



Algorithm Theoretical Basis Document (ATBD) for GEDI L4D Imputed Waveforms

Eugene Seo¹, Sean Healey², Zhiqiang Yang², John Armston³

1. Oregon State University, Corvallis, OR
2. US Forest Service, Ogden, UT
3. University of Maryland, College Park, MD

Version 2.0
Release date:
Goddard Space Flight Center, Greenbelt, MD

Authors:

Eugene Seo

Sean Healey

Zhiqiang Yang

John Armston

Principal Investigator:

Ralph Dubayah

Abstract

The Global Ecosystem Dynamics Investigation (GEDI) produces high-resolution laser ranging observations of Earth's three-dimensional structure, providing precise measurements of forest vertical profiles. These observations are processed to generate derived products, including relative height (the height above ground at which a given percentage of lidar energy is returned; GEDI L2A product), canopy cover (GEDI L2B), and aboveground biomass density (AGBD; GEDI L4A). These products are essential for advancing understanding of carbon and water cycling, biodiversity, and habitat processes. Over its operation since 2019, GEDI has collected tens of billions of individual laser shots spanning latitudes 51.6°N to 51.6°S. Despite this extensive coverage, the primary measurements represent only a fraction of Earth's surface in continuous space, which limits their direct use for local mapping with field measurements. Consequently, products that spatially extend GEDI observations at high resolution are desirable. The GEDI Level 4D (L4D) product provides imputed GEDI waveforms, along with canopy cover and aboveground biomass density (AGBD), for the year 2023 at a 30×30 m resolution globally between latitudes -51.6° and 51.6° . Imputation is the process of filling unobserved data (i.e., unmeasured locations) with substituted values estimated from available information. This document describes the theoretical basis of the imputation algorithm used to produce the GEDI L4D data product. The key objective in developing the L4D data product is to preserve the distribution of observed data within ecologically meaningful areas. To achieve this, the single nearest neighbor algorithm (SNN) was applied within 10 km grid cells globally. For every 30 m pixel within the grid, the SNN model selects the single measured waveform (and associated cover and biomass variable) that most closely matches the pixel across feature variables developed from Landsat time series and topography. The performance of imputation is evaluated on a randomly selected 20% holdout of high-quality GEDI observations. Standard uncertainty products relate both the pixel-level RMSE of key imputed variables and the statistical similarity of the distribution of predictions to the distribution of actual GEDI observations at the 10km scale. The SNN imputation approach used to generate GEDI's L4D product emphasizes maintenance of the distribution and covariance of structure of variables observed directly by GEDI's lidar sample.

Foreword

This document is the Algorithm Theoretical Basis Document for the GEDI Waveform Imputation for L4D Products. The first L4D product release is labeled Version 2.1 to maintain versioning consistency with the Version 2.1 L2 products that are imputed. The GEDI Science Team and Science Operations Center team assumes responsibility for this document and updates it, as required, as algorithms are refined. Reviews of this document are performed when appropriate and as needed updates to this document are made.

This document is a GEDI ATBD controlled document. Changes to this document require prior approval of the project. Proposed changes shall be noted in the change log, as well as incrementing the document version number.

Questions or comments concerning this document should be addressed to:

Eugene Seo
Oregon State University, Corvallis OR 97331
seoe@oregonstate.edu
+1 (541) 829 2562

Sean Healey
US Forest Service, Riverdale UT 84405
sean.healey@usda.gov
+1 (801) 391 7536

Zhiqiang Yang
US Forest Service, Riverdale, UT 84405
zhiqiang.yang@usda.gov
+1 (541) 750 7491

John Armston
University of Maryland, MD 20742
armston@umd.edu
+1 (301) 405 8444

Change History Log

Revision Level	Description of Change	Date Approved
2.0	Initial version	

Table of Contents

Abstract	2
Foreword	3
Change History Log	4
Table of Contents	5
List of Tables	7
List of Figures	8
1.0 INTRODUCTION	10
1.1 GEDI Data Products Overview	10
1.2 GEDI Configuration	11
1.3 Document Overview and Objective	12
1.4 Related Documentation	12
1.4.1 Parent Documents	12
1.4.2 Applicable Documents.....	12
2.0 THEORETICAL BACKGROUND	12
3.0 GEDI WAVEFORM IMPUTATION ALGORITHM	14
3.1 <i>k</i>-Nearest Neighbor Algorithm	14
3.2 Geospatial Domain of Single Nearest Neighbor Models	16
3.3 Environmental Covariates for GEDI Waveforms	18
3.4 High-Quality GEDI Waveforms as Training Data	19
3.4.1 GEDI Level-2A Shot Quality Filtering.....	19
3.4.2 GEDI Level-4B Granule Quality Filtering	19
3.4.3 Filtering GEDI Shots Outside Grid Tiles.....	19
3.4.4 Filtering GEDI Shots Without Covariates	20
3.4.5 Local Outlier Detection-Based Filtering.....	20
3.5 <i>k</i>-Nearest Neighbor Distance Metric	22
3.6 GEDI Algorithm Setting Group Selection	22
4.0 IMPUTED GEDI WAVEFORMS EVALUATION	22
4.1 Validation GEDI Waveform Samples	23
4.2 Evaluation Metrics	23
4.2.1 Root Mean Square Error (RMSE).....	23

4.2.2 Kolmogorov–Smirnov Test (KS Test).....	23
4.3 Validation Results	24
5.0 EXAMPLES OF GEDI WAVEFORM IMPUTATION MAPS	24
6.0 REFERENCES.....	26
GLOSSARY/ACRONYMS.....	27

List of Tables

Table 1. GEDI Data Products.....	10
Table 2. CCDC-Driven Covariates for GEDL Waveform Imputation.....	18
Table 3. Scenarios addressed by the Local Outlier Detection Algorithm.....	20

List of Figures

Figure 1. GEDI beam ground-track configuration..... 11

Figure 2. The distribution of RH98 values (Right) from qualified GEDI waveforms within UTM zone 10 (Left). The distribution of heights measured by GEDI footprints in this area is not normally distributed, and it contains an ecologically important tail of extremely tall trees..... 13

Figure 3. (Left) The height distribution of 2,979 randomly selected GEDI waveforms (101 discrete RH metrics). Sample area (blue outline) from which these waveforms were selected in Olympic National Park, Washington State, USA. (Right). “Elevation” is the height above ground, as represented in GEDI L2A products..... 13

Figure 4. The distribution of predicted waveforms (shown here as a distribution of heights from RH0 to RH100) diverges from the distribution observed by GEDI (left) as more neighbors are aggregated to determine each value of relative height (i.e., as k increases from 1 to 5 to 10 neighbors)..... 14

Figure 5. The distribution of predicted RH98 values for the study area shown in Figure 1 at three values of k (orange). The distribution of GEDI measurements of RH98 is shown in blue for reference. As more “neighbors” are averaged, predictions tend toward the mean. Kolmogorov–Smirnov (KS) Test evaluates the similarity between the observed distribution and the predicted distribution. The test statistic represents the maximum distance between the cumulative distribution functions; the p -value shows whether this difference is likely due to chance. A smaller p -value suggests the difference is unlikely to be random, implying the distributions are probably not the same..... 15

Figure 6. Scatter plots comparing observed GEDI RH98 values with k -NN predictions ($k=1, 5, 10$) in the study area. Increasing k lowers RMSE but compresses values toward the mean, leading to overestimation of short trees and underestimation of tall trees..... 15

Figure 7. RMSE and KS-Test results for the study area displayed in Figure 2. While increasing k reduces RMSE, only imputation using the single-nearest neighbor ($k=1$) produced a predicted distribution of RH98 that was not statistically different (i.e., $p > 0.05$) from the distribution measured in an independent sample of GEDI shots..... 16

Figure 8. Example of 10 km x 10 km tiles across Oregon, USA 16

Figure 9. Illustration of a focal tile (yellow) and its support tiles (orange) with a tile buffer of 1. 17

Figure 10. Imputation maps for two adjacent 10 x 10 km modeling tiles. (Top) Using training data exclusively from the focal tile produces visible predictions discontinuities at tile boundaries. (Bottom) Including training data from both the focal tile and its adjacent tiles yields a smoother transition..... 17

