



OPEN ACCESS

EDITED BY

Ting Yun,
Nanjing Forestry University, China

REVIEWED BY

Jian Li,
Nanjing Forestry University, China
Yuxuan Lin,
Nanjing Forestry University, China

*CORRESPONDENCE

Qingtai Shu
✉ shuqt@swfu.edu.cn

RECEIVED 06 May 2024

ACCEPTED 15 August 2024

PUBLISHED 11 December 2024

CITATION

Xu L, Yu J, Shu Q, Luo S, Zhou W and Duan D (2024) Forest aboveground biomass estimation based on spaceborne LiDAR combining machine learning model and geostatistical method. *Front. Plant Sci.* 15:1428268. doi: 10.3389/fpls.2024.1428268

COPYRIGHT

© 2024 Xu, Yu, Shu, Luo, Zhou and Duan. This is an open-access article distributed under the terms of the [Creative Commons Attribution License \(CC BY\)](https://creativecommons.org/licenses/by/4.0/). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Forest aboveground biomass estimation based on spaceborne LiDAR combining machine learning model and geostatistical method

Li Xu¹, Jinge Yu¹, Qingtai Shu^{1*}, Shaolong Luo¹, Wenwu Zhou¹ and Dandan Duan²

¹Faculty of College of Soil and Water Conservation, Southwest Forestry University, Kunming, China,

²Information Technology Research Center, Beijing Academy of Agriculture and Forestry Sciences, Beijing, China

Estimation of forest biomass at regional scale based on GEDI spaceborne LiDAR data is of great significance for forest quality assessment and carbon cycle. To solve the problem of discontinuous data of GEDI footprints, this study mapped different echo indexes in the footprints to the surface by inverse distance weighted interpolation method, and verified the influence of different number of footprints on the interpolation results. Random forest algorithm was chosen to estimate the spruce-fir biomass combined with the parameters provided by GEDI and 138 spruce-fir sample plots in Shangri-La. The results show that: (1) By extracting different numbers of GEDI footprints and visualize it, the study revealed that a higher number of footprints correlates with a denser distribution and a more pronounced stripe phenomenon. (2) The prediction accuracy improves as the number of GEDI footprints decreases. The group with the highest R^2 , lowest RMSE and lowest MAE was the footprint extracted every 100 shots, and the footprint extracted every 10 shots had the worst prediction effect. (3) The biomass of spruce-fir inverted by random forest ranged from 51.33 t/hm² to 179.83 t/hm², with an average of 101.98 t/hm². The total value was 3035.29×10^4 t/hm². This study shows that the number and distribution of GEDI footprints will have a certain impact on the interpolation mapping to the surface information and presents a methodological reference for selecting the appropriate number of GEDI footprints to derive various vertical structure parameters of forest ecosystems.

KEYWORDS

GEDI, spaceborne LiDAR, inverse distance weighting, biomass, spruce-fir

1 Introduction

As the energy base and material source of forest ecosystem operation, forest aboveground biomass (AGB) is a key indicator to evaluate the health status of forest ecosystem and the sustainable utilization of vegetation resources (Assiri et al., 2023). It is also the basis for the study of ecosystem carbon cycle and carbon storage. Spruce-fir forest is the most widely distributed forest type in Chinese cold-temperate evergreen coniferous forest, which has high economic and ecological value (Liu et al., 2023a). Spruce-fir has important ecological functions in water retention and soil and water conservation. It can not only maintain biodiversity in high altitude and high latitude areas, but also affect the sustainable development of social economy (Chen, 2023). Chinese spruce-fir resources are extremely rich, with a stand volume of 2.2 billion m³, accounting for about 24.23% of the national stand volume, and ranking first among all tree species (Huang et al., 2004). The spruce-fir ecosystem is one of the most productive ecosystems on earth, and its carbon storage capacity is outstanding. Because of its important carbon storage capacity, it plays an important role in mitigating climate change (Lang et al., 2019). Therefore, accurate and large-scale monitoring of the AGB of the Shangri-La spruce-fir is of great significance for ensuring the safety of the alpine ecosystem and maintaining its carbon balance. Remote sensing technology has the potential to quickly obtain large-scale vegetation growth conditions (Yun et al., 2024). However, accurate regional AGB data still requires detailed field surveys on a finer scale (Liu et al., 2023b). Combining field survey data with RS technology has become a common method for estimating regional AGB. According to the different ways of obtaining forest vertical structure and sensors, remote sensing data can be divided into optical, synthetic aperture radar (SAR) and laser radar data. At present, medium-resolution optical data is still the most commonly used data source for large-scale acquisition of forest AGB, and its rich spectral information can effectively reflect the growth of vegetation (Jiang et al., 2021). Landsat data can provide multi-spectral images with medium resolution for decades, which is the most widely used data for forest researchers to solve various practical tasks. However, cloud cover and the saturation effect of dense forests are its main limitations (Wang et al., 2021). SAR obtains information by actively emitting energy, also known as active remote sensing. Its wavelength can penetrate the vegetation canopy and obtain more detailed structural information. It has obvious advantages in obtaining the vertical structure of forests (Cartus et al., 2012). However, SAR must work in a specific range of electromagnetic spectrum, but the characteristics of these bands are not necessarily suitable for biomass estimation.

As early as the mid-1980s, LiDAR technology has been applied in the field of forestry. As an active remote sensing technology, LiDAR has strong penetrability and can overcome the signal saturation problem in SAR and optical remote sensing data (Gao et al., 2022). According to its carrying platform, it can be divided into Terrestrial Laser scanner, Airborne Laser Scanner and Space-borne Laser. Terrestrial Laser scanner is usually used for the acquisition of single target or small-scale fine 3D data (Comesaña-Cebal et al., 2021). Its top-down working model allows individual trees to be segmented to

provide accurate estimates of DBH and tree position. However, due to its top-down scanning method, the information at the top of the canopy may be missing, and it is limited by terrain, scanning field of view and distance, so it is not suitable for inverting continuous large-scale forest vertical structure parameters. Airborne LiDAR is the best choice for forest AGB estimation at single tree scale due to its low cost, flexible operation and centimeter-level spatial image resolution. In recent years, it has been more and more widely used to estimate forest canopy height and biomass. However, the point cloud data of airborne LiDAR has the characteristics of large density and inhomogeneous distribution, which increases the difficulty of data processing, and the application in large areas and the acquisition of data are limited by high cost. Compared with airborne LiDAR, spaceborne LiDAR has the characteristics of large observation range and regular repeated observation. It has great advantages in quantitative inversion of forest parameters at large regional scale. It can not only reduce the consumption of manpower and time in field investigation, but also ensure accuracy, spatial integrity and time consistency (Yuan, 2022). At present, because the spaceborne LiDAR data can be used to detect the forest vertical structure in a large area, some scholars have applied it to the estimation of forest canopy height (Wang et al., 2023), forest height (Zhu, 2021; Lin et al., 2023; Luo et al., 2023), closure (Zhang et al., 2021) and biomass (Xu et al., 2023a; Song et al., 2022). ICESat-1 is the world's first spaceborne LiDAR altimetry system. After its retirement in 2009, the United States launched ICESat-2 satellite in 2018. The photon counting laser altimeter on board adopts micro-pulse multi-beam photon counting LiDAR technology. These two satellites have been successfully applied to measure forest structure parameters, lake water level, glacier change and sea ice surface classification (Zhu, 2021; Song et al., 2022; Liu et al., 2023c; Petty et al., 2021; Cheng et al., 2023).

GEDI is the only full-waveform LiDAR system designed specifically to measure the vertical structure of vegetation that NASA launched at the end of 2018 (Xie et al., 2018). The difference between GEDI and other spaceborne LiDAR is its penetration ability in dense vegetation. The GEDI system consists of three lasers, one of which is divided into two beams (coverage beam), and the other two lasers maintain full power (full power beam). The coverage beam and the full power beam can penetrate 95% and 98% of the forest canopy to the ground, respectively (Liu et al., 2022). It is equipped with the world's first multi-beam linear system laser altimeter for high-resolution forest vertical structure measurement, which is mainly used for accurate measurement of forest canopy height, vertical structure and ground elevation in tropical and temperate regions. From the GEDI waveform, four types of structural information such as landform, canopy height, canopy coverage and vertical structure can be extracted (Hang, 2022). GEDI is a full-waveform LiDAR data that can generate a forest canopy profile with a diameter of 25 meters. However, due to the fact that the GEDI footprints are distributed along the track and there is a certain distance between the footprints, the GEDI data is discrete and sparse. In order to overcome this limitation, previous studies have integrated continuous optical or SAR images with GEDI to infer GEDI samples spatially (Potapov et al., 2021). This extrapolation technique has been used in several studies to generate large-scale CHM and AGB maps. For example, Francini et al.

(2022) integrated GEDI data and four decades of optical images to analyze forest disturbance and biomass changes in Italy, demonstrating that GEDI can capture changes in forest biomass caused by disturbance. Tamiminia et al. (2024) integrates the GEDI footprint with the Sentinel-2 image to generate a 10-meter wall to wall CHM of New York State in 2019, and then uses the generated CHM, Sentinel-1 and Sentinel-2 data to create a 10mAGB map of New York State in 2019. Guo et al. (2023) proposed a new method to integrate field measurement data, GEDI LiDAR, sentinel and terrain data to construct multi-source data-driven forest CHM and AGB models with a resolution of 30 m. Firstly, RFE-SVM method was used to determine the characteristics sensitive to forest height and AGB, and then three regression models were used to construct the CHM model. The GEDI point data was extended to the wall to wall CHM map, and finally the common features and measured data were selected to construct the model for estimating AGB. Although many scholars have done a lot of research on GEDI in forest attribute estimation, most of them are to evaluate forest canopy height or jointly estimate forest attributes with other remote sensing images. At present, no scholars have studied the effect of the number of GEDI light spots on the biomass estimation results.

The objective of this study is twofold. First, GEDI footprint of different orders of magnitude are evaluated and the optimal data set is selected. Secondly, the biomass was estimated by using the selected optimal footprint combined with the random forest model. Specifically, this paper aims to address the following research questions: 1) investigate the interpolation accuracy of GEDI footprint on different orders of magnitude. 2) explore the feasibility of using inverse distance weight interpolation to interpolate footprint spaceborne LiDAR samples. 3) and determine the optimal GEDI footprint as the input predictor of accurate AGB mapping.

2 Study area and research data

2.1 Study area

Shangri-La is located in the northwest of Yunnan Province and the eastern part of Diqing Tibetan Autonomous Prefecture, and characterized by a rugged internal terrain, mainly comprising a series of high mountains that stretch from the middle to the northeast (Figure 1). The altitude varies in the region, being higher in the northwest and lower in the southeast, with an overall altitude range of 1503 to 5545 meters, and an average altitude of approximately 3459 meters. The region experiences a mountainous cold temperate monsoon climate, with an average annual temperature of 5.4°C and an average annual precipitation of 618.4 mm. The frost-free period typically lasts from 129 to 197 days, with the rainy season occurring from July to September. This microclimate is attributable to its location in the transition zone between the subtropical evergreen broad-leaved forest vegetation area of Yunnan and the alpine vegetation area of the Qinghai-Tibet Plateau. As a result, the vegetation distribution varies significantly between the north and the south, with distinct vertical structure characteristics prevailing in the region. The distribution of the

species varies based on altitude. At 1500 ~ 2800 m, its primary habitat is found within warm-temperate coniferous forests, where *Pinus armandii* and *Pinus yunnanensis* dominate. As the altitude ranges from 2800~3500 m, the species tends to thrive in warm and cool coniferous forests, with *Tsuga yunnanensis* and *Pinus densata* as the dominant species. Moving further up, at an altitude of 3200 ~ 4200 m, the species is predominantly distributed in cold-temperate coniferous forests, where *larix potaninii var.macrocarpa*, *picea brachytyla var.complanata*, and *longbract fir* are the dominant species. Above 4200 m, the distribution extends to cold-temperate shrubs, particularly *rhododendron delavayi*, and meadows.

