Assessing age-height relationships using ICESat-2 heights and Landsat time series products for southern pine in the southeastern region

Sonia Sharma Banjade

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Science

In

Forestry

Valerie A. Thomas, Chair

Randolph H. Wynne, Co-Chair

Patrick C. Green

December 1, 2023

Blacksburg, Virginia

Keywords: ICESat-2, site-index, southern pine, remote sensing, productivity

Assessing age-height relationships using ICESat-2 ATL08 heights and Landsat time series products for southern pine in the southeastern region

Sonia Sharma Banjade

Abstract (Academic)

This study investigates pine heights by age for actively managed stands in the southeastern U.S. using ICESat-2 ATL08-derived height data and maps derived from the Landsat time series. We intersected ICESat-2 ground tracks with locations of pine plantations and the Landscape Change Monitoring System (LCMS) Fast Loss product to identify previously clear-cut pine plantations. We subtracted the LCMS Fast Loss year from the date of the ICESat-2 acquisition to determine plantation age at the time of the height measurement. We stratified the data for management intensity, where stands that experience both thinning and harvesting were considered actively managed. The goal was to develop age-height relationships across the region to characterize better the impact of management on productivity and site index.

This research involved the analysis of over 137,998 ICESat-2 ATL08 segments in actively managed pine stands in the U.S. Southeast. We compared a subset of ICESat-2 heights with heights derived from airborne laser scanning acquisitions (ALS) available through the USGS 3D Elevation Program. The resulting R^2 was 0.82, giving us confidence in the ICESat-2 ATL08-derived forest heights. Then, through data processing and analysis, we successfully stratified the spatial patterns of ICESat-2 ATL08 heights in the southeastern region. These patterns provided insights into the distribution and variability of forest heights across the region, contributing to informed decisions in forest

management. We identified some challenges in predicting pine stand age through Landsat-derived disturbance products. We found that LCMS Fast Loss labels some heavy thins as a 'Fast Loss,' in addition to stand-clearing disturbances like clear-cuts, adding noise to our estimation of stand age. To overcome this issue, we employed a robust model of the logarithm of heights with a reciprocal of age using a random sample consensus (RANSAC) model to calculate site indices at base age 25 (years). Our results showed the site index for the region at a base age of 25 years is 20.1 m with a model R^2 of 0.91. We compared the ICESat-2-derived site index with the FIA-derived site index to see the robustness of our results. Then, the modeled site index values were used to produce a map at a base age of 25 years for the U.S. Southeast, offering insights into spatial differences in regional forest productivity.

The results of this study have important implications for ecological research, forest management, and well-informed decision-making. Insights into the distribution and trends of actively managed forest heights in the Southeast are gained from studying the vast dataset, allowing for more efficient land management and conservation initiatives. In actively manage stands, our site index equation improves the ability to anticipate site productivity and estimate future timber outputs. The difficulties with age estimation that have been observed highlight the need for better methods for mapping disturbances using remote sensing in forests that use thinning as a silvicultural prescription.

Assessing age height relationship using ICESat-2 ATL08 heights and Landsat time series products for southern pine in the southeastern region

Sonia Sharma Banjade

Abstract (General Audience)

This study investigates pine productivity in the southeastern U.S. using ICESat-2 ATL08-derived height data and maps derived from the Landsat time series. We intersected ICESat-2 ground tracks with locations of pine plantations and the Land Change Monitoring System (LCMS) Fast Loss product to identify previously clear-cut pine plantations. We subtracted the LCMS Fast Loss year from the date of the ICESat-2 acquisition to determine plantation age at the time of the height measurement. We stratified the data according to management intensity, where stands that experience both thinning and harvesting were most intensively managed. The site index for the region at base age 25 is 20.1 m with a model R^2 of 0.91.

Table of Contents

List of figures

List of tables

ATTRIBUTION

Valerie A. Thomas and Randolph H. Wynne are the advisors to this study. They oversaw the process, were thesis editors, and contributed to the study design, implementation, and analysis.

Patrick C. Green helped review and edit the thesis. He also contributed to the study design, implementation, and analysis.

1. Introduction

Forests are essential, providing both ecological and economic advantages as they harbor approximately 80% of Earth's terrestrial biodiversity (Aerts & Honnay, 2011). Recently, forests' significance in addressing climate change has grown since forests and other land surfaces absorb about 30% of carbon emissions (Watson et al., 2000) from the atmosphere. Likewise, the actively managed southern pine forests in the southeast region play a dual role, offering a significant amount of timber in the global market (Sahoo et al., 2019) and absorbing roughly 13% of greenhouse gas emissions (Turner et al., 1995). However, the potential impact of climate change on these forests suggests a decline in their future sustainability (Forest, 2012).

Quantifying forest productivity is essential for effective forest management, as it provides valuable insights into various aspects of forestry, including silvicultural practices, carbon dioxide sequestration, biomass estimation, and understanding the impacts of climate change. The site index is the key indicator for quantifying forest productivity, representing the relationship between tree age and height in specific site conditions (Burkhart & Tomé, 2012b).

Traditionally, site index has been measured through parametric and non-parametric methods based on field measurements of forest attributes relating to the bio-physical site factors (Coops, 2015; Diéguez-Aranda et al., 2006; C. Li et al., 2020; Sabatia & Burkhart, 2014). While these methods have been reliable, they have limitations, especially when dealing with extensive geographical areas. Conventional field data collection for measuring site index is time-limited, cost-prohibitive, and labor-intensive. Additionally, field measurements may not provide the spatial specificity required for

comprehensive forest management strategies (Eid et al., 2004). Field measurements are inherently localized and may not capture the full extent of a forested area. Old measurements may become outdated and do not accurately represent the current situation, especially in dynamic ecosystems where environmental conditions change over time. This lack of up-to-date and spatially comprehensive data can hinder effective forest management decision-making.

Recent advancements in remote sensing technology, like airborne laser scanning and earth resource satellite imaging, give a chance to estimate forest attributes, including tree height and age, over vast areas with repeatable and consistent observations on a global scale (Hall et al., 2011). The forest productivity maps that can be generated through remotely sensed data have enhanced spatial specificity and cover large geographical areas for continuous monitoring of forest productivity.

Only a limited number of studies in the southern yellow pines have used field and remote sensing data to generate site index maps, with two noteworthy examples (Gopalakrishnan et al., 2019; Tompalski et al., 2015). Notably, both of these studies relied on ALS predominately. Despite its effectiveness in measuring forest structure and topography, ALS has several drawbacks. ALS may be prohibitively expensive for larger study areas, limiting its accessibility for comprehensive measurements. The acquisition and maintenance of ALS equipment and skilled personnel can impose a significant financial burden. In addition, the inconsistency in ALS data for a heterogeneous management landscape over a larger region is challenging. Variations in data quality and coverage can hinder the creation of reliable site index maps, particularly when confronted with diverse land management practices, tree species, and age classes. Additionally, the

temporal variability of ALS data, often due to factors like weather conditions and when data were acquired, is a significant obstacle in conducting long-term and seasonality-sensitive data analysis.

Stand age is the fundamental factor for site index determination in understanding forest dynamics and productivity. Assessing the age of individual trees on a large scale can present several challenges. The process of measuring the age of each tree can be resource-intensive and time-consuming. Determining stem-specific ages is particularly problematic when dealing with vast forested areas. Additionally, the availability of standardized planting records, crucial for accurately determining stand age, is often lacking (Packalén et al., 2011; Wulder et al., 2010). Nevertheless, predictions of forest ages can be made using Landsat disturbance products (Gopalakrishnan et al., 2019). In addition, when combined with other spectral information, these time series data can be used to identify planted pine trees (Fagan et al., 2018).

We used spaceborne lidar and available Landsat-derived disturbance products to generate a site index curve and forest productivity map for southern yellow pines in the southeastern region. We used ICESat-2 ATL08 heights to derive forest heights and predicted stand age using the Landsat time-series disturbance map.

1.1. Objective and the research questions

The objective of the research was to assess the spatially specific and aggregate actively managed southern pine height-by-age relationship across the southeastern U.S. using ATL08 height and Landsat-derived age. Specifically, we tried to answer the following research questions:

1. How do ATL08 h_canopy heights compare to those determined from airborne laser scanning?
2. What is the distribution of heights by age across actively managed pine stands in the southeastern United States?
3. What is the average site index for actively managed pine stands compared to FIA in the southeastern United States?

1.2. Background

1.2.1. Forest productivity and site index

Forest productivity is a complex concept in forest science, so various assessment approaches exist for its quantification. The prominent method to quantify forest productivity is to establish a relationship between tree growth and environmental factors.

Site index has been a critical factor in the development of forest growth models since the 18th century when stand height was observed to vary according to site characteristics (García, 2005; Hennigar et al., 2017; LD De Laillevault, 1802). When forest height is plotted by age, forest growth is represented as a sigmoidal curve through time. Site index represents the maximum height of a stand at a specific base age (e.g., at 25 years old). It is typically expressed as a numerical value and is determined by measuring the average height of dominant or codominant trees in a stand at a specific reference age, often 25, 50, or 100. The higher the site index, the more productive the stand.

Site index curves are mainly constructed based on three general methods: the guide curve method, parametric prediction model, and difference equation method

(Clutter et al., 1983; Elfving & Kiviste, 1997). The guide-curve method is the most widely used to construct a site index. If the curves are proportional to each other, they are said to be anamorphic curves. These curves are constructed to establish a relationship between the logarithm of height and the reciprocal of age. Depending on the regeneration methods, i.e., plantation or natural stands, either age at breast height or total age is used as an independent variable. For the site index of plantation stands, the age since planting is used as an independent variable. These curves are usually the best fit for the same site quality classes. However, in many species of trees, the height of trees varies with site quality. The families of curves exhibiting shape variation depending on the site qualities are called polymorphic curves (Burkhart and Tome, 2012).

The second method used to construct site index is the parameter prediction method, which involves fitting a base height-age model, assigning site index values, and relating parameter estimates to the site index. This method detects the whole population trend, not the sample of trees (Burkhart and Tennent, 1977).

