Paper:
Chen, B., Pang, Y., Li, Z., Lu, H., Liu, L., North, P. & Rosette, J. (2019). Ground and Top of Canopy Extraction From Photon-Counting LiDAR Data Using Local Outlier Factor With Ellipse Searching Area. *IEEE Geoscience and Remote Sensing Letters, 16*(9), 1447-1451.
http://dx.doi.org/10.1109/LGRS.2019.2899011

Ground and Top of Canopy Extraction from Photon Counting LiDAR Data Using Local Outlier Factor with Ellipse Searching Area

Bowei Chen, *Student Member, IEEE*, Yong Pang, Zengyuan Li, Hao Lu, Luxia Liu, P.R.J. North, and J.A.B. Rosette

***Abstract*—The future ICESat-2 is the next generation of NASA's ICESat (Ice, Cloud and land Elevation Satellite) mission scheduled to be launched in 2018. The new photon counting LiDAR onboard ICESat-2 introduced new challenges to the estimation of forest parameters and their dynamics, the greatest being the abundant photon noise appearing in returns from the atmosphere and below the ground. To identify the potential forest signal photons, we proposed an approach by using a local outlier factor (LOF) modified with ellipse searching area. Six test data from two kinds of the photon counting LiDAR data in the USA are used to test and evaluate the performance of our algorithm. The classification results for noise and signal photons showed our approach has a good performance not only in lower noise rate with relatively flat terrain surface but also works even for a quite high noise rate environment in relatively rough terrain. The quantitative assessment indicates that the horizontal ellipse searching area gives the best result compared with the circle or vertical ellipse searching area. These results demonstrate our methods would be useful for future ICESat-2 vegetation study.**

***Index Terms*—LOF, ICESat-2, photon classification, photon counting LiDAR.**

I. Introduction

ACCURATE estimation of forest height and biomass is critically important for understanding the regional and global carbon cycle and dynamic changes [1]. Successful mapping of critical forest parameters using NASA's GLAS (the Geoscience Laser Altimeter System) onboard the ICESat (Ice, Cloud and land Elevation Satellite) mission [2] showed the potential in vegetation studies [3]–[6]. ICESat-2, which is the next generation of ICESat missions, is scheduled to be launched in 2018. In contrast to the previous waveform LiDAR system, ICESat-2 will adopt a newly designed LiDAR system named ATLAS (Advanced Topographic Laser Altimeter System), a micro-pulse, multi-beam photon counting LiDAR system working at 532 nm [7]. To pre-validate the capability of this new sensor, NASA designed several airborne and micro-pulsed laser instruments, including SIMPL (the Slope Imaging Multi-polarization Photon-counting LiDAR) and MABEL (the Multiple Altimeter Beam Experimental LiDAR), and used in flight campaigns over the past few years [8].

From the currently released data products [9], the photon counting approach introduced abundant noise appearing in the atmosphere and even below the ground, making it difficult to extract the correct canopy and ground surface in vegetation area [10]. A few studies have been done to detect the noise and separate the signal, such as a spatial statistical detection algorithm based on discrete mathematical concepts [11], an ellipse search area based on DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [12] and an automated algorithm using the cumulative density of photons to identify cut off points of canopy top and ground [13]. In addition, a recent novel methodological framework to retrieve ground and canopy height for both MABEL and simulated ICESat-2 data achieved good results for various nighttime and daytime scenarios [14]. These studies showed good performance for MABEL and simulated data, but further development of noise filtering is still necessary to explore vegetation applications for future genuine ATLAS data.

In this letter, an approach that is capable of identifying potential forest signal photons by using local outlier factor (LOF) algorithm modified with an ellipse searching area is proposed for MABEL and MATLAS data. LOF is an unsupervised outlier detection method which computes a score for a point which indicates the local density around the given point to its near neighbors [15]. Points which have substantially higher scores will be considered as outliers. Our modified ellipse searching area uses the different density between horizontal and vertical directions, where the outliers will be detected more accurately.

