*Type of the Paper (Research Article.)*

# *PhotonLabeler:* An Inter-disciplinary Platform for Visual Interpretation and Labeling of ICESat-2 Geolocated Photon Data

**Lonesome Malambo[1,*] and Sorin Popescu [1]**

[1] Department of Ecology and Conservation Biology, Texas A&M University, College Station, TX;
* Correspondence: Correspondence mmoonga@tamu.edu;

**Abstract:** NASA's ICESat-2space-borne photon-counting lidar mission is providing global elevation measurements that will provide significant benefits to a variety of bio-geoscience research applications. Given the novelty of elevation and the derived data products from the ICESat-2 mission, the research community needs software tools that can facilitate photon-level analyses to support product validation and development new analysis methods. Here, we describe *PhotonLabeler*, a free graphic user interface (GUI) for manual labeling and visualization of ICESat-2 Geolocated Photon data (ATL03). Developed in MATLAB, the GUI facilitates the reading and display of ATL03 Hierarchical Data Format (HDF) files, the manual labeling of individual photons into target classes of choice using a number of point selections tools and enables eventual saving of labeled data in ASCII format. Other capabilities include saving and loading of labeling sessions to manage labeling tasks over time. We expect labeled data generated using the application to serve two main purposes. First, serve as ground truth for validating various products from ICESat-2 mission, especially for study sites around the world that do not have existing reference datasets such as airborne lidar. Second, serve as training and validation data in the development of new algorithms for generating various ICESat-2 data products. We demonstrate the first use case through a validation case study for the land and vegetation product (ATL08), which provides canopy and terrain height estimates, over two sites. For the first site, located in northwestern Zambia, we used ICESat-2 ATL03 data acquired at night and for our second site in Texas, US, we used ATL03 data acquired during the day. The *PhotonLabeler* application is freely available as a compiled MATLAB binary to enable free access and utilization by interested researchers.

**Keywords:** ICESat-2; Photon-counting lidar, Photon Labeling, Visualization, ATL03, ATL08, Visual Interpretation, Solar-induced noise

## 1. Introduction

NASA'S ICESat-2 (Ice, Cloud, and land Elevation Satellite-2) space-borne lidar mission launched in September 2018. A follow-up spaceborne lidar mission to the first ICESat mission, ICESat-2 carries the Advanced Topographic Laser Altimeter System (ATLAS) lidar instrument, a photon-counting lidar that uses a green (532 nm) laser for range measurement. The ATLAS instrument has notable advantages over its predecessor GLAS sensor including a reduced laser footprint (~14 -17 m diameter reduced from ~70 m diameter) and an increased along-track sampling through higher pulse-repetition rate (10 kHz or about 0.70 m footprint spacing). Other improvements include the increased across-track sampling by partitioning the emitted laser into multiple profiling strong and weak beams [1,2]. Since its launch, ATLAS has provided unprecedented high-resolution three-dimensional along-track measurements of ice sheets, sea ice,

vegetation canopies, inland water and clouds providing opportunities for transformational Earth science and vegetation science research [3-5].

The single-photon counting technique utilized by ATLAS differs in a fundamental way from conventional lidar sensing as ranging is achieved at an individual photon scale as opposed to sensing energy peaks of backscattered laser energy [6]. Using photon time of flight values and measurements from ancillary systems such as GPS each photon is precisely geolocated by determining its time, latitude, longitude, and height to generate the Geolocated Photon Data (ATL03)[7]. Photons reflected from the surface along with those from the atmosphere or solar radiation are detectable which usually results in a combination of signal and noise photons. Thus, ATL03 data tends to exhibit higher levels of noise for daytime acquisition given the higher impact of solar background illumination [6,8]. Notwithstanding the presence of noise in the data, ICESat-2 mission, being a space-based instrument, has improved the possibility to characterize the Earth's surface from local to global scales. Utilizing the 532 nm wavelength on the ATLAS instrument has also created opportunities for both surface and bathymetric mapping as the green energy can also penetrate water and interact with the sub-surface in addition to its surface and vegetation interactions [4,9,10]. ICESat-2 ATL03 photon data is enabling the generation of various standard data products for land ice, sea ice, the atmosphere, vegetation and land, oceans and inland water applications, which are available from the National Snow & Ice Data Center (NSIDC) website (https://nsidc.org/data/icesat-2/data-sets).

The availability of ICESat-2 products has spurred various studies to validate the products [11], develop alternative algorithms to generate similar datasets [2] or derive other products based on available products [4]. In those studies, the availability of reference data is a critical component to enable comparisons with existing products or assessment of developed algorithms, and airborne lidar data predominantly played this vital role. In Wang et al. [10] airborne lidar data were used to validate ground elevation and vegetation heights from the ATL08 product. Similar assessments have been carried out for land and vegetation height metrics [11]. Some studies also used airborne lidar data to simulate ICESat-2 data prior to launch of the mission to enable development of noise filtering algorithms [2]. However, airborne lidar data are not be available in all areas or might be outdated to allow for an objective evaluation of current ICESat-2 products. In addition, there are assessments that may require a photon level understanding such as noise filtering, photon classification or assessing sources of error that airborne lidar data might not adequately support. In such cases, expert manually labeled photon data developed by direct interpretation of raw ATL03 photon data can facilitate such analyses. Manually labeled data also eliminates positional errors due the reference data, which enhances the overall analysis. Visual interpretation of data is not new in remote sensing and has supported various studies where ground truth data were not available [12-14]. Manually labeling photons is analogous to aerial photo interpretation consisting in identification of features in remote sensing images through visual interpretation. Scientific software tools allowing the display and labeling of photon data would enhance the development of labeled datasets to support product validation and development of analysis methods.

In this paper, we describe *PhotonLabeler*, a new free software tool for labeling and visualization of ATLAS geolocated photon data. Other software tools for working with ICESat-2 data are available including PhoReal (https://github.com/icesat-2UT/PhoREAL), the ICESat-2 Visualizer (https://icesat-2-scf.gsfc.nasa.gov/) and web-based OpenAltimetry platfom (https://openaltimetry.org/). These tools have focused mainly on the visualization of ICESat-2 data, data sub-setting and format conversion but do not offer labeling functionality as designed in the *PhotonLabeler*. The ICESat-2 Visualizer also has access limitations – from our check of the website, one needs an approved account to access it. The *PhotonLabeler* can support development of labeled reference data for a variety of assessments including ice, land and vegetation elevation worldwide. The tool also serves as a platform for development of new or alternative algorithms for deriving various height measurements from ICESat-2 geolocated photon data by facilitating the collection

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 August 2020 doi:10.20944/preprints202008.0293.v1

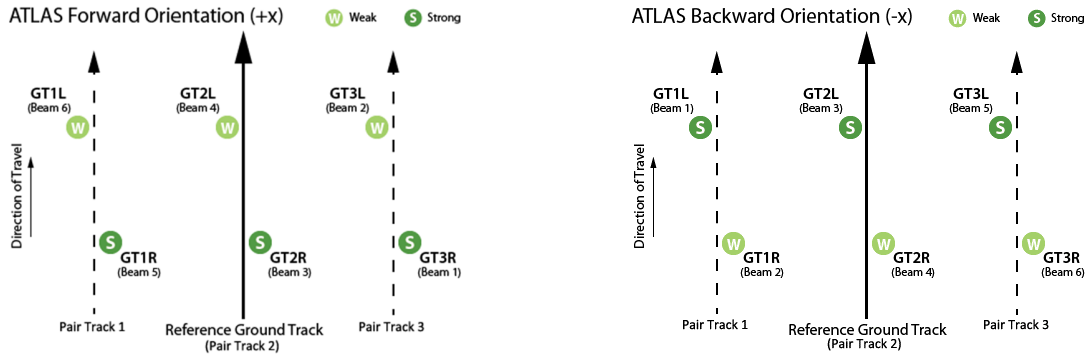
of needed training data. We describe the general functionality of the application including reading and display of ATL03 data, labeling individual photons and exporting labeled data for further analyses. To demonstrate applicability of this software to suggested assessments, we also include a case study for validating canopy and terrain heights from the ATL08 product. We collected labeled photon data using PhotonLabeler from ICESat-2 ATL03 data over two sites, one in northwestern Zambia and the other in eastern Texas in the United States and compared height estimates generated from them with ATL08 height estimates.

