintz, 2008), including a possible weak ground peak. Therefore, a large width may not be appropriate in waveforms acquired in dense canopy areas. As the LVIS waveform had a larger signal–noise ratio than the waveform from GLAS, a less intense Gaussian smoothing filter would be more appropriate to remove noise and keep signal details at the same time. After the smoothing step, the mean and standard deviation of the intensity for the first 100 time bins was calculated. These two values represent the intensity variability of the background noise left in the smoothed waveform and are used in later steps.

2.1.2. Filtering and clustering

The peak identification procedure included two steps: filtering by a set of multi-scale second derivative filters and clustering by the k-means algorithm. A second derivative filter (e.g. Laplacian filter) was originally designed for edge detection and signal sharpening (Gonzalez and Wintz, 2008). The development of FICA is based

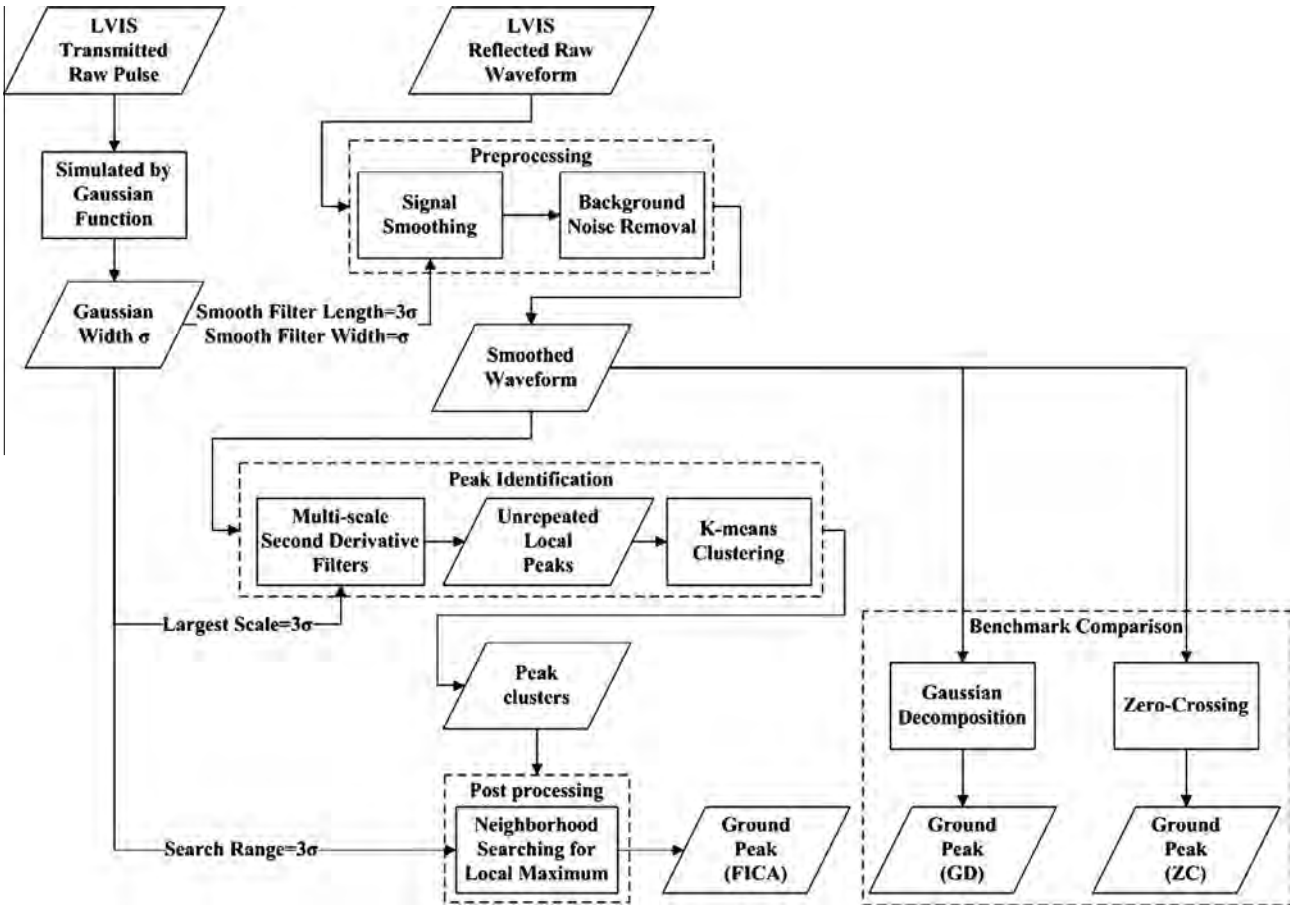

Fig. 1. Flowchart of the algorithm for ground peak identification in LVIS raw waveform data.

on second derivative filters, but overcomes known limitations by simultaneously examining the signal at multiple scales followed by a clustering procedure on the identified peak candidates. As shown in Eq. (3), FICA implements a second derivative filter (SDF) at different scale sizes $i = 1, 2, \ldots, N$ until a predefined maximum scale N around center time bin x:

$$SDF(x,i) = \frac{2 * f(x) - (f(x-i) + f(x+i))}{i^2} \quad i = 1, 2 \ldots N \tag{3}$$

where i is the scale for each second derivative filter, x is the center time bin, and N is the maximum scale size. The maximum scale used in this study was three times the width of the transmitted pulse of the LVIS waveform, since filters with larger scales would identify already selected peak candidates at smaller scales. In the edge bins, backwards version of waveform bins (e.g. the length of the backwards bins is i) was created in order to be symmetric for the second derivative calculation.

After the multi-scale filters were applied to the smoothed waveform, an intensity threshold was applied on the second derivative results to identify possible peak candidates at each scale. If any of the multi-scale second derivative filters for a given time bin was larger than this threshold, then that time bin would be identified as a peak candidate.

After all possible peak candidates were extracted, a k-means clustering algorithm was applied to the ground peak candidates. The input space of the k-means was a two dimensional space for time bin and intensity. The k-means algorithm was set to 100 iterations and the number of clusters was tested during the training process. The output of this process assigned a cluster group to each of the peak candidates.

2.1.3. Post-processing

A post-processing procedure was carried out in order to select a single final ground peak. The peak candidates belonging to the cluster with the lowest height were selected and the candidate with the largest intensity within that cluster was identified as the ground peak. The ground peak location identified from an LVIS waveform was then converted into elevation according to elevation of the first bin and the last bin.

2.2. Algorithmic calibration/validation, benchmark comparison and assessment metrics

Two existing GPIAs, the Gaussian decomposition (GD) and the Zero-Crossing (ZC) methods were applied to the LVIS waveform data in order to compare with the performance of FICA in terms of accuracy and computational efficiency. During the training process the same subset of waveform plots was used for calibrating all three algorithms. This subset consisted of 500 plots (i.e. 100 plots for each land cover type). Calibrated parameters and testing ranges for the three GPIAs are shown in Table 1.

In FICA, the width of the Gaussian smoothing filter (σ), the threshold for the peak identification (T_p) and the number of clusters (N_c) in k-means were the calibrated parameters. In GD, besides σ, the maximum iteration (N_{max}) for the Levenburg–Maquardt (LM) algorithm was the parameter determining the accuracy and the efficiency of the optimization algorithm. During the initialization of the LM algorithm the three Gaussian function parameters (amplitude, center and width) were allocated by identifying the peak using a derivative function and setting the width to a fixed initial value. The RMSE for the GD curve fitting was fixed at

Table 1
Parameters calibrated in different GPIAs.

GPIAs	Calibration parameters	Testing range (start:step:end)
FICA	σ: smoothing filter width (m[a])	0:0.1:0.8
	T_p: threshold for peak identification	0.1:0.05:2
	N_c: number of clusters in *K*-means	2:1:8
GD	σ: smoothing filter width (m)	0:0.1:0.8
	N_{max}: maximum iterations for LM algorithm	10:10:100
ZC	σ: smoothing filter width (m)	0:0.1:0.8
	T_p: threshold for peak identification	0.1:0.05:2

[a] Each waveform bin represents 0.3 m.

