A Comparison of Machine Learning and Geostatistical Approaches for Mapping Forest Canopy Height over the Southeastern US Using ICESat-2

Kasip Tiwari * and Lana L. Narine

Abstract: The availability of canopy height information in the Ice, Cloud, and Land Elevation Satellite-2’s (ICESat-2’s) land and vegetation product, or ATL08, presents opportunities for developing full-coverage products over broad spatial scales. The primary goal of this study was to develop a 30-meter canopy height map over the southeastern US, for the Southeastern Plains ecoregion and the Middle Atlantic Coastal Plains ecoregion. More specifically, this work served to compare well-known modeling approaches for upscaling canopy information from ATL08 to develop a wall-to-wall product. Focusing on only strong beams from nighttime acquisitions, the h_canopy parameter was extracted from ATL08 data. Landsat-8 bands and derived vegetation indices (normalized difference vegetation index, enhanced vegetation index, and modified soil-adjusted vegetation index) along with National Land Cover Database’s canopy cover and digital elevation models were used to extrapolate ICESat-2 canopy height from tracks to the regional level. Two different modeling techniques, random forest (RF) and regression kriging (RK), were applied for estimating canopy height. The RF model estimated canopy height with a coefficient of determination ($R^2$) value of 0.48, root-mean-square error (RMSE) of 4.58 m, mean absolute error (MAE) of 3.47 and bias of 0.23 for independent validation, and an $R^2$ value of 0.38, RMSE of 6.39 m, MAE of 5.04 and bias of $-1.39$ when compared with airborne lidar-derived canopy heights. The RK model estimated canopy heights with an $R^2$ value of 0.69, RMSE of 3.49 m, MAE of 2.61 and bias of 0.03 for independent validation, and an $R$ value of 0.68, $R^2$ value of 0.47, RMSE of 5.96 m, MAE of 4.52 and bias of $-1.81$ when compared with airborne lidar-derived canopy heights. The results suggest feasibility for the implementation of the RK method over a larger spatial extent and potential for combining other remote sensing and satellite data for future monitoring of canopy height dynamics.

Keywords: canopy height; ecoregion; ICESat-2; lidar; Landsat; random forest; regression kriging

1. Introduction

Forest ecosystems covering 30% of the land surface are important because they sequester atmospheric carbon dioxide, which helps in the mitigation of global warming [1,2] and serves as a reservoir of biodiversity [3]. The southeastern United States (US) landscape comprises 60% of the forested areas [4] and these forests are key carbon sinks and are part of a global-warming-mitigation strategy. These forests play an important role in sequestering about 10% of U.S. anthropogenic carbon inputs to the atmosphere in biomass and soils [5]. Additionally, southeastern forests have been providing a steady supply of timber and fiber. Despite climatic and land-use changes, forests in the southeast United States are expected to continue to serve as a carbon sink [6]. Canopy height is the main parameter to indicate forest carbon content [7] and is also important for studying the carbon cycle and biodiversity [8]. The spatial distribution of canopy height is essential for the estimation of carbon sources and sinks [9] and for the sustainable management of ecosystems [10]. Canopy height is an important ecological parameter and a spatially explicit representation...
of this parameter can be useful for understanding and modeling a wide range of ecological, hydrological, and atmospheric processes in forest ecosystems [11]. Forest and land managers can use comprehensive canopy height maps to assess habitat suitability for a variety of species, assess fire threats, predict potential fire behavior, and estimate timber volume [11].

Although, field-based forest inventories can provide accurate estimates of canopy height, it is very difficult to obtain this information over regional scales because of limited field measurements in remote areas [12] and time and cost constraints [9]. Remote sensing (RS) is considered a practical tool, as we can obtain data at multiple spatial scales and some RS systems also provide repeated observations [13]. Light detection and ranging (lidar) is an active remote sensing method based on laser ranging, which calculates the distance based on the time for the pulse to return to the lidar sensor from the target [14,15]. Lidar has been demonstrated as a robust tool for estimating canopy cover and height because it can penetrate the forest canopy [16].