Another method to compute site index is using a difference equation. These are dynamic equations based on a continuous four-variable prediction system (Cieszewski & Bailey, 2000). Within the difference equation approach, (Bailey & Clutter, 1974) defined the algebraic difference approach (ADA), a technique to develop site index curves independent of specific base age. This method of developing site index curves makes it more flexible for application on different plots across different stand ages. In this approach, the difference in the logarithms of height and age is modeled as a function of age. Cieszewski 2000, extended ADA to the generalized algebraic approach (GADA). These equations made the concept of passing the site index curve through zero or origin,

ensuring a logical starting point for tree growth modeling with a predicted height of zero at age zero. The concept of base-age invariance, introduced by Bailey and Clutter, ensures that the predictions from dynamic equations remain unaffected by changes in the base age while GADA allows polymorphism and varying asymptotes, contributing a comprehensive and flexible representation of tree growth patterns (Burkhart & Tomé, 2012a).

1.2.2. ICESat-2 ATL08 canopy heights - general data description

ICESat-2 was launched on 15 September 2018 with a single instrument called the Advanced Topographic Laser Altimeter System (ATLAS) (Zwally et al., 2002) and provides global measurements of vegetation canopy height.

ICESat-2 contains three pairs of beams. ATLAS splits a single laser signal into six beams such that each pair of beams is divided into strong beams with high-pulse energy and vice-versa (Neumann et al., 2019). The two beams, part of a set of six beams, are separated by a distance of 90 meters and cover an area of around 3 km cross-track. Each of the six beams has a footprint with a diameter of approximately 17 m, and the along-track sampling interval is 0.7 m (Markus et al., 2017). Each pair of beams consists of high energy and low energy with a ratio of 1:4 in terms of energy and is transmitted at a frequency of 10 Hz with a wavelength of 532 nm (Malambo & Popescu, 2021; A Neuenschwander et al., 2019). From its orbital altitude of 500 km, it can capture a laser pulse of 70 cm. Over highly reflective surfaces like ice and snow, it can detect about ten signal photons per footprint. Nevertheless, it can only detect a limited number of photons, from 0 to 4 per footprint over land and vegetation (Amy Neuenschwander & Pitts, 2019).

ATL ground tracks followed by consecutive footprints are defined as GT1L, GT1R, GT2L, GT2R, GT3L, and GT3R based on the direction of travel. When ATLAS is traveling in the forward orientation, the weak beams are on the left side, defined as GT1L, GT2L, and GT3L. When the ATLAS is in backward orientation, the stronger beams are on the left side, defined as GT1R, GT2R, and GT3R (Neumann et al., 2019).

The ATL08 data product contains along-track ground and canopy surface heights, providing a global sample of canopy structure data with repeated measures every 91 days, cloud cover permitting (Abdalati et al., 2010; Markus et al., 2017; Amy Neuenschwander et al., 2020). The ATL08 data product is derived from ATL03 (geolocated photons) using DRAGANN (Differential, Regressive, and Gaussian Adaptive Nearest Neighbor), an algorithm filtering noise and signal from the vegetative land surface to determine the canopy height (Amy Neuenschwander & Pitts, 2019). ATL08 provides canopy height measurement consecutively at every 100 m step size along-track (Figure 1), which includes estimates of relative heights (RH) relating to the canopy and statistical information about the elevation of terrain (Li et al., 2020). The data given by the instrument is dependent on the energy transmitted energy back from the reflected object. The number of photons that reach the ground is affected by the energy of a beam, atmospheric conditions, and noise levels. Acquisition time is another crucial element, as solar radiation energy may interfere with the lidar system in the day, leading to much more background noise than at night (Neumann et al., 2019; Popescu et al., 2018).

To determine canopy height metrics, ICESat-2 ATL08 provides three measurements: h_max_canopy (the highest point of the canopy in relation to the ground, h_canopy (RH98), which indicates the height from which 98% of the energy is reflected,

and canopy_h_metrics, which are cumulative relative height calculated at 5 % intervals at the percentage range of 10 to 95. The number of photons in the segment (n_seg_ph) and signal-to-noise ratio (SNR) influence the precise label of ground and canopy photons classified by ICESat-2 algorithms. The main challenge for the algorithm is the impact of solar background noise, which makes it difficult to determine the top of canopy photons. Therefore, it is recommended to use the RH98 as a measurement of the top of the canopy, as it is less affected by SNR uncertainty at the top of the canopy (Malambo & Popescu, 2021; A Neuenschwander et al., 2019; Popescu et al., 2018).

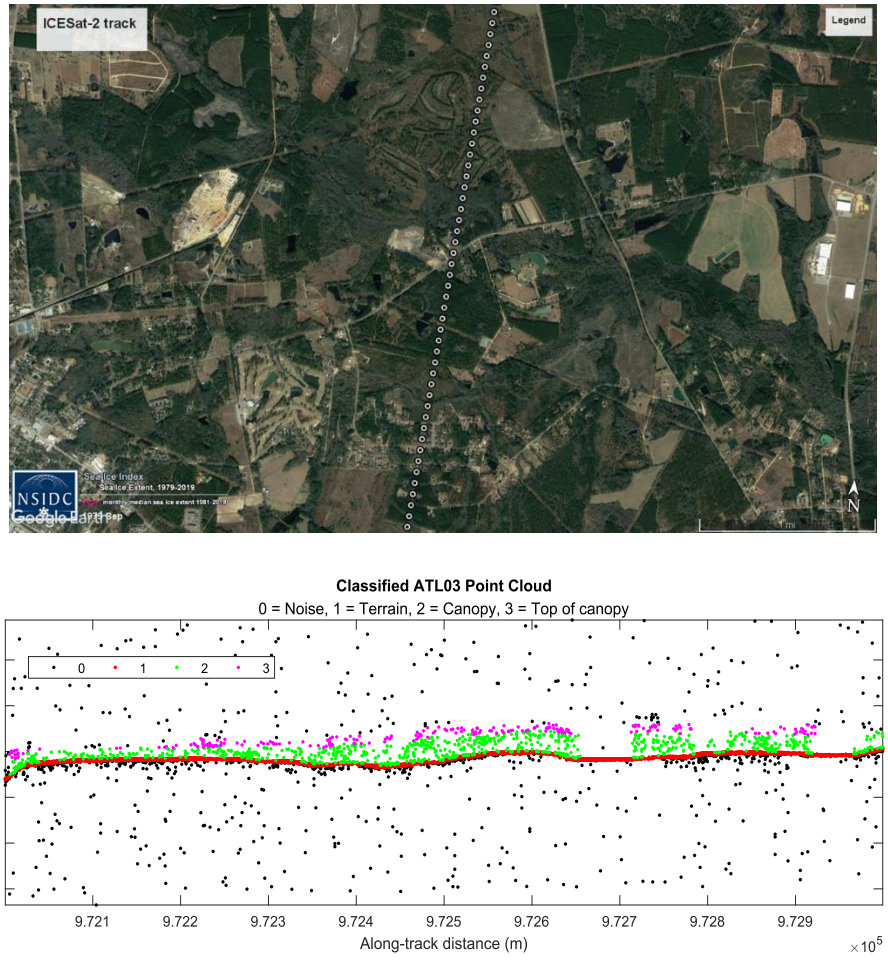

Figure 1: The top panel shows vegetated segments along the ICESat-2 track for the 'gt3r' beam, overlaid on high-resolution satellite imagery. The bottom panel shows the photon classification profile for the same track in southern pines in 2019.

Several previous studies that have used the ICESat-2 ATL08 product to evaluate its performance in retrieving canopy and terrain heights have demonstrated the validity of the algorithm and resulting data sets. A study by (Neumann et al., 2019) demonstrated the accuracy of ICESat-2 canopy heights, with a root mean squared error (RMSE) of less than 3.2 m in a vegetated region of Finland. Likewise, (Liu et al., 2021) compared the performance of ICESat-2 and GEDI, with ICESat-2 outperforming GEDI in a densely vegetated area with an RMSE of 2.23 m.

2. Methods

2.1. Study area

The study area is defined by the historic natural range of *Pinus taeda* (loblolly pine) as defined by (Little and Viereck, 1971) which includes parts of 14 states extending through southern New Jersey to central Florida and from eastern Texas to the Appalachian Highlands (Figure 2). While loblolly pines are predominant in the region, the area is also a habitat for *Pinus palustris* (longleaf pine), *Pinus echinata* (shortleaf pine), and *Pinus elliottii* (slash pine). Collectively, these species are referred to as southern pines or southern yellow pines. The region covers several geographical provinces, including the Atlantic Plain, the Piedmont Plateau, the southern extremities of the Cumberland Plateau, the Highland Rim, and the Valley and Ridge provinces. Humid and warm temperatures with long, hot summers and mild winters characterize the climate

in this region. The average annual rainfall ranges from 1020-1520 mm, while the mean annual temperature varies between 13^0 C - 15^0 C. The dominant soil type within southern pine forests is Ultisols (Baker & Langdon, 1990).

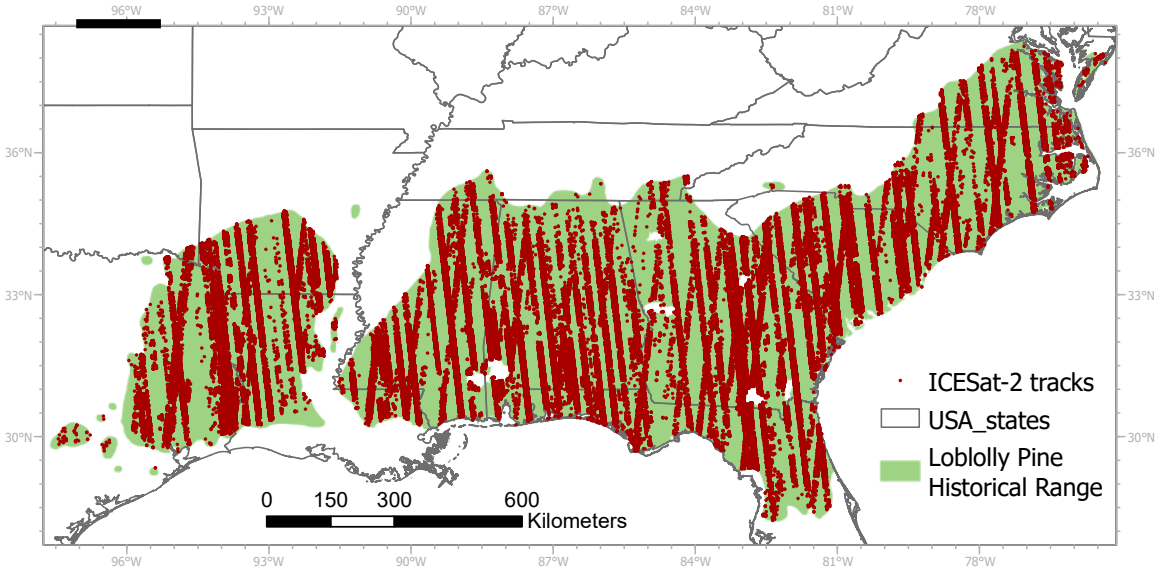

Figure 2: The green shaded area shows the natural, historic loblolly pine range of the southeastern U.S. as defined by (Little and Viereck, 1971). The red dots are the coordinates of the filtered ICESat-2 ATL08 centermost photons used for assessing the heights of southern pine.