## 2. Materials and Methods

### *2.1 ATL03 data*

#### 2.1.1 Data description and organization

ICESat-2 ATL03 data, in HDF format, are the main input to the *PhotonLabeler* application for visualization and labeling. ATL03 data (https://nsidc.org/data/atl03) provide time, latitude, longitude, and ellipsoidal height for each detected photon in the WGS-84 reference frame. The ATL03 data are organized by ground track, with ground tracks 1L and 1R forming the first pair, ground tracks 2L and 2R forming middle pair, and ground tracks 3L and 3R forming last or third pair [7]. The designation of which track is weak or strong depends on the direction of travel of ATLAS (Figure 1) - in forward orientation strong beams are mapped to the right (R) beams and weak beams to the left (L) tracks, and vice versa [7]. All the point data for each respective ground track can be accessed from the heights sub-group (/gtx/heights). For more details about the configuration of the ground tracks and data attributes in the ATL03 HDF file, we refer an interested reader to the ATL03 Algorithm Theoretical Basis document (ATBD)[7].



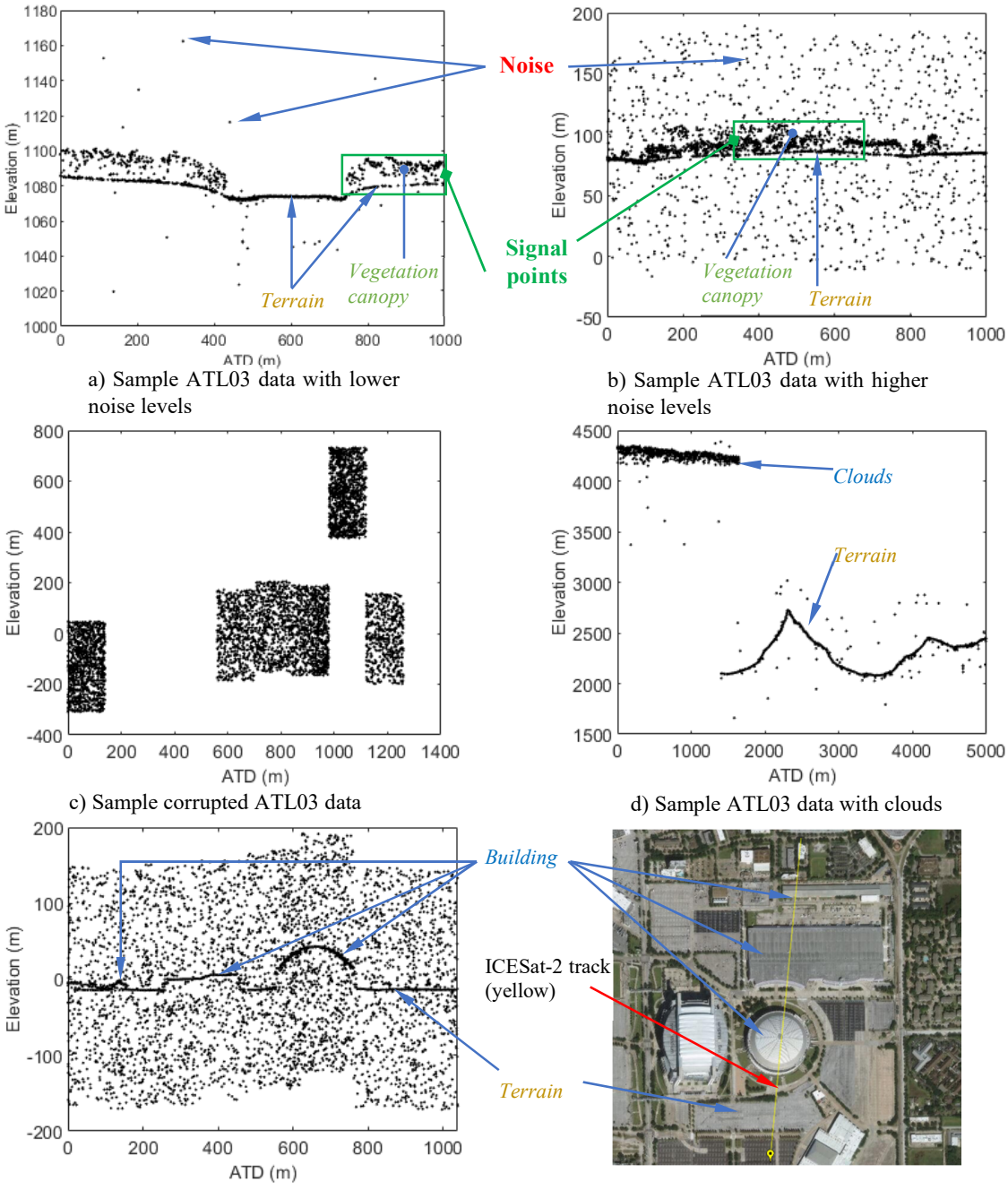

a) Forward orientation b) Backward orientation

**Figure 1**: Ground track (GT) naming convention: a) ATLAS oriented in the forward (instrument coordinate +x) direction, b) ATLAS oriented in the backward (instrument coordinate +x) direction. Image credit: (Neumann et al. 2019).

#### 2.1.2 Visualizing and interpreting ATL03 data

ATL03 data are profiling data, thus provide sample cross-sectional views of the observed area. As such, ATL03 data are usually visualized by plotting photon elevation values (y-axis) against time or along-track distance (ATD) values (x-axis) (Figure 2). ATL03 data could also be visualized in 3D space by plotting latitude, longitude and elevation values. For interpretation and labeling purposes, 2D views are generally sufficient.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 August 2020 doi:10.20944/preprints202008.0293.v1




a) Sample ATL03 data with lower noise levels

b) Sample ATL03 data with higher noise levels

c) Sample corrupted ATL03 data

d) Sample ATL03 data with clouds

e) Sample ATL03 data over built area

f) Using ancillary data to aid interpretation

**Figure 2**: Interpretation of ATL03 data. Observing point distribution differences is useful in interpreting noise (random) and signal (clustered) points: a) Sample ATL03 data with lower noise levels over a site in north-western Zambia; b) Sample ATL03 data with higher noise level over a site in eastern Texas, United States; a) and b) also show further classification of signal points into terrain (lower local elevation variation) and vegetation (above terrain and exhibit higher local elevation variation); c) Sample ATL03 data corrupted by instrument errors, which could be interpreted as noise. d) Samples data with cloud points, which are identifiable by their significant elevation values compared to terrain points; e) Interpretation of ATL03 data over a built area in Houston, Texas. Shape attributes and regularity of surface is key to differentiating them from terrain; f) Using ancillary image data to aid interpretation of data in e). ATD is along-track distance

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 August 2020 doi:10.20944/preprints202008.0293.v1

Plotting ATL03 data provides a convenient way for interpreting them into classes of interest. It is noteworthy to state that no studies have published formal guidelines for interpreting ATL03 or profiling lidar data as the case for image interpretation. Here, we provide some tips borne mostly from our experience with ATL03 data and knowledge from similar profiles generated across airborne lidar data. Unlike scanning-based lidar data with spatially complete coverage, the 3D visual cues such as shape, texture and color that users rely on to readily interpret various objects in the data are diminished for ATL03 data. Nevertheless, it is still feasible to interpret individual or group of points in ATL03 data into a variety of classes including terrain, vegetation canopy, building roof, clouds and so on. As in image interpretation, recognizing various objects or surfaces in plotted ATL03 data and an understanding of how the sensing energy interacts with various target objects (water, clouds, vegetation etc.) are critical for successful interpretation and labeling. Foremost, an interpreter should be able to recognize signal and noise points given that observing differences in the spatial distribution of points is one of the convenient ways of differentiating them. When ATL03 data are plotted on a Euclidean space, signal points (real data) tend to cluster together compared to noise points which tend to be randomly distributed in space (Figure 2a and b). Sometimes, ATL03 data are corrupted by instrument or geolocation errors, resulting in unnatural discontinuities and offsets in the data (Figure 2c). Such data sections are usually unusable and could also be interpreted as noise.