0.00001. In addition, the number of Gaussian functions was automatically determined by the number of inflection points existing in a smoothed waveform (Hofton et al., 2000). A maximum of six Gaussian functions was allowed, a number shown sufficient in previous research (Chen, 2010a,b). For the ZC algorithm σ was an important parameter for the final accuracy as this method is highly sensitive to background noise (Wagner et al., 2004). In addition, numerous T_p values were tested for identifying peaks in ZC.

Algorithmic accuracy validation was conducted by comparing the detected ground peak elevation with the reference dataset not used during training (2900 plots). In these 2900 plots, there were 900 coniferous plots, 900 deciduous plots, 900 shrub plots, 100 developed area plots and 100 grass plots. Statistical metrics were used to evaluate accuracy and computational efficiency. Accuracy was evaluated with the Root Mean Square Error (RMSE), bias and correlation coefficient. Statistics for computational efficiency assessment included percentile of the processing time for each land cover types.

3. Study area and data

3.1. Study area

The study area is located in the central region of New York State surrounding the city of Syracuse (shown in Fig. 2). The elevation of the study area ranges from 80 m to 398 m above mean sea level, and the slope varies from 0 to 30°. Various land cover types exist including coniferous forest, deciduous forest, grass, shrub, and developed area. Forests at different successional stages (e.g. abandoned pasture, early forest, secondary forest and old-growth forest) can be found in the study area. The major tree species include sugar maple (*Acer saccharum*), white ash (*Fraxinus americana*), American beech (*Fagus grandifolia*), Norway spruce (*Picea abies*), red pine (*Pinus resinosa*) and eastern white pine (*Pinus strobus*).

3.2. Data sources

The purpose of this study was to detect ground elevation from large footprint waveform LiDAR data. Discrete return LiDAR data from a different flight was used as reference for accuracy assessment of the ground elevation. The accuracy assessment was conducted in different land cover types, which were manually delineated using high resolution aerial photography. The large footprint waveform data was acquired in August 24–26, 2009 using NASA's Land, Vegetation, Ice Sensor (LVIS) sensor (Blair

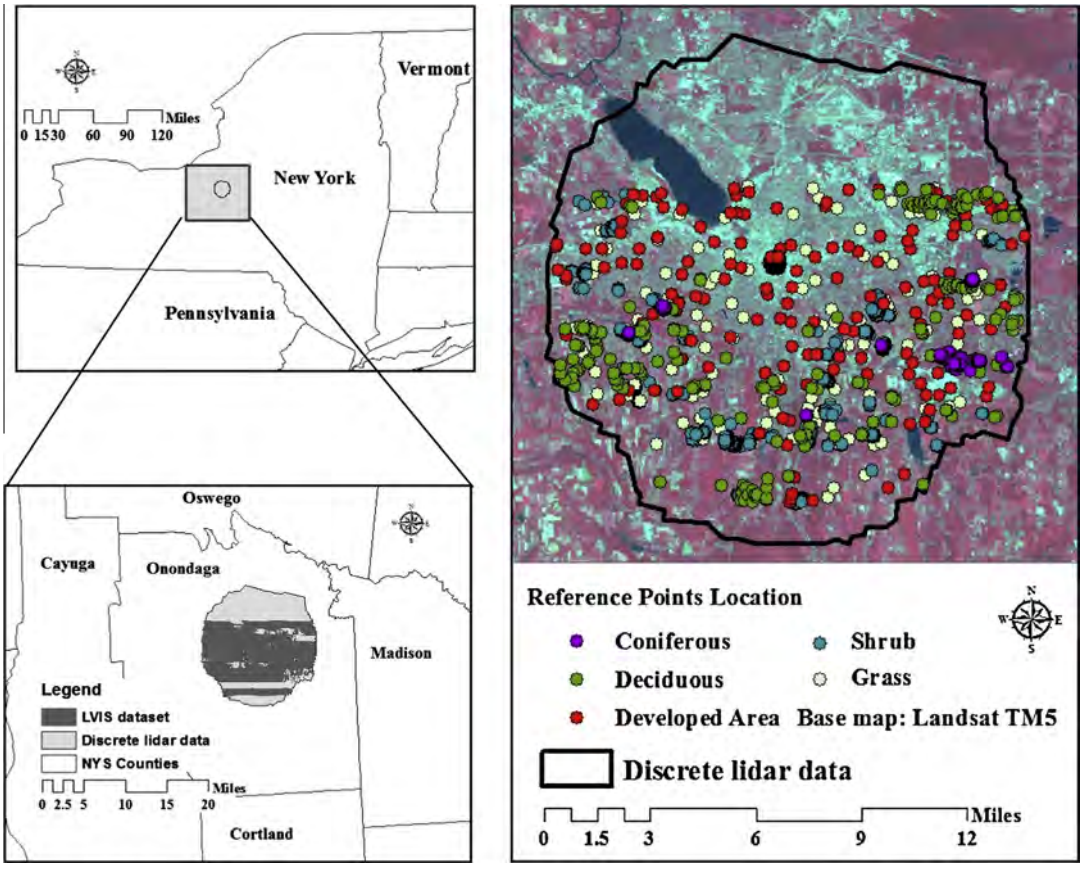

Fig. 2. Study area located in central New York state.

et al., 1999), as shown in Fig. 1. As the LVIS dataset was acquired in a leaf-on season, the ground peak detection from the waveform data was challenging due to strong canopy presence. The LVIS circular pulse is at a wavelength of 1064 nm. The diameter of the footprint is nominally 20 m on the ground for this study. The vertical resolution of the waveform data was approximately 0.3 m. Scanning angle varied from 0 to 5°. Within each waveform, a total of 432 time bins were recorded. Spatial metadata recorded included the longitude/latitude of the pulse and elevation at the first time bin and last time bin, azimuth, incident angle, distance from the sensor to the ground, mean background noise, and transmitted pulse (Blair et al., 2006). The horizontal coordinate system was Universal Transverse Mercator (UTM) 18 N on the North American Datum of 1983 (NAD83). The vertical datum was transformed into the North American Vertical Datum of 1988 (NAVD88).

Airborne discrete return LiDAR data was acquired on April 1st, 2010 in Syracuse, NY using Airborne Laser Scanner (ALS60) containing two returns per pulse (Fig. 1). The ground location in the discrete dataset can be accurately extracted as the acquisition time was in a leaf-off season. The discrete return LiDAR data was used as reference data as suggested in previous studies (Chen, 2010b; Wasser et al., 2013). The wavelength of the pulse was 1064 nm, which had strong vegetation reflectance. The average laser point density is 2.1 laser points per m^2. The footprint size was 0.37 m. The horizontal coordinate system of the discrete return LiDAR data was converted to the same coordinate system as the LVIS data. The vertical coordinate system was the same as that of LVIS and no conversion was required. According to the accuracy assessment provided by the data vender, the horizontal RMSE of the LiDAR point was less than 1 m; and the vertical RMSE was 0.033 m evaluated using 24 ground control points. The discrete return LiDAR points were also classified into ground and non-ground points using Terrasolid software.

3.3. Sampling and reference data

In order to evaluate ground peak identification from the LVIS signal, mean ground elevation was calculated from the discrete LiDAR dataset. The sampled areas were selected based on the visual interpretation of aerial images and discrete return LiDAR data. Four vegetation cover types, deciduous plots, coniferous plots, shrub plots and grass plots, and a fifth land cover expressing developed area were sampled. The aerial images (Digital Ortho) were acquired in leaf-off season allowing accurate differentiation of the coniferous, deciduous and grass plots. The delineation between shrub and deciduous classes was based on the maximum canopy height from the discrete return LiDAR data. Shrub plot was defined as the plots where the maximum height was less than 10 m and the number of LiDAR points above 4 m was less than 50% of the total number of LiDAR points within an LVIS footprint. A total of 3400 plots (i.e. LVIS footprints) were sampled within the areas occupied by the five aforementioned land cover types using stratified random sampling throughout the entire study area. Among the 3400 plots (shown in Fig. 1), there were 200 plots for grass and developed area, respectively, and 1000 plots for deciduous and coniferous forest and shrub since the canopy signal exhibited higher complexity and opportunity for algorithmic improvements than the other two land cover types.

