Algorithm Theoretical Basis Document (ATBD)

for

GEDI Level-4B (L4B) Gridded Aboveground Biomass Density

Sean P. Healey[1], Paul L. Patterson[1], John Armston[2]

[1] Inventory and Monitoring, United States Department of Agriculture (USDA) Forest Service, 1400 Independence Ave, SW, Washington, DC 20250-1111, USA

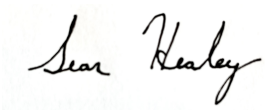
[2] Department of Geographical Sciences, University of Maryland, College Park MD, USA

Version 2.0

Release date: September 18th, 2023

University of Maryland, College Park MD

Authors:



Principal Investigator:

Abstract

The Global Ecosystem Dynamics Investigation (GEDI) lidar is a multibeam laser altimeter on the International Space Station. GEDI is the first spaceborne instrument designed specifically to measure vegetation structure and estimate aboveground carbon stocks in temperate and tropical forests and woodlands. This document describes the algorithm theoretical basis underpinning the development of the GEDI Level-4B (L4B) 1-km gridded aboveground biomass density (AGBD) data product. GEDI's primary L4B algorithm is a closed-form statistical estimation approach called hybrid inference (Ståhl et al., 2011; Patterson et al., 2019), in which mean biomass is estimated from an incomplete sample of modeled biomass values (available via the GEDI Level 4A product), and the variance of the estimated mean is a function of both L4A model uncertainty and uncertainty from GEDI's sampling pattern. Conventionally, the kind of linear observations collected by GEDI, or other air- or space-borne platforms, is treated as a cluster sample. Since at least two clusters are required to calculate a variance under hybrid estimation, GEDI's primary algorithm may only be applied to 1-km cells containing samples from at least two ground tracks that intersect the grid cell. At the end of the mission, mean biomass in cells without sufficient clusters will be estimated instead with Generalized Hierarchical Model-Based inference (GHMB). This approach makes use of a second-level model that extends biomass to a surface predicted from wall-to-wall imagery. The output grids generated by this process represents our best understanding of the spatial distribution of tropical and temperate forest biomass, and may support a broad range of scientific, policy, and management applications.

Foreword

This document is the Algorithm Theoretical Basis Document for the GEDI Level-4B (L4B) Gridded Aboveground Biomass Density product. The GEDI Science 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:

Sean P. Healey / Paul L. Patterson
Inventory and Monitoring, United States Department of Agriculture (USDA) Forest Service, 1400 Independence Ave, SW, Washington, DC 20250-1111, USA
sean.healey@usda.gov
+1 (801) 391 7536

John Armston
2181 Lefrak Hall, Department of Geographical Sciences
University of Maryland, College Park MD 20742
armston@umd.edu
+1 (301) 405 8444

Ralph Dubayah
2181 Lefrak Hall, Department of Geographical Sciences
University of Maryland, College Park MD 20742
dubayah@umd.edu
+1 (301) 405 4069

Change History Log

Revision Level	Description of Change	Date Approved
1.0	Initial version	Feb 22, 2022
2.0	Updated references, updated description of sub-orbit granule filtering, and added the method used to assign zero biomass to shots over bare ground and permanent snow/ice.	Sep 18, 2023

Table of Contents

List of Figures

List of Tables

1.0 INTRODUCTION

This document is the Algorithm Theoretical Basis Document (ATBD) for the GEDI Level 4 Gridded Biomass Product (L4B).

1.1 Document Overview and Objective

This document is designed to provide both: (1) a general theoretical overview of the algorithms, processing steps and procedures required to provide Level 4 gridded biomass products and (2) a detailed gridding algorithm implementation and processing flow specifically designed for the GEDI mission level 4 products.