Figure 11. Illustration of cases for identifying outlier GEDI waveforms using the RH98 metric of a target tree as a proxy for tree height, alongside contextual information from surrounding tree heights..... 21

Figure 12. Global map of root mean square error (RMSE) for the RH98 metric, evaluated using the validation dataset. The color scale is capped at 10 m..... 24

Figure 13. Global map showing the results of the Kolmogorov–Smirnov (KS) test comparing observed and imputed RH98 distributions. Using a significance level of 0.05, 70.8% of tiles show statistically similar distributions, while the remaining 29.2% exhibit significant different, indicating that the imputed RH98 distribution deviates from the observed distribution in those regions. Both test outcomes are broadly distributed across the globe..... 24

Figure 14. Global map showing imputed RH98 metrics..... 25

Figure 15. Illustration of imputed waveforms represented by 12 RH metrics (left) with ancillary bands for canopy cover (middle) and aboveground biomass density (right) within a 10 km x 10 km tile..... 25

1.0 INTRODUCTION

1.1 GEDI Data Products Overview

The GEDI data products are noted in Table 1. The GEDI Level 1 data products are developed in two separate products, a Level 1A (L1A) and a Level 1B (L1B) product. The GEDI L1A data product contains fundamental instrument engineering and housekeeping data as well as the raw waveform and geolocation information used to compute higher level data products. The GEDI L1B geolocated waveform data product, while similar to the L1A data product, contains specific data to support the computation of the higher level 2A and 2B data products. These L1B data include the corrected receive waveform, as well as the receive waveform geolocation information. Level 2 (L2) products contain ground and vegetation metrics derived from the L1 data. Level 3 (L3) products are grids of some of those L2 metrics. Level 4A (L4A) products provide footprint-level aboveground biomass density (AGBD) measured using L2A relative height (RH) metrics. Level 4B (L4B) products are gridded above ground biomass density derived from L4A data. Level 4C (L4C) products offer footprint level waveform structural complexity index (WSCI). Level 4D (L4D) products contain imputed GEDI metrics at 30-meter resolution, including RH metrics (L2A), gridded canopy cover (L2B), footprint-level AGBD (L4A).

Table 1. GEDI Data Products

Product	Description	Resolution	Archive Site
Level 1B	Geolocated Waveforms	25 m diameter	LP DAAC
Level 2A	Elevation and Relative Height Metrics	25 m diameter	LP DAAC
Level 2B	Canopy Cover and Vertical Profile Metrics	25 m diameter	LP DAAC
Level 3	Gridded Land Surface Metrics	1 km grid	ORNL DAAC
Level 4A	Footprint Aboveground Biomass Density	25 m diameter	ORNL DAAC
Level 4B	Gridded Above ground Biomass Density	1 km diameter	ORNL DAAC
Level 4C	Waveform Structural Complexity Index	25 m diameter	ORNL DAAC

Level 4D	Imputed Relative Height, Canopy Cover, and Aboveground Biomass Density	30 m grid	ORNL DAAC
----------	--	-----------	-----------

1.2 GEDI Configuration

The GEDI instrument is a geodetic-class, full-waveform light detection and ranging (lidar) laser system comprised of 3 lasers producing a total of eight beam ground transects that are spaced approximately 600 m apart on the Earth’s surface in the cross-track direction relative to the flight direction, and approximately 735 m of zonal (parallel to lines of latitude) spacing. Each beam transect consists of ~25 m footprint (surface return) samples approximately spaced every 60 m along track. The “coverage” laser is split into two transects that are then each dithered producing four ground transects. The other two lasers are dithered only, producing two ground transects each. The configuration of the ground tracks is shown in Figure 1. The ranging points from each footprint’s waveform are geolocated to produce geolocation data groups (“geolocation” and “geophys_corr”) provided in the L1 and L2 data products.

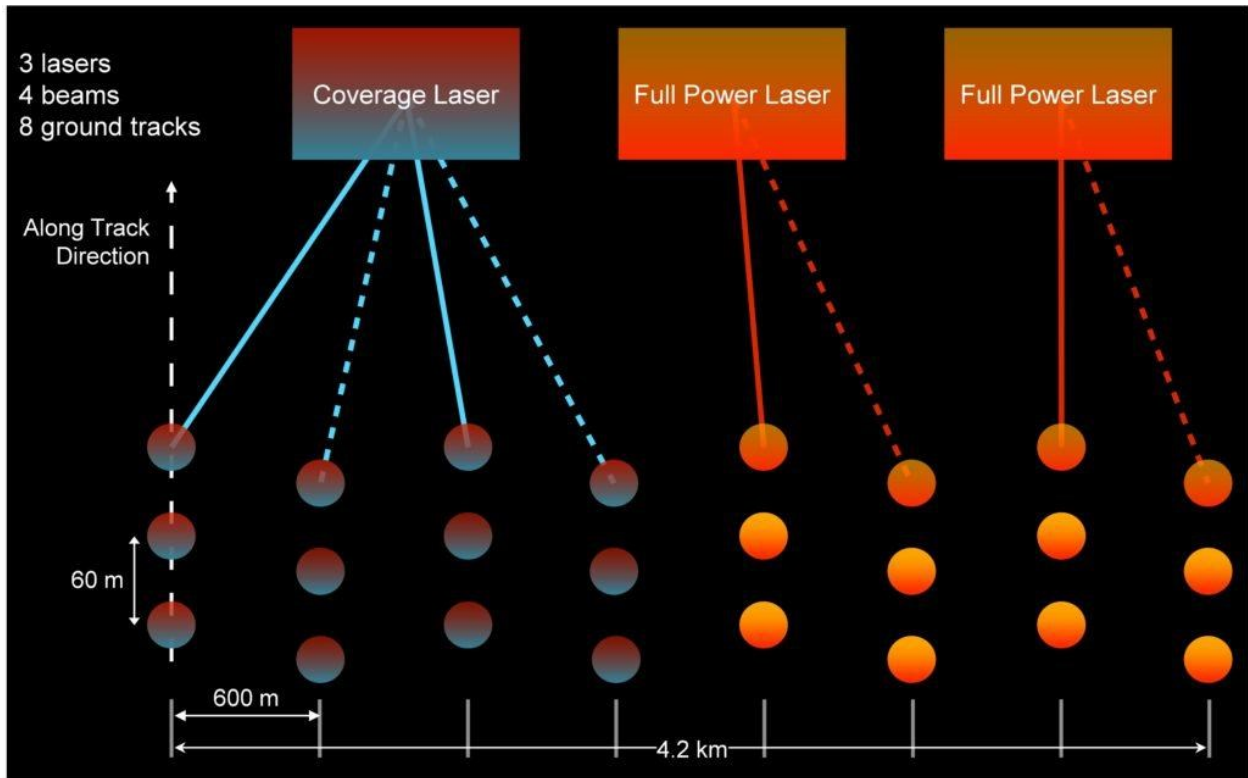


Figure 1. GEDI beam ground-track configuration

1.3 Document Overview and Objective

This document provides an overview of the algorithms and procedures required to provide GEDI Level 4D imputing waveforms, canopy cover, and AGBD at 30-meter resolution. The imputed waveforms are represented in a sequence of selected 11 RH metrics (e.g., RH 10, RH 20, ..., RH 95, RH 98).

This document is arranged in the following manner:

- **Section 1** provides a brief introduction and related documentation
- **Section 2** outlines key considerations for imputing GEDI waveforms
- **Section 3** details the GEDI waveform imputation algorithm
- **Section 4** presents the evaluation results of the imputation algorithm
- **Section 5** illustrates examples of imputed products
- **Section 6** lists the references
- An acronym glossary can be found at the end of this document

1.4 Related Documentation

Related documents include parent documents and applicable documents, and information documents, including and L4D User Guide.