2.2 Field inventory data and calculation of AGB

Forest resources survey data used in our study. For the purpose of this study, we focused on the two dominant tree species, *Picea asperata* and *Abies fabri*, totaling 138 plots, in order to compute the biomass. To determine the total biomass of each sample, it is necessary to calculate the biomass of individual trees. Consequently, the cumulative biomass of all individual trees within the sample represents the total forest biomass of the sample. Common standing tree biomass models mainly include one-dimensional, binary and multivariate biomass models. The one-dimensional biomass model mostly selects the DBH of the tree as an independent variable, and the binary biomass model selects the DBH and tree height (Wang et al., 2018). In this study, the biomass in the plot was calculated by referring to the binary biomass model provided by the forestry industry standards such as 'Tree biomass models and related parameters to carbon accounting for *Picea*' (Zeng et al., 2016a), 'Tree biomass models and related parameters to carbon accounting for *Abies*' issued by the State Forestry Administration (Zeng et al., 2016b). Based on the data of tree height and DBH of forest resource survey data, the aboveground biomass values of spruce-fir in the plot were calculated by equations, respectively. The AGB of the sample plot is divided by the area, which can be converted into the AGB per hectare, and the unit biomass information of the sample plot is obtained as shown in Table 1.

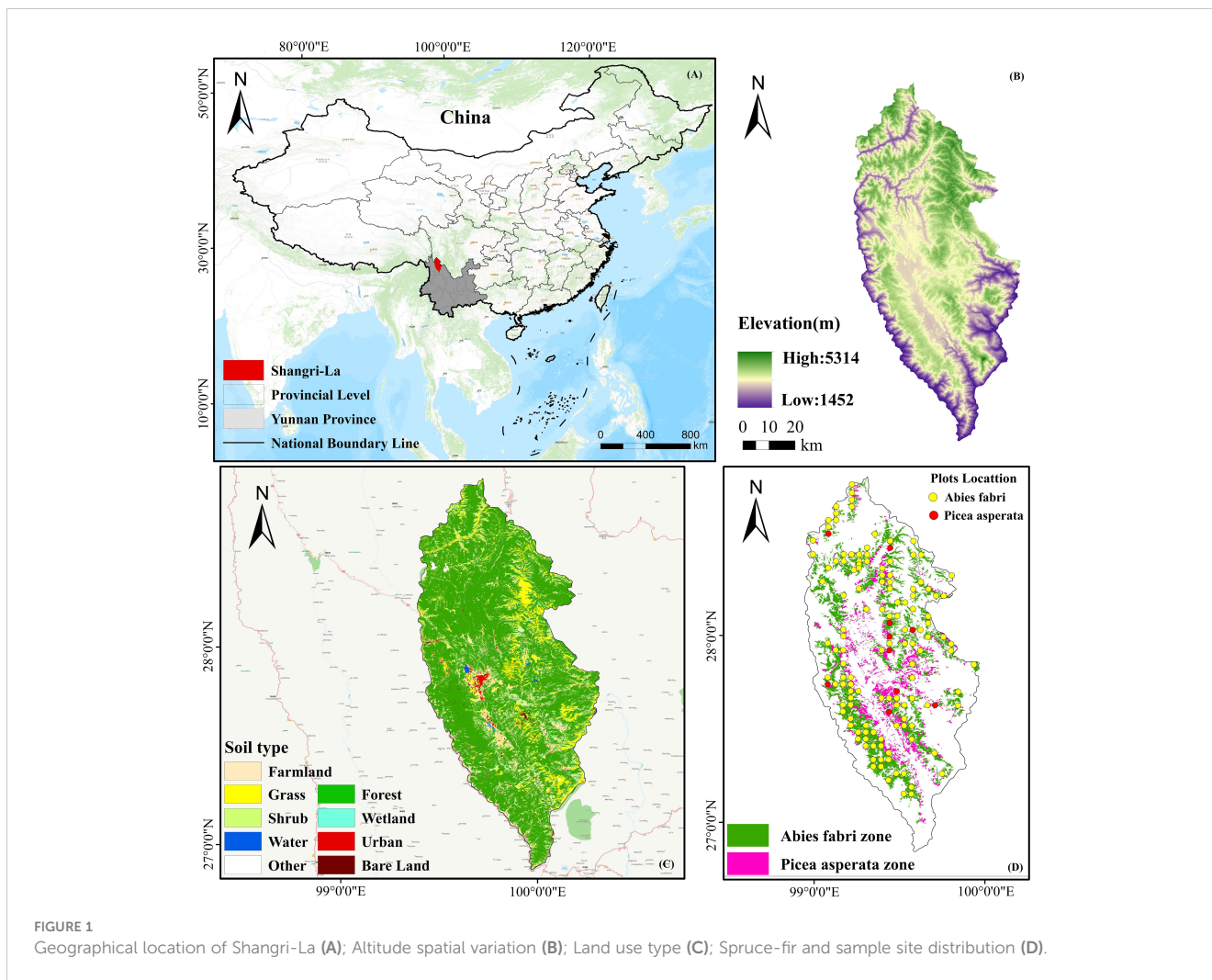
$$M_{\text{spruce}} = 0.09152D^{2.21060}H^{0.25663}$$

$$M_{\text{fir}} = 0.06127D^{2.05753}H^{0.50839}$$

Where M_{spruce} the model of aboveground single is tree biomass of *spruce*, and M_{fir} is the model of the aboveground single tree biomass of *fir*. M is the AGB; D is the diameter at breast height; H is the single tree height.

2.3 GEDI data

GEDI spaceborne LiDAR is a form of active remote sensing technology that captures extensive data by emitting short-wavelength laser pulses to penetrate the forest canopy and



retrieve precise three-dimensional forest structure information (Saarela et al., 2022). Comprising three lasers, two of which are full-power lasers and the remaining one functioning as a covering laser, GEDI utilizes beam jitter to split into two beams, resulting in a total of eight beams. The footprint diameter is approximately 25 m, with a footprint spacing of 60 m along the track and 600 m spacing between adjacent ground tracks. The overall width of the laser trajectory is 4200 m (Potapov et al., 2021). GEDI data is composed of discrete footprints, each containing various echo indicators for feature extraction and input into forest biomass models. These echo indicators are crucial for subsequent GEDI data analysis and interpretation. In this study, the data product of the spaceborne LiDAR GEDI Level 2B Version 2 is utilized. GEDI Version 2 enhances geo-positioning accuracy, updates metadata to include spatial coordinates, and enables querying in NASA Earth data retrieval. The average geolocation error of version 2 data product is 10.3 m, showing a 50% improvement in geolocation accuracy compared to Version 1. GEDI contains four levels of products: Level 1 encompasses geolocated return energy waveform, Level 2 includes footprint level canopy cover and vertical profile metrics, Level 3 comprises gridded land surface metrics, and Level 4 consists of

footprint level and gridded aboveground biomass density (Tian, 2023). All GEDI data used in this study are sourced from Earthdata (<https://www.earthdata.nasa.gov>), with the data acquisition time ranging from April 23, 2019 to December 04, 2019.

3 Methods

The implementation of this method encompasses five steps, as described in the flowchart of Figure 2: 1) GEDI Data Filtering; 2) GEDI Footprint Distribution and Interpolation Results Analysis; 3) Accuracy Evaluation of GEDI Footprint Interpolation Results; 4) Correlation Analysis between GEDI Variables and Biomass; 5) Biomass Modeling.

3.1 Inverse distance weighting model

The inverse distance weighting method (IDW) is based on the assumption that nearby things are more similar to each other than those that are farther apart. It considers that each measurement

TABLE 1 Spruce-fir plot information summary.

Number of plots	Maximum (t/hm ²)	Minimum (t/hm ²)	Average (t/hm ²)	Median (t/hm ²)	Standard error (t/hm ²)
139	296.09	5.19	101.42	92.73	65.09

point has a local influence, and that this influence gradually diminishes as the distance between point's increases. In essence, IDW assigns a weight to each interpolation point by taking the reciprocal of the distance between the interpolation point and the measured point to perform a weighted average. This method is particularly suitable for situations where measurement sites are evenly distributed and of high density.

$$Z^*(x_0) = \sum_{i=1}^N \lambda_i Z(x_i) = \sum_{i=1}^N \left(\frac{d_{i0}^{-p}}{\sum_{j=1}^N d_{j0}^{-p}} \right) Z(x_i)$$

where $Z^*(x_0)$ is the interpolation result of the interpolation point x_0 ; $Z(x_i)$ is the measured value of the measured point x_i ; n is the number of measured points involved in the calculation; λ_i is the weight coefficient; d_{i0} is the distance between the measured point x_i and the interpolation point x_0 ; p is the power of distance, generally $p = 1$ or $p = 2$.

3.2 Evaluation of interpolation accuracy

The accuracy of **interpolation** was assessed by comparing the root mean square error (RMSE) and the coefficient of determination (R^2) between the measured and predicted values at both the interpolated footprints and the verification sample points. R^2 quantifies the correlation between the predicted and estimated values, with a value closer to 1 indicating higher model accuracy. On the other hand, RMSE primarily elucidates the estimation's sensitivity to extreme effects within the sample data. The mean absolute error (MAE) calculates the average difference between predicted and measured values. A smaller MAE indicates a more accurate prediction model.

$$R^2 = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}}$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i|$$

where n is the number of footprints, \hat{y}_i is the predicted value on the footprint i , y_i is the observed value on the footprints, and \bar{y} is the arithmetic mean of the observed value.

3.3 Biomass estimation model

3.3.1 Support vector machine

Support Vector Machine (SVM) is an efficient machine learning algorithm, which is based on the principle of dividing hyperplanes. The core idea of this classifier is to divide the data into several possible categories by constructing a hyperplane. By continuously adjusting and optimizing this hyperplane, the overall deviation is minimized (Chen et al., 2023). For the known training data set, the SVM algorithm will find the hyperplane that maximizes the classification interval, and use these intervals to predict the new data set. A hyperplane is a character space one dimension less than its surrounding planes. In the two-dimensional data set, the straight line is a hyperplane. The key to its core idea is to map complex nonlinear data onto a high-dimensional array, a mapping process called kernel learning (Qiu et al., 2020). In this way, the data can be regarded as a linear function in the high-dimensional feature space,

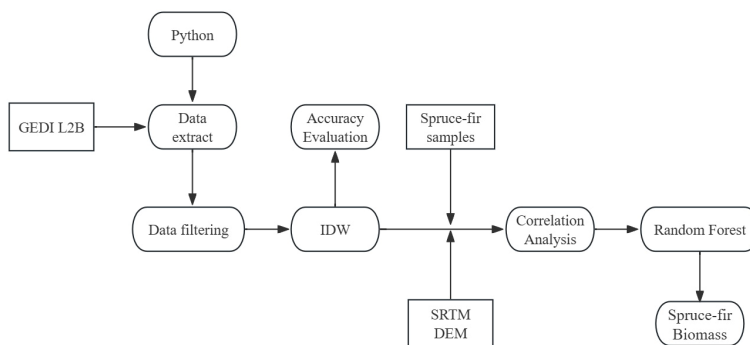


FIGURE 2 Steps used to estimate AGB using GEDI data.

while the nonlinear features are mapped to the low-dimensional space. In this study, the `e1071` package in R language was used, the kernel function was radial basis, the C of spruce-fir was 1, and the g was 0.92.

3.3.2 K-nearest neighbor algorithm

K nearest neighbor regression (KNN) is a typical nonparametric algorithm. It is a univariate or multivariate prediction method based on the spatial similarity between observation points and prediction points. It can be used for classification and parameter estimation problems, especially for remote sensing data with non-normal or unknown probability density distribution function (Zhang et al., 2022). As a non-parametric learning method widely used in scientific research and engineering technology, its core is to use the observation point data in the nearest k neighborhoods to predict the value of unknown points. This algorithm can effectively deal with remote sensing data that are affected by the non-normal distribution of the probability density distribution function (PDF) or unknown distribution, and realize the classification task and parameter estimation of the data by analyzing the correlation between the observed value and the predicted point (Wu et al., 2024). In this paper, the `knn` function in the `class` package of R language is used to implement the algorithm, and the K value of spruce and fir is set to 9 respectively.

3.3.3 Random forest regression

Random forest is a machine learning algorithm that exhibits high accuracy and robustness, making it suitable for analyzing high-dimensional and highly correlated data sets (Wei et al., 2024). The algorithm operates on the principle of utilizing the bootstrap method to iteratively and randomly sample from the original data, constructing decision trees for each sample. Through the combination of predictions from multiple decision trees, the final prediction results are determined by collective voting (Lourenço et al., 2021). This approach eliminates the necessity for feature selection and enables the handling of high-dimensional data. Moreover, the random sampling of instances leads to the utilization of different training sets for each decision tree, thereby mitigating overfitting to a certain extent (Yang et al., 2021a). Notably, the Random Forest package within R software facilitates predictive regression using the random forest algorithm, necessitating adjustments to the number of decision trees (`n tree`) and the number of variables extracted during tree splitting (`m try`). The `n tree` is set to 300 and the `m try` is set to 2.