This topic is arranged in the following manner:

- **Section 1** presents a brief introduction and related documentation
- **Section 2** presents an overview of theoretical background
- **Section 3** presents details of the retrieval algorithm
- **Section 4** presents detailed Cal/Val procedures for GEDI L4B products
- **Section 5** contains references
- An acronym glossary can be found at the end of this document

1.2 Gridded Biomass Product Overview

Forest ecosystems store globally significant amounts of carbon (Houghton et al., 2009), and the forests of the United States currently offset approximately 15% of the country's fossil fuel emissions through net addition of biomass (Woodall et al., 2015). While ground-based forest inventories can provide authoritative information about current carbon stores, they are discontinuous across space and are not designed to explain how climatic and disturbance trends affect ecosystem services such as climate change mitigation and fiber production. Moreover, such inventories are only operational in a handful of countries. Spaceborne platforms, when combined with fine-scale ecosystem models, offer the opportunity to study forest structure and carbon dynamics in a consistent way across the globe. However, large-area biomass mapping activities to date have relied upon ad hoc estimation techniques that can result in significant understatement of uncertainty (Mitchard et al., 2014).

NASA's GEDI mission has installed a lidar instrument on the International Space Station (ISS) to study forest structure in Earth's tropical and temperate biomes between 51 degrees north and south latitude. Lidar's active signals penetrate closed forest canopies and return data have been shown to be sensitive to different levels of biomass across a variety of forest types (Zolkos et al., 2013). GEDI is a sampling instrument that acquires discontinuous, along-track full-waveform footprints that are well correlated with local biomass. Multiple overpasses, each composed of several GEDI footprints, will ultimately intersect most cells in a global 1-km grid 51° N and 51°S. The challenge of GEDI's gridding process is to combine biomass modeled at each footprint in a way that approximates the true biomass of every 1-km cell. Unless stated otherwise, biomass here refers to aboveground tree biomass **density** in Mg ("megagrams" or "metric tonnes") per hectare. The 1-km grid cells for which population mean biomass is to be estimated are often simply called "cells" here.

GEDI's L1A Science Requirement states that the aboveground biomass of 80% of 1-km cells must be estimated to within a standard error of either 20 Mg/ha or 20% of the estimate, whichever is greater.

This requirement imposes important constraints:

- Accuracy is assessed at a 1-km scale, globally (in this document, the term "global" is used to indicate the temperate and tropical latitudes overflown by the ISS). The largest existing field plots are 0.5 km^2, which means that GEDI's performance cannot be directly evaluated post-launch using existing plot networks. There is a possibility of flying airborne lidar data at the 1-km scale and treating the resulting biomass prediction as truth but ignoring field-to-airborne model error can hide significant uncertainties (Saarela et al., 2016), which would make results of such a comparison ambiguous.
- Precision is defined by way of standard error of the estimate, a statistical term implying how much one would expect the estimate to vary if many more samples were taken. Addressing uncertainty formally in terms of expected estimate variation requires a theoretical basis not present with ad-hoc "error budgeting" approaches.
- Error is assessed at the 1-km cell scale, and it must account for both: 1) uncertainty due to the fact that GEDI is making discontinuous measurements of biomass on only a fraction of the cell, and 2) uncertainty due to the fact that GEDI is not measuring biomass directly but is instead providing waveform-derived correlates for biomass.

The first two constraints are shared with large-area forest inventory operations such as the Forest Service's FIA (Forest Inventory and Analysis) Program (Bechtold and Patterson, 2005). FIA maintains a systematic sample of field plots across the country and must estimate biomass (and other forest characteristics) over areas far too large to exhaustively measure for the purpose of empirical uncertainty assessment. FIA data are also frequently the basis for both private and public forest planning processes, and estimates must be accompanied by theoretically clear measures of uncertainty. Both constraints argue for an approach based upon sampling theory. GEDI has the added challenge of relying upon sample observations that are modeled (at the footprint level) instead of measured directly. This challenge can nevertheless be addressed within a formal sampling framework, using so-called "hybrid" model-based inference (Ståhl et al., 2016), which accounts for both model and sampling uncertainty. Hybrid estimation was developed with airborne lidar data (Ståhl et al., 2010; Patterson et al., 2019), and has been applied to return data from GLAS (NASA's Geoscience Laser Altimeter System aboard ICESat1), but those few efforts (Healey et al., 2012; Nelson et al., 2009) were limited to areas much larger than GEDI's 1-km cells because of coverage limitations over forests.