1.4.1 Parent Documents

- GEDI Science Data Management Plan

1.4.2 Applicable Documents

- GEDI ATBD for GEDI Waveform Geolocation for L1 and L2 Products
- GEDI ATBD for Footprint Canopy Cover and Vertical Profile Metrics
- GEDI ATBD for GEDI Footprint Above Ground Biomass Density
- GEDI L2A Product Data Dictionary ([gedi_l2a_product_data_dictionary.html](#))
- GEDI L2B Product Data Dictionary ([gedi_l2b_product_data_dictionary.html](#))

2.0 THEORETICAL BACKGROUND

Since GEDI is a sampling instrument, it provides no direct measurements for much of the Earth's surface. May et al. (2024, 2025) addressed this limitation by applying spatial interpolation to the collocated GEDI footprints and National Forest Inventory (NFI) plot locations to generate wall-to-wall GEDI waveform predictions across the contiguous United States. Pursuing a similar goal at the global scale, the GEDI L4D Waveform Imputation product is comprised of global predictions (also called imputations here) of a high-quality GEDI shot identified using multiple filtering metrics (see Section 3.4 for details). The data was collected nearby for each 30m pixel across the latitudes where GEDI operates. The imputed waveforms are derived from qualified GEDI waveforms collected between 2019-04-18 to 2023-03-16. A key design criterion for the imputed waveforms is ensuring that the distribution of heights within each RH metric and the covariance among the observed RH metrics, as measured in GEDI's sample, are preserved.

Applications such as GEDI Demonstration Products derived from the Ecosystem Demography model rely upon these emergent properties of map accuracy as much as they rely upon low average errors. Even though the distribution of forest structure can be irregular with a large dynamic range (Figure 2 & 3), many common approaches such as random forests designed to minimize residual prediction error also return simplified modeled populations with few extreme values. For the GEDI L4D Waveform Imputation product, the single nearest neighbor (SNN) technique described below was chosen to achieve our primary goal of conserving the population distribution of measured values. As a non-parametric supervised learning method that makes no assumptions about the underlying data distribution, using a single nearest neighbor minimizes bias and better preserves the true distributional characteristics of the original data, though this comes at the cost of higher variance and reduced accuracy in noisy conditions (Hastie et al., 2009; Hudak et al., 2008). Since all waveform properties (including all shot level L2 and L4) are inherited together, this approach maintains realistic covariance between imputed properties.

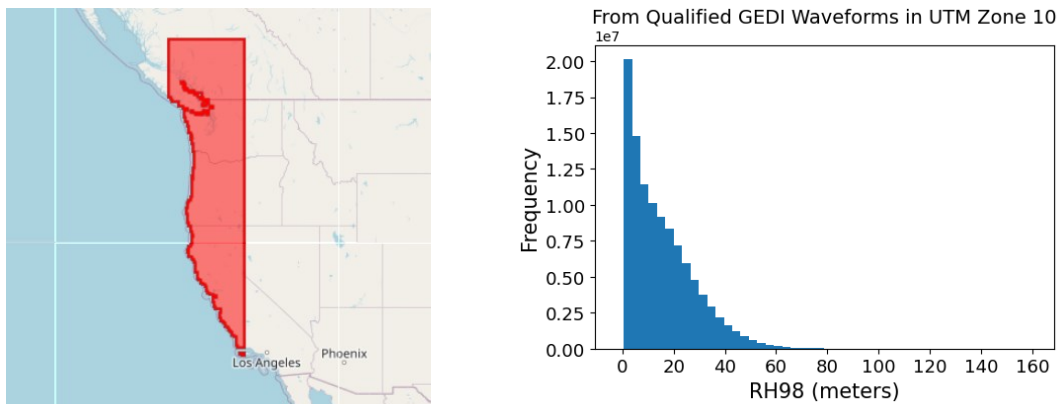


Figure 2. The distribution of RH98 values (Right) from qualified GEDI waveforms within UTM zone 10 (Left). The distribution of heights measured by GEDI footprints in this area is not normally distributed, and it contains an ecologically important tail of extremely tall trees.

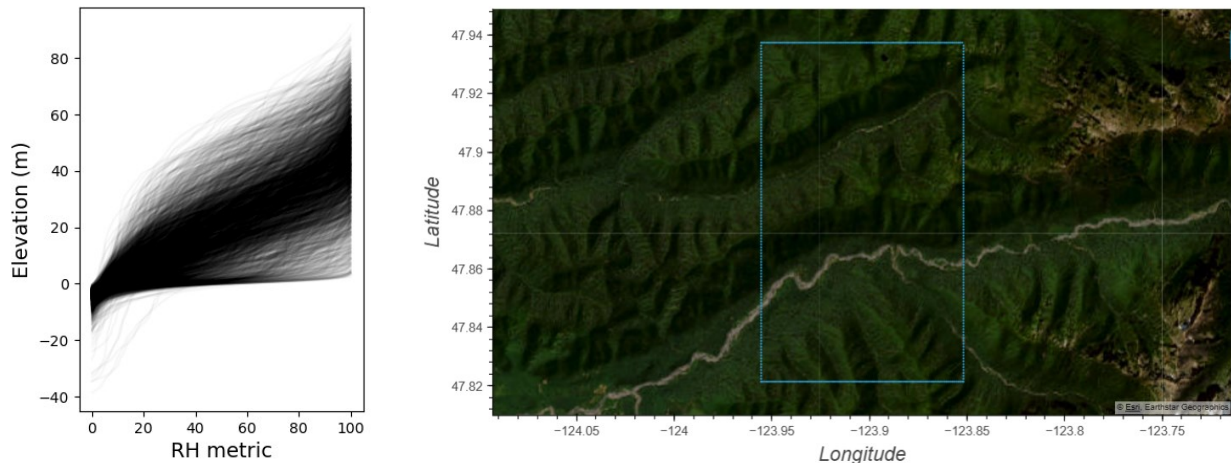


Figure 3. (Left) The height distribution of 2,979 randomly selected GEDI waveforms (101 discrete RH metrics). Sample area (blue outline) from which these waveforms were selected in Olympic National Park, Washington State, USA. (Right). “Elevation” is the height above ground, as represented in GEDI L2A products.

3.0 GEDI WAVEFORM IMPUTATION ALGORITHM

The GEDI waveform imputation is the process of filling in the whole spatial domain (Earth’s land surface between $\pm 51.6^\circ$ latitude) at 30-meter resolution with predicted waveforms. We used the single nearest neighbor (SNN) algorithm to impute GEDI waveforms in a way that preserves the properties of the observed GEDI waveforms.

3.1 k -Nearest Neighbor Algorithm

The k -NN algorithm is a non-parametric machine learning algorithm used for classification and regression tasks. It works by finding the k training samples closest in ancillary feature space (defined here by Landsat time series reflectance bands; see Section 3.3 for details) to the area for which a prediction is being made and aggregating the target values of the selected training samples. Generally, for classification, the prediction is made by majority voting on the labels of the k training data points, while for regression, it is common to use the mean or weighted average of the target values of the k training data points, considering the similarity (or distance) between the training points and a query point. In our approach, the optimal value of k was determined based on its ability to preserve the observed distribution of each RH metric and was found to be 1. The supporting empirical examinations are presented in Figures 4–7 as examples for the study areas shown in Figure 3. Figures 4-5 illustrate how the distribution of predicted waveforms diverges from the observed distribution across RH values as k increases from 1. It is evident that increasing k tends to not only simplify predicted forest structure (Figure 4) but also narrow the distribution (Figure 5). To quantify these differences, the similarity between two distributions can be formally evaluated using the Kolmogorov–Smirnov Test (KS Test), while point-wise prediction accuracy can be assessed with the Root Mean Square Error (RMSE) (See Section 4 for details). Using these evaluation metrics, Figures 6 and 7 illustrate that increasing k lowers RMSE but at the expense of distributional detail, leading to overestimation of short trees and underestimation of tall trees. Taken together, these results indicate that only $k=1$ (i.e., mapping the single nearest neighbor, SNN, in Landsat feature space) produced a predicted distribution that was statistically similar to the observed distribution of GEDI RH98. For SNN model development, we used the *KNeighborsRegressor* classier from the scikit-learn Python library.