Lidar data may be acquired from several platforms, primarily terrestrial, airborne and spaceborne [17]. Terrestrial lidar provides complete and accurate information ($R^2 = 0.48–0.99$, for different forest parameters) of forest structure with several scans and a high scanning resolution, but may only cover limited areas [18,19]. Airborne lidar also provides highly accurate estimates of forest vegetation structure and canopy height. To exemplify, pit-free canopy height models of the Bois Noir Forest, Barcelonette basin were generated from airborne lidar, which improved the accuracy of tree detection [20], and airborne lidar was used to describe and predict forest structural attributes such as canopy height and aboveground biomass [21]. However, the high-volume data and the high operation costs hinder the use of airborne lidar for regional- and global-scale studies [22,23]. The necessity to analyze the vertical structure at a global scale prompted the development of spaceborne lidar, which is placed on satellites and delivers unique near-global- or global-scale information on a range of atmospheric and terrestrial features [24]. Spaceborne lidar benefits from the satellite platform’s high orbit and wide-field perspective, making it a feasible solution for gathering broad-scale canopy height data [25]. For example, a global canopy height map and global forest aboveground biomass map of 1 km was prepared from spaceborne lidar, i.e., ICESat (Ice, Cloud, and Land Elevation Satellite) data using random forest (RF) modeling [26,27].

ICESat-2 is a follow-up mission to ICESat, which operated from 2003 to 2009. The National Aeronautics and Space Administration (NASA) launched ICESat-2 in September 2018 with the primary mission objective of observing changes in the cryosphere [28]. The Advanced Topographic Laser Altimeter System (ATLAS) instrument on-board ICESat-2 utilizes photon-counting technology for altimetry observations. A diffractive optical component of ATLAS splits the single outgoing laser beam into three pairs of beams that are spaced roughly 3.3 km apart on the surface at a laser repetition rate of 10 kHz. Each pair consists of one strong and one weak beam. Strong beams are high-energy beams that are useful for detecting surfaces with poor reflectance, such as vegetation-covered terrain [28], 2019). ATLAS has an 11-meter footprint, allowing it to produce a variety of geophysical products over various surface types, i.e., land ice (ATL06), sea ice (ATL07), land/vegetation (ATL08), atmosphere (ATL09), oceans (ATL12), and inland water (ATL13) [28].

Since ICESat-2 data are obtained along transects at discrete intervals, wall-to-wall products, such as satellite imagery and ancillary data, may be examined in conjunction with ICESat-2 data to obtain full-coverage information [29]. Undoubtedly, spaceborne lidar remote sensing offers a cutting-edge and exciting prospect for the predictive mapping of canopy height. Several studies have estimated and mapped canopy height using spaceborne lidar. For instance, a global canopy height map at a 1 km resolution was developed from ICESat’s Geoscience Laser Altimeter System (GLAS) data using the random forest regression tree method, which showed a reasonable correspondence with 66 FLUXNET sites ($\text{RMSE} = 6.1 \text{ m, } R^2 = 0.49$; $\text{RMSE} = 4.4 \text{ m and } R^2 = 0.69$ without outliers) [26].

Global mapping of mean forest canopy height with a spatial resolution of 500 m was per-
formed by combining GLAS data with MODIS by applying a balanced random forest (BRF) algorithm and validating the map with field survey data and showed $R^2 = 0.63$ and RMSE = 4.68 m [30]. Canopy height maps of Maryland were produced at a 30 m resolution by combining GLAS and Landsat spectral images using principal component analysis (PCA) and machine-learning models (BPANN, SVR, and RF) [23]. Currently, ICESat-2 data may be applied to generate canopy height maps using synergistic approaches with satellite imagery. High-resolution (250 m and 30 m) forest canopy height mapping was achieved by combining ICESat-2 lidar data with satellite imagery using machine-learning models [29].

RF has been extensively used for mapping canopy height. ICESat-2 canopy height, Sentinel-1 backscatter, and Sentinel-1 texture values were combined to map the forest canopy height in Doon Valley, Uttarakhand, India using RF, which was further used to map the spatial distribution of AGB [31]. Similarly, the RF algorithm was used to estimate the grassland height in the Tibetan Plateau [32]. RF performed better than the gradient boosting decision tree when mapping canopy height combining ICESat-2, Sentinel-1 and Sentinel-2 data in Hua’nan, China [33]. Additionally, RF has been used in extrapolating ICESat-2-derived AGB estimates to develop an AGB map with a 30 m spatial resolution using vegetation indices, land cover and canopy cover [34].

Regression kriging (RK), i.e., ordinary cokriging of the residuals of canopy height from ordinary least-square regression, was found to be the best method for canopy height estimation and mapping by combining the airborne lidar (Aeroscan) and Landsat ETM+ in western Oregon [35]. A canopy height map of French Guiana, which is situated on the northern coast of South America, was created using the ICESat–GLAS dataset and the results showed that employing RK instead of RF improved the RMSE from 6.5 to 4.2 m [36]. The RK model also produced higher accuracy for estimating lidar-derived canopy cover and AGB using SPOT-6 variables [37]. Additionally, for farmers and agricultural service providers interested in precision agriculture, RK has been demonstrated as a reliable method for combining UAV pictures with data from ground-based sensors [38].