II. Data and Methods

A. MABEL and MATLAS Data

Two kinds of the photon counting LiDAR data we used are one from MABEL and five from MATLAS. MABEL is a dual-wavelength (532 nm and 1064 nm) high-altitude system that was specifically developed as a demonstrator and validation instrument for ICESat-2 [16]. MATLAS data is generated by adjusting existing MABEL data to be more similar to the data

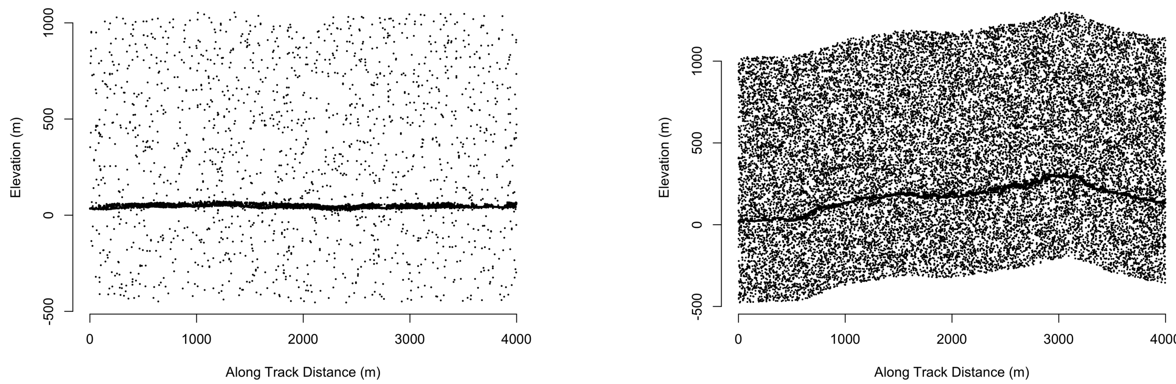
This work was supported by the National Natural Science Foundation of China under Grant 41871278 and Grant 31570546, and the China Scholarship Council. *(Corresponding author: Yong Pang.)*

Bowei Chen, Yong Pang and Zengyuan Li are with the Institute of Forest Resource Information Techniques, Chinese Academy of Forestry, Beijing 100091, China (e-mail:rs.cbw@foxmail.com, pangy@ifrit.ac.cn, lizengyuan@ifrit.ac.cn).

Hao Lu is with College of Information Science and Technology, Beijing Forestry University, Beijing 100083, China (e-mail:luhao@bjfu.edu.cn).

Luxia Liu is with School of Forestry and Landscape Architecture, Anhui Agricultural University, Hefei 230036, China (e-mail:luxia.liu@ahau.edu.cn).

P.R.J. North and J.A.B. Rosette are with Global Environmental Modelling and Earth Observation (GEMEO), Department of Geography, Swansea University, SA2 8PP, United Kingdom (e-mail:p.r.j.north@swansea.ac.uk, j.a.rosette@swansea.ac.uk).



(a) MABEL data in Virginia

(b) MATLAS data in Oregon

Fig. 1. Two examples of the photon counting LiDAR data used in this study: (a) the MABEL data collected in Virginia, USA, on September 20, 2012, with relatively flat terrain along the transect and a lower noise rate. (b) the MATLAS data collected in Oregon, USA, on September 27, 2012, with relatively rough terrain and a higher noise rate.

expected from the ATLAS system [17]. To produce MATLAS data, first, the signal, solar noise and instrument noise levels are adjusted based on NASA's vegetation design case model. Next, the spatial variation of signal and noise from the original MABEL is preserved. Finally, a large footprint size is formed by combining adjacent channels from the original MABEL data.

Fig. 1a showed the MABEL data collected in Virginia, USA, on September 20, 2012, with relatively flat terrain along the transect and a lower noise rate. In the meantime, we collected five MATLAS data from Oregon and West Coast flight campaigns. Fig. 1b showed the MATLAS data collected in Oregon, USA, on September 27, 2012, with relatively rough terrain and a higher noise rate. Both data are converted to the along track distance accordingly.