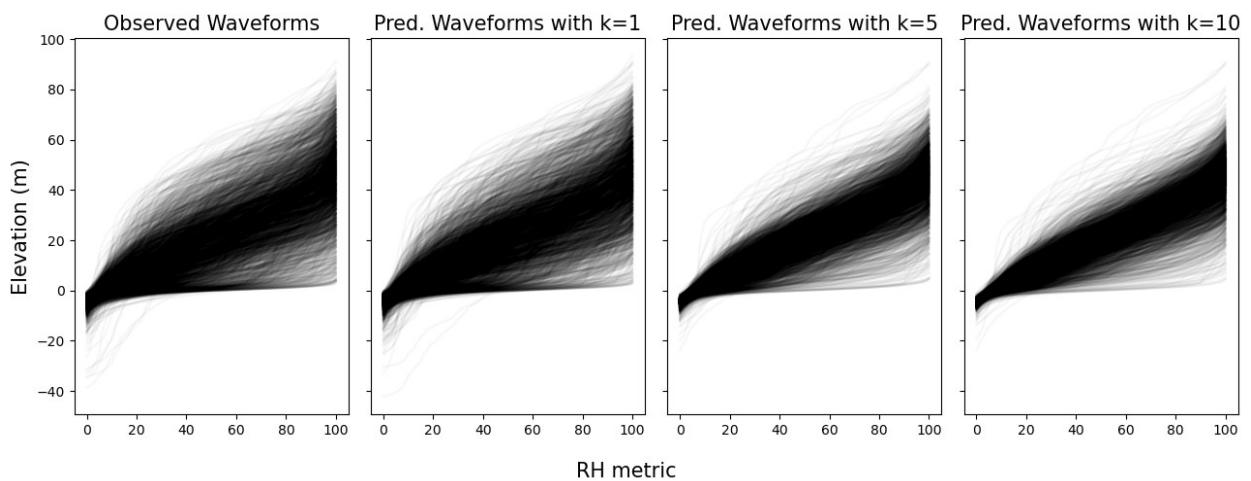


Figure 4. The distribution of predicted waveforms (shown here as a distribution of heights from RH0 to RH100) diverges from the distribution observed by GEDI (left) as more neighbors are aggregated to determine each value of relative height (i.e., as k increases from 1 to 5 to 10 neighbors).

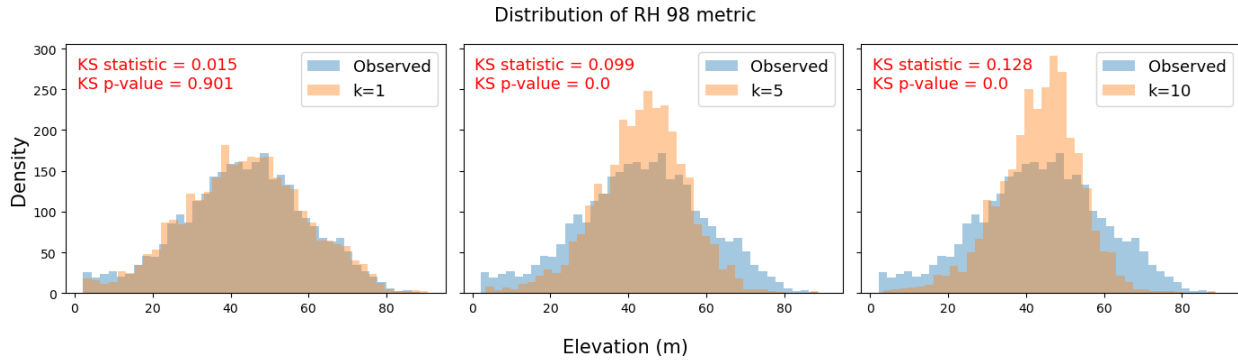


Figure 5. The distribution of predicted RH98 values for the study area shown in Figure 1 at three values of k (orange). The distribution of GEDI measurements of RH98 is shown in blue for reference. As more “neighbors” are averaged, predictions tend toward the mean. Kolmogorov–Smirnov (KS) Test evaluates the similarity between the observed distribution and the predicted distribution. The test statistic represents the maximum distance between the cumulative distribution functions; the p-value shows whether this difference is likely due to chance. A smaller p-value suggests the difference is unlikely to be random, implying the distributions are probably not the same.

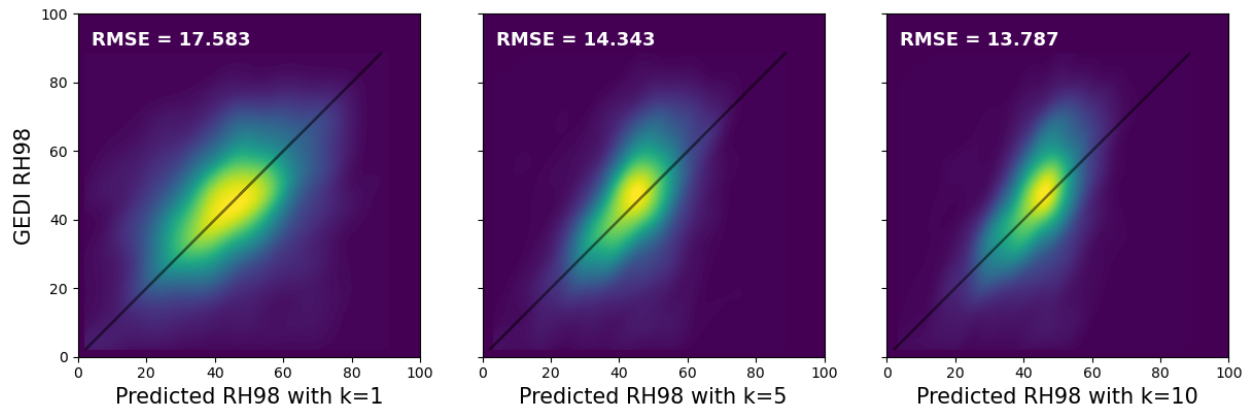


Figure 6. Scatter plots comparing observed GEDI RH98 values with k -NN predictions ($k=1, 5, 10$) in the study area. Increasing k lowers RMSE but compresses values toward the mean, leading to overestimation of short trees and underestimation of tall trees.

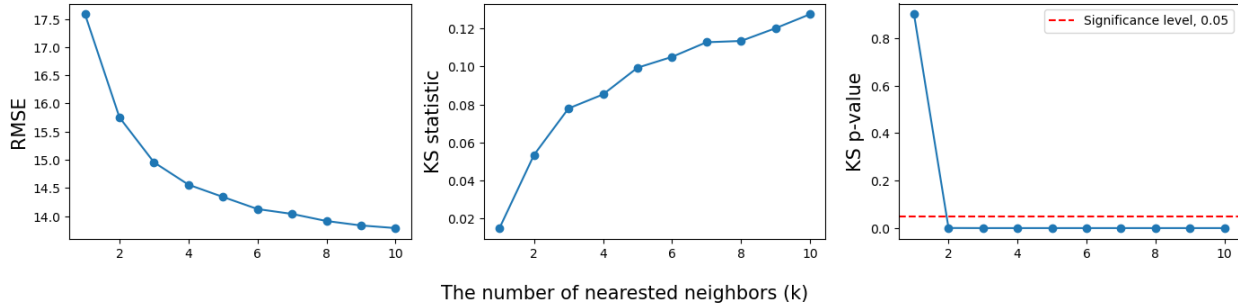


Figure 7. RMSE and KS-Test results for the study area displayed in Figure 2. While increasing k reduces RMSE, only imputation using the single-nearest neighbor ($k=1$) produced a predicted distribution of RH98 that was not statistically different (i.e., $p > 0.05$) from the distribution measured in an independent sample of GEDI shots.

3.2 Geospatial Domain of Single Nearest Neighbor Models

The SNN algorithm identifies the one nearest neighbor from the available observed samples. We developed a tiling system composed of $10 \text{ km} \times 10 \text{ km}$ tiles, with an independent SNN model applied to each tile. Tiles were defined using the EASE-Grid 2.0 projection, starting from the center coordinate (0.0, 0.0). Tiles entirely over water were excluded from processing, resulting in a total of 1,085,943 tiles between $\pm 51.6^\circ$ latitude (Figure 8). Each tile is represented by its row and column index.