Signal points could further be interpreted into terrain (e.g. land or sea ice terrain), above-terrain point classes (e.g. forest canopies, buildings and clouds) and below-terrain surfaces (e.g. bathymetry) by leveraging attributes specific to the data and using ancillary data. For instance, elevation relationships among points or point clusters (e.g. points on vegetation canopies should lie above terrain points, cloud points usually cluster together but have very high elevations above terrain, Figure 2a, b and d), shape (regularity of building roof surfaces, Figure 2e) and point distribution information (terrain points show lower local elevation variance than points on vegetation, Figure 2a, b and d) could be applied for effective interpretation. Reduction in point density for weak beams can pose interpretation challenges, but observing data in a corresponding strong beam could aid interpretation even though they two beams do not sample the same area. While interpretation of ATL03 data could largely be done based on ATL03 data themselves, using ancillary data such as high-resolution imagery could enhance interpretation by applying the visual cues normally applied in image interpretation (Figure 2f).

## *2.2 PhotonLabeler User Interface (UI) and software capabilities*

### 2.2.1. UI layout overview and display control

We developed the *PhotonLabeler* UI and associated functionality using the MathWorks MATLAB App Designer (MATLAB R2020a). Figure 3 shows the three-panel layout of the *PhotonLabeler* UI. The left panel shows ground track details such as the beam level (weak or strong), the number of points in the track and the distance spanned. The panel also contains tools for defining or loading saved label definitions and displaying label statistics. The central panel contain two plot axes: the Overview and the Detail plots. We adopted the two-plot view to enhance the interpretation and labeling of individual photons by providing both an overview of the loaded data across a larger extent, and a more detailed view across a smaller extent. The left-most panel contains tools for data display control such as the size of the delta time or along-tract distance (ATD) span, the *zFactor*, which controls the portion of data to display in the Detail plot and an option that controls whether to use delta time or ATD values on the x-axis. This panel also contains number of point selection tools to facilitate labeling. In the following sections, we outline the general workflow and critical parameters that enable the reading, labeling and export of labeled data. For details on various functionality in the software, please refer to the supplementary material (SS).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 August 2020 doi:10.20944/preprints202008.0293.v1

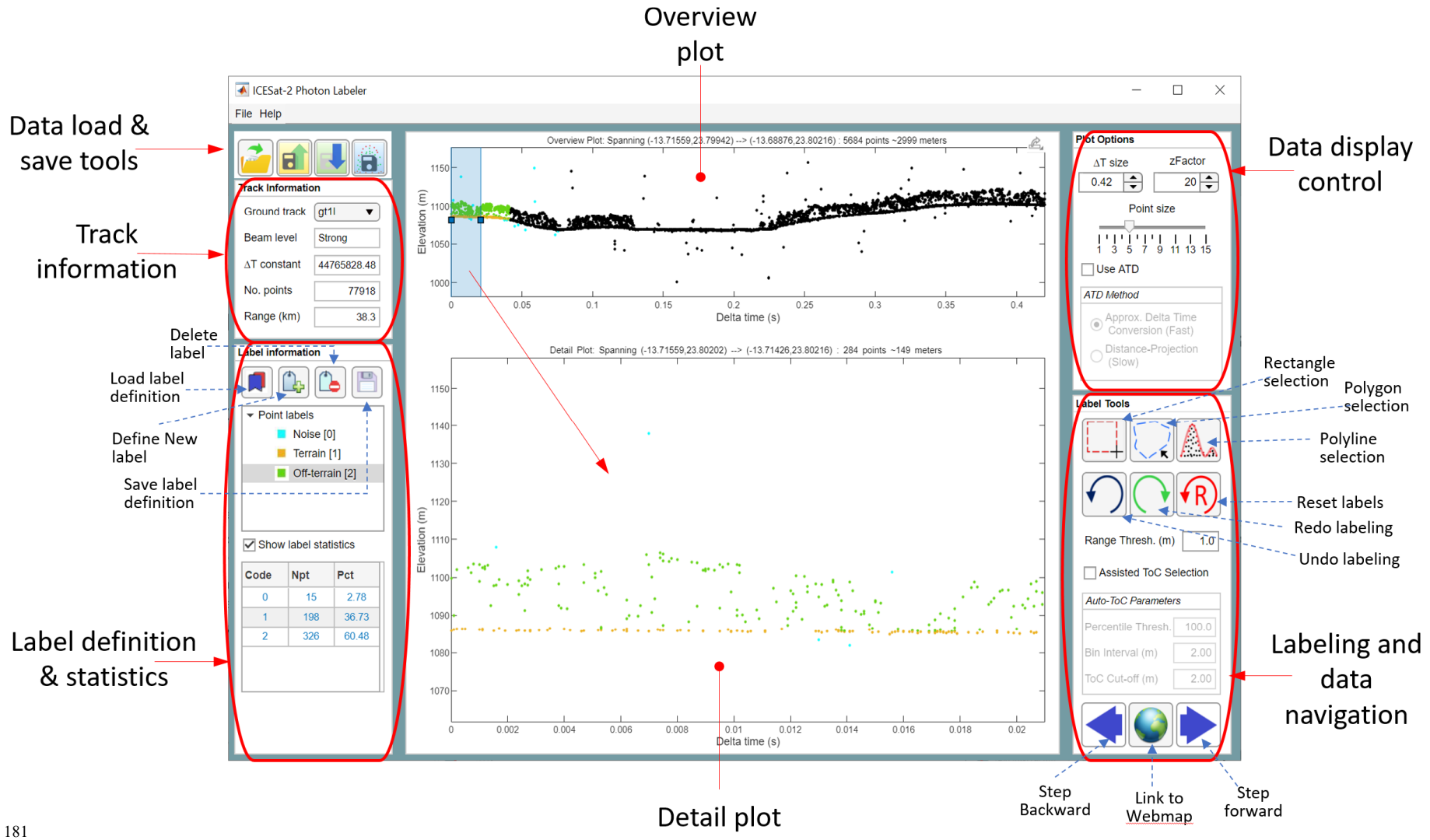



**Figure 3:** PhotonLabeler graphical user interface showing track and label information and tools on the left panel, plot areas in middle and data display control and labeling and navigation tools on the right panel.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 August 2020 doi:10.20944/preprints202008.0293.v1

### 2.2.2 Reading and visualization of ATL03 data in PhotonLabeler

The *PhotonLabeler* application reads and displays ATL03 HDF data by ground track and displays the data by plotting either the photon delta time values or the calculated along-track distance (ATD) values on the x-axis against the photon elevation values on the y-axis. To manage large file sizes, the PhotonLabeler only loads a portion of the data at a given time. Two parameters control how much data is loaded in the *Overview* and *Detail* plots: $\Delta T$ *size* parameter, which defines span of the data on x-axis and the *zFactor* parameter or zoom factor, which is a factor by which $\Delta T$ *size* is equally divided. All data that lies within a specified $\Delta T$ *size* range are loaded in the Overview plot. From the data loaded in the Overview plot, a sub-portion equal to $\Delta T$ *size/zFactor* is then loaded in the *Detail* plot, allowing a user to step through using the Next and Back buttons to view all data loaded in the *Overview* plot. At any point, the blue semi-transparent region in the Overview plot highlights the data displayed in the Detail plot as a guide when exploring or labeling the data. A user may also drag the blue region to positions of interest instead of using the Next and Back buttons to navigate through the data.

*PhotonLabeler* application provides two methods for calculating ATD values. The first approach, which is the faster approach, uses an approximate relationship between delta time and the distance (1 sec ≈ 7000 distance m) covered by the ICESat-2 satellite on the ground. Differenced delta time values are multiplied by the conversion factor (7000) and accumulated from the first point to generate along-track distances. This approach treats all points as lying on a common track, thus neglects across variation within the footprint. The other approach generates a best-fit line among the points in the dataset and projects all points to this common line. Inter-point distances are then calculated and accumulated to get along-track distances. Due to the higher number of computations, this approach can is usually slow. For visualization and labeling purposes, the first approach is adequate and is recommended.

As an aid to photon labeling or general data exploration in different part of the world, the *PhotonLabeler* application provide the means to link to base maps through MATLAB's web maps as illustrated in Figure 4. In the figure, the yellow line shows the portion of the track covered by data in the *Detail plot* with the yellow marker indicating the direction of the track. Using visual cues from ATL03 data as displayed in the Overview and Details plots and high-resolution imagery provide a powerful tool for effective photon interpretation and labeling.