B. Methods

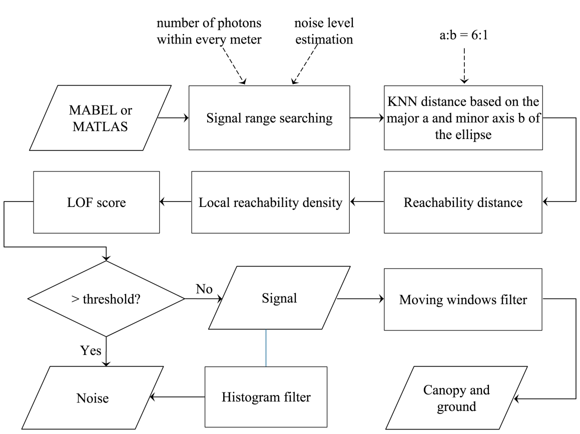


Fig. 2. Flowchart of the proposed LOF modified with ellipse searching area method.

It is noticeable that the density of the signal photons is different in terms of horizontal and vertical directions, the method proposed in this letter is to utilize the unbalanced distribution using range searching and a multi-window size histogram filter to distinguish the noise and signal. It involves the following three stages:

1) Signal Range Searching: Despite the numerous noise returns randomly scattered above and below the canopy, Fig. 3 shows the unbalanced signal density in the vertical direction, which will be used to get rid of the noise which is far away from the signal center for faster and easier calculation.

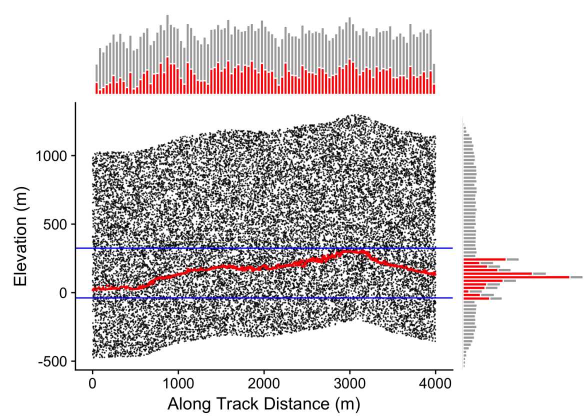


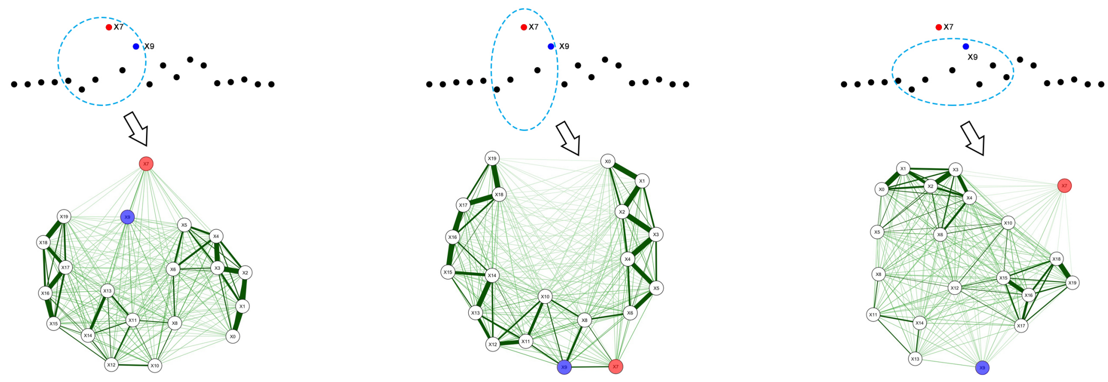
Fig. 3. Histogram showed the unbalanced signal distribution in vertical and horizontal directions of MATLAS data

We first count the number of photons within a 1 m interval based on the histogram along the elevation. Next, we calculate the mean value and standard deviation of the number of photons at the beginning and end of 50 meters. Therefore the background noise level is defined by the following equation:

$$N = \frac{\mu_1 + 2\sigma_1 + \mu_2 + 2\sigma_2}{2} \quad (1)$$

where N represents the background noise level, μ_1 and σ_1 represent the mean value and standard deviation for the first 50 records along the elevation. μ_2 and σ_2 represent the mean value and standard deviation for the last 50 records.