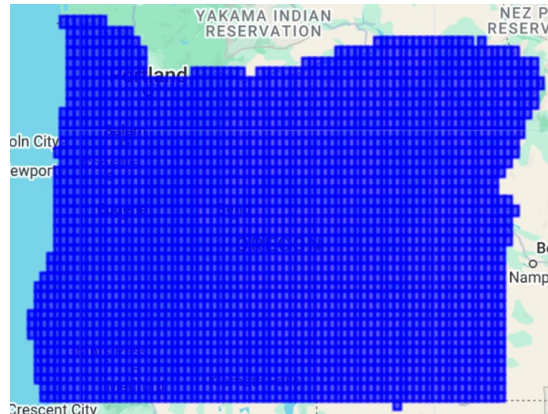


Figure 8. Example of $10 \text{ km} \times 10 \text{ km}$ tiles across Oregon, USA.

While a new model is created for every $10 \times 10 \text{ km}$ tile, GEDI training samples are drawn from both the focal tile and its adjacent tiles (Figure 9). Internal testing showed that using only the samples available in a $10 \text{ km} \times 10 \text{ km}$ tile did not always result in stable predictions, a problem that was often evident at the seamlines between tiles (Figure 10). We targeted a minimum of 2,500 training samples per model. If a tile and its support area contain fewer than 2,500 high-quality GEDI samples, we expanded the tile buffer progressively outward until this threshold was met.

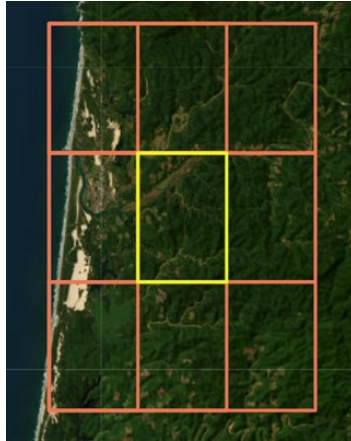


Figure 9. Illustration of a focal tile (yellow) and its support tiles (orange) with a tile buffer of 1.

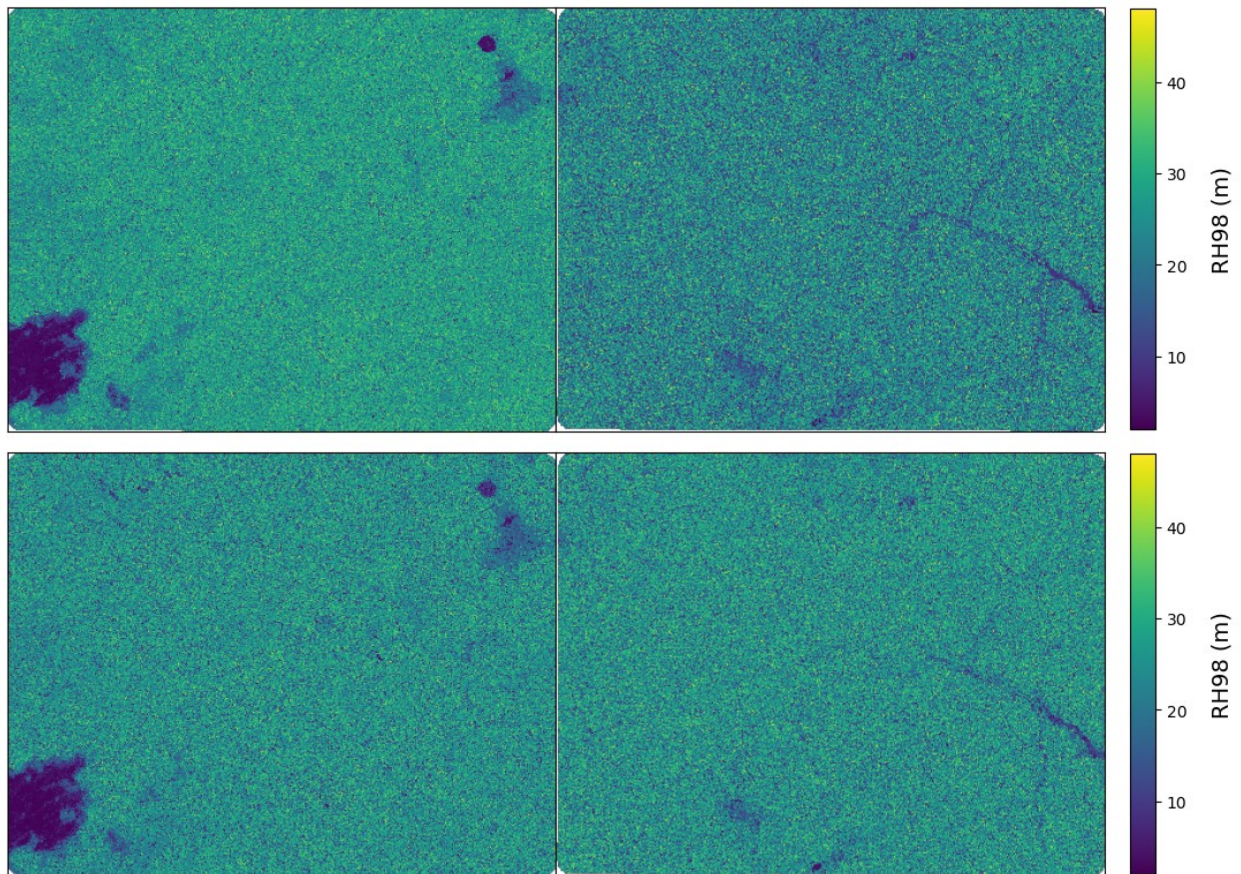


Figure 10. Imputation maps for two adjacent 10 x 10 km modeling tiles. (Top) Using training data exclusively from the focal tile produces visible predictions discontinuities at tile boundaries. (Bottom) Including training data from both the focal tile and its adjacent tiles yields a smoother transition.

3.3 Environmental Covariates for GEDI Waveforms

For each 30-meter resolution pixel, the SNN approach identifies the most similar location from among candidate sites where high-quality GEDI waveforms are available. Similarity is determined based on environmental covariates rather than geographical proximity, as the distribution of high-quality GEDI waveforms is uneven, and the geographically nearest site may not share similar land characteristics. We utilized synthetic Landsat data derived from Continuous Change Detection and Classification (CCDC) (Zhu and Woodcock 2014; Zhu et al., 2015; Gorelick et al., 2023) as a set of environmental covariates for the SNN algorithm. The CCDC algorithm fits a time series model using a mixture of sine and cosine functions to all cloud-free observations of each Landsat reflectance band (Blue, Green, Red, Near Infrared (NIR), Shortwave Infrared 1 (SWIR1), Shortwave Infrared 2 (SWIR2), Thermal) to detect land cover change. Using band-specific fitted functions, the CCDC algorithm generates synthetic spectral bands for any date. We extracted synthetic values for two dates within the target year—chosen based on phenology data from the Moderate Resolution Imaging Spectroradiometer (MODIS). Specifically, we identified the ‘onset’ and ‘peak’ greenness stages in the MODIS phenology dataset to characterize each tile and derive synthetic spectral values for both the observed GEDI waveforms and every 30-meter pixel within that tile. The Thermal Band was excluded due to its sensitivity to different physical processes (Zhu and Woodcock 2014; Zhu et al., 2015). The environmental covariates from CCDC therefore consist of 12 variables: Landsat bands Blue, Green, Red, NIR, SWIR1, and SWIR2 for both the onset (‘greenup’) and peak greenness dates. Additionally, we added one more variable representing landforms and physiographic patterns using Global SRTM CHILI (Continuous Heat-Insolation Load Index). The environmental covariates used are summarized in Table 2.