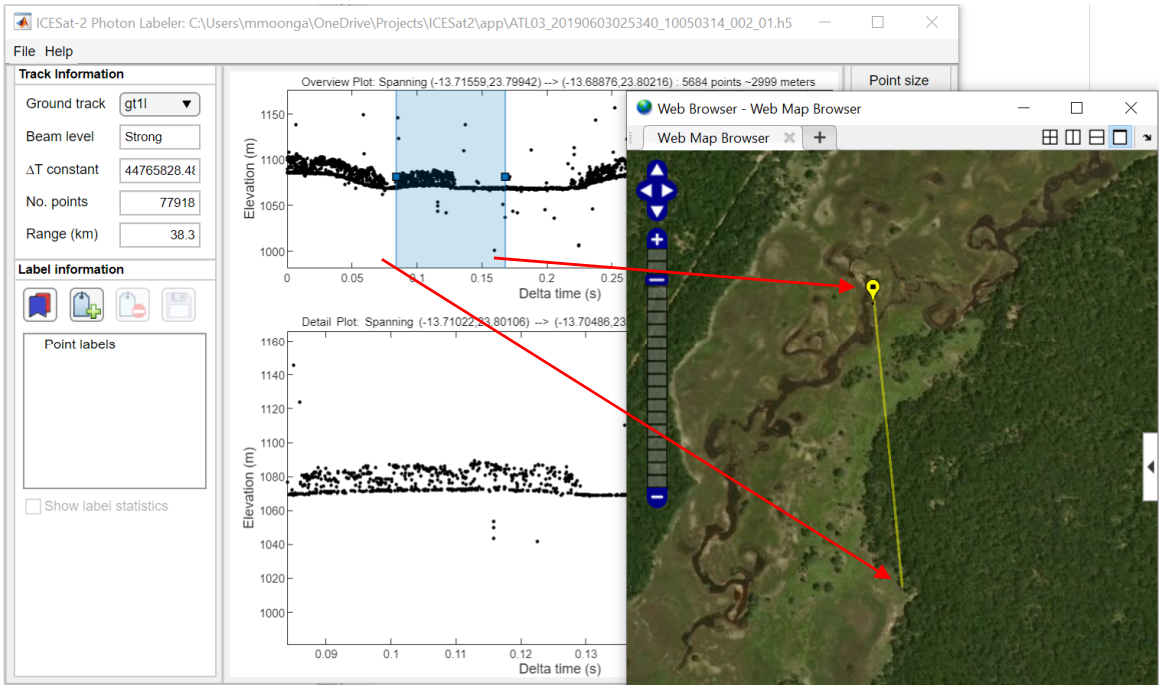



**Figure 4:** Web map integration. The section of an ICESat-2 track under consideration is displayed on a web-based map to aid interpretation of the ATL03 data.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 August 2020 doi:10.20944/preprints202008.0293.v1

### 2.2.3 Creating a point label definition, selecting and labeling points

To manage labeling tasks, *PhotonLabeler* provides tools for defining point label definitions. A point label definition or scheme defines point classes and their respective representation for a target application. A user can define individual point class representation by specifying the label name, label code and picking a representation color as illustrated in Figure 5. *PhotonLabeler* also offers tools for deletion of defined point labels, saving of and loading of saved point label scheme. A saved label definition could be loaded when specifying session options or through the Load label definition button (Figure 5). Figure 5 shows an example of a classification scheme for vegetation study were target groups include noise, terrain, canopy and top of canopy. Defining a point label definition is only critical to labeling – visualization and general exploration of the loaded dataset could be done without it.

The *PhotonLabeler* implements photon labeling on a track-by-track basis. A number of point selection tools including rectangle, polygon and polyline-based tools are available to facilitate the labeling of individual or group of photons. Normally, a user selects a target class and uses any of the available tools to select points of interest. Once the user has selected the points, the selected points are labeled and visualized based on a current active class. The user has the option to undo or reset recently labeled or all labeled data. The *PhotonLabeler* also provides an option to define top of canopy (ToC) points automatically, through the Assisted ToC Selection option (Figure 3), for vegetation related assessments. Based on the region of interest defined by the user, top of canopy points are defined by binning (see Bin Interval in Figure 3) the data along the x-axis and taking a point(s) above a specified percentile height (See Percentile Thresh in Figure 3) as top of canopy. The tool also estimates a ground level to allow elimination of points closer to the ground using a set ToC cut-off threshold. The automatic top of canopy tool relies on a user defining a region without noise above the canopy; otherwise, the tool could identify outlying noise points as ToC points. We also recommend using this tool over distances of 100 - 300 m, as longer distances tend to result in poor ground fits.

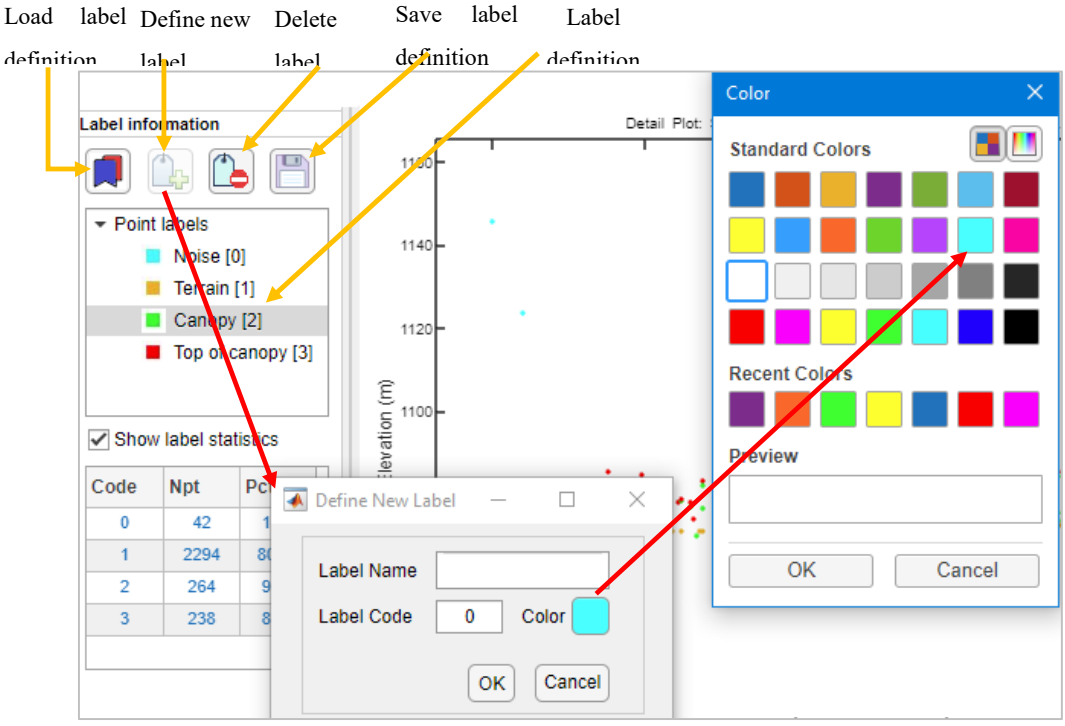



**Figure 5**: Label definition tools in the PhotonLabeler. The red arrow illustrates the process of defining a new point label

### 2.2.4 Exporting labeled data, saving and loading labeling sessions

The *PhotonLabeler* allows the saving of labeled data in ASCII formats (comma separate values (.csv) and tab delimited text files (.txt). Labeled data are saved by ground track and includes point labels, point codes, section ID, longitude, latitude, elevation, delta time and if calculated, the along-track

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 August 2020 doi:10.20944/preprints202008.0293.v1

distances. A section ID represents contiguous sections of labeled data along an ICESat-2 track and can facilitate separation of labeled data according phenomena of interest e.g. if one is interested in labeling data in sparse and dense forested regions. Section IDs are defined automatically by examining point index sequences and are more meaningful if contiguous sections of labeled data exist in the loaded dataset. To manage labeling tasks over time, the PhotonLabeler enables the saving and loading of labeling sessions. A saved session file contains the state of the application at the time of saving and stores input files path and parameters to enable one to pick up labeling from where they left.