Along the elevation, the histogram showed that noise is evenly distributed apart from the center, so we can use thresh-



(a) Circle searching area ($a = b$) (b) Ellipse searching area ($a < b$) (c) Ellipse searching area ($a > b$)

Fig. 4. Illustration of distance matrix from different searching areas, where a represents the semi-major axis and b represents the semi-minor axis. The red dot stands for point X7 and the blue dot stands for point X9. Figure (a) gives a circle searching area, (b) gives a vertical ellipse searching area and (c) gives a horizontal ellipse searching area.

olds to get the signal range. The upper boundary threshold is considered to be the first bin which reaches the condition that the number of photons in the following five bins is continuously higher than the background noise level we defined. Similarly, the lower boundary threshold is considered to be the last bin which reaches the condition that the number of photons in the previous five bins is continuously higher than the background noise level. From the upper and lower boundary threshold, we can remove the noise which is far from the signal center.

2) Implementation of LOF algorithm with ellipse searching area: Here we implement our modified LOF algorithm with ellipse searching area and assign the class tag based on the score which was returned. The basic idea for the LOF algorithm is to compute a score for a point which indicates the local density among the given point to its nearby neighbors, where the outliers are considered to be the points which are substantially lower than a threshold score compared with the density level among their neighbors. Here we introduce an ellipse searching area instead of the circle one due to the higher spatial photon density in the horizontal direction. For any given point p and q in our data, the ellipse searching area is defined by the following equation:

$$dist_k(p, q) = \sqrt{\frac{(x_p - x_q)^2}{a^2} + \frac{(h_p - h_q)^2}{b^2}} \tag{2}$$

where x and h represent the along track distance and photon height, a and b represent the major and minor axis of the ellipse respectively. The different searching shapes and the distance matrix are shown in Fig. 4. The searching shape is determined by the ratio of the major and minor axis. In this letter, we used an empirical ratio which is a:b = 6:1.

Next, a reachability distance from point p to q is estimated using equation 3, which is the maximum value between the KNN distance of point q and the distance from point p and q.

$$reachdist(p, q) = \max\{(dist_k(q), dist(p, q)\} \tag{3}$$

Next, the inverse of the average reachability distance of point p from its neighbors is used to calculate the local reachability density, and then the LOF score is defined as the average local reachability density of the neighbors divided by point p's own local reachability density. A point with a lower value of LOF score indicates that this point is closer to its neighbors, so it would not be an outlier; on the contrary, a point with a higher value would be identified as the outlier. Here the threshold to separate the signal and outlier is based on the distribution of the LOF score values.

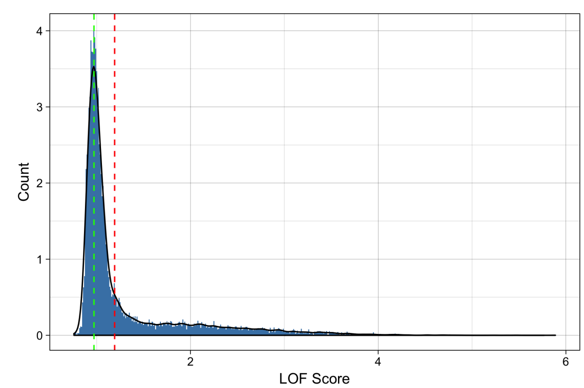
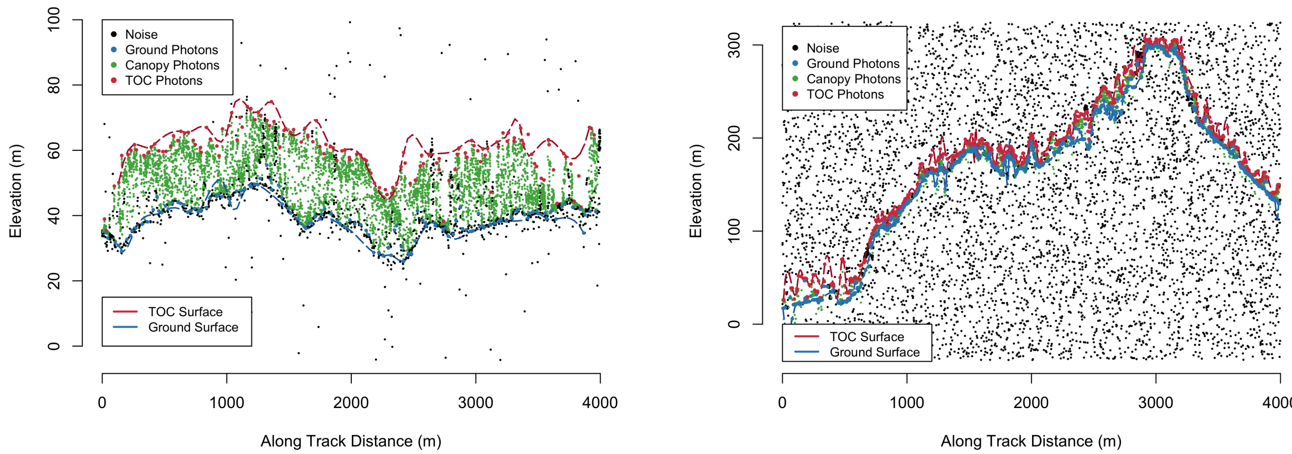