2.2. Datasets used in the research.

This analysis uses height information derived from spaceborne lidar (specifically, the ICESat-2 ATL08 product) and derived products describing forest change based on multitemporal Landsat data from 1984 to the present. The ICESat-2 ATL08 product (described in Section 1) provides heights and is central to this research. The derived products include the U.S. Forest Service Landscape Change Monitoring System (LCMS) Fast Loss Product and a map depicting the location of plantations from 1987-2019

(Thomas et al., 2021). ICESat-2 canopy heights were used with age derived from LCMS Fast Loss disturbance year to find the relationship between the age and height of the southern pine stands in the southeastern region. Brief descriptions of the LCMS product, which is used to identify stand age, and the pine plantations map, which ensures the stands are southern pines, are given below.

2.2.1. Land change monitoring system (LCMS) product

The Land Change Monitoring System (LCMS) uses a remote sensing-based dense time series to monitor continuous changes in vegetation and other classes across the conterminous United States and Southeastern Alaska from 1985 to the present. It was developed by the United States Department of Agriculture Forest Service (USFS). It was developed using Collection 1 Tier 1 Landsat 4,5,7 and 8 and Sentinel-2a and -2b level 1C top of atmosphere reflectance data as predictor layers for modeling. The LandTrendR (Kennedy et al., 2010) and Continuous Change Detection and Classification (Oliver & Webster, 1990; Zhu & Woodcock, 2014) algorithms are used for temporal segmentation of the composites derived from predictors layers. Further, a random forest model is employed to produce change detection, land use, and land cover class products. We used the fast loss year product, which provides the year of LCMS modeled change classes for stand-replacing disturbances (e.g., clear-cutting) and non-stand-replacing disturbances (thinning, damage by fire, wind, and ice) (Housman et al., 2021). The reported year of disturbance can be used as a proxy for the age of pine stands by subtracting the disturbance year from the reference date.

2.2.2. Pine plantations map- general data description

We used an active forest management map described by (Thomas et al., 2021) to identify the pine stands and management intensities. This dataset is a classification of areas that are (or were) pine plantations at some point between 1987-2019, as well as an indicator of management level (i.e., non-managed, extensively managed, and actively managed). The overall accuracy of plantation versus non-plantation is 82% (Thomas et al., 2021). With respect to management gradients, the primary class is the 'actively managed plantations,' which includes areas that have been thinned and repeatedly harvested in that period, with a class accuracy of 90%. The largest source of error in this map is stands, which are extensively managed and do not follow a characteristic harvesting pattern. Overall accuracy, including management gradients, is 72%. We used this product to constrain our ICESat-2 samples to actively managed stands.

2.3. Data processing and analysis

2.3.1. Developing a height, age, and management dataset for pines

To process ICESat-2 data, we first downloaded all ATL08 files for 2019 through the NASA EarthData website (https://search.earthdata.nasa.gov) for the southeastern US. We used Land and Vegetation Height data product (ATL08) version 5, Level 4 to obtain tree height metrics above the WGS84 ellipsoid. We extracted the following metrics for each segment: h_canopy, h_canopy_uncertainty, night_flag, latitude, and longitude (Table 1). ICESat-2 data for the southeastern U.S., prior to our data reduction processing, included 10,817 files in hierarchical data format (HDF5), requiring approximately 250 GB of data storage.

The processing was done using the "*icepyx"* package in Python 3.9 (Van Rossum & Drake, 2009), which was created for processing the HDF5 files of ICESat-2 (code shown in Appendix A). Metrics were extracted from the binary HDF5 format and stored in CSV files while maintaining an identifier to the original file (Table 1). We then removed daytime data (i.e., night_flag = 0) to avoid known issues with solar background noise in the strong beams (Liu et al., 2021; Popescu et al., 2018). We also used the h_canopy_uncertainty flag to remove invalid data (flag = 3.4028E+38). This flag considers all possible sources of systematic uncertainty, such as timing, geolocation, orbits, and many more, as well as uncertainty in photon classification. A value of 3.4028E+38 is reported if the proportion of ground photons in an ICESat-2 segment is less than or equal to 5 % of the total signal photons per 100 m segment. The computed terrain height will likely have low confidence if there are too few ground photons in a section. Since they could not give precise measurements of topography and canopy height, segments with an 'invalid' value in the canopy height uncertainty index are removed for this study. After these filtering steps, we had 2,360 files of ICESat-2 segments for further processing.

Table 1: Definition of metrics extracted from ICESat-2 ATL08 products (https://nsidc.org/sites/default/files/documents/technicalreference/icesat2_atl08_data_dict_v006_0.pdf)

Metric	Definition
h_canopy	98th percentile of height of all the individual canopy relative heights above ground for the 100 m segment
h_canopy_uncertainty	Uncertainty related to relative canopy heights incorporating all systematic uncertainties, as well as uncertainty from errors of photons
night_flag	Flag indicating data were acquired at night; 0 indicates day, and 1 indicates night.
latitude	Latitude for the center-most signal within each segment
longitude	Longitude for center-most signal within each segment

Some additional filtering was done to remove segments with unrealistic heights exceeding the highest possible heights of southern pines (but which may represent an anthropogenic feature). All segments with height observations greater than 60 m were excluded (1,470 segments).

Finally, we used the latitude and longitude metrics to intersect with the LCMS Fast Loss product to extract the year of major disturbance. This analysis was done in the R programming language (Ihaka & Gentleman, 1996) using packages "*rgdal*", "*sp*", and "*raster*"(code shown in Appendix B). The extracted forest disturbance year was subtracted from the ICESat-2 data acquisition year to calculate the age of the stands at the time of lidar acquisition. Since the LCMS product contains all forests (including deciduous species), we further intersected the dataset with the pine plantation map to extract actively managed stands that undergo a repeated pattern of harvest and thinning (137,998 segments).

2.3.2. Validation of ICESat-2 ATL08 heights using ALS point clouds

There are two ways to quantify the uncertainty of ICESat-2 heights for loblolly pine. The first is to validate directly with field measurements. The challenge with field measurements is the low probability of the ICESat-2 tracks intersecting with enough established field plots to develop a robust statistical comparison. This inability to validate with field plots is a well-known challenge with spaceborne lidars (all of which are samples across the landscape). Given this constraint, an alternative approach used by several authors has been to compare spaceborne lidar heights with airborne lidar, which has long been used to map forest height (Fernandez-Diaz et al., 2022; Su et al., 2017)

To develop a dataset of ALS heights, we downloaded over 2000 tiles (1 km^2) of high-density airborne lidar data (8 points/m^2) from the USGS 3-DEP Lidar Data Exploration Tool (https://apps.nationalmap.gov/lidar-explorer/#/). Specifically, we used the CPRA LADOTD Lidar 2019 mission, conducted from March 27, 2019, to November 04, 2019. The metadata included with the dataset indicates the root mean square error (RMSE) in the vertical dimension is 10 cm for Q1 LiDAR point clouds (USGS 2023), which is lower than the error in canopy height for ICESat-2. Hence, this can serve as the base for the quality assurance of ICESAT-2 heights. From this dataset, 83 airborne lidar tiles intersected our ICESat-2 samples.

We used the "*lidR"* package in R (Roussel et al. 2020) to process "*laz"* files downloaded from USGS LiDAR explorer to have canopy height models for each tile. This process involved creating a normalized canopy height model using a TIN interpolator on ground returns (already classified by USGS) to create a 1 m digital surface model (dsm_point2raster Digital Surface Model Algorithm in R). The ground

elevation was then subtracted from every lidar point to create a normalized height dataset. The canopy surface was interpolated using the "*rasterize_canopy"* algorithm to generate a 1 m canopy height grid (code shown in Appendix C).

To approximate the footprint of the ICEsat-2 segment on the ground, we created a 100 x 13 m box around our center coordinate to sample our 1 m cells, following (Amy Neuenschwander et al., 2020). We then extracted the 98th percentile of heights represented by the approximately 1,300 cells within the box. We also determined the dominant land cover by intersecting this box with NLCD and extracting the model. We used the Python packages *'gdal*', *'pandas*', and *'numpy'* to directly compare the heights across the land covers in the 83 tiles (code shown in Appendix D).

To assess the agreement between the two height products, we performed a linear fit and calculated the percent of the dataset with the 95% confidence interval, the RMSE and R^2 of the fit, and the mean deviation.

2.3.3. Height by age for managed southern yellow pines

We used violin plots as the visualization tool to see the distribution of ICESat-2 ATL08 heights over ages to see the trend of growth patterns for actively managed southern pines in the southeastern region. Violin plots were used to see the patterns as they effectively combined the basic summary statistics and kernel density, making it easier to see the trends (Hintze & Nelson, 1998).

2.3.4. Generating a site index curve for actively managed stands and the spatial distribution of site indices

As described in Section 1 above, the guided curve method is widely used to fit a site index curve. We used the (Schumacher, 1939) model to generate the age-height relationship using the logarithm of height and the reciprocal of age, given by,

Equation:

$$ln(h_{dom}) = b_0 + b_1(1/a) \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots . \ldots \ldots \ldots (1)$$

where,

$ln(h_{dom})$ = logarithm of dominant height

a = age of the stands

This model was used to generate a site index curve for the actively managed pine stands in the southeast, which included 137,998 ICESat-2 segments. Initial analysis revealed significant challenges associated with the miscalculation of age for some stands that have been heavily thinned. Since stand age was calculated as the ICESat-2 data acquisition year minus the forest disturbance year, heavy thins identified as fast loss resulted in many outliers (71,504 segments) that were far too tall for their age. To address this issue, we adopted a repeated random sub-sampling strategy (i.e., the RANSAC algorithm), which does not include noisy data when fitting the site-index model. For RANSAC, 25,000 segments were randomly selected to separate noise from signal (inliers) to fit the site index curve. RANSAC is an algorithm introduced by (Fischler & Bolles, 1981) which performs well with noisy data. It can effectively smooth data even when it contains a substantial amount of error (Derpanis, 2010).