Table 2. CCDC-Driven Covariates for GEDL Waveform Imputation

Index	Name	Description
1	Syn_BLUE_Greenup	Synthetic Blue band at a greenup date
2	Syn_GREEN_Greenup	Synthetic Green band at a greenup date
3	Syn_RED_Greenup	Synthetic Red band at a greenup date
4	Syn_NIR_Greenup	Synthetic NIR band at a greenup date
5	Syn_SWIR1_Greenup	Synthetic SWIR1 band at a greenup date
6	Syn_SWIR2_Greenup	Synthetic SWIR2 band at a greenup date
7	Syn_BLUE_Peak	Synthetic Blue band at a peak greenness date
8	Syn_GREEN_Peak	Synthetic Green band at a peak greenness date
9	Syn_RED_Peak	Synthetic Red band at a peak greenness date
10	Syn_NIR_Peak	Synthetic NIR band at a peak greenness date

11	Syn_SWIR1_Peak	Synthetic SWIR1 band at a peak greenness date
12	Syn_SWIR2_Peak	Synthetic SWIR2 band at a peak greenness date
13	CHILI	SRTM-derived Continuous Heat-Insolation Load Index

3.4 High-Quality GEDI Waveforms as Training Data

The recorded GEDI waveforms serve as training data in the prediction process. Including noisy or erroneous data can lead to less reliable predictions and, consequently, to lower-quality imputed GEDI waveform maps. Therefore, filtering out low-quality waveforms is essential. However, overly strict filtering may reduce spatial coverage, which can also degrade the quality of the imputation results. To balance data quality and spatial representativeness, we applied a series of filtering steps with discretion to ensure a sufficient number of high-quality shots.

3.4.1 GEDI Level-2A Shot Quality Filtering

The initial quality filtering is based on V002 GEDI Level-2A flags. The following criteria are used to filter out low-quality GEDI waveforms:

- ‘quality_flag’ should be equal to 1 (valid surface returns with sensitivity > 0.9). We did not need to apply the same sensitivity thresholding as described in Dubayah *et al.* (2022) because the GEDI L4D *k*-NN algorithm did not have the same assumptions as the GEDI L4B algorithm.
- ‘degrade_flag’ should be one of 0, 10, 20, 30, 3, 13, 23, 33 (not acquired under significantly degraded geolocation conditions). The degrade flag values were selected by examining geolocation offsets between colocated on-orbit and simulated GEDI waveforms (Hancock *et al.*, 2019).
- ‘leaf_off_flag’ should be either 0 (leaf-on) or 255 (unknown)

3.4.2 GEDI Level-4B Granule Quality Filtering

The second filtering is based on a list of sub-orbit granules excluded as part of quality filtering applied in the GEDI L4B product. This algorithm is detailed in the L4B Algorithm Theoretical Basis Document (ATBD). Any GEDI waveforms belonging to the list of excluded sub-orbit granules for a given 72x72 km GEDI L4B product tile were omitted as imputation candidates.

3.4.3 Filtering GEDI Shots Outside Grid Tiles

GEDI waveforms falling in 10 x 10 km tiles completely covered by water were excluded.

3.4.4 Filtering GEDI Shots Without Covariates

The CCDC time series models used to create synthetic Landsat data (Gorelick et al., 2023) failed to achieve adequate fits in large parts of Coastal West Africa and in more isolated areas throughout the tropics and over permanent ice/snow in the Andes. This was primarily a function of low availability of cloud-free acquisition of surface reflectance. Cloud cover in these areas also results in relatively low densities of GEDI observations. Future L4D versions may implement alternative processing of Landsat data for these areas, but GEDI waveforms are not imputed for these areas in Version 2.0.

3.4.5 Local Outlier Detection-Based Filtering

The final filtering step employs a local outlier detection algorithm designed to remove anomalous values that deviate markedly from their surroundings. Unlike L4B’s sub-orbit granule filtering, which excludes erroneous sub-orbits within 72x72 km GEDI L4B product tiles, this method identifies individual GEDI waveforms for exclusion. Many of the anomalies eliminated through this process would otherwise degrade the spectral relationship used to model tree height. This filter is aggressive with respect to eliminating spectral outliers that are also spatial outliers. The cost of losing some valid model training data was considered less than the cost of degraded spectral models.

The algorithm is applied to each 10 km x 10 km tile, where it fits a linear regression relating geo coordinates and three environmental variables to the RH98 metric. The fitted linear regression is then used to identify waveforms with unusually abrupt or protruding RH98 metric values relative to their local context. The three environment variables include Near Infrared (NIR) reflectance on the onset of greenness date, NIR on the peak greenness date, and the SRTM-derived Continuous Heat-Insolation Load Index (CHILI) that represent landforms and physiographic patterns. The two greenness dates are derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) data.

- $RH98 = f(\text{Longitude}, \text{Latitude}, \text{CHILI}, \text{NIR}_{\text{greenup}}, \text{NIR}_{\text{peak}})$
- $\text{residual}_i = RH98_i - \widehat{RH98}_i$
- $RMSE_{RH98}^{(t)} = \sqrt{\frac{\sum_{i \in I_t} \text{residual}_i^2}{|I_t|}}, I_t = \{i \mid \text{sample } i \text{ lies in the tile } t\}$

Common scenarios involving spatial outliers (high local RMSE) are enumerated in Table 3 and Figure 11. Cases 1 and 2 are not removed, while Cases 3-5 are excluded from the population making up the imputable neighbors for the 10 x 10 km area model.

Table 3. Scenarios addressed by the Local Outlier Detection Algorithm

Case #	Criteria	Description	Outlier Flag
--------	----------	-------------	--------------

Case 1	$residual_i < n \times RMSE_{RM98}^{(t)}$	Low canopy height surrounded by tall canopy heights. Not a problem.	No
Case 2	$residual_i > n \times RMSE_{RM98}^{(t)}$ $40 < RH98_i < 75$ $40 < \overline{RH98}_i$	Valid canopy height from emergent trees (40-75m) surrounded by shorter mature trees (<40m)	No
Case 3	$residual_i > n \times RMSE_{RM98}^{(t)}$ $40 < RH98_i < 75$ $\overline{RH98}_i \leq 40$	The surrounding canopy heights are too small to classify the target as a canopy height from an emergent tree.	Yes
Case 4	$residual_i > n \times RMSE_{RM98}^{(t)}$ $RH98_i \geq 75$	It is an unrealistically tall canopy height, exhibiting a significant height disparity with the surrounding canopy heights.	Yes
Case 5	$residual_i > n \times RMSE_{RM98}^{(t)}$ $RH98_i \leq 40$	The target canopy height is too small relative to the surrounding canopy heights, resulting in a significant height disparity.	Yes

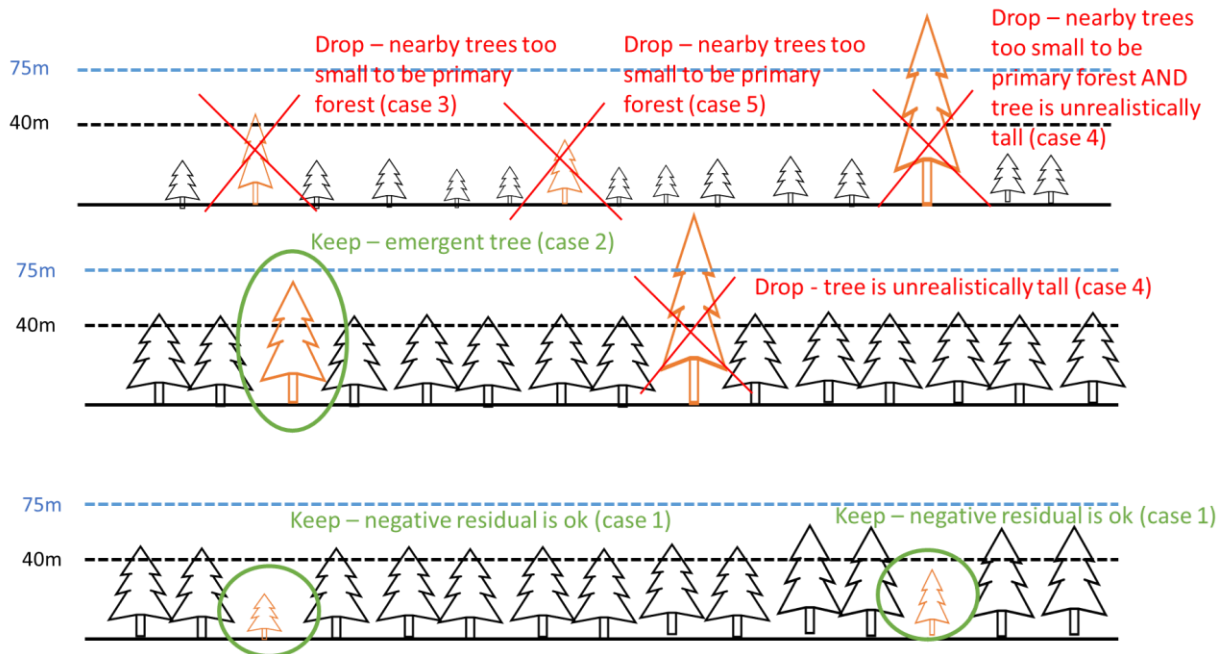


Figure 11. Illustration of cases for identifying outlier GEDI waveforms using the RH98 metric of a target tree as a proxy for tree height, alongside contextual information from surrounding tree heights.