The overall goal of this study was to develop a regional-scale canopy height product at a 30 m spatial resolution in the southeastern US for the: (i) Southeastern Plains ecoregion and (ii) Middle Atlantic Coastal Plains ecoregion. By comparing two different modeling approaches for upscaling canopy information, i.e., random forest (RF) and regression kriging (RK), this work serves to define a methodology for developing a wall-to-wall map using ICESat-2’s vegetation product, ATL08. There are no studies regarding canopy height mapping in the southeastern US at the ecoregion level from ICESat-2. This study provides comprehensive canopy height information over two ecoregions in the southeastern US. The outputs, consisting of gridded canopy height and a suitable upscaling algorithm, are intended to contribute to biomass estimation and carbon mapping with ICESat-2, and a basis on which to study vegetation changes in the region.

2. Materials and Methods

2.1. Study Area

The study was carried out in two ecoregions: (1) the Southeastern Plains ecoregion and (2) the Middle Atlantic Coastal Plains ecoregion (Figure 1). The Southeastern Plains ecoregion encompasses much of the coastal plains of the southeastern US, stretching from Maryland, southwest to Mississippi, with small portions extending to Louisiana and Tennessee. This ecoregion has a mild, mid-latitude, humid subtropical climate, i.e., hot, humid summers and mild winters. The mean annual temperature ranges from 13 °C in the north to 19 °C in the south. The annual precipitation ranges from 1140 to 1520 mm with a mean annual precipitation of 1358 mm. The precipitation is evenly distributed throughout the year. Much of the area has pine forests, with the longleaf pine (*Pinus palustris*) a dominant species in much of the area, and the loblolly pine (*Pinus taeda*) extending farther north. These areas also consist of mixed oaks–hickory–pine forests throughout the region. The southern part of this region has mixed forests, with a mixture of broadleaf evergreens, deciduous, and pines [39,40]. The Middle Atlantic Coastal Plains ecoregion encompasses
the outer coastal plain, from New Jersey to the South Carolina/Georgia border. This ecoregion has a mild, mid-latitude, humid subtropical climate, i.e., hot, humid summers and mild winters. The mean annual temperature ranges from 14 °C in the north to 17 °C in the south. The annual precipitation ranges from 1020 to 1420 mm with a mean annual precipitation of 1229 mm. The northern part mostly consists of the loblolly pine, shortleaf pine, oak, sweetgum, and cypress, whereas the southern part consists of live oak, sand laurel oak, and loblolly pine [40].

![Figure 1. Study area in southeastern US (A). Southeastern Plains ecoregion (B). Middle Atlantic Coastal Plains ecoregion (C). Map coordinate system is WGS 1984. World Imagery is used as basemap in (B,C).](image)

2.2. Data
2.2.1. ICESat-2 ATL08 Data

The ICESat-2 mission’s ATL08 data product provides estimates of terrain height and canopy height along 100 m segments in the along-track direction [41]. The ATL08 data product, Version 004 (released on 13 April 2021), was downloaded from the National Snow and Ice Data Center in the HDF5 data format [41]. The ATL08 data product consists of ground track (GT) groups; each GT contains subgroups for land and canopy heights as well as for beam and reference parameters. Photons are classified as ground, canopy, or noise in the ATL08 method, with ground photons determining the interpolated ground surface, and canopy height is computed by differencing the height between canopy photons and the interpolated ground surface [41]. The “h_canopy” parameter, i.e., 98% height of all the individual relative canopy height for the segment, is recommended for ICESat-2 as a measurement of the top canopy height [41], and was therefore extracted from the ATL08 data product. Green energy (500–600 nm) from solar radiation is strongest during the day, and given ICESat-2’s transmitted green laser pulses (~532 nm), there is higher background noise for day acquisitions than for night acquisitions [42]. While estimating forest height, studies have indicated that daytime acquisitions and weak-beam data perform worse than strong beams and nighttime data [43]. Hence, only strong beams of the night data were used. Additionally, version 004 performs better than older versions at high vegetation density, contributing to more accurate canopy estimates [41].
2.2.2. Landsat-8 Data