In this context, RANSAC provided resilience to the impact of miscalculation of age. Traditional least squares methods can be susceptible to these noisy data, leading to skewed results and inaccurate model fits. On the other hand, RANSAC addresses this challenge by iteratively fitting the model to random subsets of the data, identifying the inliers that conform to the model, and disregarding the noise. Consequently, developing a reliable site index curve better represents the growth patterns of the actively managed pine stands.

The site index values from the model were used to develop a spatial distribution at a base age of 25 years.

2.3.5. Comparison of ICESat-2 derived site index curve with FIA-derived site index

To compare our results to a trusted independent dataset, we also processed Forest Inventory and Analysis (FIA) data provided by the Southern Research Station (SRS) of the USFS. We used the maximum height (MaxofHT) and stand age (STDAGE) of southern yellow pines of the most recent cycle of annual FIA plots in the southeastern region. The data included one observation for each plot location, 601 samples in total. We then used linear regression develop a smoothed site index curve for direct comparison to the ICESat-2 derived curve.

3. Results

3.1. Statistical summary of heights dataset for actively managed pine stands

The statistical summary of processed ATL08 heights 25-year-old southern yellow pines in the southeastern region is in the table below (Table 2).

Table 2: Statistical summary of ATL08 heights of age 25 for actively managed pine stands

Summary	Statistics
Count	1987
Mean	20.06
Standard deviation	4.73
Minimum	2.21
25%	18.00
50%	20.433
75%	22.60
Maximum	55.39

The histogram of heights demonstrates a normal distribution across the region (Figure 4).

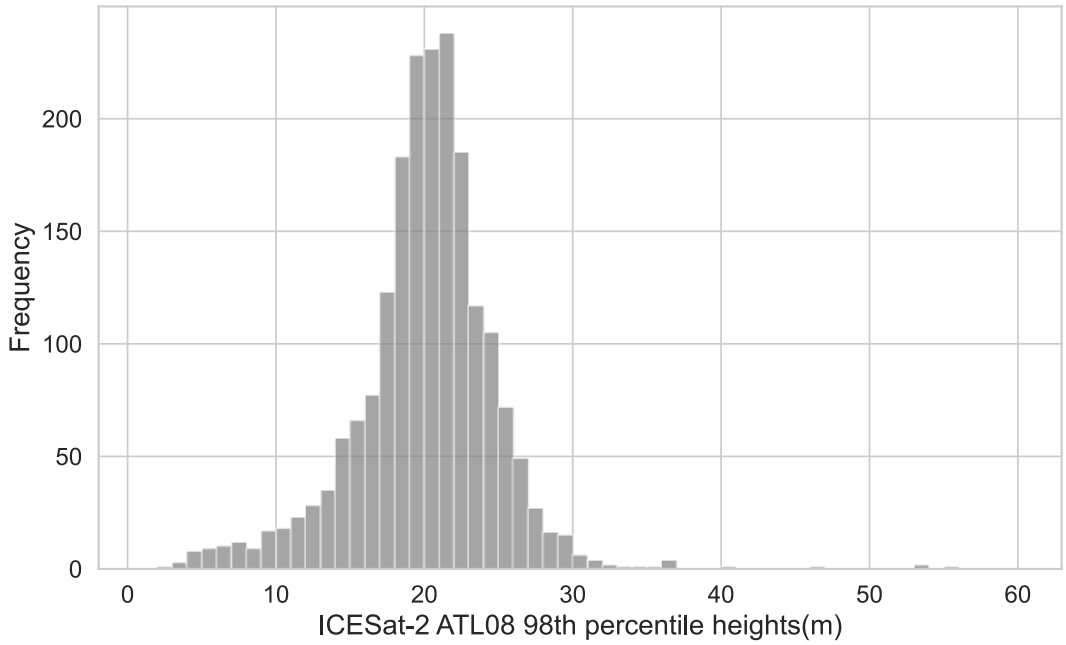

Figure 3: Histogram of processed ATL08 heights for 25-year-old actively managed stands (1,987 segments).

3.2. Validation of ICESat-2 ATL08 heights using ALS point clouds

The linear fit (Figure 5) between ICESat-2 ATL08 98th percentile heights and ALS 98th percentile heights showed the heights were strongly correlated, with an R^2 of 0.82. According to their National Land Cover Database (NLCD) classification, multiple land cover classes were present, with the majority being evergreen (coniferous), woody wetlands, and grassland/herbaceous.

Approximately 96% of ICESat-2 RH98 heights were within the computed 95% confidence interval of ALS Rh98 heights (Table 3). The average difference between ALS and ICESat-2 ATL08 heights is about 1.8 m, while the overall difference between dataset is 2.66 m.

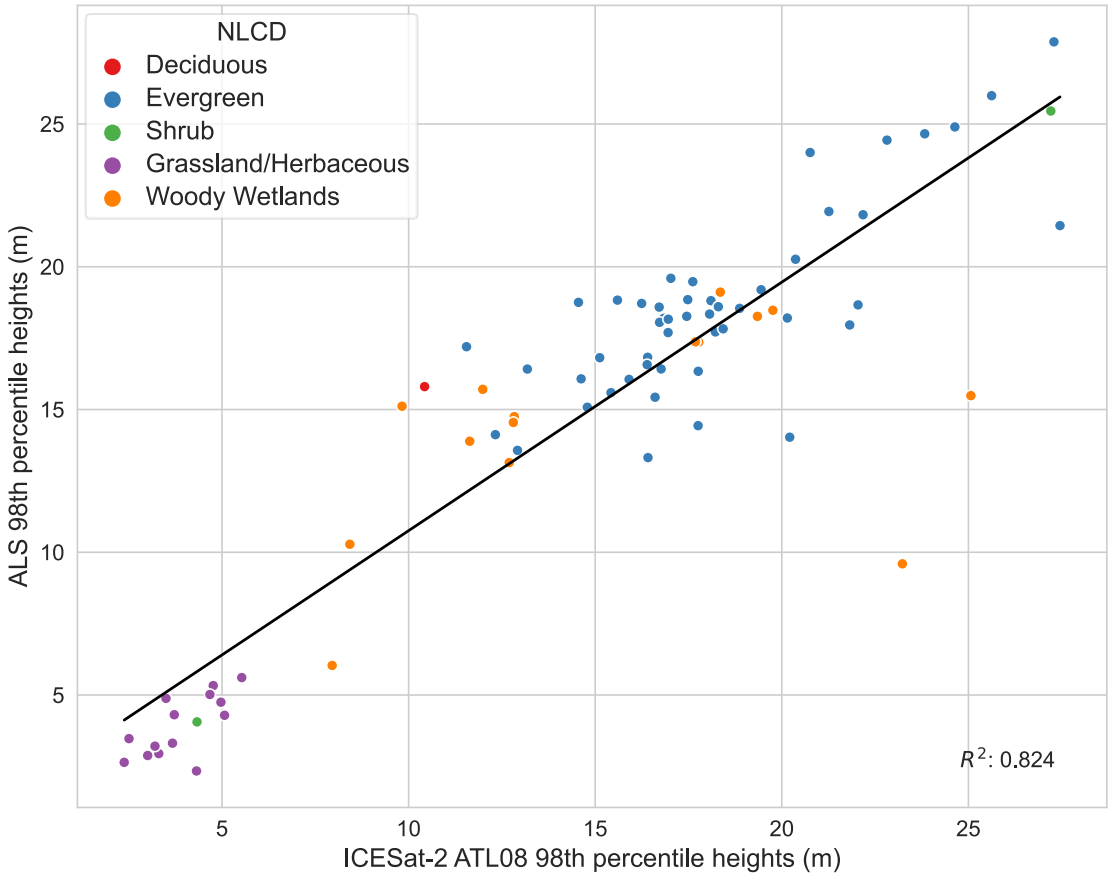

Figure 4: Linear fit showing the relationship between ICESat-2 ATL08 98th percentile heights in meters with ALS 98th percentile heights in meters. The color code represents the various NLCD land cover classes that fall within the region of ICESat-2 ATL08 tracks.

Table 3: Statistical value of the comparison of ATL08 heights and ALS point clouds

Condition	Sample size	%within 95% CI	MD	RMSE	R^2
ATL08 98th vs ALS 98th	83 segments	96.29	1.8	2.66	0.82

3.3. Height by age for southern yellow pines

Figure 7 represents the ATL08 height (y-axis) distribution for each age (x-axis) category for actively managed southern yellow pines. A distinct trend in the violin plot is the accelerated growth rate in the younger stands. The older stands, comparatively, had a higher density than the younger stands. The presence of outliers in the younger stands is caused by heavy thinning, which confounds our approach to estimating age. The disturbance due to thinning can be seen in the Fast Loss product. In our estimation of age, this has the effect of setting the thinning year to "Age=0".

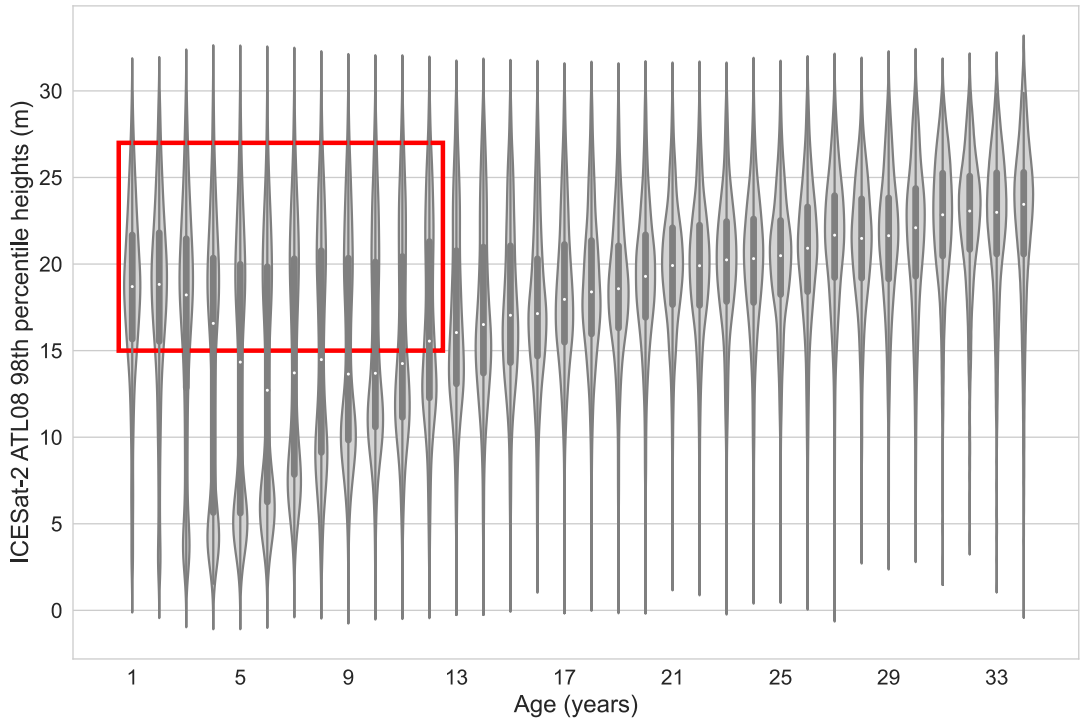

Figure 5: The violin plot of ICESat-2 ATL08 98th percentiles heights in meters over ages in years for actively managed stands. The red bounding box indicates the noisy data detected, showing higher heights in the younger stands ranging from 1-12 years.