3.5 k -Nearest Neighbor Distance Metric

We explored various distance metrics available in the Python SciPy library for measuring similarity between target samples and training data, and selected the Bray–Curtis distance, which yielded the lowest RMSE across six sample tiles: an agricultural site and a woodland in Brazil; woodlands in the Democratic Republic of the Congo and Tanzania; and woodlands in Utah and Oregon, USA. The Bray-Curtis distance between two feature vectors $X = (x_1, x_2, \dots, x_m)$ and $Y = (y_1, y_2, \dots, y_m)$ is defined as

$$d_{bc}(X, Y) = \frac{\sum_{i=1}^m |x_i - y_i|}{\sum_{i=1}^m |x_i| + |y_i|}$$

The Bray–Curtis distance in Python’s SciPy library differs from the traditional formulation for species-composition dissimilarity. The SciPy’s Bray–Curtis distance expresses the L1 (Manhattan) distance as a proportion of the total magnitude by dividing by the sum of absolute values, yielding a relative measure of dissimilarity, i.e.,

$$d_{bc}(X, Y) = \frac{d_{l1}(X, Y)}{\sum_{i=1}^m |x_i| + |y_i|}, \text{ where } d_{l1}(X, Y) = \sum_{i=1}^m |x_i - y_i|$$

The Bray–Curtis distance is bounded between 0 and 1: a value of 0 indicates identical samples, while a value of 1 indicates complete dissimilarity.

3.6 GEDI Algorithm Setting Group Selection

In GEDI Release 2 (V002), the L2 and L4 data product algorithms use different schemes for selecting waveform processing algorithm setting groups. This discrepancy arose because calibration and validation (cal/val) updates to the L2 algorithm were completed after the V002 L2 products were released but were incorporated into the implementation of the V002 L4 product. For the GEDI L4D product, the only relevant difference is that the V002 L2 algorithm did **not** include algorithm setting groups 5 and 10 for the *Evergreen Broadleaf Tree* stratum (land_cover_data/pft_class == 2) in *South America* (land_cover_data/region_class == 6), whereas the V002 L4 algorithm did. The GEDI L4D product algorithm therefore uses GEDI metrics derived from the V002 L4 setting group selection scheme. This approach ensures that imputed AGBD predictions are consistent with the imputed RH metrics used as predictors. The definition of the minimum detectable pulse varies across algorithm setting groups; therefore, the interpretation of beam sensitivity also differs among groups. To maintain consistency with Dubayah *et al.* (2022), beam sensitivity for the L4D product was extracted using only algorithm setting group 2.

4.0 IMPUTED GEDI WAVEFORMS EVALUATION

The performance of the single nearest neighbor model (SNN) applied to each individual 10 km x 10 km tile is evaluated on 20% randomly-held out validation samples using two different evaluation metrics: RMSE and Kolmogorov–Smirnov (KS) test.

4.1 Validation GEDI Waveform Samples

By default, validation samples are drawn exclusively from the focal tile, since the fitted SNN model is applied only to that tile's area. However, validation scores derived from small sample sizes may not provide meaningful insights. To ensure robust validation, we impose a minimum of 100 validation samples drawn from the focal tile. If the focal tile contains fewer than 500 high-quality GEDI shots—yielding fewer than 100 validation samples—the 20% validation subset is randomly selected from the entire training dataset, including adjacent tiles. Tiles for which validation data are drawn from adjacent tiles are flagged in the validation file as part of the L4D product.

4.2 Evaluation Metrics

4.2.1 Root Mean Square Error (RMSE)

RMSE is defined as

$$RMSE_{RH(j)} = \sqrt{\frac{\sum_{i=1}^N (X_i^{(j)} - \widehat{X}_i^{(j)})^2}{N}}$$

where N is the size of validation data, $X_i^{(j)}$ is the observed value of the j th RH metric for sample i , and $\widehat{X}_i^{(j)}$ is the corresponding predicted value.

4.2.2 Kolmogorov–Smirnov Test (KS Test)

The KS test yields two outputs: the KS statistic and the corresponding p-value. The KS statistic of the j th RH metric, which measures the maximum vertical distance between two cumulative distribution functions, is defined as:

$$D = \sup_{x=RH(j)} |F(x) - \widehat{F}(x)|$$

where $F(x)$ is the cumulative distribution function (CDF) of the observed j th RH metric in the validation data, and $\widehat{F}(x)$ is the CDF of the imputed j th RH metric in the validation data.

The approximate KS p-value for the j th RH is defined as

$$p = 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2 \lambda^2}$$

where $\lambda = D \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$, n_1 and n_2 is the sample size of the two CDFs, respectively. The null hypothesis that the two distributions are identical is rejected at a significance level of α .

4.3 Validation Results

The L4D product provides a validation file that includes RMSE and KS test evaluation metrics for all global 10 km x 10 km grids (see GEDI L4D User Guide). Higher RMSE values are typically observed in regions with dense populations of tall trees, such as tropical rainforests (Figure 12). Figure 13 shows the KS test evaluation of RH98, with a binary indication of whether the null hypothesis—that the predicted distribution is the same as the observed distribution—is rejected. At a significance level of 0.05, about 71% of global grids show a similar distribution of imputed RH98 to that of measured RH98, increasing to 83% at the 0.01 level.

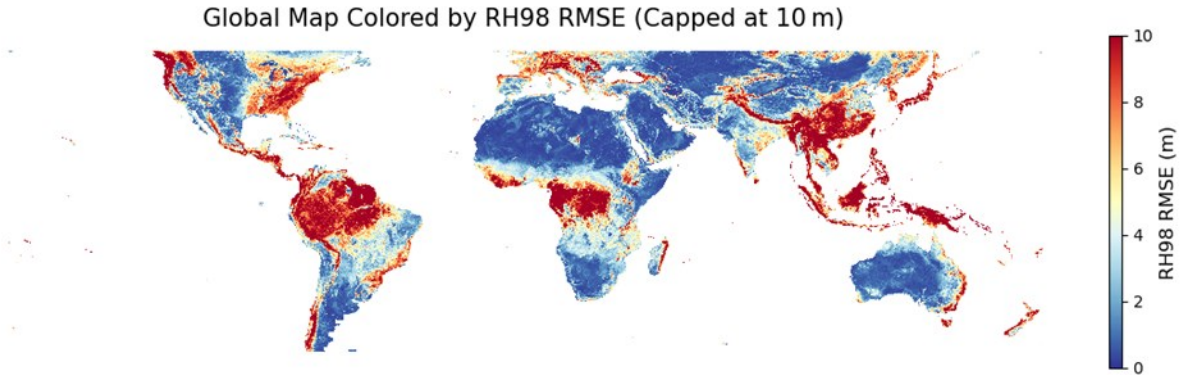


Figure 12. Global map of root mean square error (RMSE) for the RH98 metric, evaluated using the validation dataset. The color scale is capped at 10 m.