3.4 Biomass model validation

To improve the accuracy of the biomass inversion model, it is essential to select the optimal variables that are conducive to biomass inversion while avoiding data redundancy, which can diminish the accuracy of model inversion. This process not only reduces the running time of the algorithm, but also prevents over-fitting caused by model complexity. Through an analysis of the correlation between independent variables and the biomass of spruce-fir, the first five

variables with the highest correlation were identified and selected as the determinants of the model. These variables were utilized to estimate the biomass of spruce-fir within the study area, and their accuracy was subsequently evaluated through cross-validation (Vaithyanathan and Sudalaimuthu, 2022). This method ensures that the model is based on the most relevant data, leading to an improved accuracy in biomass estimation.

To ensure the stability of the model and reduce accidental error caused by the division of training and verification samples, we employed the ten-fold cross-validation method in this study. The data set, containing 138 ground observation samples, was divided into 10 groups with an equal number of samples. Each group was subsequently utilized as the verification set in turn to assess the model's estimation ability, while the remaining 9 groups served as the training set for constructing the model. For each cross-validation, we calculated the determination coefficient (R^2), the root mean square error (RMSE) and mean absolute error (MAE) between the predicted and measured values. This process was repeated 10 times, and the accuracy of each model was evaluated based on the average R^2 , RMSE and MAE values. The calculation formulas for R^2 , RMSE and MAE can be found in 3.2 sections.

4 Results

4.1 GEDI preprocessing and data filtering

4.1.1 GEDI data pre-processing

GEDI is a geolocation waveform product, which contains 8 beams and 1 metadata. Each beam group stores the transmitted and received waveforms, geolocation parameters and geophysical correction parameters obtained by the beam after geolocation correction (Dubayah et al., 2021). GEDI uses 1064nm laser pulse to measure the vertical structure of the forest, and the obtained waveform needs a series of processing to obtain the vertical structure parameters of the forest. According to the theoretical basis file of GEDI official algorithm, a Gaussian filter (smoothing width) with a width of 6.5ns is first used to smooth the waveform to reduce the noise in the signal without affecting the signal in the waveform (Schwartz et al., 2024). After smoothing, the two positions represented as `searchstart` and `searchend` in the waveform are determined. `searchstart` and `searchend` are the first and last positions of the signal with a signal strength higher than the following thresholds:

$$\text{threshold} = \text{mean} + c \cdot \text{std}$$

Where 'threshold' is the background threshold, 'mean' is the average noise level, 'std' is the standard deviation of the smoothed waveform noise, and 'c' is a given constant, which is stored in the data-aided group of waveform processing as `back_threshold` or `front_threshold`.

After determining the positions of `searchstart` and `searchend`, the region between them (i.e., waveform range) is extended to the expected length, and the highest (`toploc`) and lowest (`botloc`) detectable echoes are determined within the waveform range.

toploc and botloc represent the highest and lowest positions in the waveform range when the two adjacent intensities are higher than the threshold, respectively. At this time, the threshold equations used to determine toploc and botloc are the same as the above equations. As shown in Figure 3, the black waveform is the waveform before Gaussian fitting, and the red waveform is the waveform after Gaussian fitting. After fitting, the burr part of the waveform is removed, and a smoother waveform is obtained, while reducing the influence of noise on the waveform. Subsequently, the second filtering is performed on the waveform part whose intensity exceeds the threshold to determine the position of different peaks in the waveform, such as the ground peak or the peak at the top of the crown. The width of the second Gaussian filter is expressed as smoothwidth cross. Finally, the position of the final detected peak is used to determine the position of the ground echo in the waveform. After smoothing and filtering the waveform data stored in the L1B product, the results such as noise mean, noise standard deviation, waveform processing-related parameter settings, and geolocation parameters are output and stored in the L2B product.

4.1.2 GEDI data filtering

In the range of Shangri-La, python software is used to write code to extract and filter the footprints according to the position and parameter name of each parameter in the GEDI L2 B data source file, and convert it into a Shapefile format that can be processed by ArcGIS. In this study, 14 modeling parameters were extracted from GEDI_L2B data products, which are cover, dem, fhd_normal, Landsat_treecover, leaf_off_doy, leaf_on_doy, modis_treecover, modis_nonvegetated, pai, pgap_thea, rg, rh100, rv and sensitivity. The specific names of these 14 parameters and their corresponding descriptions are listed in Table 2. Due to the lack of information or poor quality of some footprints, not all waveform data can meet the forest resource survey. Therefore, when processing data, three values (sensitivity, degrade_flag, quality_flag) of GEDI's own quality characteristics are used to screen out the spots that meet the conditions. A quality_flag value of 1 indicates the laser shot meets criteria based on energy, sensitivity, amplitude, real-time surface tracking quality and difference to a DEM. Sensitivity is defined as

the difference between the ratio of the ground echo area to the total echo area and 1, and its value range is 0 ~ 1. Different surface coverage types have different values, and the general forest is set to 0.9. Degrade_flag indicates the degradation flag of the satellite pointing to the positioning state, and the footprint with degrade_flag of 0 needs to be retained (Jiang et al., 2021).

4.2 GEDI data distribution analysis

To verify the effect of footprints interpolation of different data and achieve a random and uniform distribution of footprints, our study utilized a Python algorithm to extract representative footprints at intervals of 10, 30, 50, 70, and 100 shots when extracting GEDI data. The extracted data is presented in Table 3, which reveals that 13,127 footprints were extracted at intervals of 10 shots, and 4,343 footprints were extracted at intervals of 30 shots. Similarly, 2,646 footprints were obtained at intervals of 50 shots, while 1,871 and 1,309 footprints were obtained at intervals of 70 and 100 shots, respectively.

The footprints distribution map can be obtained by visualizing the extracted footprints. Figure 4 illustrates the distribution of footprints extracted at intervals of every 10, 30, 50, 70, and 100 shots (as denoted by a-e in Figure 4). The characteristics of GEDI's laser scanning are evident in the distribution of GEDI footprints, demonstrating the formation of footprints lattices as a result of orbit overlap. Dense footprints are primarily concentrated around the lattice, with no footprints inside. It is apparent from Figure 4 that an increase in the number of footprints results in closer distances between footprints, leading to a more pronounced lattice effect, especially in Figure 4A. Conversely, a decrease in the number of footprints significantly reduces the density of footprints on the same track, as depicted in Figure 4E, subsequently weakening the lattice effect and evoking a more random distribution pattern.

4.3 GEDI footprint interpolation results

The spatial interpolation prediction of independent variables extracted from different numbers of GEDI footprints is depicted in Figures 5–9. Overall, the spatial distribution of the independent variable prediction does not exhibit significant differences across the five orders of magnitude of footprints. However, variations are evident in the specific areas where the footprints are located. The prediction map (Figures 5–7) derived from the interpolation of the independent variables with shot intervals of 10, 30, and 50, demonstrates a lack of smoothness, characterized by the presence of a “bull’s-eye” phenomenon, which allows for clear visibility of the footprint positions (For example, the local magnification of sensitivity and Landsat_treecover in Figures 5–7). Moreover, the distribution of GEDI orbit reveals discernible staggered bands in the prediction results of four specific independent variables – cover, Landsat_treecover, pgap_thea, and pai. This observation may be attributed to the dense distribution and substantial number of footprints along the track. In contrast, the four independent variables of dem, modis_nonvegetated, leaf_on_doy, and

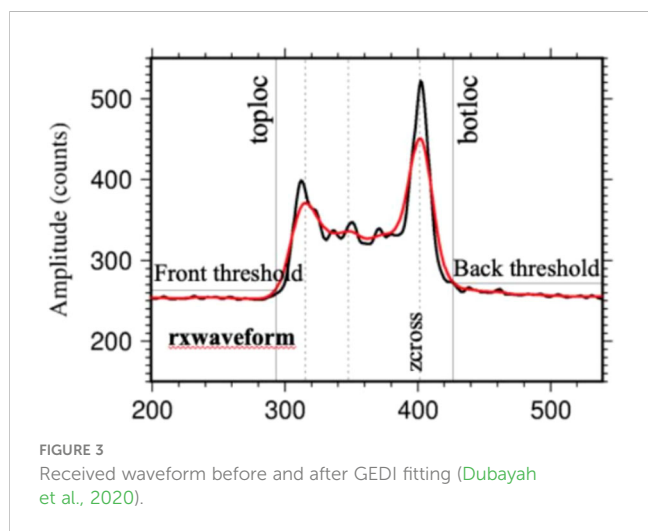


FIGURE 3
Received waveform before and after GEDI fitting (Dubayah et al., 2020).

TABLE 2 GEDI parameters information.

Parameters name	description	Parameters name	description
cover	Total canopy cover,	DEM	Digital elevation model
fhd_normal	Foliage Height Diversity index calculated by vertical foliage profile normalized by total plant area index	landsat_treecover	Landsat tree canopy cover
leaf_off_doy	Leaf off day of year	leaf_on_doy	Leaf on day of year
modis_treecover	Percent tree cover from MODIS data	modis_nonvegetated	Percent non-vegetated from MODIS data
pai	Total plant area index	pgap_thea	Estimated Pgap(theta) for the selected L2A algorithm.
rg	Integral of the ground component in the RX waveform	rh100	Height above ground of the received waveform signal start
rv	Integral of the vegetation component in the RX waveform	sensitivity	Maximum canopy cover that can be penetrated considering the SNR of the waveform
degrad_flag	Non-zero values indicate the shot occurred during a degraded period. A non-zero tens digit indicates degraded attitude, a non-zero ones digit indicates a degraded trajectory.	quality_flag	Flag simplifying selection of most useful data for Level 2B

leaf_off_doy yield relatively smooth prediction results with minimal error, regardless of the magnitude of the prediction, compared to other variables.

The estimation results of different independent variables have changed as the number of predicted footprints decreases. Figures 8, 9 demonstrates that the previously observed staggered bands in the prediction results of each variable have notably reduced, albeit the persisting presence of point-like ‘salt and pepper’ high-value points in the figure (Such as the local magnification of sensitivity and Landsat_treecover in Figures 8, 9). A comparison of the spatial prediction maps of the five quantities of independent variables reveals smoother interpolation results for the footprints extracted every 100 shots, with a relatively continuous spatial distribution map. These results exhibit more gradients and smoother edges compared to the other five respective results. Notably, among the interpolation results of different numbers of footprints, those extracted every 100 shots yield the best outcomes, followed by 70 and 50, with 30 and 10 producing the least favorable results.

4.4 Accuracy evaluation of GEDI footprint interpolation results

To evaluate the accuracy of the GEDI footprint extraction, the data was divided into two parts: 70% for interpolation and 30% for accuracy evaluation. The study area’s accuracy of spatial prediction for independent variables at different sampling densities is measured through the R², RMSE and MAE of the verification set (consisting of 3938, 1303, 794, 561, and 275 footprints). It can be

TABLE 3 The Number of Spots with different shot intervals.