3.4. Site index curve and spatial distribution

We generated the following parameters for our site index equation:

$$ln(h_{dom}) = b_0 + b_1(1 \,/\, age) \ldots\ldots\ldots (2)$$

where,

$b_0 = 3.3$ and $b_1 = -8.99$ using the RANSAC algorithm. The curve fit had an $R^2 = 0.90$ and RMSE = 0.13 m (Figure 7).

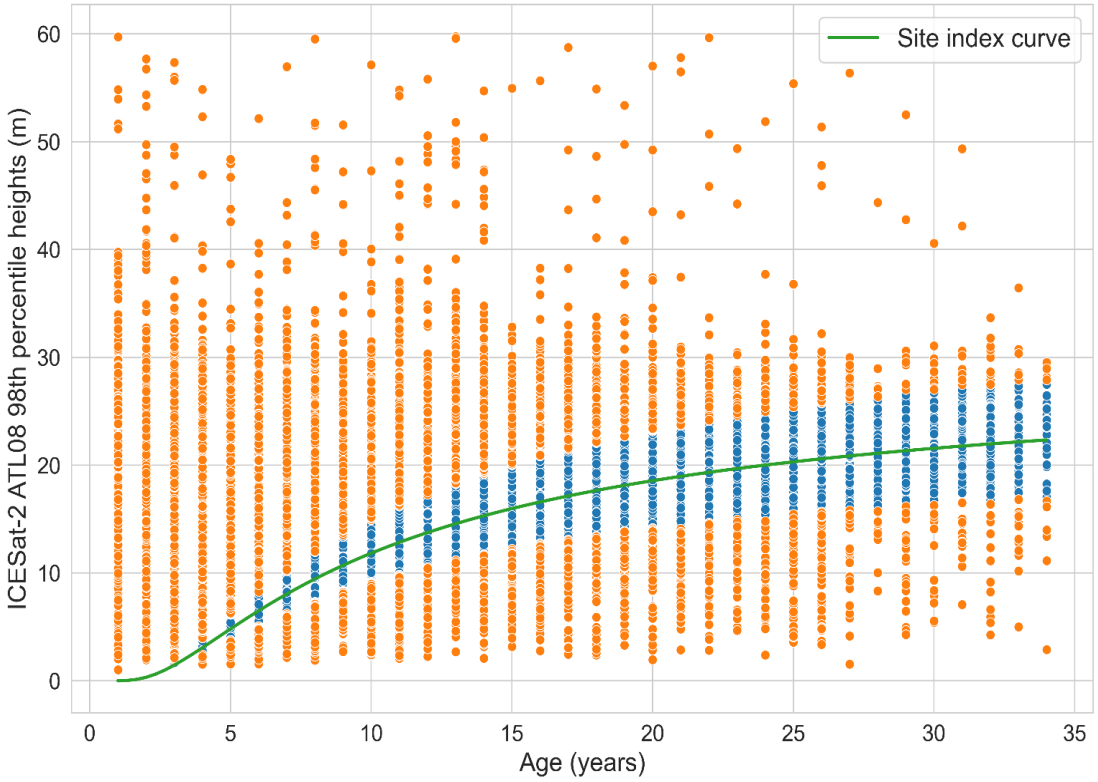

Figure 6: The green fitted line indicates the site index curve generated using ICESat-2 ATL08 heights and Landsat-derived age for the southeastern region using the RANSAC algorithm and the (Schumacher, 1939) model. The blue points are identified as inliers, and the orange points are identified as noise by RANSAC.

The site indices at a base age of 25 years show pockets of high productivity throughout the natural historic range (Figure 8).

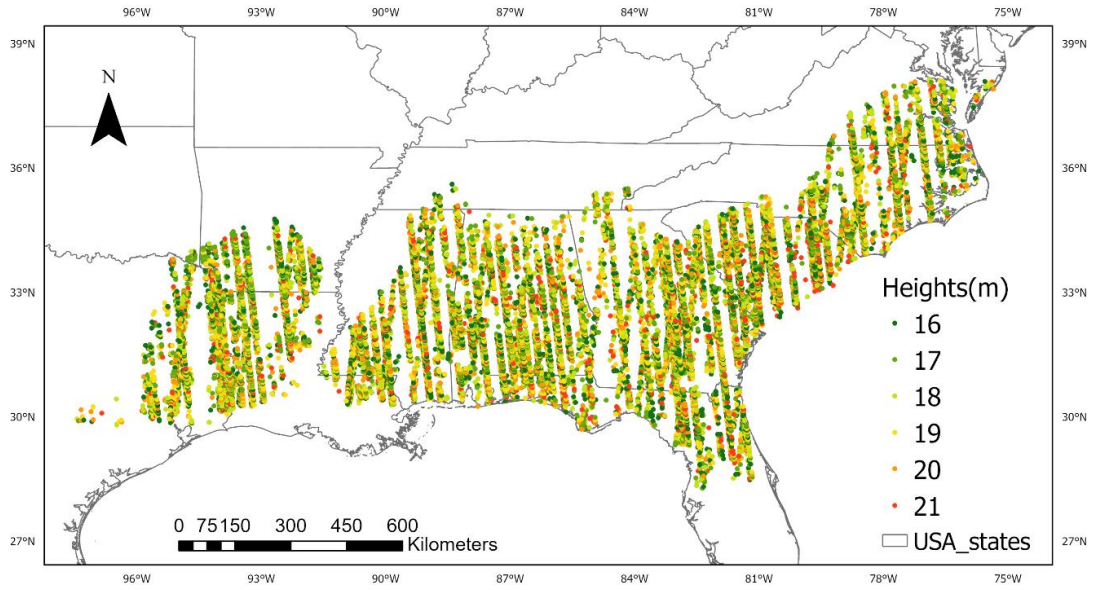

Figure 7: The site indices at base age 25 years using the inliers detected from RANSAC in the natural historical loblolly pine range.

3.5. Comparison of ICESat-2 derived site index with FIA-derived site index

Fitting the FIA data points with the same site index equation:

$$ln(h_{dom}) = b_0 + b_1(1 / age) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

we get the following parameters for the FIA points: $b_0 = 3.4$, $b_1 = -8.35$, $R^2 = 0.81$ and RMSE = 0.30 m. The FIA-derived site index value for base age 25 years was calculated as 21.5 m using ground plot data (compared to 20.1 m for ICESat-2) (Figure 9).

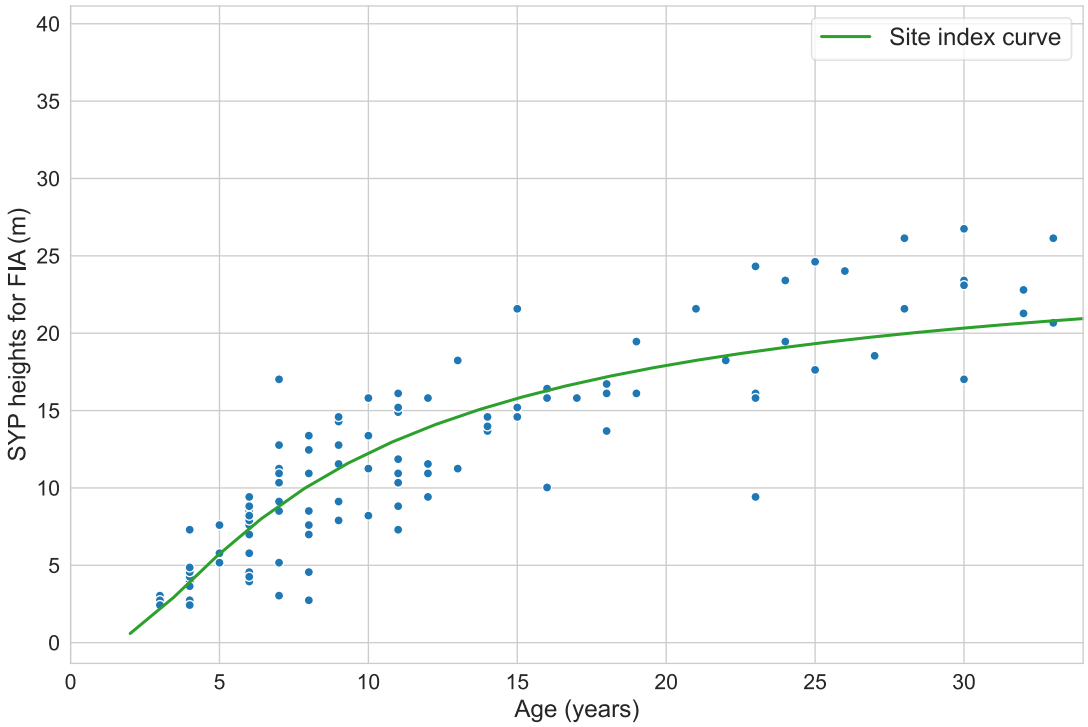

Figure 8: The green fitted line indicates the site index curve generated using FIA plot data for the southeastern region using the linear regression and the (Schumacher, 1939) model.

FIA ground plot-derived site index is compared with ICESat-2 derived site index in Figure 10. There is a strong agreement between the two curves, with the FIA-derived heights being slightly higher at every age.