Fig. 5. Histogram of the LOF score from MATLAS data. The green line presents the peak value; the red line presents the threshold to define noise and signal.

Finally, the cutting point is defined as twice the distance between the beginning and peak of the histogram shown in Fig. 5. All points with LOF scores less than the threshold are tagged as signal photons, and the rest are labeled as noise photons instead.

3) Histogram filter and surface detection: Although most of the noise could be removed after the two steps discussed above, there can still be some dense cluster centers within the noise photons, resulting in some noise misclassified as signal photons. In order to assign the correct classification label to the remaining noise clusters, a histogram filter was implemented to detect these noises. Here we divide the whole area into small parts from the along track distance, then the signal center could be calculated, and any photons above or below the defined distance from the center will be considered as noise. The



(a) MABEL data in Virginia

(b) MATLAS data in Oregon

Fig. 6. The results of forest signal extraction from two example datasets used in the study. The red points stand for the top of canopy photons (TOC), blue points stand for ground photons, the green points stand for photons between the canopy and the ground, and the black ones stand for noise.

last part is to detect the points which belong to the top of the canopy (TOC), ground and within the canopy. Finally, we use a 20 m moving window to find the local maximum and minimum value within the signal photons detected as signal and ground photons respectively, then apply a 50 m moving window to obtain the canopy surface and ground surface.

C. Results Assessment

To quantitatively assess our results, four statistical indicators known as accuracy, kappa coefficient, specificity, and F1 are computed based on the confusion matrix. They are defined as follows: accuracy is the proportion of the total number of photons which are identified as signal and noise correctly; kappa coefficient measures the instances classified by our algorithm matching the data labeled as ground truth; specificity is the proportion of photons considered as noise photons that are correctly identified as such, and the F1 score is the harmonic mean of precision and sensitivity, in which the precision is the fraction of true signal photons from all points identified as photons and sensitivity is photons considered as signal photons that are correctly identified as such.

We use two kinds of reference data here to evaluate our classification results. The first one is the photon classification flags from the data products themselves given by NASA; the second one is our manually assigned class labels, which are visually adjusted and corrected based on the classification flags from the data products.

III. Results and Evaluation

The result of forest signal extraction from MABEL data is shown in Fig. 6a and the result of MATLAS data is in Fig. 6b. It can be seen from the results that our method could separate the forest signal from the noise effectively. In the meantime, it is noticeable that our approach could do a proper classification for signal and noise photons, and extract the canopy and ground surface from a quite high noise rate environment in relatively rough terrain for the MATLAS data here. It can be seen that the photons that belong to the TOC, ground and within the canopy are well detected.

In addition, we quantitatively assessed the sensitivity regarding different shapes of searching areas as demonstrated in Table I. The exact ratios of the major and minor axis for the horizontal, vertical and circular searching areas are 6:1, 1:6 and 1:1 respectively. It can be seen from the table that for assessment based on manual labeling, the overall mean accuracies are 0.89, 0.87 and 0.84 for the horizontal ellipse, circle, and vertical ellipse searching area respectively. The overall mean kappa coefficients are 0.76, 0.71 and 0.67. These two indicators show that our approach has good results to detect the signal from noise, especially the kappa coefficient reported the good agreement between the two classes. Furthermore, the overall mean specificity values are 0.87, 0.82 and 0.78, while the F1 measures are 0.85, 0.83 and 0.81 for the vertical ellipse, circle, and horizontal ellipse searching area respectively. From the results based on the classification flags from the data products themselves, we can see that there is a good consistency with results from manually assigned labels.