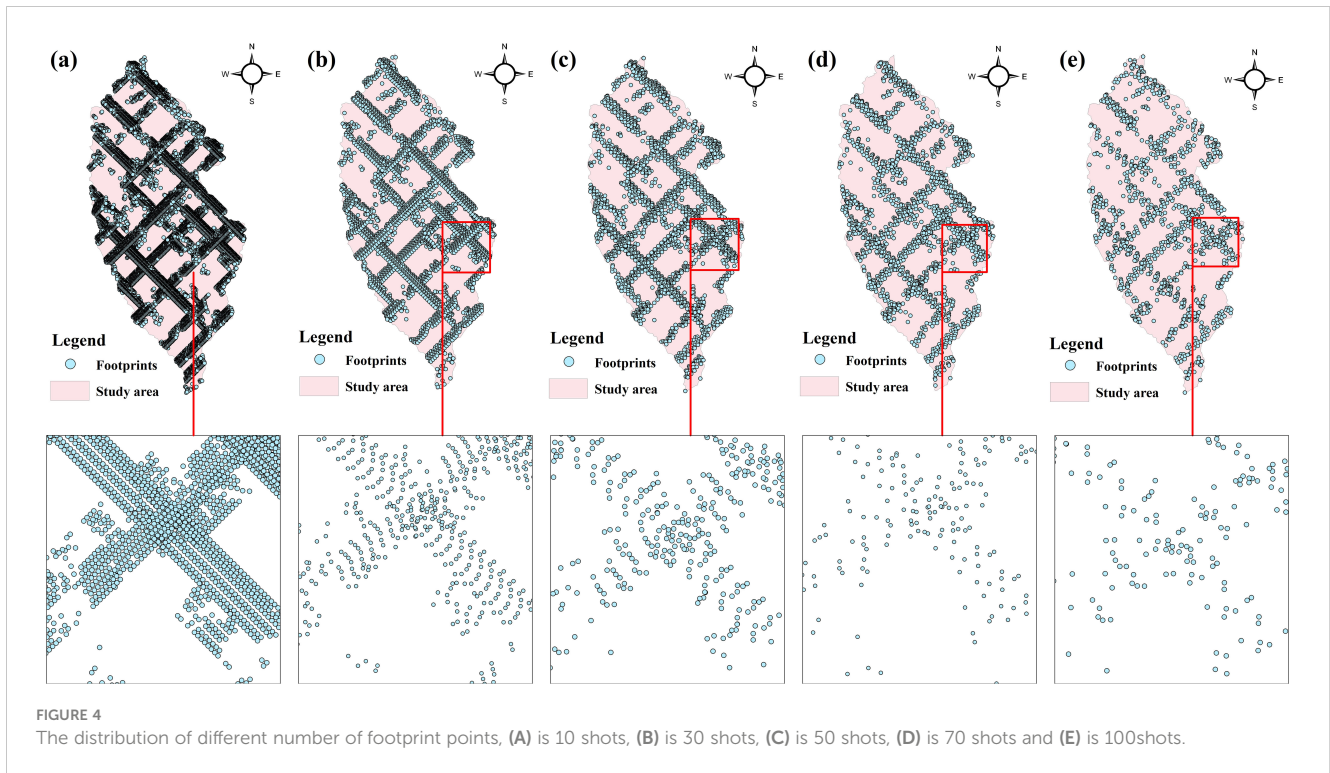
Shot interval	10	30	50	70	100
Number of footprints	13127	4343	2646	1871	1309

seen from Table 4 that the R² of the independent variable increases with the decrease of the sampling density, while the values of RMSE and MAE generally decrease with the decrease of the sampling density, indicating that the prediction accuracy increases with the decrease of the number of light spots. This trend suggests that prediction accuracy improves with decreasing footprints. Notably, the group with the highest R² and lowest RMSE and MAE is the one with extractions made every 100 shots, followed by 70 and 50 shots. The prediction effect for footprints extracted every 10 shots is the poorest, with the lowest R² observed for pai, Landsat_treecover, pgap_thea, and cover recorded at 0.47, 0.47, 0.49, and 0.51, respectively. These accuracy results align with the interpolation findings in Figure 4 to Figure 8, demonstrating that lower accuracy in variable interpolation leads to a more pronounced band phenomenon along the footprint distribution.

4.5 Correlation analysis between GEDI variables and biomass

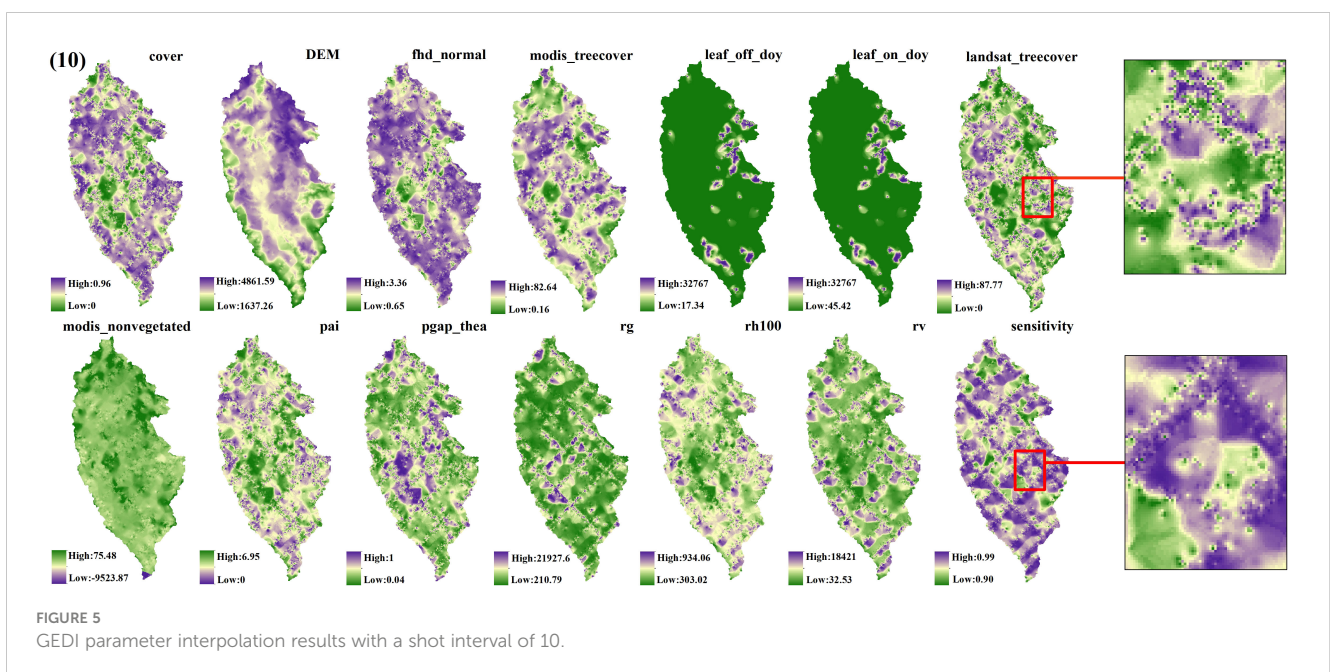
In section 3.2, the footprint extracted from every 100th shot is found to have the best interpolation effect of GEDI independent variables. Consequently, the 916 footprint independent variables are chosen for biomass estimation. In conjunction with Shangri-La’s topographic factors, a correlation analysis is performed with the biomass of Shangri-La spruce-fir, and the variables showing the highest correlation are then identified as the modeling indicators. The correlation coefficient matrix presented in Figure 10 displays the Pearson correlation coefficients between all remote sensing variables and forest aboveground biomass (AGB).

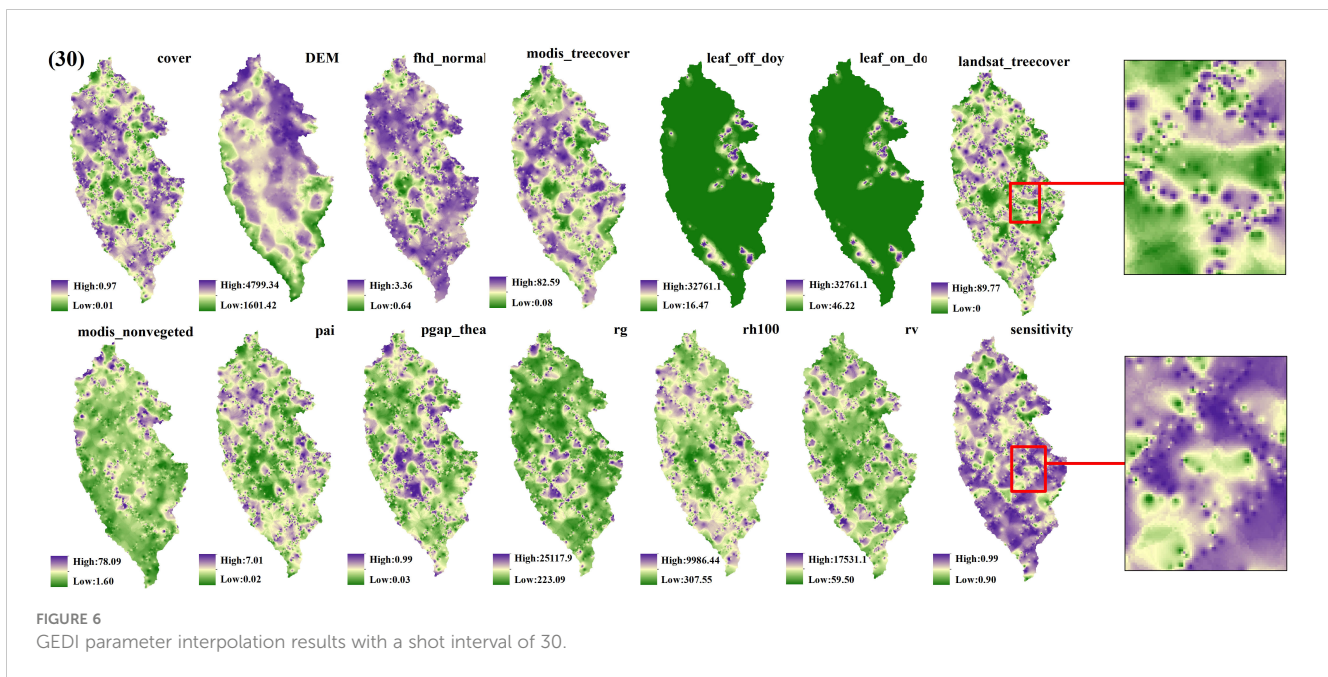
Upon examination of Figure 11, it is evident that seven characteristic factors of the GEDI independent variables, derived from 100 shots, exhibited significant correlations with the biomass of the plot. These factors, in descending order of correlation



strength, are: pai, pgap_thea, cover, fhd_normal, rh100, and DEM. Notably, DEM, rh100, pai, fhd_normal, and cover were found to be significantly positively correlated with AGB, with correlation coefficients falling within the range of 0.174 to 0.224. Conversely, rg and pgap_thea were significantly negatively correlated with AGB, yielding correlation coefficients ranging from -0.208 to -0.236. Notably, among the 14 factors, rg and pai demonstrated the highest correlation, being significantly correlated at the 0.01 level

with correlation coefficients of -0.236 and 0.224, respectively. The impact of GEDI L2B provided vegetation leaf area index, leaf height diversity index, and vegetation coverage area on biomass estimation for Shangri-La spruce-fir becomes evident through the analysis of variables. On the other hand, the remaining 10 factors show no significant correlation with the biomass of Shangri-La spruce-fir. Hence, based on the correlation analysis results of each characteristic factor and plot biomass, the top 5 factors with the