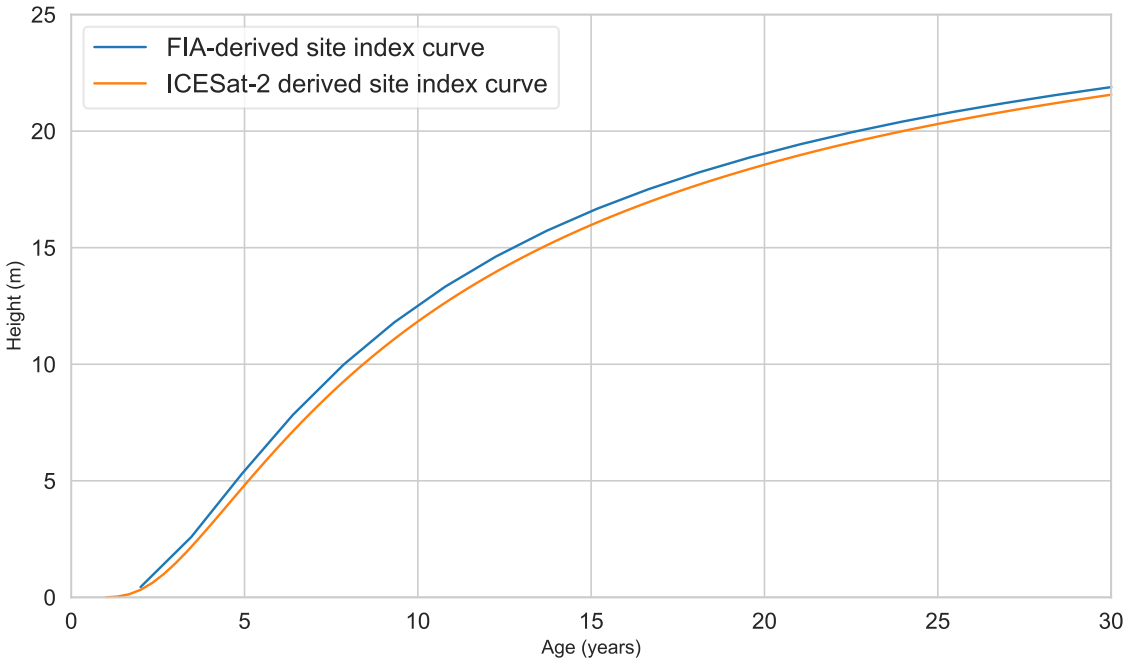

Figure 9: The blue fitted line indicates the site index curve generated using FIA plot data, and the orange line shows the site index curve generated by remote sensing data for the southeastern region

4. Discussion

The validation of ICESat-2 data in our study is crucial to our research findings. We found a strong correlation between ICESat-2 ATL08 heights and ALS point clouds, signifying the reliability and accuracy of ICESat-2 data in estimating tree heights. The quality of ICESat-2 data was maintained using data acquired at night, as mentioned by (Neumann et al., 2019; Popescu et al., 2018), to avoid solar background noise. With a small mean difference of only 1.8 m between the ICESat-2 and ALS data, our results underscore the precision of ICESat-2 as a valuable tool for height estimation. One of the key strengths of utilizing ICESat-2 data is its consistency in estimating tree heights across a vast region, which is impractical with ground measurements and ALS data. This

consistency allows for a more comprehensive assessment of the age-height relationship in southern yellow pine stands across the southeastern United States.

In our study, we utilized Landsat time series disturbance maps to estimate the age of trees, a crucial factor in site index determination. However, the noisy data detected as exceptionally tall canopy heights in younger stands were the limitation. These noisy data may be associated with potential errors in age predictions using Landsat disturbance products. These findings are consistent with prior research (Cohen et al. 2013; Desrochers et al. 2022), which identified limitations in the Vegetation Change Tracker (VCT) in detecting lower-intensity disturbances. LCMS has the highest accuracy in detecting clear-cuts but may struggle to capture subtle changes in forest cover resulting from low-intensity disturbances (Desrochers et al., 2022). These challenges may occur from the varying intensity, duration, and frequency of disturbances, influenced by factors such as regional differences, forest composition, and management practices. These challenges emphasize the need for more accurate methods to identify and characterize such disturbances.

One of the key strengths of this research is using the RANSAC algorithm for detecting and handling noisy data in a huge dataset. The RANSAC algorithm (Olofsson et al., 2014) to eliminate noisy data in tree stem and height measurements. We followed a similar approach to find the exceptionally tall canopy heights in our research. Traditional regression methods could have been influenced by the noise, leading to biased and skewed results. RANSAC significantly improved the robustness of our study's results, enhancing the reliability of our analysis. RANSAC's effectiveness in identifying and managing noisy data is a noteworthy contribution to our research methodology.

The site index curve and productivity map we generated hold practical significance for forest management and policymaking. Our research conducted a comparison between the ICESat-2-derived and FIA-derived site index curves. As demonstrated in Figure 10, the results affirm the reliability of remote sensing data for site index determination. The curve created using heights from ICESat-2 ATL08 and ages from LCMS nearly matches the FIA-derived site index curve. This comparison highlights the potential of ICESat-2 and similar remote sensing technologies in accurately and efficiently assessing site index, further underlining the utility of these data sources in forest management and ecological research.

Furthermore, the productivity map generated from the site indices provides both spatial and temporal specificity. The spatial specificity allows for sub-regional comparison of productivity based on climatic and ecological factors. This information is crucial for forest managers and policymakers, enabling them to identify areas of both high and low productivity within the study region. Also, the temporal specificity helps capture the variations and changes over time, providing dynamic insights into forest growth patterns. For low-productivity areas, targeted management interventions, such as fertilization or thinning, can be implemented to enhance tree growth. In contrast, high productivity areas can be managed to maximize carbon sequestration, contributing to environmental conservation efforts and maintaining a balanced carbon balance over time.

5. Conclusion

In conclusion, our research validates ICESat-2 data for tree height measurement, showing a strong correlation with ALS point clouds. Despite challenges with Landsat disturbance maps, the study's strength lies in applying the RANSAC algorithm to handle a noisy, large dataset of 71,504 segments enhancing the robustness of our analysis. The site index curve and productivity map have practical applications for forest management. The methodologies showcase the potential of remote sensing technologies for site index assessment representing various ecological and climatic factors across a vast geographical range. Our study integrates advanced technologies, addresses challenges, and provides actionable insights for sustainable forest management and ecological research.

References

Abdalati, W., Zwally, H. J., Bindschadler, R., Csatho, B., Farrell, S. L., Fricker, H. A., Harding, D., Kwok, R., Lefsky, M., Markus, T., Marshak, A., Neumann, T., Palm, S., Schutz, B., Smith, B., Spinhirne, J., & Webb, C. (2010). The ICESat-2 Laser Altimetry Mission. *Proceedings of the IEEE*, *98*(5), 735–751. https://doi.org/10.1109/JPROC.2009.2034765

Aerts, R., & Honnay, O. (2011). Forest restoration, biodiversity and ecosystem functioning. *BMC Ecology*, *11*, 29. https://doi.org/10.1186/1472-6785-11-29

Bailey, R. L., & Clutter, J. L. (1974). Base-age invariant polymorphic site curves. *Forest Science*, *20*(2), 155–159.

Baker, J. B., & Langdon, O. G. (1990). Pinus taeda L. Loblolly pine. *Silvics of North America*, *1*, 497–512.

Burkhart, H. E., & Tomé, M. (2012a). Evaluating Site Quality. In *Modeling forest trees and stands* (pp. 131–173). Springer Netherlands. https://doi.org/10.1007/978-90-481-3170-9_7

Burkhart, H. E., & Tomé, M. (2012b). *Modeling forest trees and stands*. Springer Netherlands. https://doi.org/10.1007/978-90-481-3170-9

Cieszewski, C. J., & Bailey, R. L. (2000). Generalized Algebraic Difference Approach: Theory Based Derivation of Dynamic Site Equations with Polymorphism and Variable Asymptotes. *Forest Science*.

Clutter, J. L., Fortson, J. C., Pienaar, L. V., Brister, G. H., & Bailey, R. L. (1983). *Timber management: A quantitative approach*. John Wiley & Sons, Inc.

Desrochers, M. L., Tripp, W., Logan, S., Bevilacqua, E., Johnson, L., & Beier, C. M. (2022). Ground-Truthing Forest Change Detection Algorithms in Working Forests of the US Northeast. *Journal of Forestry*, *120*(5), 575–587. https://doi.org/10.1093/jofore/fvab075

Eid, T., Gobakken, T., & Næsset, E. (2004). Comparing stand inventories for large areas based on photo-interpretation and laser scanning by means of cost-plus-loss analyses. *Scandinavian Journal of Forest Research*, *19*(6), 512–523. https://doi.org/10.1080/02827580410019463

Elfving, B., & Kiviste, A. (1997). Construction of site index equations for Pinus sylvestris L. using permanent plot data in Sweden. *Forest Ecology and Management*, *98*(2), 125–134. https://doi.org/10.1016/S0378-1127(97)00077-7

Fagan, M. E., Morton, D. C., Cook, B. D., Masek, J., Zhao, F., Nelson, R. F., & Huang, C. (2018). Mapping pine plantations in the southeastern U.S. using structural, spectral, and temporal remote sensing data. *Remote Sensing of Environment*, *216*, 415–426. https://doi.org/10.1016/j.rse.2018.07.007

Fernandez-Diaz, J. C., Velikova, M., & Glennie, C. L. (2022). Validation of ICESat-2 ATL08 Terrain and Canopy Height Retrievals in Tropical Mesoamerican Forests. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *15*, 2956–2970. https://doi.org/10.1109/JSTARS.2022.3163208

Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*(6), 381–395. https://doi.org/10.1145/358669.358692

Forest, S. U. (2012). *Future of America's forest and rangelands: forest service 2010 resources planning act assessment.*

García, O. (2005). Comparing and combining stem analysis and permanent sample plot data in site index models. *Forest Science*, *51*(4), 277–283.