It is obvious that the shape of the searching area is sensitive to the final result. For all five test sites of MATLAS data, the horizontal ellipse searching area always gives the best result compared with the circle or vertical ellipse searching area in terms with the four statistical indicators we calculated above. The possible reason behind this could be the unbalanced information distribution, which would be quite useful for the photon classification.

It should be stated that the accuracy assessment using the presented method is not a sufficiently objective strategy, as manual labeling can vary from the experience of the users. Besides, the classification flags from the data products over vegetation can still be improved although they have been visually checked. The results could be better convincing if airborne laser scanning data can be precisely registered to simulation data or real ICESat-2 data for delineating the ground and

TABLE I
FOUR STATISTICAL INDICATORS OF MATLAS DATA BASED ON THE CIRCLE, HORIZONTAL ELLIPSE, AND VERTICAL ELLIPSE SEARCHING AREA, H STANDS FOR HORIZONTAL ELLIPSE, V STANDS FOR VERTICAL ELLIPSE AND C STANDS FOR CIRCLE

Test sites	Reference	Accuracy			Kappa coefficient			Specificity			F1 score		
		H	C	V	H	C	V	H	C	V	H	C	V
MATLAS, Oregon	Manual	0.91	0.85	0.86	0.81	0.69	0.70	0.90	0.80	0.81	0.87	0.8	0.81
MATLAS, West Coast Transit 2	Manual	0.93	0.94	0.91	0.87	0.87	0.81	0.89	0.88	0.82	0.94	0.94	0.92
MATLAS, West Coast Transit 3	Manual	0.88	0.82	0.79	0.71	0.61	0.57	0.89	0.82	0.74	0.80	0.75	0.72
MATLAS, West Coast Transit 4	Manual	0.92	0.89	0.87	0.83	0.78	0.74	0.85	0.78	0.74	0.93	0.91	0.90
MATLAS, West Coast Transit 5	Manual	0.81	0.83	0.78	0.58	0.62	0.54	0.84	0.83	0.78	0.72	0.75	0.71
Mean values in 5 test sites	Manual	0.89	0.87	0.84	0.76	0.71	0.67	0.87	0.82	0.78	0.85	0.83	0.81
Mean values in 5 test sites	NASA	0.91	0.87	0.85	0.79	0.73	0.68	0.87	0.82	0.77	0.87	0.83	0.81

canopy surface. Furthermore, potential signal photons after filtering can still include some outliers in the current work, and it remains a challenge to distinguish these outliers from those photons. Another one is that the accuracy of separating photons into ground and canopy is still limited to an extent, we will further explore a method by using some adaptive algorithms to better identify the ground and canopy photons [14], and test it with various noise levels and atmospheric moisture conditions in future work.

IV. CONCLUSION

In this letter, an approach that is capable of identifying potential forest signal photons by using LOF modified with ellipse searching area is proposed for MABEL and MATLAS data. The quantitative assessment using accuracy, kappa coefficient, specificity, and F1 measurement proved our algorithm works well regarding these indicators. Also, we found that the final results are sensitive to the shape of the searching area. The horizontal ellipse searching area gives the best result compared with the circle or vertical ellipse searching area. We tested our method not only with flat terrain in low noise situations but also verified with rough surface at high noise level with MATLAS data. These results demonstrated that the method we proposed could be of use for future ICESat-2 data vegetation applications.

ACKNOWLEDGMENT

The authors would like to thank the ICESat-2 Project Science Office at NASA/GSFC for providing the MABEL and MATLAS data used in this study. We also acknowledge the importance of the constructive criticism provided by the anonymous reviewers who helped improve this manuscript.

REFERENCES

[1] S. G. Zolkos, S. J. Goetz, and R. Dubayah, "A meta-analysis of terrestrial aboveground biomass estimation using lidar remote sensing," *Remote Sensing of Environment*, vol. 128, pp. 289–298, Jan. 2013.

[2] A. W. Yu, M. A. Stephen, S. X. Li, G. B. Shaw, A. Seas, E. Dowdye, E. Troupaki, P. Liiva, D. Poulios, and K. Mascetti, "Space laser transmitter development for ICESat-2 mission," in *Proceedings of the SPIE*. NASA Goddard Space Flight Ctr., United States, Feb. 2010, p. 757809.