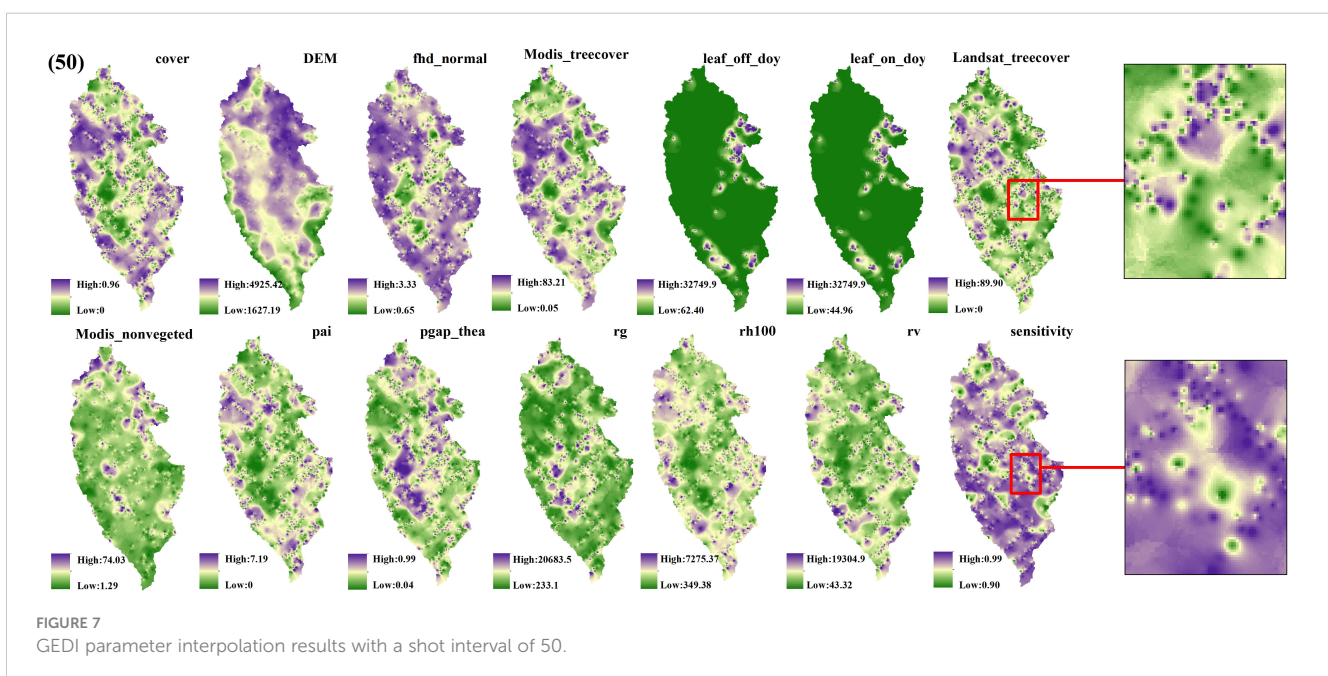


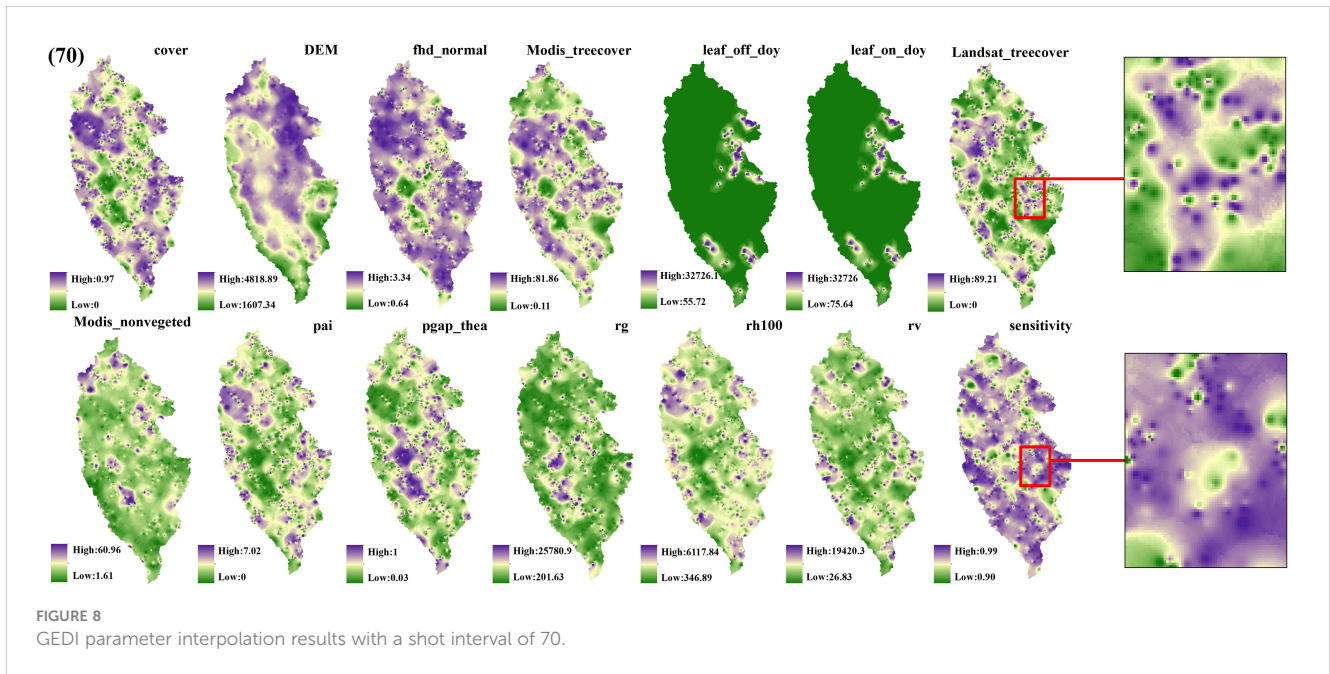
highest correlation coefficients were selected to serve as the independent variables in the subsequent random forest biomass estimation model.

4.6 Biomass modeling

In this study, the first five variables with the highest correlation were selected as the modeling factors of support vector machine, k-nearest neighbor method and random forest to estimate the biomass of spruce-fir in the study area, and the accuracy was tested by cross validation. In order to intuitively compare the

estimation performance of the three machine learning algorithm estimation models, the relationship between the estimated biomass and the measured biomass was drawn (Figure 12). It can be seen from Figure 3 that the AGB estimation accuracy based on the random forest model is higher, $R^2 = 0.87$, $RMSE=30.96t/hm^2$, $MAE=23.65t/hm^2$, followed by KNN, $R^2 = 0.45$, $RMSE=49.90t/hm^2$, $MAE=35.94t/hm^2$, while the SVM model has the lowest estimation accuracy, $R^2 = 0.31$, $RMSE=54.12t/hm^2$, $MAE=38.19t/hm^2$. From the data distribution map Figure 13, it can be seen that the distribution range of AGB predicted values of the three models is 16.19~177.40t/hm², 24.04~181.78t/hm², 27.22~199.10t/hm², respectively, which is less than the measured value range of





5.19~296.10t/hm². Among the three models, random forest has higher estimation accuracy than SVM and KNN, and the overall performance of the estimated value is closer to the measured value, and the model performance is better.

Based on the footprints provided by GEDI, the AGB of spruce-fir forest in the whole study area was inverted by using the random forest model with the highest estimation accuracy (Figure 14). The random forest estimation revealed that the above-ground biomass of spruce-fir forest in the study area was generally at a medium level, with a maximum of 179.33 t/hm², a minimum of 51.83 t/hm², and an average of 101.98 t/hm². The total biomass value was estimated to

be 3035.29×10⁴ t/hm². From the perspective of spatial distribution of biomass, the spruce-fir in the study area showed patches inhomogeneous distribution from northwest to southeast. On the whole, it is distributed longitudinally along the Hengduan Mountains, which conforms to the growth habit of spruce-fir. The biomass of spruce-fir in the northwest and east of Shangri-La is relatively higher than that in the southwest and northeast of Shangri-La. The total biomass of mountain and hilly areas mountainous and hilly areas such as Shangri-La Grand Canyon and Pudacuo National Park display the highest total biomass (Figure 14A), while Xiaozhongdian and Shangri-La urban area have the lowest total biomass (Figure 14C).

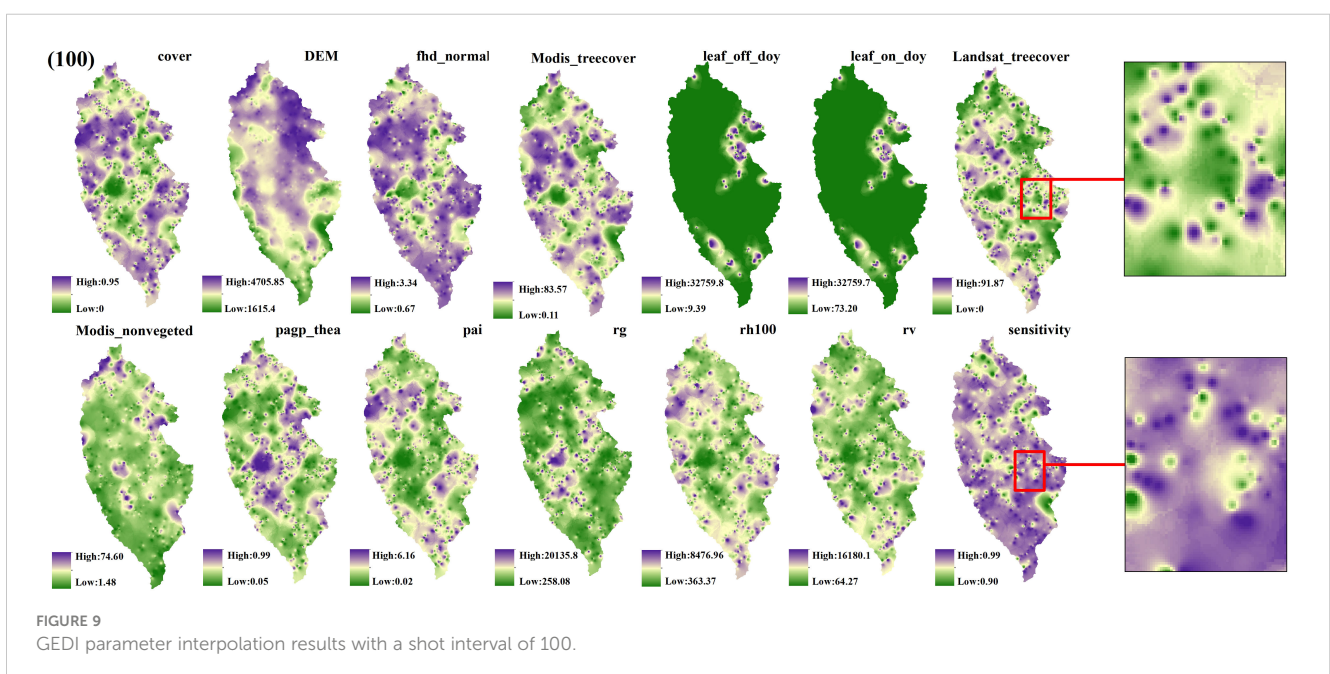


TABLE 4 Different GEDI footprints interpolation R², RMSE and MAE.

parameter name	Different GEDI footprints interpolation R ² , RMSE and MAE														
	10			30			50			70			100		
	R ²	RMSE	MAE	R ²	RMSE	MAE	R ²	RMSE	MAE	R ²	RMSE	MAE	R ²	RMSE	MAE
pai	0.47	1.23	0.90	0.61	1.04	0.70	0.62	1.03	0.67	0.63	1.02	0.64	0.94	0.44	0.26
Modis_treecover	0.88	8.90	5.60	0.79	11.56	7.76	0.7	13.14	7.68	0.77	12.10	7.98	0.96	5.34	3.10
Modis_nonvegetated	0.71	6.15	4.13	0.73	5.92	3.55	0.75	5.81	3.31	0.69	6.17	3.34	0.97	2.2	1.98
Landsat_treecover	0.47	25.09	18.45	0.6	21.92	14.08	0.6	21.35	14.66	0.66	20.42	13.43	0.93	9.44	5.39
Leaf_off_doy	0.8	3582.7	889.6	0.78	3144.4	686.25	0.8	3086.9	711.47	0.87	2618.5	715.29	0.98	1420.5	349.48
Leaf_on_doy	0.71	3590.4	893.9	0.78	3151.8	690.23	0.8	3093.0	715.45	0.87	2625.3	718.96	0.98	1425.4	351.19
Fhd_normal	0.56	0.39	0.26	0.71	0.32	0.20	0.68	0.34	0.20	0.67	0.34	0.19	0.96	0.12	0.07
DEM	0.97	108.6	76.23	0.96	118.95	73.42	0.95	129.96	77.44	0.94	146.82	83.13	0.99	55.33	35.49
cover	0.51	0.22	0.17	0.63	0.19	0.13	0.61	0.21	0.13	0.64	0.19	0.12	0.94	0.08	0.05
rg	0.52	3041.9	2010.7	0.67	2543.7	1585.2	0.6	2565.3	1564.4	0.68	2610.0	1491.8	0.95	1032.1	563.69
Pgap_thea	0.49	0.22	0.17	0.65	0.19	0.13	0.65	0.19	0.13	0.69	0.12	0.12	0.9	0.08	0.05
Sensitivity	0.57	0.014	0.01	0.61	0.013	0.01	0.65	0.012	0.01	0.68	0.011	0.01	0.92	0.004	0.00
rv	0.54	2799.5	1899.3	0.67	2340.7	1492.3	0.68	2356.7	1478.7	0.62	2361.2	1388.2	0.89	1007.8	583.94
Rh100	0.55	721.1	505.81	0.63	638.88	389.63	0.62	648.05	391.44	0.67	612.99	382.29	0.96	224.04	139.90

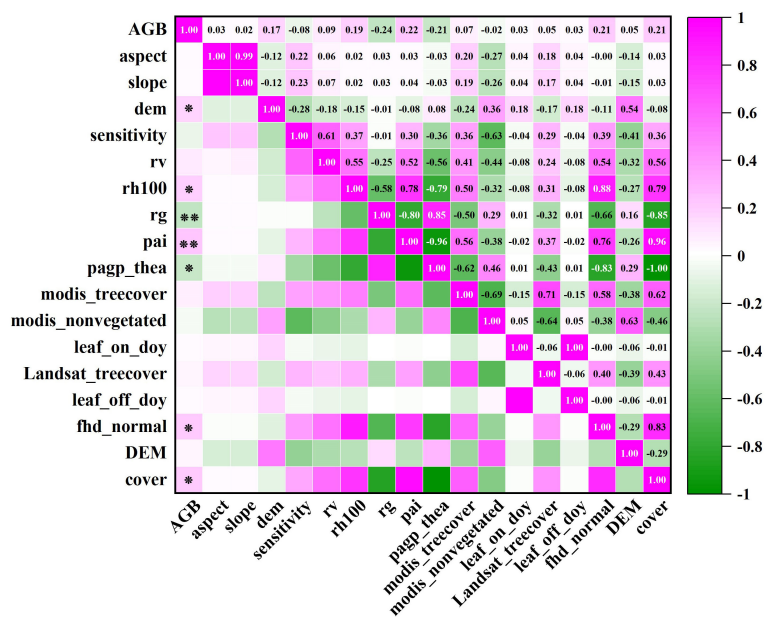


FIGURE 10 GEDI parameters and spruce-fir AGB correlation coefficient matrix. ‘**’ was significantly correlated at the 0.01 level (bilateral), and ‘*’ was significantly correlated at the 0.05 level (bilateral).