NASA and the United States Geological Survey (USGS) started the Landsat Data Continuity Mission (LCDM) on 11 February 2013 to guarantee the continuity of the unmatched Landsat record. USGS took over control of the mission’s activities on 30 May 2013, renamed it Landsat-8, and made the data available to people globally [44]. The Landsat-8 observatory is designed in a Sun-synchronous orbit for a 705 kilometer with a 16-day repeat cycle, completely orbiting Earth every 98.9 minutes [45]. Landsat-8 has two sensors, i.e., Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS). The OLI sensor collects image data for 9 shortwave spectral bands (Band1–Band 9) over a 190 km swath with a 30-meter (m) spatial resolution for all bands except the 15 m Pan band (Band-8). The TIRS sensor collects image data for two thermal bands (Band10 and Band11) with a 100 m spatial resolution over a 190 km swath [45]. The main objectives of Landsat-8 are to provide continuous data with Landsat 4, 5, and 7, to provide 16-day repetitive Earth coverage, and to build and periodically refresh a global archive of Sun-lit and mostly cloud-free land images [45].

Landsat-8 images acquired in the year 2020 were downloaded from USGS Earth Explorer. A total of 34 Landsat scenes with cloud cover 0–2% were used for this study. Six spectral bands at 30 m resolution (Band2—Blue, Band3—Green, Band4—Red, Band5—Near Infrared (NIR), Band6—Shortwave Infrared1 (SWIR1), Band7—Shortwave Infrared2 (SWIR2)) and three vegetation indices (Normalized Difference Vegetation Index (NDVI), Modified Soil Adjusted Vegetation Index (MSAVI) and Enhanced Vegetation Index (EVI)) were used in the study to map wall-to-wall forest canopy height. Various studies have demonstrated that spectral bands and vegetation indices are correlated with canopy heights [29,46,47].

The calculation of vegetation indices was performed as follows [48–50]:

\[
NDVI = \frac{(\text{Band 5} - \text{Band 4})}{(\text{Band 5} + \text{Band 4})}
\]

\[
MSAVI = \frac{(2 \times \text{Band 5} + 1 - \sqrt{(2 \times \text{Band 5} + 1)^2 - 8 \times (\text{Band 5} - \text{Band 4})})}{2}
\]

\[
EVI = 2.5 \times \frac{(\text{Band 5} - \text{Band 4})}{(\text{Band 5} + 6 \times \text{Band 4} - 7.5 \times \text{Band 2} + 1)}
\]

These Landsat-8-derived vegetation indices were obtained on-demand via the USGS Earth Resources Observation and Science (EROS) Center Science Processing Architecture (ESPA) On Demand Interface.

2.2.3. Ancillary Data

i. Land Cover data:

The National Land Cover Database (NLCD) is a real-time land-cover-monitoring project that, between 2001 and 2019, offered spatially specific and trustworthy data on the land cover and changes of the United States at intervals of two to three years [51,52].

The landcover map (30 m spatial resolution) of NLCD 2019 was downloaded to mask the forested areas. The forested areas include deciduous forest, evergreen forest, mixed forest, shrub/scrub, and woody wetland classes.

ii. Canopy Cover (CC) data:

The tree canopy cover product from the U.S. Forest Service is made up of a single percent tree canopy cover layer with values ranging from 0 to 100 percent, with each value reflecting the percentage of that 30 m cell that is covered by trees [53]. The NLCD 2016 canopy cover map at the 30 m grid size was downloaded and used as an independent variable for modeling.

iii. Digital Elevation Model (DEM):

The digital representation of the land surface elevation with respect to any reference datum is known as a DEM, which is used to determine terrain attributes such as elevation, slope, and aspect [54]. The DEM at a spatial resolution of one arc second (~30 m), down-
loaded from the USGS Earth Explorer, was resampled to a 30 m resolution and snapped to the pixels of Landsat. The DEM was used as a predictor for canopy height modeling.

2.2.4. Airborne Lidar Data

In order to offer uniform, standardized national lidar coverage, the USGS launched the 3D Elevation Program (3DEP) in 2012 in collaboration with other federal and state organizations [55]. The 3DEP was launched to address the nation’s expanding need for topographic data of the highest quality as well as a wide range of other 3D representations of natural and manmade structures [56]. The 3DEP products (lidar point clouds and DEMs) are available free of charge and without usage limitations [56]. The lidar point clouds are the primary data for 3DEP, which consist of 3D information and are used for creating the DEM products [56]. A total of 20 sites were randomly selected, and point-cloud data of these sites were downloaded via OpenTopography. These point-cloud data ranged from 2007 to 2019. The CHM data derived from the point-cloud data were used as a reference canopy height for the comparison with estimated canopy height by random forest and regression kriging.