The reference ground elevation was extracted from the discrete return LiDAR data. The corresponding airborne discrete LiDAR data within the range of a circular waveform pulse was extracted by a 10 m buffer zone (in radius) of the center location of a sampled LVIS ground pulse. The mean ground elevation within an LVIS footprint was then estimated by the elevation of the ground points labeled in the discrete data through a weighting system. As the spatial energy distribution of an LVIS pulse is Gaussian-shaped, a weighting strategy was applied when calculating the mean elevation (Chen, 2010a) considering the spatial energy distribution (Blair et al., 1999), as shown in the equation below:

$$E_M = \sum_{i=1}^{N} \frac{W_i}{\sum_{i=1}^{N} w_i} * E_i \tag{4}$$

$$W_i = \frac{1}{\sigma\sqrt{2\pi}} * \exp \frac{(x_i - x_o)^2 + (y_i - y_o)^2}{2 * \sigma^2} \tag{5}$$

where (x_i, y_i) is the coordinates for each discrete LiDAR ground point, (x_0, y_0) is the center location of a waveform pulse, σ is the estimated Gaussian width from the nominal footprint size (20 m) when considering the radius of the footprint is about 3 widths of the Gaussian pulse, W_i is the weight for a specific discrete LiDAR ground point based on the waveform distribution (Blair et al., 1999), E_i is the elevation of a discrete LiDAR ground point, E_M is the weighted mean elevation calculated from discrete LiDAR data, and N is the total number of discrete return LiDAR points falling within a waveform circular pulse. The sum of the weights in Eq. (5) was normalized to 1 before averaging, as shown in Eq. (4).

4. Results and discussion

4.1. Algorithmic sensitivity analysis using calibration data

In this section, all three GPIAs were calibrated using the parameter ranges in Table 1. As the Gaussian smoothing filter width σ was required for all the GPIAs, the RMSEs for different σ were first compared between the GPIAs. The other parameter settings of the GPIAs in the comparison used the parameters that achieved the lowest RMSEs in each GPIA. Then, accuracy of peak identification of each GPIA was further examined for the other parameters when the best width was fixed according to the previous comparison. All results reported in the entire Section 4.1 used the calibration dataset of 100 points for each of the five land cover classes.

4.1.1. Gaussian smoothing filter width

The peak identification accuracy of the GPIAs responding to the smoothing width σ is shown in Fig. 3. It can be seen that ZC was severely affected by the choice of smoothing filter; while the RMSEs of FICA and GD showed a general increasing trend as the width increased. The best smoothing widths for FICA, GD, and ZC were 0.1 m, 0 m (i.e. no smoothing procedure), and 0.6 m, respectively. This investigation provided guidance for the additional tests for FICA, GD and ZC presented later.

4.1.2. FICA-specific parameters

The sensitivity of the proposed FICA algorithm was examined with respect to the other two parameters, the threshold for peak identification (T_p) and number of clusters predefined in the k-means classifier (N_c). The smoothing width parameter was kept constant at the optimal value identified above (0.1 m). Results are depicted in Fig. 4. With respect to N_c, values larger than 3 did not lead to substantial RMSE improvement. The optimal range for T_p can be found between 1.0 and 1.3. Throughout all the experiments, the lowest RMSE (i.e. 2.37 m) was found using 1.30 for the threshold for peak identification and 7 for the number of clusters.

4.1.3. GD-specific parameters

The GD process, as a non-linear fitting procedure, required an optimization algorithm that iteratively tries to improve the fit. The maximum number of iterations N_{max} is an important parameter as it significantly affects the execution speed; the smaller the maximum iteration number the better. Fig. 5 examines the influence of this parameter to ground detection accuracy while the

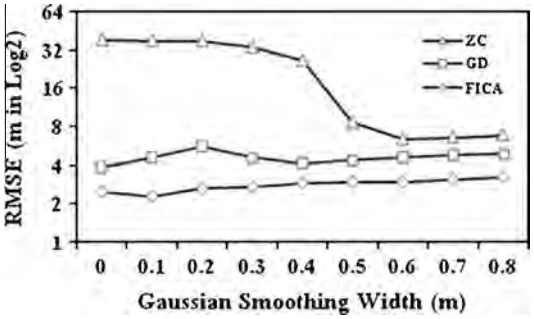

Fig. 3. RMSEs of GPIAs for different Gaussian smoothing filter width.

Gaussian filter width varied from 0 to 0.8. A general decreasing trend can be found, however the differences are very small as indicated by the short range of the *Y* axis. The Gaussian filter width also has limited influence to the algorithm performance, as indicated in Fig. 3. The minimum RMSE was found when N_{max} was 60 and no smoothing was applied (i.e. width was zero).

4.1.4. ZC-specific parameters

The ZC algorithm has two parameters, the smoothing width and the threshold for peak identification (T_p). The influence of T_p was evaluated in Fig. 6 while the smoothing width varied from 0 to 0.8. The Gaussian filter width showed significant influence in the algorithm performance.

Results indicate that a Gaussian filter helps but in order to achieve optimal results a relatively small T_p value should also be combined. A local minimum was found when T_p was 0.15 and the width was 0.6. After that, the RMSE increased sharply stabilizing at a T_p of 0.8.

4.1.5. FICA parameter influence on ground peak detection on different land cover types

Determination of ground peaks varies among land cover types due to high variability in their vertical structures. A sensitivity analysis took place using the calibration sample plots to investigate how FICA calibration parameter choices affect ground detection accuracy in each land cover type. The RMSE was calculated for a series of values for one parameter, while other parameters were fixed at their optimal values.

Three FICA parameters were examined: Gaussian smoothing filter width (σ), threshold for peak identification (T_p), and number of clusters used in *k*-means (N_c) and their influence in different land cover types (Fig. 7). Two different trends of RMSE can be found, as σ increased in Fig. 7a. In the deciduous and coniferous plots, the RMSE increased as the σ became larger; while, in the other types of plots, RMSE had a slightly decreasing trend. This may be attributed to the fact that forested plots tend to absorb a significant amount of the transmitted signal as it propagates vertically through the thick canopy leaving a weak ground return. A larger σ width would eliminate weak peaks, which may include the ground peak, and would lead to less accurate results. The identification of ground peaks present in grass, shrub and developed area plots became better as the width increased, since the smoothing filter eliminated the background noise that may lead to incorrect ground peak identification.

The two similar trends of Fig. 7a can also be found in Fig. 7b for the threshold for peak identification (T_p). Local minimum RMSEs can be found for coniferous and deciduous plots when T_p was set at 0.4 and 1.0, respectively. On the other hand, RMSE of other types of plots slightly decreased. In the tree plots, weak ground peaks would be missed by FICA, if a too large threshold for the peak identification was set. For example, the RMSE of the deciduous plots increased when the T_p changed from 1.0 to 2.0. The ground peak in the non-deciduous plot types was much stronger; therefore, the increase of T_p did not affect FICA's performance but helped reduce false peaks from background noise. Also, false ground peak may be identified, when a small threshold was set. The misidentification may be due to signal noise and/or reflection from the understory vegetation.

In Fig. 7c, a general decreasing error trend was found for the number of clusters used in *k*-means (N_c) in the deciduous, coniferous and developed area plots. On the other hand the RMSE of grass plots increased as N_c increased while the RMSE of shrub plots did not have substantial variability. A possible explanation for higher error in grass plots is that while only one large peak may exist in the signal, larger N_c value would force the algorithm to identify non-existent multiple clusters from the peak candidates and thus decrease the accuracy. Inversely, the waveform of coniferous, deciduous and developed area had multiple peaks representing different vegetation or building layers. The increased N_c assisted the algorithm to associate each cluster with a vertical layer. A significant decrease of RMSE can be seen in the developed area plots as the number of clusters increased from 2 to 3. Three clusters reflect better the complexity of the vertical layers as different ground features (e.g. vegetation, buildings and ground) may be present in the same plot.