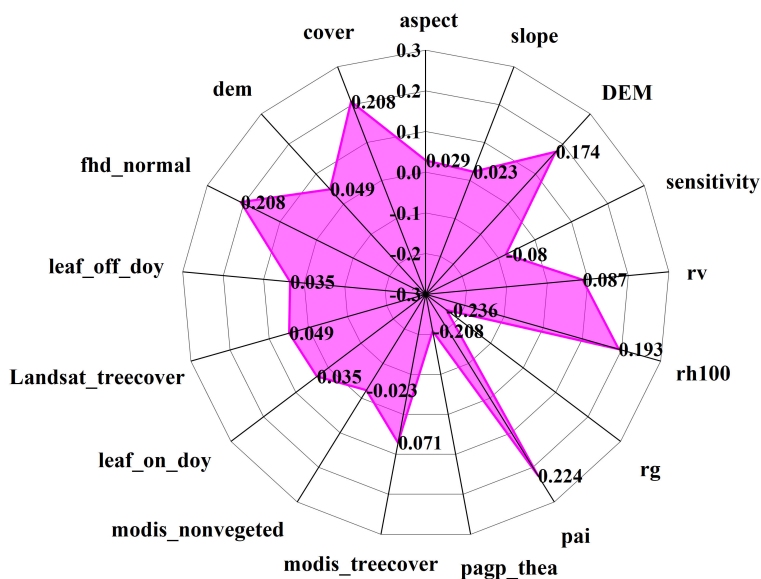


FIGURE 11 The correlation coefficient between GEDI parameters and AGB of spruce-fir.

5 Discussions

5.1 Effects of the distribution and number of GEDI footprint points on the interpolation results and biomass estimation

GEDI is the first space-borne LiDAR satellite dedicated to detecting the three-dimensional structure of vegetation. The beam emitted by GEDI can accurately obtain the vertical structure of vegetation. Its data record the energy returned by different trees at different heights on the ground, including structural information such as surface topography and canopy height, coverage and vertical profile indicators. Most of the previous studies on biomass estimation by GEDI are combined with multi-source remote sensing data extrapolation to obtain continuous forest biomass. For example, [Chen et al. \(2023\)](#) first estimated the biomass on the GEDI footprint point through random forest, and

then combined with the characteristic variables of Landsat 8 and Sentinel-1 to estimate the biomass of the entire study area. The ALS biomass extracted by [Dorado-Roda et al. \(2021\)](#) in the GEDI lens is used as an independent variable. The GEDI footprint contains a number of different echo indicators for subsequent GEDI data feature extraction and forest biomass model input. The ALS-derived biomass model is used to estimate the AGB of different forest ecosystems at the GEDI footprint level.

GEDI data are composed of discrete footprint points. After visualization, it can be seen that its distribution law is that the beam emitted along it is evenly distributed on the running track at a distance of 60 m. In our research, different echo indexes of footprints are extended to the surface by using interpolation method. Therefore, in order to verify the effect of interpolation of footprints with different amount and make the distribution of it random and uniform, representative footprints are extracted every 10, 30, 50, 70 and 100 shot when extracting GEDI data. The analysis shows that the interpolation results of the footprint (1309

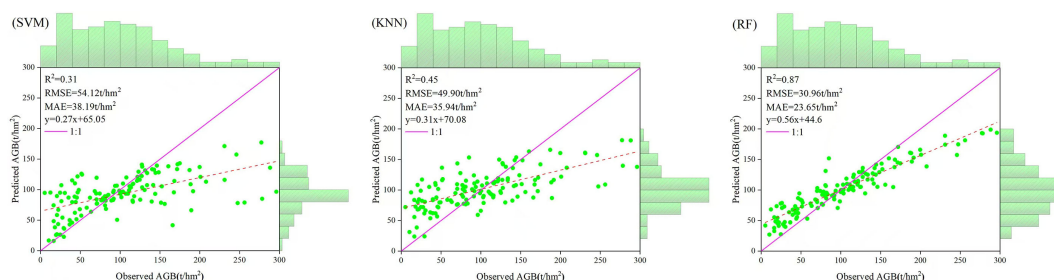
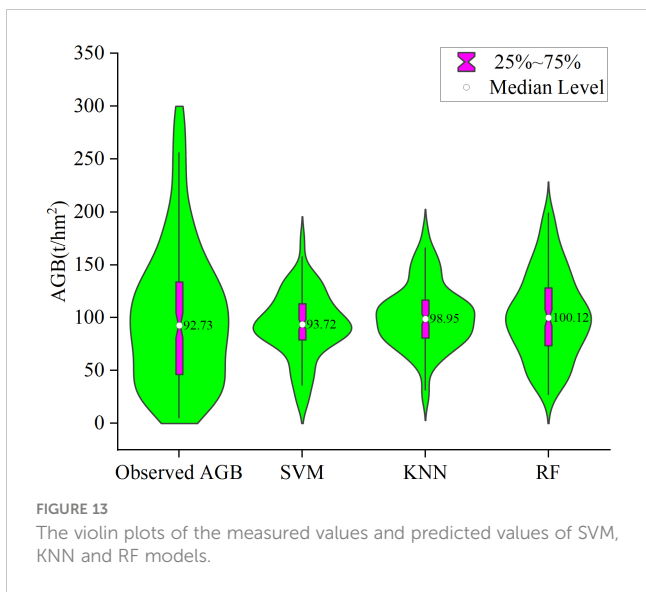


FIGURE 12 The scatter plots of measured values and predicted values of SVM, KNN and RF models.

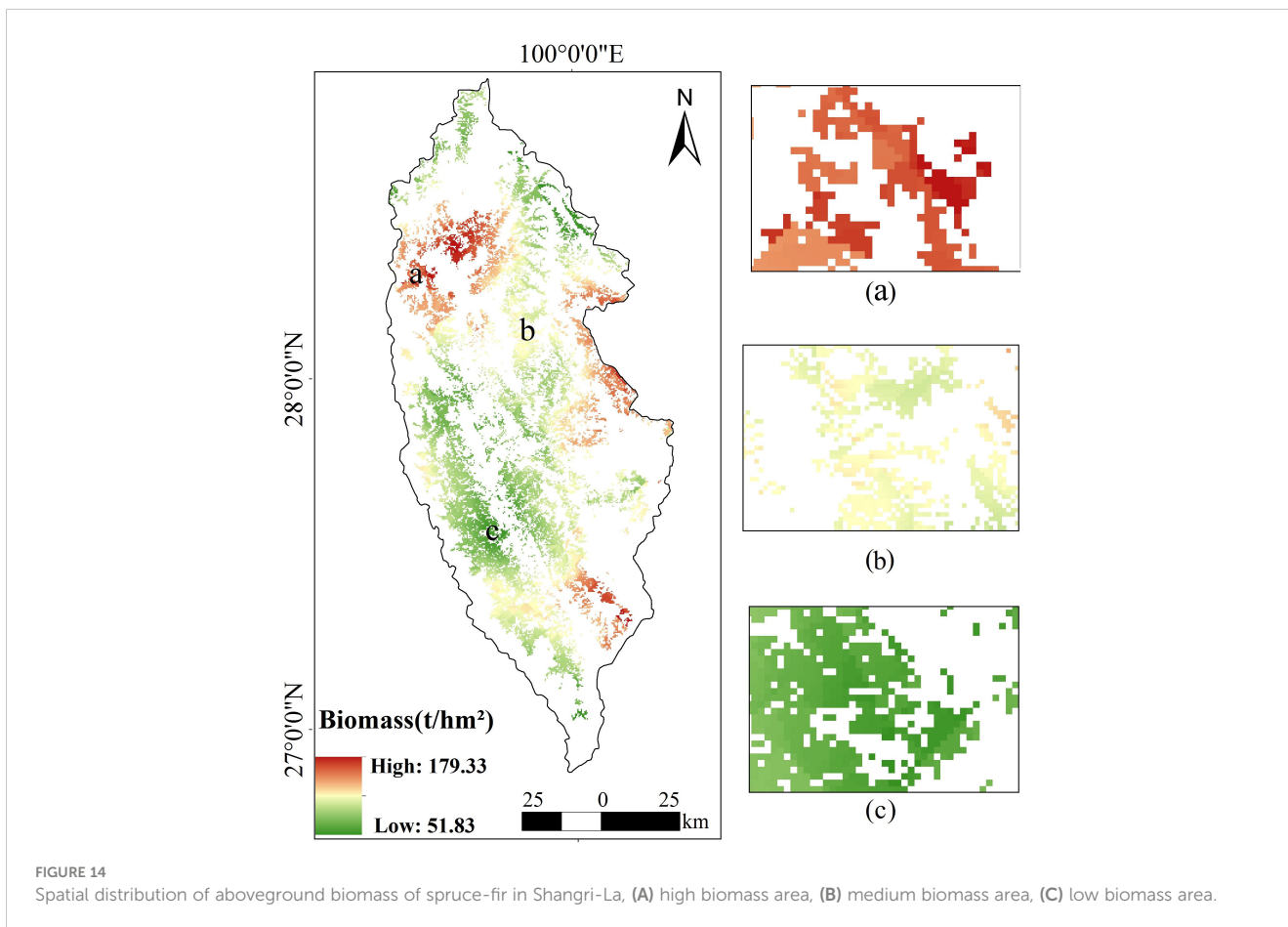


sample size has a significant impact on the uncertainty of the estimation of tree aboveground biomass at the regional scale. [Fu et al. \(2015\)](#) found that the influence of parameter variability of individual tree biomass estimation model caused by modeling sample size on the uncertainty of regional-scale tree aboveground biomass estimation increased with the decrease of modeling sample size, which eventually led to the increase of total uncertainty. Increasing the amount of modeling data can effectively improve the estimation accuracy, accuracy and efficiency of the biomass estimation model, and reduce the uncertainty. In this research, when dealing with footprint data, Python is used to randomly select its number so that it has randomness and distribution uncertainty, so as to disrupt its aggregation distribution phenomenon. In the follow-up study, we can refer to the sample optimization method proposed by [Shu et al. \(2022\)](#), which combine the variance function in geology with the value coefficient in value engineering, or the second-order and third-order spatial sampling methods to extract different numbers of footprint point data ([Duncanson et al., 2019](#)).

footprints) extracted every 100 shots are smoother than the other five results (the number of footprints is 13127, 4343, 2646, 1871, respectively). The fewer footprints there are the better the interpolation effect and the higher the prediction accuracy. It can be seen that the distribution and number of GEDI footprints will affect the spatial interpolation results, and also affect the subsequent biomass estimation. Previous studies have shown that the modeling

5.2 The influence of GEDI parameter selection on biomass estimation accuracy

In this study, we extracted 14 features from GEDI data that are highly correlated with the AGB of the Shangri-La spruce-fir, including canopy cover area, leaf area index, waveform return



ground energy value, etc. These features are used for subsequent random forests to estimate the explanatory variables of the AGB of the spruce-fir. Through correlation analysis, it was found that among all the variables, the variables with the highest correlation with the biomass of spruce-fir in Shangri-La were the integral of the ground component in the waveform (rg), leaf area index (pai), vegetation gap (pgap_thea), canopy cover (cover), leaf height diversity index (fhd_normal). These variables are related to vegetation structure and topography. Several variables with the highest correlation were input into the random forest model, and a higher inversion accuracy ($R^2 = 0.87$, RMSE = 30.96 t/hm², MAE=23.65t/hm²) was obtained. And Figure 14 shows that the biomass estimation results (3035.29×104 t/hm², 101.98 t/hm²) in the study area are similar to the results of Yang et al. (2021b) based on the second survey data of forest resources, using the biomass expansion factor method to calculate the biomass of Shangri-La spruce-fir (total value: 3665.9×104 t/hm², average value 113.02 t/hm²), indicating that the inversion algorithm based on the parameters provided by GEDI is feasible and the results are reliable.