2.3. Data Processing

The shapefiles for the two ecoregions were obtained from United States Environmental Protection Agency (USEPA) website [57] and were used to define the extent of the study area. The ICESat-2 ATL08 data parameters were extracted via R-programming, focusing on only strong beams from nighttime acquisitions. The ATL08 segment parameters extracted were h_canopy, h_min_canopy, h_max_canopy, h_mean_canopy data, along with segment geolocation information (longitude and latitude). Outlier filtering was performed to remove segments above 40 m [26,58] and below 2 m [59,60] canopy height in order to retain data representative of a reasonable canopy height. Each spectral-band and vegetation-index scene was mosaicked and masked to the extent of the study area in ArcGIS Pro [61]. Predictor variables (Blue, Green, Red, NIR, SWIR1, SWIR2, NDVI, EVI, MSAVI, CC, DEM) were imported in ArcGIS Pro and the pixel value of each variable was combined with the spatially coincident ATL08 data. The canopy height metrics of ICESat-2’s ATL08 data were assigned to Landsat-8 30 m pixels based on the geographical location (latitude and longitude) of the centroid of ICESat-2’s segment [29,62]. The attribute table containing the values of the predictor variables along with the ATL08 height data was then exported for analysis.

For the airborne lidar data, point clouds (*.las) were used to generate a raster of first returns, i.e., digital surface model (DSM), and ground returns (DEM) at the spatial resolution of 1 m. Then, the DEM was subtracted from the DSM to obtain the canopy height model (CHM) and the cells with height below 2 m and above 40 m were filtered in accordance with the ICESat-2 data. The CHM was then resampled to a 30 m grid size representing 98th-percentile height within the pixel using the zonal statistics tool in ArcGIS pro. Then, the ATL08 point data were overlaid on the raster file of canopy height estimated by RF, RK and airborne-lidar-derived CHM, and the pixel values of each raster were extracted to the point shapefile using the “extract multi values to point” tool, and the statistical values were calculated.

2.4. Data Analysis

The machine-learning model, RF, and the geostatistical model, RK, were used to extrapolate canopy height to the entire study area and the model accuracy from each method was compared. The ICESat-2-derived canopy height was used as the dependent variable and the spectral bands and vegetation indices from the Landsat imagery along with canopy cover and DEM were used as independent variables. The ICESat-2 canopy height data were randomly divided into 80% for model training and 20% for independent validation [29].
2.4.1. Canopy Height Mapping Using Random Forest

The RF technique developed by Breiman [63] is being used as a reliable regression method for estimating forest attributes, including canopy height [29,64–66] and biomass [31,34,67]. The RF algorithm is based on a decision tree that has been combined with several ensemble regressions or classification trees [68]. To create a large number of different training subsets, the decision tree employs a bagging or bootstrap algorithm [69,70]. Multiple predictor variables can be included without making assumptions about their statistical distribution or covariance structure, which is a major benefit [26].

For random forest modeling, the ModelMap package in R software with the “model.build” function was used to build the model with the argument “RF”, and the “model.diagnostics” function was used to quantify the contribution of each predictor to the model [71].

2.4.2. Canopy Height Mapping by Regression Kriging

RK is a spatial-prediction technique that combines the kriging of the regression residuals with the regression value of predictor variables [72]. The residual is the difference between the reference canopy heights and the RF-estimated canopy heights. These residuals were estimated using ordinary kriging (OK) and the prediction by the regression kriging was obtained by adding the kriged residuals to the prediction of RF.

\[
\text{CH(RK)} = \text{CH(RF)} + R_k
\]

where \(\text{CH(RK)}\) is the predicted canopy height value by RK, \(\text{CH(RF)}\) is the RF canopy height estimates and \(R_k\) is the kriged residual.

OK is a commonly used geostatistical technique that uses a semi-variogram based on regionalized variables to obtain an optimal unbiased estimated surface [73]. A semi-variogram has three main parameters: nugget, range and sill. At infinite small distances, the nugget represents the spatial variance of measurement errors. The effective distance of the spatial autocorrelation is defined by the range. When the spatial distance between two sites exceeds the range value, the sill is the maximum value of the semi-variogram [74]. The OK interpolation is denoted by:

\[
R_k = \sum_{i=1}^{n} w_i \cdot r_i
\]

where \(R_k\) is the kriged residual, \(w_i\) is the weight associated with the measured residuals of canopy height and \(r_i\) is the residual at location \(i\).

The OK of the residual was performed in ArcGIS Pro using the geostatistical wizard and the final prediction surface of the canopy height was created by adding the two raster layers of the RF-estimated canopy height and the kriged residuals in the raster calculator.