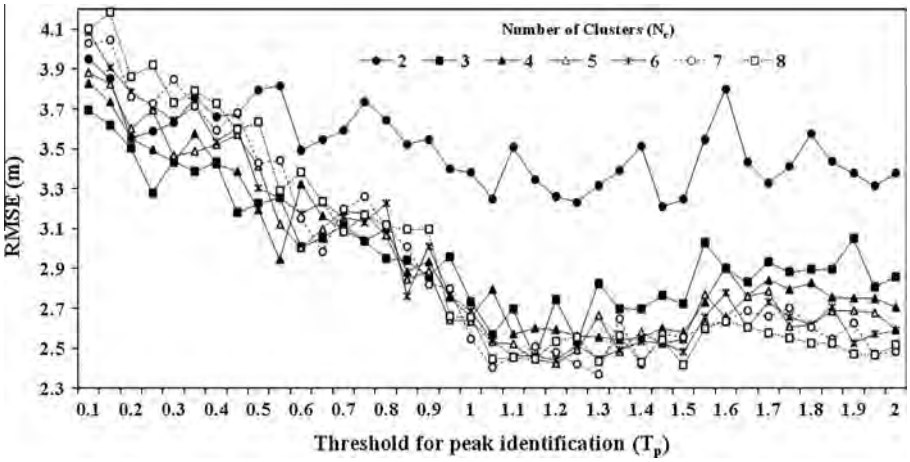

Fig. 4. FICA sensitivity analysis for parameters: threshold for peak identification and the number of clusters (smoothing width was fixed at 0.1 m).

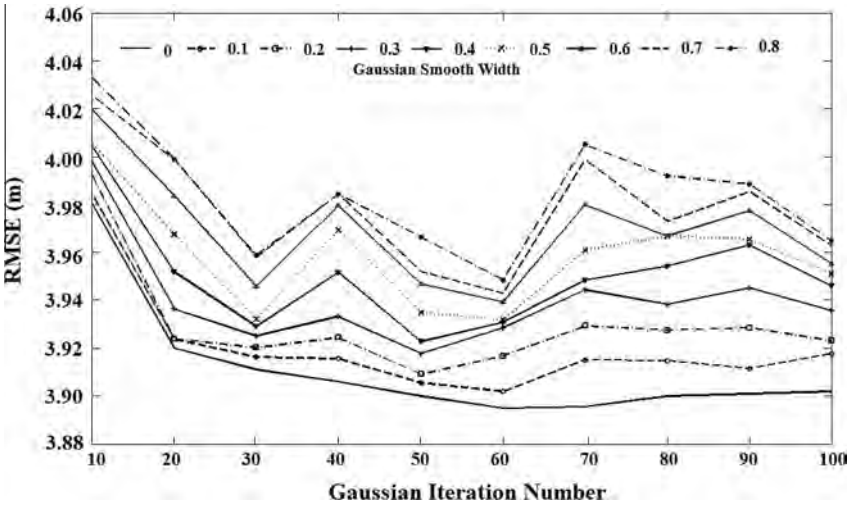

Fig. 5. GD sensitivity analysis for different maximum iteration numbers and Gaussian smoothing widths.

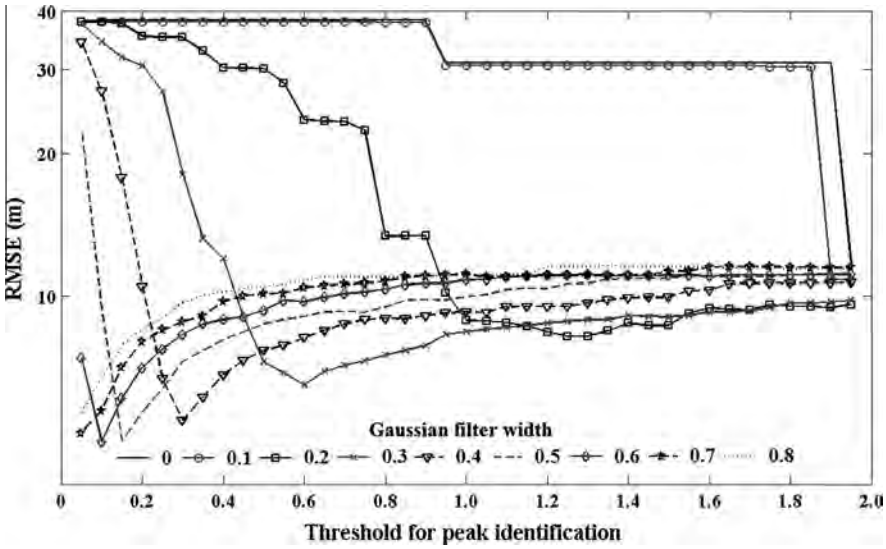

Fig. 6. ZC sensitivity analysis for the threshold for peak identification parameter and Gaussian smooth filter width.

4.2. Accuracy assessment using validation data

In this section, we used the 2900 point validation dataset to evaluate ground peak identification accuracy with the best parameter settings extracted from the calibration process (Table 2). Our assessment included accuracy distribution in different land cover types, different CC, and different slopes.

4.2.1. Ground peak identification accuracy of FICA

FICA achieved accurate results for ground peak identification using large footprint LiDAR data in five different land cover types, as shown in Fig. 8. The optimal parameter settings for FICA's initialization are shown in Table 2 above. Accuracy statistics were calculated by comparing the ground elevation detected from LVIS waveform using the FICA method and the reference elevation from the discrete LiDAR data. Correlation coefficient, bias and RMSE accuracy statistics were calculated for each land cover type. All FICA-derived ground elevations had high correlation with the reference elevation. The RMSE and bias in tree and shrub plots were higher than grass and developed area plots, results expected due to higher vertical complexity in the former plots. Negative bias can be seen in the deciduous and coniferous tree plots. This indicates that the identified ground was higher than the reference ground peak and it can be attributed to the weakening of the ground peak due to the occlusion effect from the canopy. When the ground peak is weak stronger peaks from the subcanopy or shrub layer may be identified as ground.

4.2.2. Accuracy comparison among different land cover types

A further comparison was conducted between FICA and the two benchmark algorithms, GD, and ZC. As shown in Fig. 9, FICA significantly outperformed the other two algorithms in the deciduous and coniferous plots, while in the shrub, grass and developed area plots FICA and GD had similar performance. The RMSE of ZC was much higher than both FICA and GD in all plot types. In the deciduous and coniferous plots, the vegetation structures were more complex than the other types of plots. Signal overlapping and weak ground peaks may exist in the deciduous and coniferous plots. FICA was developed based on a peak identification strategy and therefore it can better capture the signal fluctuation in a local range. The GD algorithm that adopts a curve-fitting strategy may not correctly identify the weak or overlapping ground peaks. For the ZC

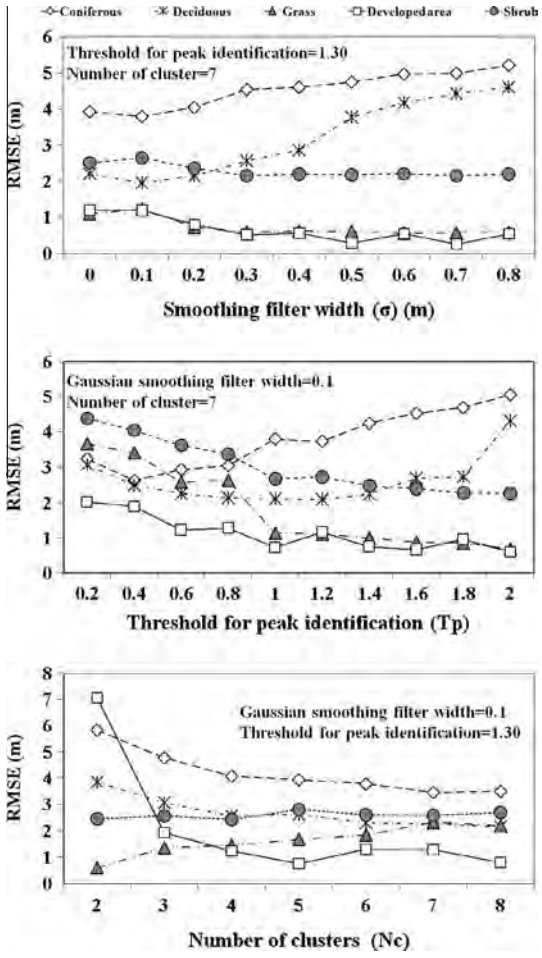

Fig. 7. Land cover type sensitivity analysis of FICA: (a) Gaussian smoothing filter width, (b) threshold for peak identification, and (c) number of cluster used in *K*-means algorithm. Other parameters settings are also shown in each figure.