[3] 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, "Estimates of forest canopy height and aboveground biomass using ICESat," *Geophysical Research Letters*, vol. 32, no. 2, p. L22S02, Nov. 2005.

[4] M. A. Lefsky, M. Keller, Y. Pang, P. B. De Camargo, and M. O. Hunter, "Revised method for forest canopy height estimation from geoscience laser altimeter system waveforms," *Journal of Applied Remote Sensing*, vol. 1, no. 1, p. 013537, 2007.

[5] L. I. Duncanson, K. O. Niemann, and M. A. Wulder, "Estimating forest canopy height and terrain relief from GLAS waveform metrics," *Remote Sensing of Environment*, vol. 114, no. 1, pp. 138–154, Jan. 2010.

[6] S. Los, J. Rosette, N. Kljun, P. North, L. Chasmer, J. Suárez, C. Hopkinson, R. Hill, E. Van Gorsel, C. Mahoney *et al.*, "Vegetation height products between 60 s and 60 n from icesat glas data." *Geoscientific Model Development*, vol. 5, pp. 413–432, 2012.

[7] T. Evans, "Optical Development System life cycle for the ICESat-2 ATLAS instrument," in *2014 IEEE Aerospace Conference*. IEEE, pp. 1–12.

[8] T. Markus, T. Neumann, A. Martino, W. Abdalati, K. Brunt, B. Csatho, S. Farrell, H. Fricker, A. Gardner, D. Harding, M. Jasinski, R. Kwok, L. Magruder, D. Lubin, S. Luthcke, J. Morison, R. Nelson, A. Neuenschwander, S. Palm, S. Popescu, C. K. Shum, B. E. Schutz, B. Smith, Y. Yang, and J. Zwally, "The Ice, Cloud, and land Elevation Satellite-2 (ICESat-2): Science requirements, concept, and implementation," *Remote Sensing of Environment*, vol. 190, pp. 260–273, Mar. 2017.

[9] H. W. Leigh, L. A. Magruder, C. C. Carabajal, J. L. Saba, and J. F. McGarry, "Development of Onboard Digital Elevation and Relief Databases for ICESat-2," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, pp. 2011–2020, Apr. 2015.

[10] M. E. Brown, S. Delgodo Arias, T. Neumann, M. F. Jasinski, P. Posey, G. Babonis, N. F. Glenn, C. M. Birkett, V. M. Escobar, and T. Markus, "Applications for ICESat-2 Data: From NASA's Early Adopter Program," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 4, pp. 24–37, Dec. 2016.

[11] U. C. Herzfeld, B. W. McDonald, T. A. Neumann, B. F. Wallin, T. A. Neumann, T. Markus, A. Brenner, and C. Field, "Algorithm for detection of ground and canopy cover in micropulse photon-counting lidar altimeter data in preparation for the icesat-2 mission," 2014.

[12] J. Zhang and J. Kerekes, "An Adaptive Density-Based Model for Extracting Surface Returns From Photon-Counting Laser Altimeter Data," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, pp. 726–730, Apr. 2015.

[13] D. Gwenzi, M. A. Lefsky, V. P. Suchdeo, and D. J. Harding, "Prospects of the icesat-2 laser altimetry mission for savanna ecosystem structural studies based on airborne simulation data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 118, pp. 68–82, 2016.

[14] S. Popescu, T. Zhou, R. Nelson, A. Neuenschwander, R. Sheridan, L. Narine, and K. Walsh, "Photon counting lidar: An adaptive ground and canopy height retrieval algorithm for icesat-2 data," *Remote Sensing of Environment*, vol. 208, pp. 154–170, 2018.

[15] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," *Sigmod Record*, vol. 29, no. 2, pp. 93–104, Jun. 2000.

[16] M. McGill, T. Markus, S. S. Scott, and T. Neumann, "The multiple altimeter beam experimental lidar (MABEL): An airborne simulator for the ICESat-2 mission," *Journal of Atmospheric and Oceanic Technology*, vol. 30, no. 2, pp. 345–352, Feb. 2013.

[17] [Online]. Available: https://icesat-2.gsfc.nasa.gov/legacy-data/matlas/