2.5. Accuracy Assessment

Based on the canopy height values of test data, statistical values were calculated to measure the accuracy of the canopy height estimated from RF and RK. The statistical measures included (i). the coefficient of determination \((R^2)\), (ii). the root-mean-square error (RMSE), (iii). the mean average error (MAE), and iv. the mean bias.

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} (x_i - y_i)^2}{\sum_{i=1}^{n} (y_i - y')^2}
\]

\[
\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2}
\]
\[ \text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |x_i - y_i| \]
\[ \text{Bias} = \frac{1}{n} \sum_{i=1}^{n} (x_i - y_i) \]

where \(x\) is the canopy height estimated from random forest and regression kriging, \(y\) is the observed canopy height from the test data of the independent validation and the airborne canopy height data when comparing with the airborne data, and \(x'\) and \(y'\) are the averages of the estimated and observed values, respectively. Additionally, to measure the performance improvement, the relative improvement (RI) index was calculated.

\[ \text{RI} = \frac{\text{RMSE}(\text{RF}) - \text{RMSE}(\text{RK})}{\text{RMSE}(\text{RF})} \]  

(1)

3. Results

3.1. Canopy Height Mapping

The RF and RK techniques were used to map canopy height at a grid size of 30 m. The withheld test data (20% of data) (Figure 2) showed that RF and RK produced an \(R^2\) value of 0.48 and 0.69, respectively. The RK model had an RMSE of 3.49, which is less than the RF model’s RMSE of 4.58. Additionally, the MAE and bias of RK were 2.61 and 0.02, respectively, whereas the MAE and bias of RF were 3.47 and 0.23, respectively. It showed that the estimation error was less in the RK model as compared to the RF model. The RI index showed that the canopy height estimation of RK was improved by 23.8% compared to that of RF. According to the RF model (Figure 3), CC contributed most to the model followed by DEM and the Green band. The resulting canopy height map (Figure 4) from the RK model ranged from 0 to 49.5 m with a mean of 20.69 m and a standard deviation of 4.1 m, whereas with the RF model, the heights ranged from 4.6 to 34.2 m with a mean of 20.51 m and a standard deviation of 3.26 m. The comparison of the observed and estimated canopy height for both RF and RK are shown in histogram (Figure 5). Hence, comparing different statistical measures between estimated and observed canopy height from ICESat-2 data shows that RK estimates canopy height better than RF.

Figure 2. Scatter plot of observed and estimated canopy height at 30 m spatial resolution from the independent validation for (a) random forest and (b) regression kriging.
Figure 2. Scatter plot of observed and estimated canopy height at 30 m spatial resolution from the independent validation for (a) random forest and (b) regression kriging.

Figure 3. Importance of the predictor variables for estimating canopy height based on random forest.

Figure 4. Cont.
3.2. Comparison of Canopy Height Estimated by RF and RK with Airborne Lidar Canopy Height Data

Pixels coinciding with the ATL08 point data were used for comparing the canopy height estimated by RF and RK with airborne lidar canopy height data. Compared with the airborne CHM data (Figure 6), the RK-estimated canopy height yielded a better $R^2$ value of 0.47 than the RF model’s $R^2$ value of 0.38. The estimation error was observed to be less in the RK model as compared to the RF model, except for the bias. The RK model had an RMSE of 5.96, which was less than the RF model’s RMSE of 6.39. Additionally, the MAE and bias of RK were 4.52 and $-1.81$, respectively, whereas the MAE and bias of RF were
5.04 and −1.39, respectively. The comparison of estimated and airborne canopy height for both RF and RK are shown in histogram (Figure 7).

Figure 6. Scatterplot of airborne-lidar-derived and estimated canopy height at 30 m spatial resolution for (a) random forest and (b) regression kriging.

Figure 7. Histogram of airborne-lidar-derived and estimated canopy height at 30 m spatial resolution for (a) random forest and (b) regression kriging.

4. Discussion

In this study, a regional-scale canopy height map at a 30 m grid size was produced for the Southeastern Plains ecoregion and the Middle Atlantic Coastal Plains ecoregion within the southeastern US using ICESat-2 data integrated with Landsat-8 and ancillary data. The modeling techniques have great potential for estimating canopy height and suggest the feasibility of implementation over larger spatial extent.