#### 2.2.5 Software availability

*PhotonLabeler* is available to interested scientists through our project website on GitHub (https://github.com/Oht0nger/PhoLabeler/releases/tag/v1.0). The application is available as a compiled binary, which one can install without a MATLAB license. The option requires a download of free MATLAB runtime environment. On our GitHub page, we also provide a detailed user manual on how to use the software.

### *2.3 Case study: Using manually labeled data to access accuracy of ATL08 data*

#### 2.3.1 ATL08 product overview

The Land and Vegetation Height product (ATL08) contains along-track terrain and canopy heights above the WGS84 ellipsoid. The height estimates are derived from Geolocated Photon (ATL03) data in fixed 100 m x 14 m (footprint size) data segments [15]. The ATL03 photon data undergoes some preprocessing to filter background noise and to classify data points into respective terrain, canopy and top of canopy points [16]. Based on the classified terrain, canopy and top of canopy points, the various height metrics including descriptive statistics (minimum, mean, maximum) and height percentiles, are calculated within each 100-m segment. For canopy heights, the ATL08 product reports the height metrics in terms of the absolute height above the reference ellipsoid and relative height above an estimated ground. The ATL08 product, also distributed in HDF format, organizes calculated height metrics by ground track. Each ground track group in the HDF model contains subgroups that hold generated canopy and terrain heights and individual photon classification flags [15]. Table 1 summarizes the ATL08 height metrics evaluated for this case study.

**Table 1:** ATL08 height metrics evaluated for the study. Height metrics include the minimum (Min), mean (Mean), maximum (Max) and height percentiles (Pxx, e.g., P25 for 25th percentile height). Metrics evaluated for a particular height type (Absolute canopy, Relative canopy and Terrain height) are marked by x

| *Metric* | *Max* | *Mean* | *Min* | *P25* | *P50* | *P60* | *P70* | *P75* | *P80* | *P85* | *P90* | *P95* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Absolute canopy height | x | x | x | x | x | x | x | x | x | x | x | x |
| Relative canopy height | x | x | x | x | x | x | x | x | x | x | x | x |
| Terrain height | x | x | x | | | | | | | | | |

#### 2.3.2 Study sites and objective

We used the *PhotonLabeler* application to collect sample validation data over two sites, one in northwestern Zambia and the other in eastern Texas, United States. Our main goal here is to demonstrate how manually labeled data from PhotonLabeler could be used to assess accuracy of derived ICESat-2 products such as the ATL08. Given the influence of solar background radiation on noise levels in ICESat-2 photon data, our objective was to assess the agreement of ATL08 height estimates with matching estimates derived from labeled data for nighttime (when noise levels are lower) and daytime (when noise levels are higher) data acquisition scenarios. The Zambia site (circa

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 August 2020 doi:10.20944/preprints202008.0293.v1

Latitude 13°49'10"S, Longitude 23°48'46" E) stretched across the Zambezian dry evergreen forest ecoregion comprising mainly of a tropical dry broadleaf forest, which rarely exceeds 25 m in height [17]. The Texas site (around Latitude 31°26'8.6"N, Longitude 95°16'38"W) lay across temperate conifer forests dominated by several species of Pine (*Pinus* spp.) as well as hardwoods including Hickory (*Carya* spp.) and Oak (*Quercus* spp.) [18].

#### 2.3.3 Data and labeling

We selected one ATL03 granule over each site – a nighttime granule for the Zambian site and a day-time granule for the Texas site. We labeled data from strong beams only due to the low number of weak beam estimates to compare with in the ATL08 data. ATL03 granule IDs for the data used are ATL03_20190603025340_10050314_003_01 for the Zambia site and ATL03_20190415130456_02640302_003_01 for Texas site. From the granule ID, the data for the Zambia site was acquired on June 3rd 2019 while the data for the Texas site was acquired on April 15th 2019. In both sites, our labeling scheme consisted of three main point classes: *Terrain*, for terrain points, *Off-terrain* comprising canopy points and top of canopy points, and *Noise,* for all background noise points. For the Zambian case, we labeled three 3000-m sections along a gt1l ICESat-2 track. In Texas, we labeled data across a 10,000 m transect, composed of two sections – one stretching about 4000-m, the other about 6000-m along a gt2l track of the data granule.

#### 2.3.4 Preparing height metrics

We extracted canopy and terrain metrics from ATL08 data corresponding with the ATL03 data granules used for labeling. The extracted data was saved as ASCII with the segment ID to serve a unique identifier when linking with estimates generated from our labeled data. To facilitate calculation of corresponding canopy and terrain height metrics, we recreated each ATL08 segment, which measures approximately 1000x14 m, based on segment attributes (begin/end delta times and the corresponding begin/end latitude/longitude) from both ATL03 and ATL08 data for the target tracks. Using the generated segment polygons as regions of interest, we calculated corresponding ATL08 metrics from the labeled point data. For absolute canopy heights, the labeled data excluding noise was used without ground normalization while relative canopy heights were calculated from normalized data, which leveraged already labeled terrain points. For terrain estimates, we used points labeled as terrain only. We applied LAStools software tools [19] for ground normalization and for calculating the corresponding ATL08 height metrics. To account for the point spread exhibited by ATL03, we imposed a 0.5 m threshold when calculating relative canopy metrics.

#### 2.3.5 Comparing ATL08 and PhotonLabeler derived height metrics

We assessed the level of agreement between ATL08 height metrics and corresponding estimates derived from manually labeled data using regression analysis. We took values calculated from manually labeled data to be the reference or observed variables and took ATL08 data as predicted values, using the regression coefficient of determination ($R^2$) as a measure of correlation. We also calculated a mean bias metric, calculated as the mean of differences between reference and predicted values, as a measure of precision and to shade light on under and over-estimation of the ATL08 metrics with respect to manually labeled data, which we took as ground truth.

## 3. Results

### *3.1 Photon Labeling*

Figure 6 shows the manually labeled sections along the selected ICEsat-2 tracks and close-up views of ICESat-2 photon data over 100 m distance over the northwestern Zambia and Texas sites. The differences in the noise levels is notable reflecting the significant impact of solar background radiation can have on photon counting lidar measurements. The close-up views of the data also show differences in the amount of noise and the number of terrain points indicating differences in the level

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 August 2020 doi:10.20944/preprints202008.0293.v1

of photon canopy penetration. Table 2 summarizes the number of points labeled along selected ICESat-2 tracks for the two sites. The distribution of the three-point classes is markedly different between the two sites. The proportion of noise points in the two sites was 1.7% for Zambia and 27.6% for Texas. The ICEsat-2 granule labeled for the Texas site was collected during the day (13:04:56 UTC) when the effect of background solar radiation is higher while the data used for the Zambia site was collected at night (02:53:40 UTC). The other notable difference is in the proportion of terrain and off-terrain points. About two-thirds of the points for the Zambian site were ground whereas terrain points only account for less than a fifth of the points labeled. The reverse is observed for off-terrain points with the Texas site showing a large proportion of points than the Zambian sites. Apart from the time the data was acquired, these differences a reflective of the different vegetation structure (sparse vs closed canopy) or phenological stages (leaf-on vs leaf-off) in the two sites. The data for the Zambia site was acquired in June, which is a leaf-off period for most deciduous forest in Zambia while Texas forests were in their leaf-off state at the time the data was acquired in April [20,21].