This document describes GEDI's use of hybrid estimation to create an exhaustive coverage of non-overlapping 1x1-km mean biomass estimates, together with a grid of standard errors for each 1-km estimate (a process also called "gridding" here). GEDI's innovative instrument and mission design allow unprecedented coverage of Earth's forests with strong predictors of biomass. The gridding approach described herein relies upon clear sampling theory, described below. The GEDI Science Team has tested this approach with simulated GEDI data (derived from airborne discrete return lidar) in 6 diverse ecosystems, as described in Section 2.3 and by Patterson et al. (2019).

1.3 Summary of GEDI's Footprint-level Biomass Modeling Strategy

Mean aboveground biomass density (ABGD) in each 1-km cell of the gridded biomass product will be estimated using a formal method of inference called "hybrid" estimation. As described in later sections of this document, the uncertainties generated around 1-km hybrid estimates of biomass account for both sample error (due to incomplete coverage of the cell) and model error (owing to the fact that biomass must be modeled at the footprint level from lidar metrics instead of directly measured).

Prior to describing hybrid estimation, it is important to summarize the methods used to assign a particular level of biomass to each footprint. This is because GEDI has taken a calibration approach that uses simulated GEDI waveforms, which has implications for the GEDI Level 4B algorithm implementation. This topic is discussed in more detail in the ATBD for the GEDI L4A Footprint Aboveground Biomass product (Kellner et al., 2023).

Specifically, the GEDI Science Team has worked in collaboration with the global research community to develop a database consisting of field-inventory plots with associated AGBD and coincident airborne laser scanning (ALS) data, referred to here as the GEDI Forest Structure and Biomass Database (FSBD). The Team is using ALS data to produce simulated GEDI waveforms using a GEDI waveform simulator (Blair and Hofton, 1999; Hancock et al., 2011). These waveforms are co-located with field measurements of biomass to generate the lidar-to-biomass models that will be used to predict biomass at the footprint level (i.e., GEDI's L4A footprint-level biomass product). The GEDI L4A AGBD models are described by Kellner et al. (2023) and Duncanson et al. (2022). Calibration of these models will be updated annually as new data are incorporated into the GEDI FSBD, which will be used to update GEDI L4B grids. The primary benefits of this approach include:

1. The availability of a much broader calibration dataset than would be feasible with a directed, post-launch GEDI field campaign
2. Elimination of the spatial mismatch, and accompanying model noise, involved with placing field plots at GEDI shots for which we have only limited pointing knowledge (up to approximately 10m positional error)

Models have been developed and applied to prediction strata defined by plant functional type (PFT) and world region. Two sets of parametric models have been developed: one representing the strongest possible model using "unnoised" waveforms; and another using "noised" lidar waveforms, accounting for realistic errors related to factors that impact the signal-to-noise ratio of waveforms (e.g., atmospheric attenuation). As described below, the GEDI L4B algorithm will use properties of both the noised and unnoised waveform simulations.

1.4 Product/Algorithm Objectives

The two products of this algorithm are: mean aboveground biomass density estimates for cells in a global (51.6°S – 51.6°N) 1-km grid; and standard errors corresponding to each of those estimates. In addition to these variables, the L4B product also reports: (i)

whether the main or the contingency mode of inference was used (discussed in Section 2.0); (ii) decomposition of the variance into either modeling error and sampling error (for the main mode of inference) or two levels of modeling error (for the contingency method); and (iii) whether a given 1-km cell meets L1 precision goals for the mission. Specifically, those goals stipulate that 80% of cells should be estimated within a standard error of 20% of the estimate or 20 Mg/ha, whichever is greater.