Two models were used, i.e., machine learning (RF) and geostatistical (RK), to model and extrapolate canopy height across the study area. The RF model estimated the canopy height with an $R^2$ value of 0.48 and an RMSE of 4.58 m from the independent validation and an $R^2$ value of 0.38 and an RMSE of 6.39 m when compared with the airborne-lidar-derived canopy height. The RK model estimated the canopy height with an $R^2$ value of 0.69 and an RMSE of 3.49 m from the independent validation and an $R^2$ value of 0.47 and an RMSE of 5.96 m when compared with the airborne-lidar-derived canopy height. The RK model performed substantially better than the RF model in estimating canopy height, indicating the strong potential of RK for mapping forest canopy height. Fayad et al. (2016) developed a canopy height map of French Guiana, which is located on the northern coast of South America, using the ICESat–GLAS dataset. The results indicated an improvement of RF. Previous studies that focused on canopy height estimation with ICESat-2 high-resolution data, e.g., for the Southeastern Plains ecoregion and the Middle Atlantic Coastal Plains ecoregion within the southeastern US using ICESat-2 data integrated with Landsat-8 and ancillary data.
America, using the ICESat–GLAS dataset. The results indicated an improvement in terms of the RMSE from 6.5 to 4.2 m using RK as compared to RF. Since the RK model in this study is based on the RF, i.e., the prediction by the RK was obtained by adding the kriged residuals to the prediction of RF, it requires additional effort as compared to that of RF. Previous studies that focused on canopy height estimation with ICESat-2 highlighted an RMSE of 3.69 m for canopy height retrievals and an MAE of 3.2 m when compared with the airborne lidar canopy height [75], whereas this study obtained an RMSE of 5.96 m and an MAE of 4.52 m when comparing the RK-estimated canopy height with the airborne-derived canopy height. The RF model was used to upscale the ICESat-2 canopy height by integrating Sentinel data, which had an $R^2$ of 0.46 [29], whereas in this study an $R^2$ value of 0.48 was obtained using the RF model. This shows similar accuracy between these studies in mapping canopy height at a 30 m spatial resolution. Similarly, canopy height was estimated with an $R^2$ of 0.77 and an RMSE of 1.96 m by integrating ICESat-2 and Sentinel data using a stacking algorithm that consisted of multiple linear regression (MLR), support vector machine (SVM), k-nearest neighbor (kNN), and RF for independent validation [76]. Additionally, Lin et al. (2020) estimated canopy height by combining ICESat-2 ATLAS data with Stereo-Photogrammetry using a backpropagation artificial neural network (BP-ANN) model with an RMSE of 3.35 m and an $R^2$ of 0.51 for independent validation, whereas the $R^2$ and RMSE values for the RK model in this study were 0.69 and 3.49 m, respectively. Additionally, canopy height in the northwest Himalaya foothills of India was evaluated by the RF model with an $R^2$ of 0.84 and an RMSE of 1.15 m using ICESat-2 and Sentinel data. The high accuracy of this study may be because of the relatively homogenous forest canopy along with less topographic variation [31]. The canopy height of a coniferous forest in Heilongjiang province, China was estimated using ICESat-2 and Sentinel data with an $R^2$ of 0.52 and an RMSE of 3.15 m by the RF model, whereas this study had an $R^2$ of 0.48 and an RMSE of 4.58 m [33]. Although, being a mixed forest, the accuracy was slightly low, but the accuracy increased in the RK model with an $R^2$ of 0.69 and an RMSE of 3.49 m. Malambo and Popescu (2021) evaluated the canopy heights of ICESat-2’s ATL08 over various US locations and compared it with that of the ALS canopy height, showing a bias of $-1.71$ m and an MAE of 3.44 m, whereas this study obtained a bias of $-1.81$ m and an MAE of 4.52 m when comparing the RK-estimated canopy height with the airborne-derived canopy height. The accuracy of this study was slightly lower than that reported in other studies, which may be due to the large and topographically varied area with changes in environmental and climatic conditions, mixed species and structural differences, surface roughness, and a lack of predictor variables. Additionally, the Landsat images and ICESat-2 data used were from throughout the year 2020 and the airborne lidar data ranged from 2007 to 2019. As such, these temporal differences might have influenced the validation of the canopy height because the natural dynamics of forests change over time, i.e., growth of trees, canopy gaps due to death and fall of trees, natural calamities, etc. The results were satisfactory for mapping canopy height using ICESat-2 data, which could create future potential for global-scale mapping, even over a broad range of vegetation at a higher resolution [29,75]. This study also showed the potential of ICESat-2 for height mapping without any kind of data calibration.