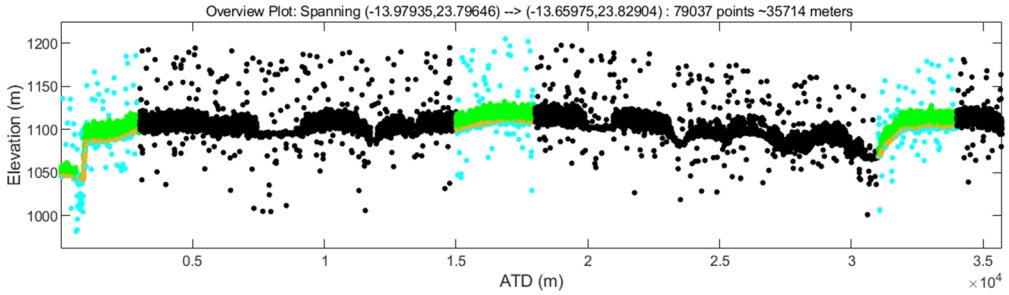
**Table 2:** Number of manually labeled points in northwestern Zambia and Texas

| | | Northwestern Zambia | | Eastern Texas, USA | |
|---|---|---|---|---|---|
| *Label* | *Label code* | *No. points* | *Proportion (%)* | *No. points* | *Proportion (%)* |
| Noise | 0 | 368 | 1.7 | 6532 | 27.6 |
| Terrain | 1 | 14198 | 65.3 | 4415 | 18.6 |
| Off-terrain | 2 | 7174 | 33.0 | 12749 | 53.8 |
| **Total** | | **21740** | **100** | **23696** | **100.0** |

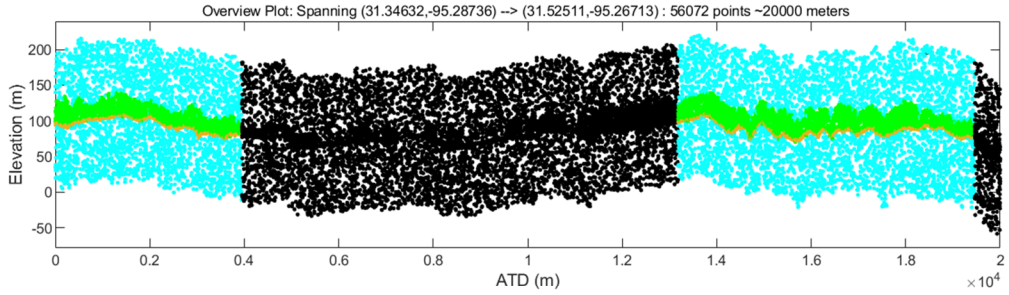
### *3.2 ATL08 height metric comparison results*

The total numbers of ATL08 segments with valid canopy height estimates matched with labeled data from the Zambian and Texas sites were 84 and 90 respectively. The number of segments with valid terrain height estimates were 92 and 88 for the Zambian and Texas site respectively. **Table 3** and **Table 4** summarize the relationships between various height metrics compared with corresponding estimates derived from manually labeled point data. In general, ATL08 height metrics were highly correlated ($R^2 > 0.8$) with corresponding metrics from labeled data especially for absolute canopy and terrain height metrics. The correlation of ATL08 relative canopy heights with PhotonLabeler (PL) metrics varied by height metric with the minimum height values for the Zambia site showing a zero $R^2$ value. We attribute the lack of correlation for minimum values to the 0.5 m threshold we applied when calculating relative height metrics, which force most of our minimum values to be 0.5 m. The maximum relative canopy values for the Texas site showed higher correlation with PL values compared to values for the Zambian site. This is somewhat surprising observation considering near-clean data for the Zambian site. The correlation for ATL08 percentile height metrics generally increased with the percentile level from P25 to P95 ($R^2 = 0.57 – 0.89$ for the Texas site, $R^2 = 0.67 – 0.86$ for Zambia site).

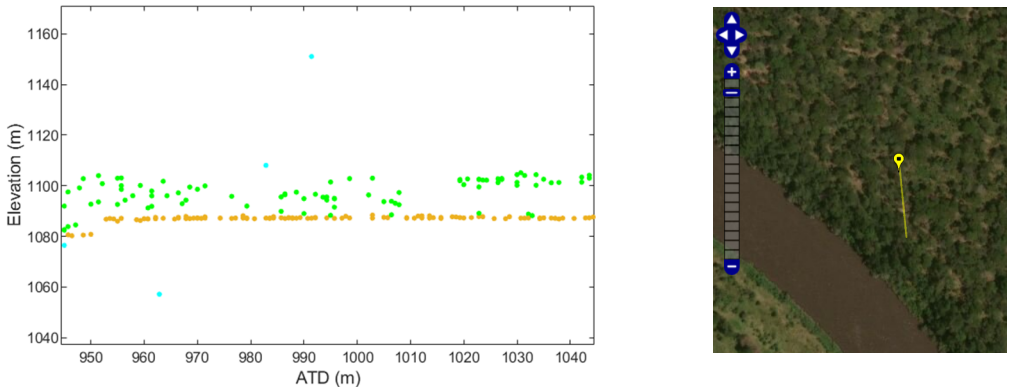
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 August 2020 doi:10.20944/preprints202008.0293.v1



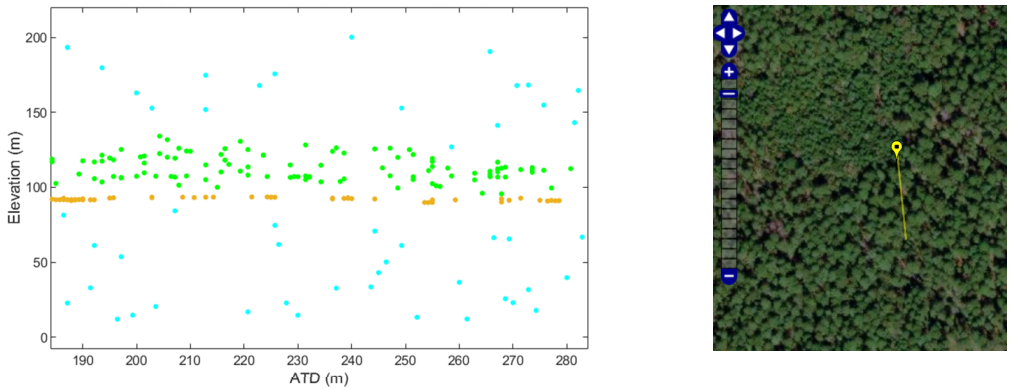

a) Label sections for northwestern Zambia site




b) Label sections for Texas site




c) Close-up view of a 100-m segment and its path on ESRI base map from Zambian site




d) Close-up view of a 100-m segment and its path on ESRI base map from Texas site

**Figure 6:** Manual Photon Labeling using the Photon Labeler. a) – b) Label sections along ICESat-2 tracks for the Zambian and Texas sites; c) –d) Close-up plot and base map views of 100-m segments for the Zambia and Texas sites

The precision (Table 4) based on calculated mean biases of ATL08 absolute canopy, relative canopy and terrain height metrics with respect to PL metrics varied from -9.76 m (over-estimation) to 4.63 m (under-estimation). The precision between the two datasets improved with percentile level. For the Texas site, we observed over-estimation at lower percentiles (P25 – P60) including the minimum and mean height, which switched to under-estimation at higher percentiles (P70 – P95, Max). For the Zambia case, there was general over-estimation across all metrics except for the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 August 2020 doi:10.20944/preprints202008.0293.v1

maximum height. Again, we expected the Zambia case to provide better results due to the lower noise levels. These variations could be attributed to the impact of different noise levels on performance of ATL08 filtering algorithms. Results for relative canopy heights were as expected with the Zambia case showing lower biases (-0.36 – 0.70 m) compared to the Texas case (0.48 – 4.63 m). Lastly, ATL08 terrain height estimates showed high precision with respect to PL metrics for both sites. Figure 7 shows graphical relationships between ATL08 absolute canopy, relative canopy and terrain data with corresponding PL data for the minimum, mean and maximum height metrics.