Table 2
Parameter choice for GPIAs after algorithmic calibration.

GPIAs	Parameter optimization after training
FICA	$\sigma = 0.1$, $T_p = 1.30$, $N_c = 7$
GD	$\sigma = 0$ (no smoothing), $N_{max} = 60$
ZC	$\sigma = 0.6$, $T_p = 0.15$

algorithm, it was shown highly sensitive to background noise despite the fact that a smoothing filter was applied.

4.2.3. Accuracy comparison among different forest canopy coverage

This section examines the occlusion effect from canopy cover on the accuracy of ground peak identification in GD and FICA. ZC was not included in further analysis as results from Fig. 9 indicated poor performance. The energy of the transmitted pulse reaching the ground may be significantly weakened due to strong occlusion from canopy. The weak ground peak can lead to a less accurate result of ground peak identification if high background noise was present and/or the peak was eliminated by the smoothing filter. Furthermore, the reflection from dense low vegetation may overlap with the ground reflection. Three different vegetation types were assessed, deciduous, coniferous and shrub.

Each plot was categorized in one of two groups based on its CC: high CC (vegetation height > 4 m) and low CC (vegetation height < 4 m) (Riaño et al., 2002, 2007). The division of canopy cover allowed detailed investigation of the influence of CC on the algorithmic performance at different height layers. For coniferous and shrub plots, the CC was estimated through the ratio between the number of discrete LiDAR points from the canopy and total number of discrete return LiDAR points within an LVIS footprint (Hall et al., 2005; Jensen et al., 2008) (shown in Eq. (6)), as the number of LiDAR points from canopy did not significantly decrease in leaf-off season (the season discrete LiDAR was collected). Unfortunately, in the deciduous plots, the number of LiDAR points from canopy would be negatively biased due to the leaf-off season acquisition for discrete LiDAR. Thus, an alternative metric was incorporated to examine the coverage percentage of vegetation in the canopy layer or the understory layer out of the whole vegetation coverage represented by the summation of the LiDAR points higher and lower than 4 m within a plot (shown in Eq. (7)).

$$\begin{cases} CC_{High} = \frac{N_{Height>4m}}{N_{allpoints}} \\ CC_{Low} = \frac{N_{Height<4m}}{N_{allpoints}} \end{cases} \quad (for\ coniferous\ and\ shrub\ plots) \tag{6}$$

$$\begin{cases} CC_{Relativehigh} = \frac{N_{Height>4m}}{N_{Vegetationpoints}} \\ CC_{Relativelow} = \frac{N_{Height<4m}}{N_{Vegetationpoints}} \end{cases} \quad (for\ deciduous\ plots) \tag{7}$$

CC ranges were defined so that a statistically meaningful sample size larger than 5% of the total number (Kotrlik and Higgins, 2001) can be reached in most groups. RMSEs were calculated from the plots within each canopy cover interval. The RMSEs corresponding to different relative high and low CC for deciduous plots are shown in Fig. 10a and b and coniferous in Fig. 10c and d, respectively. In both graphs FICA showed stable performance, as opposed to clear trends present in the GD algorithm. The shrub plots (Fig. 10e and f) plots showed a slight trend for GD in the high CC, with no obvious trend in the low CC.

The difference in performance patterns between FICA and GD demonstrates the results of different peak identification strategies applied in the two algorithms. In GD, the occlusion effect from high CC became the major limiting factor. This is because the generalized criterion (i.e. global fitting error) incorporated by the GD for searching the optimized parameters for the Gaussian functions (Hofton et al., 2000; Wagner et al., 2006) forces the algorithm to put more attention in accurately simulating stronger rather than weaker peaks. The canopy peak was stronger as the high CC increased, and consequently the ground peak became weaker and thus easier to miss.

This is consistent with Jutzi and Stilla (2006) that pointed out that the ground peak may not be identified by the fitting strategy in GD if the distance between the ground peak and the canopy peak was smaller than $0.85\sigma_t$ (σ_t was the length of the transmitted pulse) in small footprint LiDAR data. On the other hand, the weak ground peak can still be detected using the peak detection strategy in FICA as long as it does not completely overlap with the canopy peak, since FICA only focuses on the local maximum. That is the reason behind FICA RMSEs remaining relatively stable in different CC and land cover types.

4.2.4. Accuracy comparison among different slopes

The reflected ground signal can easily overlap returns from low vegetation in a sloped area, since the temporal duration of the ground reflected signal becomes wider as slope increases (Chen, 2010b). This signal overlap may lead to misidentification of the ground peak. In this section the slope effect was examined for both GD and FICA. The deciduous (Fig. 11a), coniferous (Fig. 11b) and shrub (Fig. 11c) plots are divided into 6 slope groups with a slope interval of 4°. The number of plots for each slope group is also

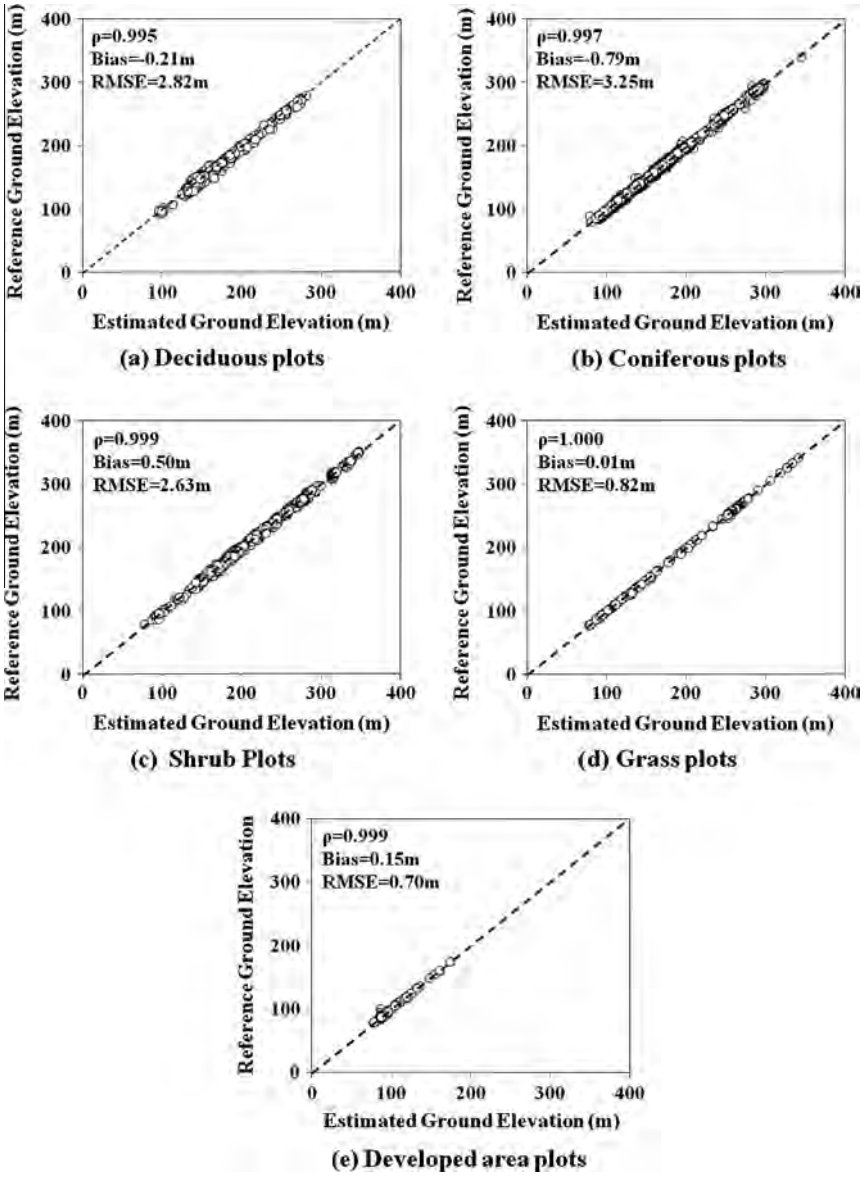

Fig. 8. Ground peak identification results for different land cover types using FICA.