Similar to the findings of mapping the biomass from ICESat-2, CC was the most important predictor variable [34,77], but it was surprising that the NDVI was the least important variable in the model. The poor importance of the Landsat-8-derived NDVI may be because of the diverse and variable forest structure, terrain shadow effects, and moisture levels [62]. Additionally, the acquisition time was throughout the year 2020, which includes all the seasons, and the vegetation index that corresponds to the same canopy height value may show different values in different images due to the varied phenophases of the forest in the different images [23]. A regional-scale canopy height map may be applied to characterize forest carbon at a similar spatial scale. For instance, a canopy height map prepared by combining ICESAT-2 and Sentinel-1 data has been used to estimate aboveground biomass using RF [31]. Similarly,
canopy height was estimated and mapped using ICESat–GLAS data and Landsat images using different machine-learning models, and a power model was developed to estimate biomass from canopy height [23]. This regional-scale map can be utilized for a wide range of tasks, including land-use and land-cover mapping, detecting changes, and other forestry, ecological, and geological applications. Additionally, it has been used to track the Kyoto Framework’s CO$_2$ sink and forest biomass buildup processes [78]. Combining the canopy height map with other remote sensing data may help in the identification of vegetation communities and enhance our understanding of the vegetation distribution across the landscape [79]. For example, to map the various vegetation community types in a severely degraded landscape in southern California, lidar-derived canopy height data were combined with a hierarchical object-based image analysis [80]. Site index (SI), which is the average height of dominating trees at a specific index age, is the most extensively used indicator of site productivity [81,82], hence this canopy height map could facilitate an indication of site quality at a regional scale.

There is another type of spaceborne lidar, the Global Ecosystem Dynamics Investigation (GEDI), which is specifically optimized for the measurement of vegetation structure and has been collecting data since April 2019 [83]. Datasets on canopy height, canopy cover, leaf-area index, gridded AGB, etc. have been created using GEDI lidar observations [83]. Additionally, the bias of the canopy-height-estimation accuracy of GEDI is less than 1 m [83,84]. Integrating ICESat-2 with GEDI may further reduce the strip effect of each individual dataset [85], which may improve the accuracy in mapping canopy height.

5. Conclusions

This study demonstrated that forest canopy height can be mapped at the regional scale by integrating ICESat-2 data and Landsat imagery, with RK as a recommended approach to achieving a 30 m full-coverage product. The utilization of two models, machine learning (RF) and geostatistical (RK), suggested that there is a strong potential for RK in estimating and mapping forest canopy height since it performed relatively better than the RF model. Future work will involve the preparation of an aboveground biomass map based on the most accurate canopy height map developed. In doing so, additional environmental predictors such as topographic roughness, precipitation and temperature will be examined. Finally, the integration of ICESat-2 and GEDI data can be investigated to further improve the estimation of canopy height at a regional scale.

Author Contributions: Conceptualization, L.L.N. and K.T.; methodology, K.T. and L.L.N.; formal analysis, K.T.; investigation, K.T.; resources, L.L.N. and K.T.; data curation, K.T.; writing—original draft preparation, K.T.; writing—review and editing, K.T. and L.L.N.; visualization, K.T.; supervision, L.L.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable.

Acknowledgments: We would like to thank Zhaofei Fan for his advice and guidance during this study. We would also like to thank Nisham Thapa, Schyler Brown, Cam Boland and Eian Davis for their support and advice during this study.

Conflicts of Interest: The authors declare no conflict of interest.

References


8. Liu, M.; Cao, C.; Dang, Y.; Ni, X. Mapping forest canopy height in mountainous areas using ZiYuan-3 stereo images and Landsat data. Forests 2019, 10, 105. [CrossRef]


64. Fayad, I.; Baghdadi, N.; Bailly, J.-S.; Barbier, N.; Gond, V.; El Hajj, M.; Fabre, F.; Bourgine, B. Canopy height estimation in French Guiana with LiDAR ICESat/GLAS data using principal component analysis and random forest regressions. *Remote Sens.* 2014, 6, 11883–11914. [CrossRef]
70. Tian, S.; Zhang, X.; Tian, J.; Sun, Q. Random forest classification of wetland landcovers from multi-sensor data in the arid region of Xinjiang, China. *Remote Sens.* 2016, 8, 954. [CrossRef]
84. Wang, C.; Elmore, A.J.; Numata, I.; Cochrane, M.A.; Shaogang, L.; Huang, J.; Zhao, Y.; Li, Y. Factors affecting relative height and ground elevation estimations of GEDI among forest types across the conterminous USA. *GIScience Remote Sens.* 2022, 59, 975–999. [CrossRef]