**Table 3:** Correlation ($R^2$) between ATL08 absolute canopy, relative canopy and terrain height metrics with metrics derived from manually labeled point data for the Zambia and Texas sites. Height metrics include the minimum (Min), mean (Mean), maximum (Max) and percentiles (Pxx, e.g., P25 for 25th percentile height).

| | Absolute Canopy Height | | Relative Canopy Height | | Terrain Height | |
|---|---|---|---|---|---|---|
| *Metric* | *Texas* | *Zambia* | *Texas* | *Zambia* | *Texas* | *Zambia* |
| Max | 0.94 | 1.00 | 0.82 | 0.60 | 0.99 | 1.00 |
| Mean | 0.96 | 0.99 | 0.82 | 0.82 | 1.00 | 1.00 |
| Min | 0.98 | 1.00 | 0.29 | 0.00 | 0.99 | 1.00 |
| P25 | 0.87 | 0.98 | 0.57 | 0.67 | | |
| P50 | 0.89 | 0.95 | 0.62 | 0.86 | | |
| P60 | 0.90 | 0.94 | 0.71 | 0.79 | | |
| P70 | 0.94 | 0.92 | 0.75 | 0.73 | | |
| P75 | 0.96 | 0.93 | 0.79 | 0.78 | | |
| P80 | 0.97 | 0.95 | 0.85 | 0.78 | | |
| P85 | 0.98 | 0.97 | 0.84 | 0.79 | | |
| P90 | 0.97 | 0.99 | 0.87 | 0.80 | | |
| P95 | 0.97 | 0.99 | 0.89 | 0.76 | | |

**Table 4:** Precision (mean biases) between ATL08 absolute canopy, relative canopy and terrain height metrics with metrics derived from manually labeled point data for the Zambia and Texas sites. Height metrics include the minimum (Min), mean (Mean), maximum (Max) and percentiles (Pxx, e.g., P25 for 25$^{th}$ percentile height)

| | Absolute Canopy Height | | Relative Canopy Height | | Terrain Height | |
|---|---|---|---|---|---|---|
| *Metric* | *Texas* | *Zambia* | *Texas* | *Zambia* | *Texas* | *Zambia* |
| Max | 4.29 | 0.35 | 4.63 | -0.17 | 0.45 | 0.19 |
| Mean | -1.51 | -5.89 | 2.91 | 0.26 | -0.08 | 0.02 |
| Min | -2.39 | -1.31 | 0.48 | 0.18 | -0.61 | -0.21 |
| P25 | -5.13 | -6.36 | 3.14 | 0.70 | | |
| P50 | -1.76 | -9.63 | 3.27 | 0.14 | | |
| P60 | -0.61 | -9.76 | 3.16 | 0.03 | | |
| P70 | 0.39 | -7.89 | 3.00 | -0.02 | | |
| P75 | 0.90 | -6.44 | 3.00 | -0.10 | | |
| P80 | 1.47 | -4.93 | 3.02 | -0.25 | | |
| P85 | 1.83 | -3.31 | 3.12 | -0.33 | | |
| P90 | 2.33 | -2.14 | 3.19 | -0.36 | | |
| P95 | 2.69 | -1.33 | 3.36 | -0.34 | | |

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 August 2020 doi:10.20944/preprints202008.0293.v1

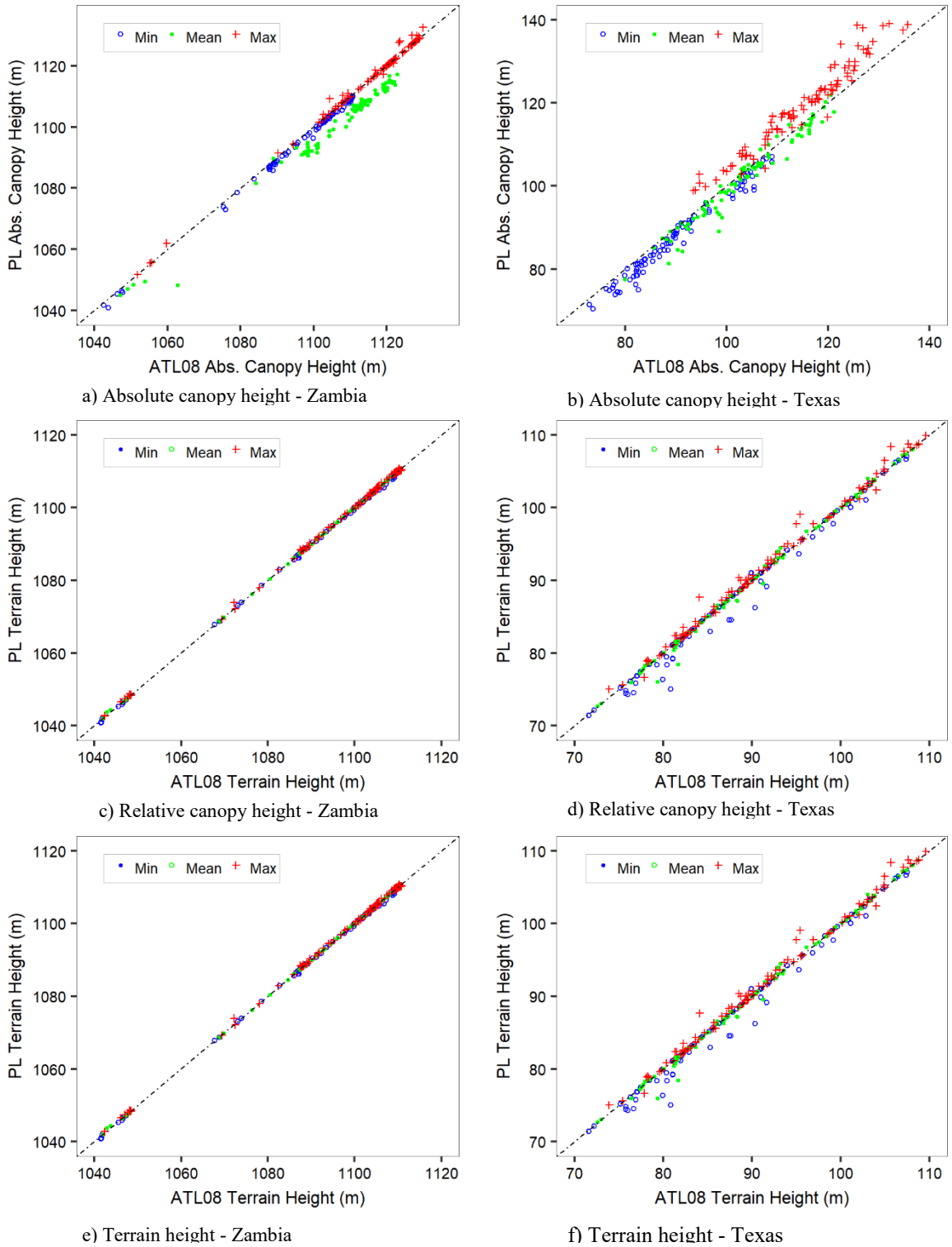



**Figure 7:** Relationships between ATL08 minimum (Min), mean (Mean) and maximum (Max) height estimates and matching height estimates from PhotonLabeler (PL) labeled data for the Zambian (first column) and Texas (second column) sites: a) - b) Absolute canopy heights; c) - d) Relative canopy heights; e) - f) Terrain heights. The dashed line shows the expected 1:1 relationship between PL and ATL08 estimates.

## 4. Discussion

Over the decades lidar remote sensing has proven to be a highly effective technique for characterizing the 3D structure of terrestrial ecosystems including forests, snow ice and topography

in general. Space-borne lidar missions such as ICESat-2 extend our capability to characterize terrestrial ecosystems from local to global scales. Critical to these endeavors are analysis methods and software tools to foster a better understanding of phenomena of interest including forest structure, bathymetry or surface topography or the understanding of the data itself. The *PhotonLabeler* application described in this paper is an effort to provide researchers with a tool for easy visualization and labeling of ICESat-2 photon data. Such capability could support a number of analyses including checking the accuracy of height estimates provided by ICESat-2 products as demonstrated in the case study or training new or re-calibrating existing algorithms with manually labeled data. Development of labeled photon datasets could motivate the development of even better algorithms for generating various products from ATL03 data. Machine learning and deep learning techniques presents one avenue for developing robust approaches [22,23], but still lack labeled point data to motivate adoption as in image-based analyses. Using applications such as PhotonLabeler could readily support such labeled data collection tasks.

The critical importance of data visualization in remote sensing research to enhancing understanding of phenomena and communication cannot be over-emphasized. We believe software tools such as the PhotonLabeler present a great avenue to understanding ICESat-2 data in diverse environments. Researchers and educators may use it as a tool for instruction to demonstrate a variety of aspects of the ICESat-2 ATL03 data including responses of photon counting lidar in various environments. For instance, the impact of solar background illumination on the data quality could be shown by displaying day and night acquisitions. The impact of canopy cover on photon penetration in environments such as the Amazon compared to sparsely forested environments in the Savannas of Southern Africa, is another example. Insight or cues on developing algorithms could also be generated from visualizing data in different ecoregions.