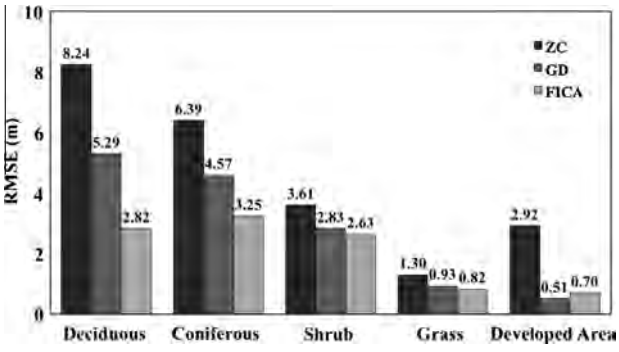

Fig. 9. RMSE comparisons in different land cover types.

shown in Fig. 11. The grass and developed area plots were not included in the analysis, since the occlusion effect was weak and the ground peaks in waveforms were sufficiently strong to allow accurate ground peak detection.

FICA showed relatively stable accuracy among different slope groups in deciduous plots compared to GD. The largest RMSE (3.32 m) in the deciduous plots (Fig. 11a) can be observed in the areas with slope between 12° and 16°; while the smallest RMSE (2.09 m) was calculated in the plots with slope between 4° and 8°. In the coniferous plots both FICA and GD showed an increasing RMSE trend (Fig. 11b) but FICA's errors were overall smaller. In the shrub plots (Fig. 11c), no clear pattern of the RMSE was found for both FICA and GD, although overall FICA had a small advantage. The significant accuracy degradation as slope increases for the GD method can be attributed to the complexity of the vertical structure of deciduous plots. In deciduous plots, the reflection of heavy subcanopy in the middle layer of the deciduous plots may overlap with the ground reflection when the slope is large. The reflection from the 'bottom heavy' canopies led to the inaccuracy of the ground peak identification.

As previously stated, FICA applied a peak detection algorithm, which considered the local shape of a waveform. FICA can detect a weak ground peak despite being masked by slope returns. On the other hand, GD may not be able to do so due to the curve-fitting strategy and its global fitting performance assessment. This

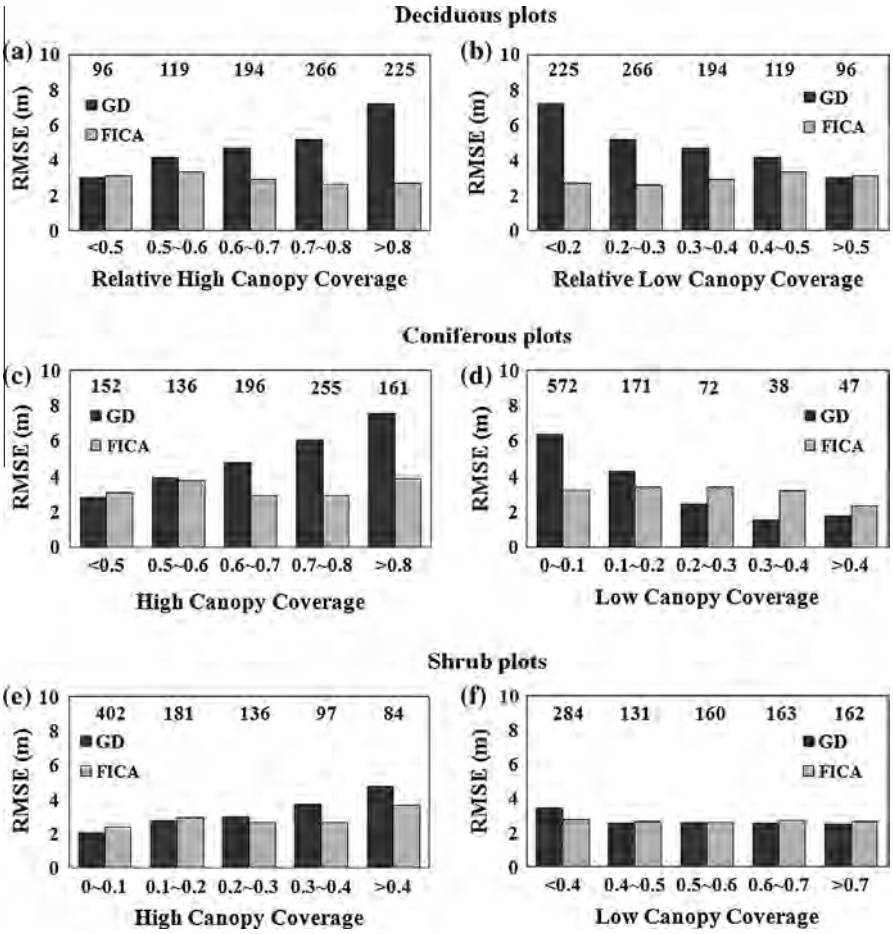

Fig. 10. Relationship between ground extraction errors and CC. Number of plots within each CC group shown on top of each bar.

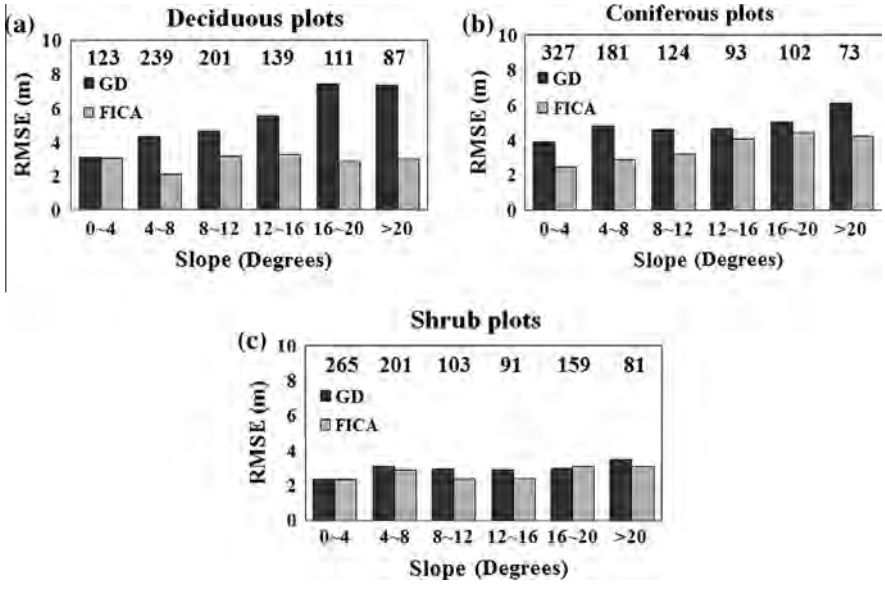

Fig. 11. Height extraction errors for different slope groups: (a) deciduous plots, (b) coniferous plots and (c) shrub plots. The number of plots used in each slope group is shown on the top of each graph.

is especially evident in the deciduous plots depicted in Fig. 11a. In Fig. 11b FICA is also more error-tolerant than GD in coniferous plots, although slope this time does influence FICA's results. This can be attributed to the larger understory coverage of low vegetation in the coniferous plots than that in the deciduous plots. In the shrub plots, GD and FICA had a similar RMSE and both had no clear trend, probably due to the fact that the ground peak may totally dissolve into the strong canopy reflection and become a single

large peak. This situation was common in the shrub waveform even in low slopes, because shrub heights are smaller than tree heights (Lefsky et al., 1999).