The predicted biomass of spruce-fir in this study averaged 101.98 t/hm². The estimated biomass at lower positions generally exceeded that at higher positions. This disparity may be attributed to the inclusion of low shrubs and grasslands in the prediction at the sparsely-covered areas beneath the forest canopy, resulting in higher estimates, while the lower predictions at higher positions could be due to the limited penetration of GEDI spaceborne lidar in dense forest areas. Although the saturation of biomass cannot be completely eliminated, the addition of effective remote sensing features can reduce the impact of signal saturation (Qian et al., 2021). Subsequently, the GEDI data can be combined with other image data with reference to previous research methods to overcome the problem of high-value underestimation or low-value overestimation. For example, Xu et al. (2023b) combined Landsat 9 data and GEDI data to invert the carbon storage of Shangri-La Quercus; Silva et al. (2021) integrated GEDI, ICESat-2 and NISAR data to estimate regional-scale biomass, and obtained high estimation accuracy. In addition, although the increase in the number of independent variables makes the estimated biomass closer to the true value, it also reduces the general validity of the biomass model. Therefore, when constructing biomass models, the theories of other disciplines (ecology, biology, meteorology, etc.) should be cross-referenced, and the balance between statistical standards and practical application requirements should be considered (Tamiminia et al., 2021).

5.3 The potential and prospect of interpolation technology combined with machine learning model in using GEDI to predict future AGB on a large scale

The discrete footprint points of GEDI lead to the fact that the acquired data are not continuously distributed like optical remote sensing or SAR data. Therefore, most studies combine GEDI and multi-source remote sensing data for large-scale AGB prediction

(Kanmegne Tamga et al., 2022; Rodda et al., 2024). In this study, the GEDI footprint is taken as the research object, and the GEDI footprint points are interpolated to obtain the information in the whole research area. By focusing on exploring the influence of the number of different footprint points, that is, the number of sample locations, on the interpolation results. Using different light spot data, the inverse distance weight method is used to interpolate the parameters in the GEDI footprint. Our research results show that the number of light spots is different, and the accuracy of interpolation results is different. The less the number of interpolation spots is, the more dispersed the distribution is, and the closer the interpolation result is to the actual observation value. The machine learning model can capture the complex nonlinear relationship between AGB and explanatory variables through flexible model structure (Pham et al., 2018). Compared with statistical models, machine learning prediction is usually more accurate, so it has been widely used in forest AGB estimation research (Hopman et al., 2021). In this study, based on the interpolation of GEDI in the early stage, GEDI parameters were used as the dependent variable of the random forest model, and spruce-fir sample sites were used as independent variables. The R^2 of the model predicted by the random forest model was 0.88, RMSE was 30.96 t/hm², and MAE was 23.65 t/hm². The results show that the model has high accuracy. Previous studies have shown that machine learning methods have great potential and advantages in predicting future forest biomass (Li et al., 2020). For example, Liu et al. (2023c) comprehensively considered the contribution of forest factors, site conditions and meteorological factors in modeling, and used BP-ANN and SVM to establish a prediction model for future forest biomass based on sample plots. In the future research, the GEDI parameters can be continuous based on the interpolation method, so as to be added to the biomass model as independent variables together with stand factors, site conditions and meteorological factors, and then the future biomass changes can be predicted by machine learning methods.

5.4 Limitations and prospects

In GEDI observations, the reflection of a portion of the laser pulse from a flat bare land or something other than vegetation may lead to biased predictions of information at the footprint level, such as waveform that intersect with buildings, low clouds, steep slopes, rough terrain, or other topographic features, including vegetation and non-vegetation waveform (Bruening et al., 2023). The presence of steep slopes (with or without vegetation) or non-vegetated objects in the waveform footprint will change the value of the relative height index, and may also lead to geodesic errors. If many such observations are used in the mixed estimation, the estimated results may be very different from the unbiased independent reference data. Although the L2B algorithm has a built-in quality marker, when a waveform clearly does not represent surface conditions (such as clouds over the earth's surface), the GEDI algorithm cannot distinguish between waveform from forest canopy and those containing buildings, low clouds, or non-vegetation terrain features, such as rock outcrops, canyon walls, or steep

slopes. Therefore, it is necessary to identify these observations and remove them from the set used in the mixed AGB estimation. In this study, GEDI's three filtering conditions were used to eliminate outliers, but the footprint points at buildings, low clouds and non-vegetated terrains were not considered. Subsequent research can refer to Zhang et al.'s (Zhang et al., 2023) filtering of low-quality observation data that is not suitable for AGBD estimation through improved data filtering and GEDI-FIA fusion AGBD model, so as to accurately estimate AGB on the basis of ensuring the quantity and quality of GEDI footprint.

At the same time, the GEDI user manual mentions that although the second version of GEDI's data has improved its geo-positioning accuracy compared to the first version, there is still a position error of 10.2m. The uncertainty of GEDI's geographical location means that the GEDI's 25m diameter footprint will move within a range of 10.2m. Occasionally, this leads to GEDI data being obtained from adjacent but spatially disjoint regions of footprint locations reported by GEDI products. Therefore, caution will be taken in the use of GEDI data on canopy layers with spatial heterogeneity, such as small stand or forest land with fragmented distribution, and at the edge of the forest (Roy et al., 2021). However, in this study, the surface information is mainly obtained by interpolation using the attribute information in the GEDI footprint point. The geolocation error has little effect on this study, so it is not corrected for geolocation. However, some researchers have proved that the geolocation error of GEDI has little effect on the estimation results (Adam et al., 2020). Therefore, whether it is necessary to do geolocation correction for GEDI application scenarios is the focus and difficulty of the next research.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

LX: Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing. JY: Investigation, Methodology, Visualization, Writing – review & editing. QS:

Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – review & editing. SL: Data curation, Investigation, Visualization, Writing – review & editing. WZ: Data curation, Investigation, Visualization, Writing – review & editing. DD: Writing – review & editing, Validation.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The work was supported by the National Key Research and Development Program of China under Grant 2023YFD2201205; the Joint Agricultural Project of Yunnan Province under Grant 202301BD070001-002 and in the part of Yunnan Provincial Education Department Scientific Research Fund Project under Grant 2023Y0728.

Acknowledgments

The authors would like to thank NASA NSIDC for distributing the GEDI data (<https://search.earthdata.nasa.gov>, accessed on 16 October 2023), and the reviewers and members of the editorial team for their constructive comments. We thank the reviewers and the guest editor for having provided precious comments and suggestions that noticeably improved the quality of the paper.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

- Adam, M., Urbazaev, M., Dubois, C., and Schullius, D. (2020). Accuracy assessment of GEDI terrain elevation and canopy height estimates in European temperate forests: Influence of environmental and acquisition parameters. *Remote Sens.* 12, 3948. doi: 10.3390/rs12233948
- Assiri, M., Sartori, A., Persichetti, A., Miele, C., Faelga, R. A., and Blount, T. (2023). Leaf area index and aboveground biomass estimation of an alpine peatland with a UAV multi-sensor approach. *GI Sci. Remote Sens.* 60, 2270791. doi: 10.13292/j.1000-4890.2011.0274
- Bruening, J., May, P., Armston, J., and Dubayah, R. (2023). Precise and unbiased biomass estimation from GEDI data and the US forest inventory. *Front. Forests Global Change* 6. doi: 10.3389/fc.2023.1149153
- Cartus, O., Santoro, M., and Kellndorfer, J. (2012). Mapping Forest aboveground biomass in the Northeastern United States with ALOS PALSAR dual-polarization L-band. *Remote Sens. Environ.* 124, 466–478. doi: 10.1016/j.rse.2012.05.029
- Chen, W. S. (2023). Forest biomass estimation in alpine mountain based on GEDI: the case of Qilian Mountain National Park. Lanzhou University, Lan Zhou.