2.0 THEORETICAL BASIS OF ALGORITHM

2.1 Sampling Theory and Lidar

GEDI is not a wall-to-wall sensor; it does not provide the kind of single-date images generated from platforms like Landsat or MODIS. Unlike these sensors, though, GEDI's canopy-penetrating measurements provide high-quality structure and biomass information. The discontinuous measurements provided by airborne and spaceborne lidar sensors such as GEDI's has led to interest in application of sampling methods and theory (Wulder et al., 2012; Gregoire et al., 2016).

Sample theory allows inferences to be drawn regarding both population parameters (mean biomass density in this case) and the uncertainty around those parameters under clearly defined probability distribution assumptions. It is sometimes useful to test estimators proposed for a given problem by using real data to compare analytical versus empirical results. This kind of test for GEDI is described in Section 2.3, following the description of the estimators provided here.

Like traditional field inventories, GEDI's spaceborne lidar system only observes a spatially distributed portion of the landscape. GEDI's sampling pattern is projected to roughly equate to a lattice (Figure 1), with parallel ascending and descending overpasses and approximately 500-meter cross-track spacing. Given local irregularities in the pattern of GEDI's overpass schedule and unpredictable loss of signal to clouds and other atmospheric phenomena, the precise location of GEDI sample lines within any single 1x1-km cell is considered to be random. However, because the locations of GEDI observations ("shots") along any track are pre-determined once an overpass' entry point is known, they cannot be considered to be independent. Conventionally, such samples are treated as cluster samples within the context of variance calculations (Ståhl et al., 2010). Therefore, GEDI's sampling pattern may be considered a randomly allocated cluster sample.

Ståhl et al., (2011 and 2016) pioneered an approach based upon "hybrid" model-based inference, where discrete lidar observations within a study area are combined within the frame of a designed sample to infer both average biomass and variance around that estimated mean. The statistical variance estimators used in this approach are composed of terms representing both sampling uncertainty (on the basis of sampling theory applied to the spatial distribution of lidar footprints) and model uncertainty (derived from the covariance matrix produced around the terms in the parametric model that allows prediction of footprint-level biomass).

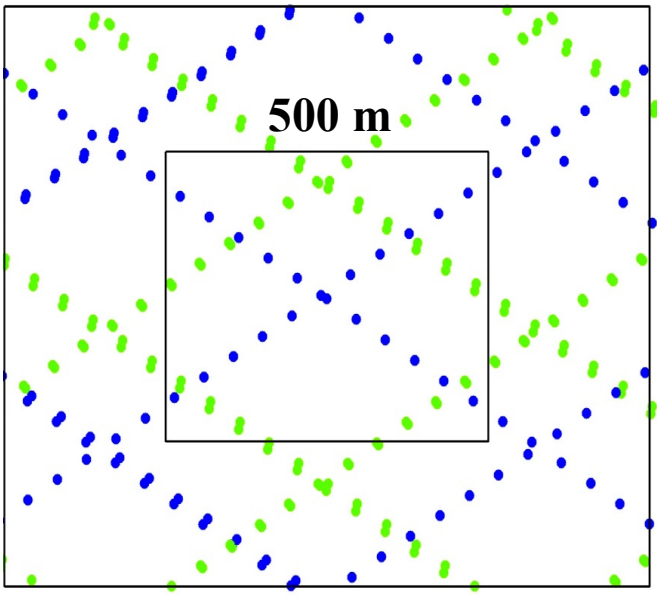


Figure 1. Generalized projected pattern of GEDI overpasses at the equator after 18 months of operation.

Assuming a linear model ($\hat{\alpha}$)linking lidar and biomass, a general estimator of biomass would be:

$$E(\hat{\tau}) = \sum_{i \in s} \frac{x_i \hat{\alpha}}{\pi_i} = \pi' X \hat{\alpha} \qquad [1]$$

where s is the sample of auxiliary data (i.e. lidar measurements), $\pi_i$is the probability of including population element *i* into the auxiliary data sample, π' is an n-length column vector of $(1/\pi_i)$ – values, and *X* is an n x p matrix of sampled auxiliary data (with n the number of footprints and p the number of lidar-based parameters). The model parameters are estimated from a sample presumed to be independent, which is the case with the methods used by the GEDI footprint-level product (see Section 1.3 and Kellner et al., 2023). The variance estimator, which is detailed in Section 3.4, likewise relies upon both model properties and probability sampling. It is crucial to appreciate that the probability of inclusion (π_i) must be known for each observation (each footprint, in the case of GEDI). As described above, GEDI's sample within any particular 1x1-km cell uses a randomly allocated cluster design, and π_i is determined accordingly within the context of finite sampling.