4.3. Algorithmic comparison in computational efficiency

A computational efficiency comparison was conducted between FICA and GD algorithms; ZC was not included due to its low accuracy. The experiments in this study were conducted in Matlab 2012a on a powerful workstation computer (four cores Intel i7 3.4 GHz CPU with 32 GB memory). Execution times were measured using the tic and toc command in Matlab and no parallel processing was implemented. As shown in Fig. 12, FICA ran much faster than GD. Because FICA took additional time to run the smoothing procedure, FICA without smoothing had better efficiency. The efficiency of FICA was also slightly affected by the number of clusters used in the *k*-means algorithm; the RMSE was increased as lower number of clusters was used in the *k*-means process. In GD, the computational execution time was severely affected by the maximum iteration number used in the LM optimization algorithm; however, the accuracy only slightly improved.

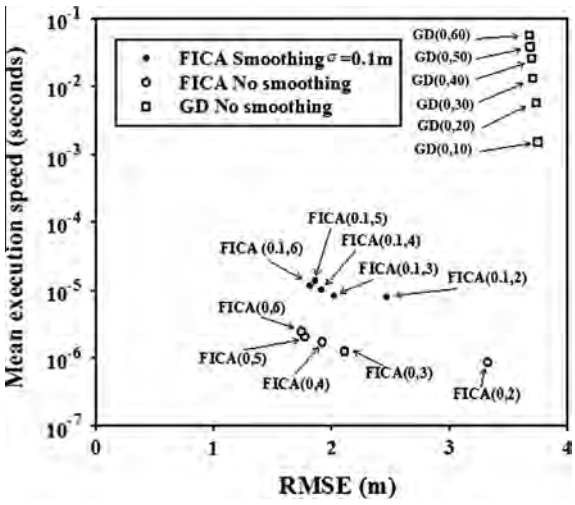

Fig. 12. Mean execution time and RMSE plots for FICA and GD. Notation is (Gaussian filter size, number of clusters) for FICA, (Gaussian filter size, maximum iteration number) for GD. Zero Gaussian filter size indicates no smoothing filtering.

The computational efficiency comparison was also conducted between different land cover types, as shown in Fig. 13. The parameter settings for FICA and GD were the same as those settings in Table 2. It can be seen in the figure that compared to FICA, the execution speed of GD had more fluctuation among the waveforms in different land cover types. In deciduous and coniferous plots, there may be more peaks in the waveforms due to multiple vegetation layers present in the forest structure. This would force GD to take more time to optimize the parameters of the Gaussian functions. FICA did not use the optimization strategy to identify the ground peak; therefore, it did not show variability in the execution time among different land cover types.

In summary, the FICA method is sufficiently fast for large-scale applications. If we take for example a dataset of 27 billion samples (roughly the NLCD national coverage at 30 m pixel size) that would take approximately (27 * 10^9 samples × 10^−5 s/sample) = 27 * 10^4 s, which is approximately 75 h. This is a very reasonable execution time that could be easily reduced further through parallel processing techniques since each sample plot processing is independent of others (e.g. a typical 4-core machine could process the entire continental U.S. in less than a day).

5. Conclusion

Ground detection is a significant bottleneck toward full utilization of large-footprint LiDAR potential. Accurate ground detection is essential for above ground height estimation, an important product for biodiversity. Ground detection is also a key factor in above ground biomass estimation, a topic that has attracted significant interest due to its carbon sequestration and climate linkages.

Accurate ground detection is also a challenging task, especially in vegetated areas. The proposed Filtering and Clustering Algorithm showed significant improvements over existing methods. In most cases it is more accurate and this accuracy holds stable for a variety of landscape characteristics such as varying slope and CC. In other words, it is a more trustworthy algorithm for large-scale implementation. Considering the possibility of placing the LVIS sensor in space, our work contributes toward unique, accurate and standardized product creation (e.g. similarly to MODIS products). In addition, FICA is significantly faster to

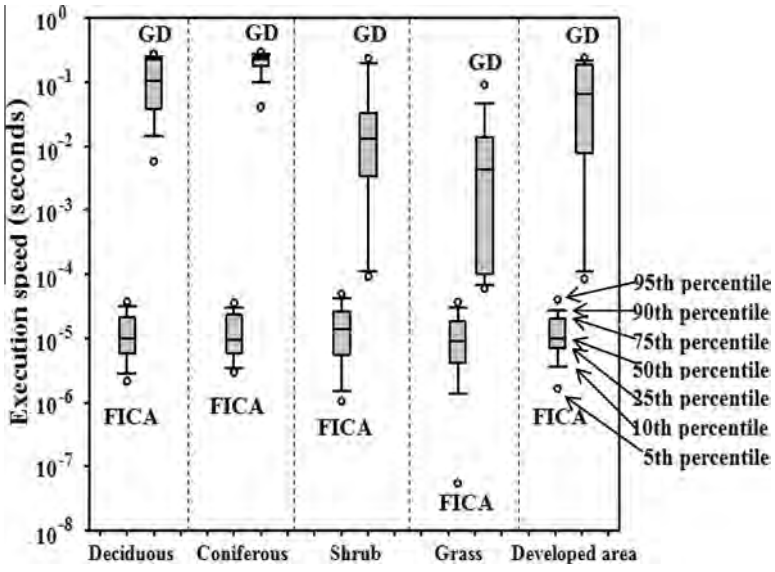

Fig. 13. Boxplot of the execution time for FICA and GD in different land cover types.

execute, providing another major advantage for large-scale mapping. It is also simple to implement in any programming language and users can easily relate to parameters definitions. Our analysis also identified two areas of future improvement, minimizing the effect of slope and specifically targeting shrubland for further algorithmic development.

Acknowledgements

This work was supported through NASA's Biodiversity Program (Grant # NNX09AK16G). We would like to thank Dr. Blair and Dr. Hofton for providing the LVIS dataset as part of that grant, Dr. Beier and Mr. Wiley for helpful discussions and the anonymous reviewers for significantly improving this manuscript.

References

Andersen, H.E., McGaughey, R.J., Reutebuch, S.E., 2005. Estimating forest canopy fuel parameters using LIDAR data. Remote Sens. Environ. 94 (4), 441–449.

Anderson, J., Martin, M.E., Smith, M.L., Dubayah, R.O., Hofton, M.A., Hyde, P., Peterson, B.E., Blair, J.B., Knox, R.G., 2006. The use of waveform LiDAR to measure northern temperate mixed conifer and deciduous forest structure in New Hampshire. Remote Sens. Environ. 105 (3), 248–261.

Anderson, J.E., Ducey, M.J., Fast, A., Martin, M.E., Lepine, L., Smith, M.L., Lee, T.D., Dubayah, R.O., Hofton, M.A., Hyde, P., Peterson, B.E., Blair, J.B., 2011. Use of waveform LiDAR and hyperspectral sensors to assess selected spatial and structural patterns associated with recent and repeat disturbance and the abundance of sugar maple (Acer saccharum Marsh.) in a temperate mixed hardwood and conifer forest. J. Appl. Remote Sens. 5 (1), 053504–053504.

Blair, J.B., Rabine, D.L., Hofton, M.A., 1999. The laser vegetation imaging sensor: a medium-altitude, digitisation-only, airborne laser altimeter for mapping vegetation and topography. ISPRS J. Photogramm. Rem. Sens. 54 (2–3), 115–122.

Blair, J.B., Hofton, M.A., Rabine, D.L., 2006. Processing of NASA LVIS Elevation and Canopy (LGE, LCE and LGW) Data Products, version 1.01. <http://lvis.gsfc.nasa.gov>.

Chauve, A., Vega, C., Durrieu, S., Bretar, F., Allouis, T., Deseilligny, M.P., Puech, W., 2009. Advanced full-waveform LiDAR data echo detection: assessing quality of derived terrain and tree height models in an alpine coniferous forest. Int. J. Remote Sens. 30 (19), 5211–5228.