- Chen, Z., Sun, Z., Zhang, H., Zhang, H., and Qiu, H. (2023). Aboveground forest biomass estimation using tent mapping atom search optimized backpropagation neural network with Landsat 8 and Sentinel-1A data. *Remote Sens.* 15, 5653. doi: 10.3390/rs15245653
- Cheng, C., Du, W., Li, J., Bao, A., Ge, W., and Wang, S. (2023). Spatiotemporal variations of glacier mass balance in the tomur peak region based on multi-source altimetry remote sensing data. *Remote Sens.* 15, 4143. doi: 10.3390/rs15174143
- Comesaña-Cebral, L., Martínez-Sánchez, J., Lorenzo, H., and Arias, P. (2021). Individual tree segmentation method based on mobile backpack LiDAR point clouds. *Sensors* 21, 6007. doi: 10.3390/s21186007
- Dorado-Roda, I., Pascual, A., Godinho, S., Silva, C. A., Botequim, B., and Rodríguez-González, P. (2021). Assessing the accuracy of GEDI data for canopy height and aboveground biomass estimates in Mediterranean forests. *Remote Sens.* 13, 2279. doi: 10.3390/rs13122279
- Dubayah, R., Luthcke, S., Blair, J. B., Hofton, M., Armston, J., and Tang, H. (2020). *GEDI L1B geolocated waveform data global footprint level V001* (Washington, DC, USA: National Aeronautics and Space Administration).
- Dubayah, R., Luthcke, S., Blair, J. B., Hofton, M., Armston, J., and Tang, H. (2021). *GEDI L1B Geolocated Waveform Data Global Footprint Level V002* (NASA EOSDIS Land Processes DAAC) (Accessed 2022-10-08).
- Duncanson, L., Armston, J., Disney, M., Avitabile, V., Barbier, N., and Calders, K. (2019). The importance of consistent global forest aboveground biomass product validation. *Surv. Geophys.* 40, 979–999. doi: 10.1007/s10712-019-09538-8
- Francini, S., D'Amico, G., Vangi, E., Borghi, C., and Chirici, G. (2022). Integrating GEDI and Landsat: Spaceborne lidar and four decades of optical imagery for the analysis of forest disturbances and biomass changes in Italy. *Sensors* 22, 2015. doi: 10.3390/s22052015
- Fu, Y., Lei, Y., and Zeng, W. (2015). Uncertainty analysis for regional-level above-ground biomass estimates based on individual tree biomass model. *Acta Ecol. Sin.* 35, 7738–7747. doi: 10.5846/stxb201405130980
- Gao, L., Chai, G., and Zhang, X. (2022). Above-ground biomass estimation of plantation with different tree species using airborne lidar and hyperspectral data. *Remote Sens.* 14 (11), 2568. doi: 10.3390/rs14112568
- Guo, Q., Du, S., Jiang, J., Guo, W., Zhao, H., and Yan, X. (2023). Combining GEDI and sentinel data to estimate forest canopy mean height and aboveground biomass. *Ecol. Inf.* 78, 102348. doi: 10.1016/j.ecoinf.2023.102348
- Hang, M. H. (2022). Study on forest structure parameters inversion based on GEDI data. Forestry. University, Northeast.
- Hopman, H. J., Chan, S. M. S., Chu, W. C. W., Lu, H., Tse, C. Y., and Chau, S. W. H. (2021). Personalized prediction of transcranial magnetic stimulation clinical response in patients with treatment-refractory depression using neuroimaging biomarkers and machine learning. *J. Affect. Disord.* 290, 261–271. doi: 10.1016/j.jad.2021.04.081
- Huang, X., Kang, X., Lin, T. M., and Zheng, X. D. (2004). Grey Analysis of Natural Regeneration of Mixed Coniferous and Broadleaf Stands of Spruce and Fir. *For. Resour. Management.* (4), 28–30. doi: 10.13466/j.cnki.lyzygl.2004.04.008
- Jiang, F., Kutia, M., Ma, K., Chen, S., Long, J., and Sun, H. (2021). Estimating the aboveground biomass of coniferous forest in Northeast China using spectral variables, land surface temperature and soil moisture. *Sci. Total Environ.* 785, 147335. doi: 10.1016/j.scitotenv.2021.147335
- Kanmegne Tamga, D., Latifi, H., Ullmann, T., Baumhauer, R., Bayala, J., and Thiel, M. (2022). Estimation of aboveground biomass in agroforestry systems over three climatic regions in West Africa using sentinel-1, sentinel-2, ALOS, and GEDI data. *Sensors* 23, 349. doi: 10.3390/S23010349
- Lang, X., Yu X. H., Shu, Q., T., Zhang, Z., Xie, F., and Zi, L. (2019). *Nonparametric Model for Remote Sensing Estimating the Volume of Spruce-Fir Forest in Shangri-La Vol.* 39 (Kun Ming: Southwest forestry university), 146–151. doi: 10.11929/j.swfu.201811019
- Li, Y., Li, M., Li, C., and Liu, Z. (2020). Forest aboveground biomass estimation using Landsat 8 and Sentinel-1A data with machine learning algorithms. *Sci. Rep.* 10, 9952. doi: 10.1038/s41598-020-67024-3
- Lin, X., Shang, R., Chen, J. M., Zhao, X., Huang, Y., and Yu, G. (2023). High-resolution Forest age mapping based on forest height maps derived from GEDI and ICESat-2 space-borne lidar data. *Agric. For. Meteorol.* 339, 109592. doi: 10.1016/j.agrformet.2023.109592
- Liu, B., Bu, W., and Zang, R. (2023a). Improved allometric models to estimate the aboveground biomass of younger secondary tropical forests. *Global Ecol. Conserv.* 41, e02359. doi: 10.1016/j.gecco.2022.e02359
- Liu, C., Hu, R., Wang, Y., Lin, H., Wu, D., and Dai, Y. (2023b). Integrating ICESat-2 laser altimeter observations and hydrological modeling for enhanced prediction of climate-driven lake level change. *J. Hydrol.* 626, 130304. doi: 10.1016/j.jhydrol.2023.130304
- Liu, L., Wang, C., Nie, S., Zhu, X. X., Xi, X. H., and Wang, J. L. (2022). Analysis of the influence of different algorithms of GEDI L2A on the accuracy of ground elevation and forest canopy height. *J. Univ. Chin. Acad. Sci.* 39, 502–511. doi: 10.7523/j.ucas.2021.0076
- Liu, J., Yue, C., Pei, C., Li, X., and Zhang, Q. (2023c). Prediction of regional forest biomass using machine learning: A case study of Beijing, China. *Forests* 14, 1008. doi: 10.3390/f14051008
- Lourenço, P., Godinho, S., Sousa, A., and Gonçalves, A. C. (2021). Estimating tree aboveground biomass using multispectral satellite-based data in Mediterranean agroforestry system using random forest algorithm. *Remote Sens. Applic.: Soc. Environ.* 23, 100560. doi: 10.1016/j.rsase.2021.100560
- Luo, Y., Qi, S., Liao, K., Zhang, S., Hu, B., and Tian, Y. (2023). Mapping the forest height by fusion of ICESat-2 and multi-source remote sensing imagery and topographic information: A case study in Jiangxi province, China. *Forests* 14, 454. doi: 10.3390/f14030454
- Petty, A. A., Bagnardi, M., Kurtz, N. T., Tilling, R., Fons, S., and Armitage, T. (2021). Assessment of ICESat-2 sea ice surface classification with Sentinel-2 imagery: Implications for freeboard and new estimates of lead and floe geometry. *Earth Space Sci.* 8, e2020EA001491. doi: 10.1029/2020ea001491
- Pham, T. D., Yoshino, K., Le, N. N., and Bui, D. T. (2018). Estimating aboveground biomass of mangrove plantation on the northern coast of Vietnam using machine learning techniques with an integration of ALOS-2 PALSAR-2 and Sentinel-2A data. *Int. J. Remote Sens.* 39, 7761–7788. doi: 10.1080/01431161.2018.1471544
- Potapov, P., Li, X., Hernandez-Serna, A., Tyukavina, A., Hansen, M. C., and Komareddy, A. (2021). Mapping global forest canopy height through integration of GEDI and Landsat data. *Remote Sens. Environ.* 253, 112165. doi: 10.1016/j.rse.2020.112165
- Qian, C., Qiang, H., Wang, F., and Li, M. (2021). Estimation of forest aboveground biomass in karst areas using multi-source remote sensing data and the K-DBN algorithm. *Remote Sens.* 13, 5030. doi: 10.3390/rs13245030
- Qiu, A., Yang, Y., Wang, D., Xu, S., and Wang, X. (2020). Exploring parameter selection for carbon monitoring based on Landsat-8 imagery of the aboveground forest biomass on Mount Tai. *Eur. J. Remote Sens.* 53, 4–15. doi: 10.1080/22797254.2019.1686717
- Rodda, S. R., Nidamanuri, R. R., Fararoda, R., Mayamanikandan, T., and Rajashekar, G. (2024). Evaluation of height metrics and above-ground biomass density from GEDI and ICESat-2 over Indian tropical dry forests using airborne LiDAR data. *J. Indian Soc. Remote Sens.* 52, 841–856. doi: 10.1007/s12524-023-01693-1
- Roy, D. P., Kashongwe, H. B., and Armston, J. (2021). The impact of geolocation uncertainty on GEDI tropical forest canopy height estimation and change monitoring. *Sci. Remote Sens.* 4, 100024. doi: 10.1016/j.srs.2021.100024
- Saarela, S., Holm, S., Healey, S. P., Patterson, P. L., Yang, Z., and Andersen, H. (2022). Comparing frameworks for biomass prediction for the Global Ecosystem Dynamics Investigation. *Remote Sens. Environ.* 278, 113074. doi: 10.1016/j.rse.2022.113074
- Schwartz, M., Ciaisi, P., Ottlé, C., Truchis, A. D., Vega, C., and Fayad, I. (2024). High-resolution canopy height map in the Landes Forest (France) based on GEDI, Sentinel-1, and Sentinel-2 data with a deep learning approach. *Int. J. Appl. Earth Observ. Geoinform.* 128, 103711. doi: 10.1016/j.jag.2024.103711
- Shu, Q., Xi, L., Wang, K., Xie, F., Pang, Y., and Song, H. (2022). Optimization of samples for remote sensing estimation of forest aboveground biomass at the regional scale. *Remote Sens.* 14, 4187. doi: 10.3390/rs14174187
- Silva, C. A., Duncanson, L., Hancock, S., Neuenschwander, A., Thomas, N., and Hofton, M. (2021). Fusing simulated GEDI, ICESat-2 and NISAR data for regional aboveground biomass mapping. *Remote Sens. Environ.* 253, 112234. doi: 10.1016/j.rse.2020.112234
- Song, H., Xi, L., Shu, Q., Wei, Z., and Qiu, S. (2022). Estimate Forest aboveground biomass of mountain by ICESat-2/ATLAS data interacting cokriging. *Forests* 14, 13. doi: 10.3390/f14010013
- Tamimnia, H., Salehi, B., Mahdianpari, M., Beier, C. M., Johnson, L., and Phoenix, D. B. (2021). A comparison of decision tree-based models for forest above-ground biomass estimation using a combination of airborne lidar and landsat data. *ISPRS Ann. photogram. Remote Sens. spatial Inf. Sci.* 3, 235–241. doi: 10.5194/isprs-annals-V-3-2021-235-2021
- Tamimnia, H., Salehi, B., Mahdianpari, M., and Goulden, T. (2024). State-wide Forest canopy height and aboveground biomass map for New York with 10 m resolution, integrating GEDI, Sentinel-1, and Sentinel-2 data. *Ecol. Inf.* 79, 102404. doi: 10.1016/j.ecoinf.2023.102404
- Tian, H. C. (2023). Remote sensing estimation of aboveground biomass in subtropical forests based on active remote sensing data, passive remote sensing data, and spaceborne LiDAR data. NanJing Forestry. Nan Jing University.
- Vaithyanathan, D., and Sudalaimuthu, K. (2022). Area-to-point regression Kriging approach fusion of Landsat 8 OLI and Sentinel 2 data for assessment of soil macronutrients at Anaimalai, Coimbatore. *Environ. Monit. Assess.* 194, 916. doi: 10.1007/s10661-022-10571-1
- Wang, J., Du, H., Li, X., Mao, F., Zhang, M., and Liu, E. (2021). Remote sensing estimation of bamboo forest aboveground biomass based on geographically weighted regression. *Remote Sens.* 13, 2962. doi: 10.3390/rs13152962
- Wang, J. H., Li, F. R., and Dong, L. H. (2018). Additive aboveground biomass equations based on different predictors for natural Tilia Linn. *J. Appl. Ecol.* 29, 3685–3695. doi: 10.13287/j.1001-9332.201811.020
- Wang, S., Liu, C., Li, W., Jia, S., and Yue, H. (2023). Hybrid model for estimating forest canopy heights using fused multimodal spaceborne LiDAR data and optical imagery. *Int. J. Appl. Earth Observ. Geoinform.* 122, 103431. doi: 10.1016/j.jag.2023.103431
- Wei, C., Qin, H., Li, W., Wang, W., Hua, Y., and Yao, Y. (2024). Estimating aboveground biomass of urban trees based on ICESat-2 LiDAR and Zhuhai-1 hyperspectral data. *Phys. Chem. Earth* 135, 103605. doi: 10.1016/j.pce.2024.103605
- Wu, Z., Yao, F., Zhang, J., and Liu, H. (2024). Estimating forest aboveground biomass using a combination of geographical random forest and empirical bayesian kriging models. *Remote Sens.* 16, 1859. doi: 10.3390/rs16111859

- Xie, D. P., Li, G. Y., Zhao, Y. M., Yang, X. D., Tang, X. M., and Fu, A. M. (2018). U.S. GEDI Space-based Laser Altimetry System and its Application. *Space Int.* (12), 39–44.
- Xu, L., Lai, H., Yu, J., Luo, S., Guo, C., and Gao, Y. (2023a). Carbon storage estimation of quercus aquifolioides based on GEDI spaceborne LiDAR data and landsat 9 images in Shangri-La. *Sustainability* 15, 11525. doi: 10.3390/su151511525
- Xu, L., Shu, Q., Fu, H., Zhou, W., Luo, S., and Gao, Y. (2023b). Estimation of Quercus biomass in Shangri-La based on GEDI spaceborne LiDAR data. *Forests* 14, 876. doi: 10.3390/f14050876
- Yang, H., Li, F., Wang, W., and Yu, K. (2021a). Estimating above-ground biomass of potato using random forest and optimized hyperspectral indices. *Remote Sens.* 13, 2339. doi: 10.3390/rs13122339
- Yang, Y., Wang, R., and Xu, H. (2021b). Optimal design of second-order sampling for forest biomass in Shangri-La city based on the forest management inventory. *J. Southw. Forest.* 41, 160–167. doi: 10.11929/j.swfu.202012009
- Yuan, H. Z. (2022). Research on forest height retrieval and monitoring method using space-borne LiDAR data. Hennan Polytech. Kai Feng University.
- Yun, T., Li, J., Ma, L., Zhou, J., Wang, R., and Eichhorn, M. P. (2024). Status, advancements and prospects of deep learning methods applied in forest studies. *Int. J. Appl. Earth Observ. Geoinform.* 131, 103938. doi: 10.1016/j.jag.2024.103938
- Zeng, W., Li, L., Xiao, Q., Dang, Y., Ma, K., and Sun, J. (2016a). *Tree Biomass Models and Related Parameters to Carbon Accounting for Picea, LY/T 2656-2016* (Beijing, China: Standards Press of China).
- Zeng, W., Li, L., Zhou, X., Wang, X., Ma, K., and Gan, S. (2016b). *Tree Biomass Models and Related Parameters to Carbon Accounting for Abies, LY/T 2655-2016* (Beijing, China: Standards Press of China).
- Zhang, G., Niu, L., Wu, M., Kaufmann, H., Li, H., and He, Y. (2023). Ecological impact patterns and temporal cycles of green tide biomass in the settlement region: based on time-series remote sensing and *in situ* data. *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.* 17), 1610–1622. doi: 10.1109/jstars.2023.3338979
- Zhang, J., Tian, J., Li, X., Wang, L., Chen, B., and Gong, H. (2021). Leaf area index retrieval with ICESat-2 photon counting LiDAR. *Int. J. Appl. Earth Observ. Geoinform.* 103, 102488. doi: 10.1016/j.jag.2021.102488
- Zhang, W., Zhao, L., Li, Y., Shi, J., Yan, M., and Ji, Y. (2022). Forest above-ground biomass inversion using optical and SAR images based on a multi-step feature optimized inversion model. *Remote Sens.* 14, 1608. doi: 10.3390/rs14071608
- Zhu, X. X. (2021). *Forest height retrieval of chine with a resolution of 30 m using ICESat-2 and GEDI data* (Bei Jing: Chinese Academy of Sciences).