The case studies on the manual validation of ATL08 data provided a glimpse into the accuracy of ATL08 data for day and night acquisition. These results were generally promising with high correlations ($R^2 > 0.8$) and precision (mean biases < 5) between ATL08 estimates and estimates generated from manually labeled data. However, these observations as the results of this assessment are limited. Further validation assessment incorporating data in various ecoregions, seasons and of different noise levels is still needed to provide a more complete view of the accuracy of the ATL08 estimates. We also acknowledged that manually labeled data is not immune to error and inconsistencies may arise between how the ATL08 algorithms define surfaces and how we manually labeled data. Given the high correlations between ATL08 and PL estimates, we think that was not a big issue for this assessment. For future studies intending to do similar assessments on a large scale, we recommended developing labeling protocols to enhance consistency among labeling experts or data in different environments.

Additional functionality in the *PhotonLabeler* application is in the works to enhance productivity and general user-friendliness. Given that some ICESat-2 products such as ATL06 and ATL08 also contain photon level classification data, one capability envisioned is to enable users start labeling based on the existing labels and updating them where necessary. In collecting ground truth data for our demonstration study, it took about 20 – 30 minutes to label a 3000-m section along an ATL03 track, which could add up for many sections of the data. Providing functionality to start labeling from some existing classification would enhance labeling efficiency from our current setup and would complement assisted labeling, such as automatic top of canopy selection, already implemented in the application. We are also working on providing wider options for saving labeled data. One option is saving the labeled data in HDF format similar to the ATL03 structure to improve the organization of the data. The other option to be included in future releases of the software is saving to industry formats such as LAS to facilitate view of the data with other software packages.

## 5. Conclusion

This study presented a user-friendly software application to support visualization and manual labeling of ICESat-2 ATL03 photon data. The application enables definition of custom labeling schemes to meet requirements of a research study and offers various point selection tools to facility

collection of labeled data. Labeled data collected using the *PhotonLabeler* can serve as ground truth for validating various ICESAt-2 elevation products especially in developing countries were airborne lidar acquisitions are not routinely done. We were able to demonstrate how manually labeled data could be applied for validating ATL08 estimates. Such an assessment could be extended to other validation studies for snow ice elevation or bathymetry. The PhotonLabeler can also facilitate the development of new algorithms by providing critical labeled data to support testing and validation.

## Acknowledgements

This research was funded by NASA ICESat-2 Science Team, Studies with ICESat-2 NNH19ZDA001N grant.

## References

1. Markus, T.; Neumann, T.; Martino, A.; Abdalati, W.; Brunt, K.; Csatho, B.; Farrell, S.; Fricker, H.; Gardner, A.; Harding, D., *et al.* The ice, cloud, and land elevation satellite-2 (icesat-2): Science requirements, concept, and implementation. *Remote Sensing of Environment* **2017**, *190*, 260-273.
2. Popescu, S.C.; Zhou, T.; Nelson, R.; Neuenschwander, A.; Sheridan, R.; Narine, L.; Walsh, K.M. Photon counting lidar: An adaptive ground and canopy height retrieval algorithm for icesat-2 data. *Remote Sensing of Environment* **2018**, *208*, 154-170.
3. Petty, A.A.; Kurtz, N.T.; Kwok, R.; Markus, T.; Neumann, T.A. Winter arctic sea ice thickness from icesat-2 freeboards. *Journal of Geophysical Research: Oceans* **2020**, *125*, e2019JC015764.
4. Narine, L.L.; Popescu, S.C.; Malambo, L. Using icesat-2 to estimate and map forest aboveground biomass: A first example. *Remote Sens-Basel* **2020**, *12*, 1824.
5. Smith, B.; Fricker, H.A.; Gardner, A.S.; Medley, B.; Nilsson, J.; Paolo, F.S.; Holschuh, N.; Adusumilli, S.; Brunt, K.; Csatho, B., *et al.* Pervasive ice sheet mass loss reflects competing ocean and atmosphere processes. *Science* **2020**, *368*, 1239-1242.
6. Harding, D.J.; Dabney, P.W.; Valett, S. In *Polarimetric, two-color, photon-counting laser altimeter measurements of forest canopy structure*, International Symposium on Lidar and Radar Mapping 2011: Technologies and Applications, 2011; International Society for Optics and Photonics: p 828629.
7. Neumann, T.; Brenner, A.; Hancock, D.; Robbins, J.; Saba, J.; Harbeck, K.; Gibbons, A. Ice, cloud, and land elevation satellite–2 (icesat-2) project: Algorithm theoretical basis document (atbd) for global geolocated photons (atl03). *National Aeronautics and Space Administration, Goddard Space Flight Center* **2019**.
8. Swatantran, A.; Tang, H.; Barrett, T.; DeCola, P.; Dubayah, R. Rapid, high-resolution forest structure and terrain mapping over large areas using single photon lidar. *Sci Rep-Uk* **2016**, *6*, 28277.
9. Parrish, C.E.; Magruder, L.A.; Neuenschwander, A.L.; Forfinski-Sarkozi, N.; Alonzo, M.; Jasinski, M. Validation of icesat-2 atlas bathymetry and analysis of atlas's bathymetric mapping performance. *Remote Sens-Basel* **2019**, *11*, 1634.
10. Wang, C.; Zhu, X.; Nie, S.; Xi, X.; Li, D.; Zheng, W.; Chen, S. Ground elevation accuracy verification of icesat-2 data: A case study in alaska, USA. *Opt Express* **2019**, *27*, 38168-38179.
11. Neuenschwander, A.L.; Magruder, L.A. Canopy and terrain height retrievals with icesat-2: A first look. *Remote Sens-Basel* **2019**, *11*, 1721.
12. Malambo, L.; Heatwole, C.D. Automated training sample definition for seasonal burned area mapping. *ISPRS Journal of Photogrammetry and Remote Sensing* **2020**, *160*, 107-123.
13. Kennedy, R.E.; Cohen, W.B.; Schroeder, T.A. Trajectory-based change detection for automated characterization of forest disturbance dynamics. *Remote Sensing of Environment* **2007**, *110*, 370-386.

14. Malambo, L.; Popescu, S.C.; Horne, D.W.; Pugh, N.A.; Rooney, W.L. Automated detection and measurement of individual sorghum panicles using density-based clustering of terrestrial lidar data. *ISPRS Journal of Photogrammetry and Remote Sensing* **2019**, *149*, 1-13.
15. Neuenschwander, A.; Popescu, S.; Nelson, R.; Harding, D.; Pitts, K.; Robbins, J.; Pederson, D.; Sheridan, R. Ice, cloud, and land elevation 1 satellite 2 (icesat-2) algorithm theoretical basis document (atbd) for land-vegetation along-track products (atl08) release 002. *National Aeronautics and Space Administration. Goddard Space Flight Centre* **2019**.
16. Neuenschwander, A.; Pitts, K. The atl08 land and vegetation product for the icesat-2 mission. *Remote sensing of environment* **2019**, *221*, 247-259.
17. Hansen, M.C.; Potapov, P.V.; Goetz, S.J.; Turubanova, S.; Tyukavina, A.; Krylov, A.; Kommareddy, A.; Egorov, A. Mapping tree height distributions in sub-saharan africa using landsat 7 and 8 data. *Remote Sensing of Environment* **2016**, *185*, 221-232.
18. Liu, C.; Neal, J.; Scofield, C.; Chang, J.; Ludeke, A.K.; Frentress, C. Classification of land cover and assessment of forested wetlands in the cypress creek watershed. *Texas Parks and Wildlife Department, Austin, Texas. 21p* **1996**.
19. Isenburg, M. Lastools: Efficient tools for lidar processing. https://rapidlasso.com/lastools/ (Accessed: January,15,2020),
20. Chidumayo, E. Climate and phenology of savanna vegetation in southern africa. *J Veg Sci* **2001**, 347-354.
21. Hill, M.J.; Millington, A.; Lemons, R.; New, C. Functional phenology of a texas post oak savanna from a chris proba time series. *Remote Sens-Basel* **2019**, *11*, 2388.
22. Malambo, L.; Popescu, S.; Ku, N.-W.; Rooney, W.; Zhou, T.; Moore, S. A deep learning semantic segmentation-based approach for field-level sorghum panicle counting. *Remote Sens-Basel* **2019**, *11*, 2939.
23. Narine, L.L.; Popescu, S.C.; Malambo, L. Synergy of icesat-2 and landsat for mapping forest aboveground biomass with deep learning. *Remote Sens-Basel* **2019**, *11*, 1503.