Chen, Q., 2010a. Assessment of terrain elevation derived from satellite laser altimetry over mountainous forest areas using airborne LiDAR data. ISPRS J. Photogramm. Rem. Sens. 65 (1), 111–122.

Chen, Q., 2010b. Retrieving vegetation height of forests and woodlands over mountainous areas in the Pacific Coast region using satellite laser altimetry. Remote Sens. Environ. 114 (7), 1610–1627.

Cheng, F., Wang, C., Wang, J., Tang, F., Xi, X., 2011. Trend analysis of building height and total floor space in Beijing, China using ICESAT/GLAS data. Int. J. Remote Sens. 32 (23), 8823–8835.

Dubayah, R.O., Sheldon, S.L., Clark, D.B., Hofton, M.A., Blair, J.B., Hurtt, G.C., Chazdon, R.L., 2010. Estimation of tropical forest height and biomass dynamics using LiDAR remote sensing at la Selva, Costa Rica. J. Geophys. Res. G: Biogeosci. 115 (G2), 2156–2202.

Duncanson, L.I., Niemann, K.O., Wulder, M.A., 2010. Estimating forest canopy height and terrain relief from GLAS waveform metrics. Remote Sens. Environ. 114 (1), 138–154.

Gong, P., Li, Z., Huang, H., Sun, G., Wang, L., 2011. ICEsat GLAS data for urban environment monitoring. IEEE Trans. Geosci. Remote Sens. 49 (3), 1158–1172.

Gonzalez, C.R., Wintz, P., 2008. Digital Image Processing, second ed. CRC, 160–162 and 276–277.

Hall, S., Burke, I., Box, D., Kaufmann, M., Stoker, J., 2005. Estimating stand structure using discrete-return LiDAR: an example from low density, fire prone ponderosa pine forests. For. Ecol. Manage. 208 (1–3), 189–209.

Harding, D.J., Carabajal, C.C., 2005. ICESat waveform measurements of within-footprint topographic relief and vegetation vertical structure. Geophys. Res. Lett. 32 (21), L21S10.

Hofton, M.A., Minster, J.B., Blair, J.B., 2000. Decomposition of laser altimeter waveforms. IEEE Trans. Geosci. Remote Sens. 38 (4), 1989–1996.

Hofton, M., Dubayah, R., Blair, J.B., Rabine, D., 2006. Validation of SRTM elevations over vegetated and non-vegetated terrain using medium footprint LiDAR. Photogramm. Eng. Rem. Sens. 72 (3), 279–285.

Hofton, M., Blair, J., Luthcke, S., Rabine, D., 2008. Assessing the performance of 20–25 m footprint waveform LiDAR data collected in ICESat data corridors in Greenland. Geophys. Res. Lett. 35 (24), L24501.

Hyde, P., Dubayah, R., Peterson, B., Blair, J., Hofton, M., Hunsaker, C., Knox, R., Walker, W., 2005. Mapping forest structure for wildlife habitat analysis using waveform LiDAR: validation of montane ecosystems. Remote Sens. Environ. 96 (3–4), 427–437.

Jensen, J.L., Humes, K.S., Vierling, L.A., Hudak, A.T., 2008. Discrete return LiDAR-based prediction of leaf area index in two conifer forests. Remote Sens. Environ. 112 (10), 3947–3957.

Jutzi, B., Stilla, U., 2006. Range determination with waveform recording laser systems using a Wiener Filter. ISPRS J. Photogramm. Rem. Sens. 61 (2), 95–107.

Kotrlik, J.W.K.J.W., Higgins, C.C.H.C.C., 2001. Organizational research: determining appropriate sample size in survey research appropriate sample size in survey research. Inf. Technol. Learn. Perform. J. 19 (1), 43–50.

Lefsky, M.A., Cohen, W.B., Acker, S.A., Parker, G.G., Spies, T.A., Harding, D., 1999. LiDAR remote sensing of the canopy structure and biophysical properties of Douglas-fir western hemlock forests. Remote Sens. Environ. 70 (3), 339–361.

Lefsky, M.A., Harding, D.J., Keller, M., Cohen, W.B., Carabajal, C.C., Del Bom Espirito-Santo, F., Hunter, M.O., de Oliveira Jr., R., 2005. Estimates of forest canopy height and aboveground biomass using ICESat. Geophys. Res. Lett. 32 (22), 1–4.

Lefsky, M.A., Keller, M., Pang, Y., De Camargo, P.B., Hunter, M.O., 2007. Revised method for forest canopy height estimation from Geoscience Laser Altimeter System waveforms. J. Appl. Remote Sens. 1 (1), 013537–013537.

Lim, K., Treitz, P., Wulder, M., St-Onge, B., Flood, M., 2003. LiDAR remote sensing of forest structure. Prog. Phys. Geogr. 27 (1), 88–106.

Mallet, C., Bretar, F., 2009. Full-waveform topographic LiDAR: State-of-the-art. ISPRS J. Photogramm. Rem. Sens. 64 (1), 1–16.

Pang, Y., Li, Z.Y., Sun, G., Lefsky, M., Che, X.J., Chen, E.X., 2007. Effects of terrain on large footprint LiDAR waveform of forests. Forest Res. 20 (4), 464–468.

Popescu, S.C., Zhao, K., Neuenschwander, A., Lin, C., 2011. Satellite LiDAR vs. small footprint airborne LiDAR: comparing the accuracy of aboveground biomass estimates and forest structure metrics at footprint level. Remote Sens. Environ. 115 (11), 2786–2797.

Riaño, D., Chuvieco, E., Salas, J., Palacios-Orueta, A., Bastarrika, A., 2002. Generation of fuel type maps from Landsat TM images and ancillary data in Mediterranean ecosystems. Can. J. For. Res. 32 (8), 1301–1315.

Riaño, D., Chuvieco, E., Ustin, S.L., Salas, J., Rodríguez-Pérez, J.R., Ribeiro, L.M., Viegas, D.X., Moreno, J.M., Fernández, H., 2007. Estimation of shrub height for fuel-type mapping combining airborne LiDAR and simultaneous color infrared ortho imaging. Int. J. Wildland Fire 16 (3), 341–348.

Soderman, U., Persson, A., Topel, J., Ahlberg, S., 2005. On Analysis and Visualization of Full-Waveform Airborne Laser Scanner Data. In: Proc. SPIE 5791, Laser Radar Technology and Applications X, pp. 184–192.

Sun, G., Ranson, K.J., Kimes, D.S., Blair, J.B., Kovacs, K., 2008. Forest vertical structure from GLAS: an evaluation using LVIS and SRTM data. Remote Sens. Environ. 112 (1), 107–117.

Sun, G., Ranson, K.J., Guo, Z., Zhang, Z., Montesano, P., Kimes, D., 2011. Forest biomass mapping from LiDAR and radar synergies. Remote Sens. Environ. 115 (11), 2906–2916.

Wagner, W., Ullrich, A., Melzer, T., Briese, C., Kraus, K., 2004. From single-pulse to full-waveform airborne laser scanners: potential and practical challenges. Int. Archives Photogramm. Rem. Sens. 35, 201–206.

Wagner, W., Ullrich, A., Ducic, V., Melzer, T., Studnicka, N., 2006. Gaussian decomposition and calibration of a novel small-footprint full-waveform digitising airborne laser scanner. ISPRS J. Photogramm. Rem. Sens. 60 (2), 100–112.

Wasser, L., Day, R., Chasmer, L., Taylor, A., 2013. Influence of vegetation structure on LiDAR-derived canopy height and fractional cover in forested riparian buffers during leaf-off and leaf-on conditions. PLoS ONE 8 (1), e54776.

Zwally, H.J., Schutz, B., Abdalati, W., Abshire, J., Bentley, C., Brenner, A., Bufton, J., Dezio, J., Hancock, D., Harding, D., Herring, T., Minster, B., Quinn, K., Palm, S., Spinhirne, J., Thomas, R., 2002. ICESat's laser measurements of polar ice, atmosphere, ocean, and land. J. Geodyn. 34 (3–4), 405–445.