Gopalakrishnan, R., Kauffman, J., Fagan, M., Coulston, J., Thomas, V., Wynne, R., Fox, T., & Quirino, V. (2019). Creating Landscape-Scale Site Index Maps for the Southeastern US Is Possible with Airborne LiDAR and Landsat Imagery. *Forests*, *10*(3), 234. https://doi.org/10.3390/f10030234

Hall, F. G., Hilker, T., & Coops, N. C. (2011). PHOTOSYNSAT, photosynthesis from space: Theoretical foundations of a satellite concept and validation from tower and

spaceborne data. *Remote Sensing of Environment*, *115*(8), 1918–1925. https://doi.org/10.1016/j.rse.2011.03.014

Hennigar, C., Weiskittel, A., Allen, H. L., & MacLean, D. A. (2017). Development and evaluation of a biomass increment based index for site productivity. *Canadian Journal of Forest Research*, *47*(3), 400–410. https://doi.org/10.1139/cjfr-2016-0330

Hintze, J. L., & Nelson, R. D. (1998). Violin Plots: A Box Plot-Density Trace Synergism. *The American Statistician*, *52*(2), 181–184. https://doi.org/10.1080/00031305.1998.10480559

Housman, I., Campbell, L., Goetz, W., Finco, M., Pugh, N., & Megown, K. (2021). *US Forest Service landscape change monitoring system methods.*

Ihaka, R., & Gentleman, R. (1996). R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics*, *5*(3), 299–314. https://doi.org/10.1080/10618600.1996.10474713

Kennedy, R. E., Yang, Z., & Cohen, W. B. (2010). Detecting trends in forest disturbance and recovery using yearly Landsat time series: 1. LandTrendr — Temporal segmentation algorithms. *Remote Sensing of Environment*, *114*(12), 2897–2910. https://doi.org/10.1016/j.rse.2010.07.008

LD De Laillevault. (1802). *Treatise on the management and restoration of the woods and forests of France*. books.google.com.

Liu, A., Cheng, X., & Chen, Z. (2021). Performance evaluation of GEDI and ICESat-2 laser altimeter data for terrain and canopy height retrievals. *Remote Sensing of Environment*, *264*, 112571. https://doi.org/10.1016/j.rse.2021.112571

Li, W., Niu, Z., Shang, R., Qin, Y., Wang, L., & Chen, H. (2020). High-resolution mapping of forest canopy height using machine learning by coupling ICESat-2 LiDAR with Sentinel-1, Sentinel-2 and Landsat-8 data. *International Journal of Applied Earth Observation and Geoinformation*, *92*, 102163. https://doi.org/10.1016/j.jag.2020.102163

Malambo, L., & Popescu, S. C. (2021). Assessing the agreement of ICESat-2 terrain and canopy height with airborne lidar over US ecozones. *Remote Sensing of Environment*, *266*, 112711. https://doi.org/10.1016/j.rse.2021.112711

Markus, T., Neumann, T., Martino, A., Abdalati, W., Brunt, K., Csatho, B., Farrell, S., Fricker, H., Gardner, A., Harding, D., Jasinski, M., Kwok, R., Magruder, L., Lubin, D., Luthcke, S., Morison, J., Nelson, R., Neuenschwander, A., Palm, S., … Zwally, J. (2017). The Ice, Cloud, and land Elevation Satellite-2 (ICESat-2): Science requirements, concept, and implementation. *Remote Sensing of Environment*, *190*, 260–273. https://doi.org/10.1016/j.rse.2016.12.029

Neuenschwander, A, Pitts, K., Jelley, B., Robbins, J., Klotz, B., Popescu, S. C., Nelson, R. F., & Harding, D. (2019). Ice, Cloud, and Land Elevation Satellite 2 (ICESat-2) algorithm theoretical basis document (ATBD) for land-vegetation along-track products (ATL08). *National Aeronautics and Space Administration: Washington, DC, USA*.

Neuenschwander, Amy, Guenther, E., White, J. C., Duncanson, L., & Montesano, P. (2020). Validation of ICESat-2 terrain and canopy heights in boreal forests. *Remote Sensing of Environment*, *251*, 112110. https://doi.org/10.1016/j.rse.2020.112110

Neuenschwander, Amy, & Pitts, K. (2019). The ATL08 land and vegetation product for the ICESat-2 Mission. *Remote Sensing of Environment*, *221*, 247–259. https://doi.org/10.1016/j.rse.2018.11.005

Neumann, T. A., Martino, A. J., Markus, T., Bae, S., Bock, M. R., Brenner, A. C., Brunt, K. M., Cavanaugh, J., Fernandes, S. T., Hancock, D. W., Harbeck, K., Lee, J., Kurtz, N. T., Luers, P. J., Luthcke, S. B., Magruder, L., Pennington, T. A., Ramos-Izquierdo, L., Rebold, T., … Thomas, T. C. (2019). The Ice, Cloud, and Land Elevation Satellite - 2 Mission: A Global Geolocated Photon Product Derived From the Advanced Topographic Laser Altimeter System. *Remote Sensing of Environment*, *233*. https://doi.org/10.1016/j.rse.2019.111325

Oliver, M. A., & Webster, R. (1990). Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information Systems*, *4*(3), 313–332. https://doi.org/10.1080/02693799008941549

Olofsson, K., Holmgren, J., & Olsson, H. (2014). Tree Stem and Height Measurements using Terrestrial Laser Scanning and the RANSAC Algorithm. *Remote Sensing*, *6*(5), 4323–4344. https://doi.org/10.3390/rs6054323

Packalén, P., Mehtätalo, L., & Maltamo, M. (2011). ALS-based estimation of plot volume and site index in a eucalyptus plantation with a nonlinear mixed-effect model that accounts for the clone effect. *Annals of Forest Science*, *68*(6), 1085–1092. https://doi.org/10.1007/s13595-011-0124-9

Popescu, S. C., Zhou, T., Nelson, R., Neuenschwander, A., Sheridan, R., Narine, L., & Walsh, K. M. (2018). Photon counting LiDAR: An adaptive ground and canopy height retrieval algorithm for ICESat-2 data. *Remote Sensing of Environment*, *208*, 154–170. https://doi.org/10.1016/j.rse.2018.02.019

Sabatia, C. O., & Burkhart, H. E. (2014). Predicting site index of plantation loblolly pine from biophysical variables. *Forest Ecology and Management*, *326*, 142–156. https://doi.org/10.1016/j.foreco.2014.04.019

Sahoo, K., Bergman, R., Alanya-Rosenbaum, S., Gu, H., & Liang, S. (2019). Life Cycle Assessment of Forest-Based Products: A Review. *Sustainability*, *11*(17), 4722. https://doi.org/10.3390/su11174722

Schumacher, F. X. (1939). A new growth curve and its application to timber yield studies. *Journal of Forestry*.

Su, Y., Ma, Q., & Guo, Q. (2017). Fine-resolution forest tree height estimation across the Sierra Nevada through the integration of spaceborne LiDAR, airborne LiDAR, and optical imagery. *International Journal of Digital Earth*, *10*(3), 307–323. https://doi.org/10.1080/17538947.2016.1227380

Thomas, V. A., Wynne, R. H., Kauffman, J., McCurdy, W., Brooks, E. B., Thomas, R. Q., & Rakestraw, J. (2021). Mapping thins to identify active forest management in

southern pine plantations using Landsat time series stacks. *Remote Sensing of Environment*, *252*, 112127. https://doi.org/10.1016/j.rse.2020.112127

Tompalski, P., Coops, N. C., White, J. C., Wulder, M. A., & Pickell, P. D. (2015). Estimating forest site productivity using airborne laser scanning data and landsat time series. *Canadian Journal of Remote Sensing*, *41*(3), 232–245. https://doi.org/10.1080/07038992.2015.1068686

Turner, D. P., Koerper, G. J., Harmon, M. E., & Lee, J. J. (1995). A carbon budget for forests of the conterminous united states. *Ecological Applications*, *5*(2), 421–436. https://doi.org/10.2307/1942033

Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual:(Python Documentation Manual Part 2). *Scotts Valley, CA: CreateSpace*.

Watson, R., Noble, I., Bolin, B., & Ravindranath, N. H. (n.d.). *Summary to policy makers: Land use, land-use change, and forestry*.

Wulder, M. A., White, J. C., Stinson, G., Hilker, T., Kurz, W. A., Coops, N. C., St-Onge, B., & Trofymow, J. A. T. (2010). Implications of differing input data sources and approaches upon forest carbon stock estimation. *Environmental Monitoring and Assessment*, *166*(1–4), 543–561. https://doi.org/10.1007/s10661-009-1022-6

Zhu, Z., & Woodcock, C. E. (2014). Continuous change detection and classification of land cover using all available Landsat data. *Remote Sensing of Environment*, *144*, 152–171. https://doi.org/10.1016/j.rse.2014.01.011

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. *Journal of Geodynamics*, *34*(3–4), 405–445. https://doi.org/10.1016/S0264-3707(02)00042-X

USGS 2023. https://www.usgs.gov/3d-elevation-program/topographic-data-quality-levels-qls