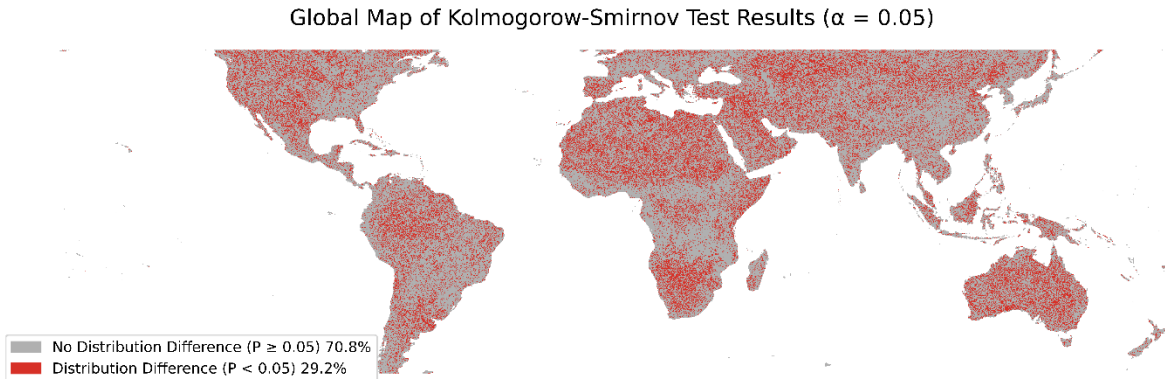


Figure 13. Global map showing the results of the Kolmogorov–Smirnov (KS) test comparing observed and imputed RH98 distributions. Using a significance level of 0.05, 70.8% of tiles show statistically similar distributions, while the remaining 29.2% exhibit significant differences, indicating that the imputed RH98 distribution deviates from the observed distribution in those regions. Both test outcomes are broadly distributed across the globe.

5.0 EXAMPLES OF GEDI WAVEFORM IMPUTATION MAPS

An overview of the GEDI L4D product is provided in Figures 14 and 15. Figure 14 shows the spatial distribution of imputed RH98, which can serve as a proxy for canopy height. Beyond a

single metric such as canopy height, the L4D product provides imputed forest vertical structure at 30 m resolution, represented by 11 RH metrics (RH10, RH20, ..., RH95, RH98; Figure 15). These metrics capture key observed forest traits, enabling a more detailed representation of forest structure and heterogeneity, while also supporting biodiversity conservation and ecological studies by facilitating integration with local field measurements. We note that beyond the standard imputed maps provided by L4D (Figure 15), the shot number is also provided. With that index, a user could generate a map of any of GEDI's shot-level variables.

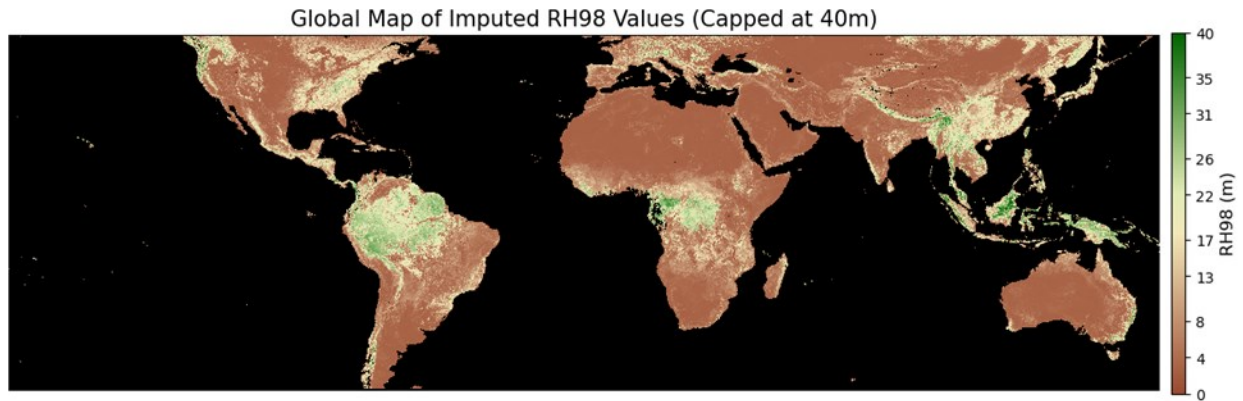


Figure 14. Global map showing imputed RH98 metrics.

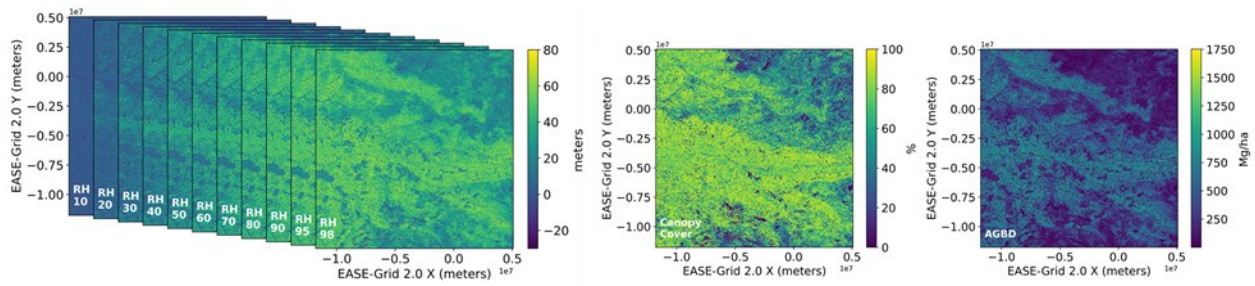


Figure 15. Illustration of imputed waveforms represented by 12 RH metrics (left) with ancillary bands for canopy cover (middle) and aboveground biomass density (right) within a 10 km x 10 km tile.

6.0 REFERENCES

Dubayah, R., J.B. Blair, S. Goetz, L. Fatoyinbo, M. Hansen, S. Healey, M. Hofton, G. Hurtt, J. Kellner, S. Luthcke, J. Armston, H. Tang, L. Duncanson, S. Hancock, P. Jantz, S. Marselis, P.L. Patterson, W. Qi, and C. Silva. 2020. The Global Ecosystem Dynamics Investigation: High-resolution laser ranging of the Earth's forests and topography. *Science of Remote Sensing* 1:100002. <https://doi.org/10.1016/j.srs.2020.100002>

Gorelick, N., Z. Yang, P. Arévalo, E. L. Bullock, K. P. Insfrán, and S. P. Healey. 2023. A global time series dataset to facilitate forest greenhouse gas reporting. *Environmental Research Letters* 18(8):084001. <https://doi.org/10.1088/1748-9326/ace2da>

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The elements of statistical learning: Data mining, inference, and prediction*. 2nd ed. Springer, New York.

Hudak, A. T., N. L. Crookston, J. S. Evans, D. E. Hall, and M. J. Falkowski. 2008. Nearest neighbor imputation of species-level, plot-scale forest structure attributes from LiDAR data. *Remote Sensing of Environment* 112(5):2232–2245. <https://doi.org/10.1016/j.rse.2007.10.009>

May, P. B., R. O. Dubayah, J. M. Bruening, and G. C. Gaines. 2024. Connecting spaceborne lidar with NFI networks: A method for improved estimation of forest structure and biomass. *International Journal of Applied Earth Observation and Geoinformation* 129:103797. <https://doi.org/10.1016/j.jag.2024.103797>

May, P. B., R. O. Dubayah, J. M. Bruening, and G. C. Gaines. 2025. GEDI-FIA Fusion: Training lidar models to estimate forest attributes. ORNL Distributed Active Archive Center. <https://doi.org/10.3334/ORNLDAAC/2417>

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

Zhu, Z., C. E. Woodcock, C. Holden, and Z. Yang. 2015. Generating synthetic Landsat images based on all available Landsat data: Predicting Landsat surface reflectance at any given time. *Remote Sensing of Environment* 162:67–83. <https://doi.org/10.1016/j.rse.2015.02.009>

GLOSSARY/ACRONYMS

ATBD	Algorithm Theoretical Basis Document
CCDC	Continuous Change Detection and Classification
CDF	Cumulative Distribution Function
CHILI	Continuous Heat-Insolation Load Index
GEDI	Global Ecosystem Dynamics Investigation
KS	Kolmogorov–Smirnov
L2A	Level 2A
L2B	Level 2B
L4A	Level 4A
L4B	Level 4B
L4D	Level 4D
MODIS	Moderate Resolution Imaging Spectroradiometer
NIR	Near-Infrared
RH	Relative Height
RMSE	Root Mean Square Error
SNN	Single Nearest Neighbor
SRTM	Shuttle Radar Topography Mission
SWIR	Short-Wave Infrared
UTM	Universal Transverse Mercator