Appendix A: Python code to process the ICESat-2 HDF5 files into csv files

```
#import necessary libraries
import os
import icepyx as ipx
import numpy as np
import pandas as pd
import h5py
import os,json
from glob import glob
os.chdir('Z:/Users/sonia/icesatdata/smallfiles/f')   #change directory
files=glob('*.h5')   #list the files having extension .h5
for f in files:
    group = [ '/gt1l','/gt2l','/gt3l,'/gt1r','/gt2r','/gt3r'  ]  #extracting each beam and the
ground track
    for g in (group):   #iterating through the group
      with h5py.File (f, 'r') as fi:     #reading the h5 files
         #extracting the values of the files through the each beam and ground track
          lat = fi[g+'/land_segments/latitude'][:]
          lon = fi[g+'/land_segments/longitude'][:]
          canopy_h_metrics = fi[g+'/land_segments/canopy/canopy_h_metrics'][:]
          canopy_openness = fi[g+'/land_segments/canopy/canopy_openness'][:]
          h_canopy = fi[g+'/land_segments/canopy/h_canopy'][:]
          h_mean_canopy=fi[g+'/land_segments/canopy/h_mean_canopy'][:]
```

```
    h_dif_canopy=fi[g+'/land_segments/canopy/h_dif_canopy'][:]
    h_min_canopy=fi[g+'/land_segments/canopy/h_min_canopy'][:]
    h_max_canopy=fi[g+'/land_segments/canopy/h_max_canopy'][:]
    toc_roughness=fi[g+'/land_segments/canopy/toc_roughness'][:]
    h_canopy_quad=fi[g+'/land_segments/canopy/h_canopy_quad'][:]
    n_ca_photons=fi[g+'/land_segments/canopy/n_ca_photons'][:]
    n_toc_photons=fi[g+'/land_segments/canopy/n_toc_photons'][:]
    centroid_height=fi[g+'/land_segments/canopy/centroid_height'][:]
    canopy_flag=fi[g+'/land_segments/canopy/canopy_flag'][:]
    landsat_flag=fi[g+'/land_segments/canopy/landsat_flag'][:]
ofile=f.replace('.h5','_'+g[1:]+'.h5') #making h5 file of extracted values
ofilecsv = f.replace('.h5', '_'+g[1:]+'.csv') #making csv file of h5 files
result = pd.DataFrame() #creating dataframe
#assigning values to dataframe
result['lon'] = lon
result['lat'] = lat
result['canopy_h_metrics_0'] = canopy_h_metrics[:,0]
result['canopy_h_metrics_1'] = canopy_h_metrics[:,1]
result['canopy_h_metrics_2'] = canopy_h_metrics[:,2]
result['canopy_h_metrics_3'] = canopy_h_metrics[:,3]
result['canopy_h_metrics_4'] = canopy_h_metrics[:,4]
result['canopy_h_metrics_5'] = canopy_h_metrics[:,5]
result['canopy_h_metrics_6'] = canopy_h_metrics[:,6]
```

```
result['canopy_h_metrics_7'] = canopy_h_metrics[:,7]
result['canopy_h_metrics_8'] = canopy_h_metrics[:,8]
result['canopy_h_metrics_9'] = canopy_h_metrics[:,9]
result['canopy_h_metrics_10'] = canopy_h_metrics[:,10]
result['canopy_h_metrics_11'] = canopy_h_metrics[:,11]
result['canopy_h_metrics_12'] = canopy_h_metrics[:,12]
result['canopy_h_metrics_13'] = canopy_h_metrics[:,13]
result['canopy_h_metrics_14'] = canopy_h_metrics[:,14]
result['canopy_h_metrics_15'] = canopy_h_metrics[:,15]
result['canopy_h_metrics_16'] = canopy_h_metrics[:,16]
result['canopy_h_metrics_17'] = canopy_h_metrics[:,17]
result['canopy_openness'] = canopy_openness
result['h_canopy_quad'] = h_canopy_quad
result['h_canopy']= h_canopy
result['h_mean_canopy'] = h_mean_canopy
result['h_dif_canopy'] = h_dif_canopy
result['h_min_canopy'] = h_min_canopy
result['h_max_canopy'] = h_max_canopy
result['toc_roughness']= toc_roughness
result['n_ca_photons'] = n_ca_photons
result['n_toc_photons'] = n_toc_photons
result['centroid_height']= centroid_height
result['canopy_flag']= canopy_flag
```

```
        result['landsat_flag'] = landsat_flag
       result.to_csv(ofilecsv,index=None)
```

Appendix B: R code to intersect ICESat-2 tracks with LCMS and Plantations map

```
library(rgdal)
library(sp)
library(maptools)
library(raster)
library(ggplot2)
library(dplyr)
library(tidyverse)  #library to read the multiple files
fnames <- list.files("C:\\Users\\CNRCommonAccess\\Desktop\\PLANT_CSV", pattern =
"*.csv") #list file names
fnames
#loop through the files
for (f in fnames) {
  cordinates <- read.csv(f)
  str(cordinates)
  Lcms <- raster(paste0("Plantation_reproj.tif"))   #read raster file
  Lcms@crs
  cordinates_spdf <- SpatialPointsDataFrame(                #convert to spatial dataframe
    cordinates[,1:2], proj4string=Lcms@crs, cordinates)
  #chnage cordiante syetem as raster
```

```
point_extract <- raster::extract(Lcms,            # raster layer
                                 cordinates_spdf,      # SPDF
                                 method = 'simple',    #method
                                 fun = NULL,
                                 buffer = NULL,        # buffer size
                                 df=TRUE)
point_extract
merge_file <- cbind (cordinates, point_extract)   #column bind of the two data frame
no_na <- na.omit(merge_file)          #omitting the na values
write.csv(no_na, gsub(".csv","_Plantation.csv",f))   # write csv in different folder
```

Appendix C : R code to process ALS laz files into canopy height models

```
library(raster)
library(lidR)
#####Set working directory and other directory information #########
wd <- "/C:/3dep_lidar/"
setwd(wd)
filepath=wd
outpath="/3dep_lidar_output/"
###### Get a list of the lidar files#################
filenames <- list.files(path <- filepath, pattern="*.laz", full.names=TRUE)
```

```
#Read lidar and process lidar data #########
##need to loop through the filenames, using "i" as the counter
tot<-length(filenames)-1
for (i in tot){
  ## read the lidar data
  laz_2019 <-readLAS(filenames[i],select="*",filter="")

   ############### Look at data specs for gut check (not in loop) ######
   #check out the extent, coordinate system
  summary(laz_2019)

   #######plot the lidar point cloud
   #plot(laz_2019)
   #######Generate DTM and normalized points
  dtm <- rasterize_terrain(laz_2019,res= 1,algorithm=tin()) #the res has to be smaller if
in DD
  # plot_dtm3d(dtm)
  nlas<-laz_2019-dtm
  # plot(nlas, size=4, bg="white")

  #############Generate Canopy Height Model
  #chm <- rasterize_canopy(nlas,res=1,algorithm=p2r(),)
```

```
  chm <- rasterize_canopy(nlas,res=1,algorithm=p2r())
  #plot_dtm3d(chm)

  #########ASSIGN projection and write raster
  ##Note. Since they don's seem to write the projection where lidR expects it in the laz
files from 3DEP, we might just need to see the projection from metadata
  #Use the same naming conventions so we can link back to original tile
  lidname1 <- strsplit(filenames[i],"/")
  lidname2 <- strsplit(lidname1[[1]][7],".laz")
  outdtm <- paste(outpath,lidname2,"_dtm.tif")
  outchm <- paste(outpath,lidname2,"_chm.tif")
  # Set the projection. You can use epsg codes
  crs(chm)<- "epsg:4326"
  crs(dtm)<- "epsg:4326"

  #Save both the dtm and chm to tiff files in outdir
  x <- writeRaster(chm, outchm, overwrite=TRUE)
  x1<- writeRaster(dtm, outdtm, overwrite=TRUE)
}
```

Appendix D: Python code to extract 98th percentile from canopy height models

```
#import necessary libraries
from osgeo import gdal
```

```
import pandas as pd
import numpy as np

# Open the raster dataset
raster_path = "projected.tif"
dataset = gdal.Open(raster_path)

# Define the dimensions of the bounding box in pixels
bbox_width_pixels = 100
bbox_height_pixels = 13

# Function to calculate the 98th percentile height within a bounding box
def calculate_98th_percentile(latitude, longitude):
    geo_transform = dataset.GetGeoTransform()
    x_pixel = int((longitude - geo_transform[0]) / geo_transform[1])
    y_pixel = int((latitude - geo_transform[3]) / geo_transform[5])

    # Read the pixels within the bounding box
    band = dataset.GetRasterBand(1)  # Assuming a single-band raster
    pixels = band.ReadAsArray(x_pixel, y_pixel, bbox_width_pixels, bbox_height_pixels)

    # Calculate the 98th percentile height
    percentile_98 = np.percentile(pixels, 98)
```

```
    return percentile_98

# Read the CSV file containing coordinates
csv_path = "nlcd_extract.csv"
coordinates_df = pd.read_csv(csv_path)

# Create a list to store 98th percentile values
percentile_98_list = []

# Iterate through each row in the CSV and calculate 98th percentile height
for index, row in coordinates_df.iterrows():
    latitude = row['lat']
    longitude = row['lon']
    percentile_98 = calculate_98th_percentile(latitude, longitude)

    percentile_98_list.append(percentile_98)

# Create a DataFrame from the percentile values
percentile_df = pd.DataFrame({'percentile_98': percentile_98_list})

# Save the DataFrame to a new CSV file
percentile_csv_path = "percentile_98_results.csv"
```

```
percentile_df.to_csv(percentile_csv_path, index=False)
```

Appendix E: Python code to employ a guided curve equation and have a site index curve using RANSAC algorithm

```
#import necessary libraries
import pandas as pd
import numpy as np
from sklearn.linear_model import RANSACRegressor, LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm

# Load the data from your CSV file
data = pd.read_csv('active_man_60.csv')

# Set a random seed for reproducibility
np.random.seed(23)

# Select a random sample of the data
sample_size = 25000  # Set the desired sample size
data_sample = data.sample(n=sample_size)
```

```
# Fit the RANSAC model
X = 1 / data_sample['age'].values.reshape(-1, 1)
y = np.log(data_sample['h_canopy_x']).values.reshape(-1, 1)

ransac = RANSACRegressor()
ransac.fit(X, y)

# Obtain the inliers and their corresponding X and y values
inliers_mask = ransac.inlier_mask_
X_inliers = X[inliers_mask]
y_inliers = y[inliers_mask]

# Fit a linear regression model to the inliers
linear_model = LinearRegression()
linear_model.fit(X_inliers, y_inliers)

# Obtain the slope and intercept for inliers
slope_inliers = linear_model.coef_[0][0]
intercept_inliers = linear_model.intercept_[0]
print('Slope for inliers:', slope_inliers)
print('Intercept for inliers:', intercept_inliers)
# Calculate the R-squared for inliers
y_pred_inliers = linear_model.predict(X_inliers)
```

```
r2_inliers = r2_score(y_inliers, y_pred_inliers)
print('R-squared for inliers:', r2_inliers)

# Calculate the RMSE for inliers
rmse_inliers = np.sqrt(mean_squared_error(y_inliers, y_pred_inliers))
print('RMSE for inliers:', rmse_inliers)

# Obtain p-value for the linear regression model
X_inliers_with_intercept = sm.add_constant(X_inliers)
ols_model = sm.OLS(y_inliers, X_inliers_with_intercept).fit()
p_value = ols_model.pvalues[1]  # Assuming age is the independent variable (index 1)
print('P-value for the linear regression model:', p_value)

# Calculate the predicted heights for all data points
age_range = np.linspace(data['age'].min(), data['age'].max(), num=100)
X_all = 1 / age_range.reshape(-1, 1)
h_canopy_x_pred = np.exp(slope_inliers * X_all + intercept_inliers)

# Obtain inliers and outliers
inliers = data_sample[inliers_mask]
outliers = data_sample[~inliers_mask]
inliers.to_csv("inliers.csv")
```

```
# Set the white grid style
sns.set_style('whitegrid')

# Plot the scatter plot of inliers and outliers
plt.figure(figsize=(12, 8))
sns.scatterplot(data=inliers, x='age', y='h_canopy_x', color='blue')
sns.scatterplot(data=outliers, x='age', y='h_canopy_x', color='red')

# Plot the trend line
plt.plot(age_range, h_canopy_x_pred, color='green', linewidth=2, label='Site index
curve')

plt.xlabel('Age')
plt.ylabel('Height')
plt.legend()
plt.show()
```