Section 3.4 gives details for the hybrid estimators to be used for mean biomass within each 1-km cell and for the variance of those estimates.

While hybrid model-based biomass estimation methods were developed using airborne lidar (e.g. Ståhl et al., 2011; Gobakken et al., 2012; Nelson et al., 2012; Corona et al., 2014; Saarela et al., 2015), the approach has also been extended to spaceborne

applications, primarily with the GLAS instrument aboard ICESat (e.g. Healey et al., 2012; Neigh et al., 2013; Margolis et al., 2015; Nelson et al., 2017). An alternative approach that also uses lidar data in a sample-survey context is called model-assisted estimation (Gregoire et al., 2011 and 2016; Magnussen, 2015). Model-assisted approaches condition variance estimates upon the sample's design instead of the biomass model's properties (and are thus "design-based" instead of "model-based"). However, such approaches require a designed sample of field plots within the domain of estimation (1-km grid cell in GEDI's case), whereas hybrid estimators (like all model-based approaches) can make use of spatially remote field samples, even purposive or opportunistic samples. Since designed samples of field plots clearly do not cover most of the Earth's forests, hybrid estimation's flexibility in this regard is critical. There is, however, an assumption that the footprint-level model applies to the population to which it is applied. This is discussed further in Section 4.

2.2 Population and Sample Properties

The population of interest is defined as the area within a single 1-km GEDI grid cell. The first step in developing a statistical estimator is to define the population of interest in terms relevant to a pre-defined sampling strategy.

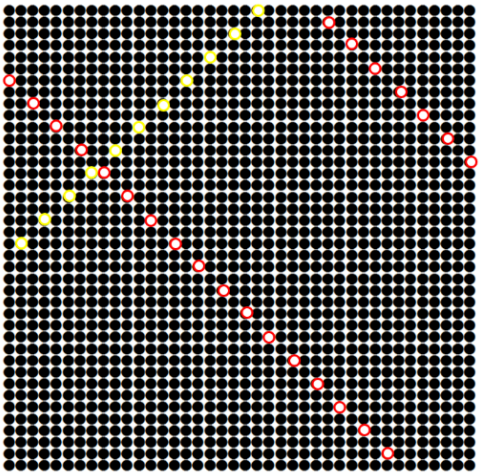


Figure 2. Population units (20x20-meter) within a 1-km grid cell, as sampled by three GEDI overpasses.

Each 1-km cell may be viewed as gridded into a finite number of approximately 20x20-meter population elements (herein called "pixels"), with the same diameter of GEDI's footprint: approximately 20m (see Figure 2). The GEDI shots will sample this population along tracks determined by the daily path of the Space Station and the mission's pointing plan. In Figure 2, we represent the heading of the GEDI tracks as 45 and 315 degrees and the GEDI shots as equidistant apart. We assume that difference in overpass angle, which will vary by latitude, will not fundamentally change the sampling patterns or properties shown in Figure 2. In each population element we have the aboveground biomass

prediction, $g(\boldsymbol{x}_{it}, \widehat{\boldsymbol{\alpha}})$, in Mg/ha (Equation 1). The population attribute of interest is the average of the GEDI predictions over the N population elements. This average needs to be expressed in terms of GEDI sample tracks, which will be viewed as clusters of population elements (GEDI shots). This is consistent with similar applications in the airborne lidar-assisted inventory literature.

Since footprints along a GEDI track may be viewed as alternating (one footprint is sampled while the next is not due to GEDI's beam-dithering approach), each laser path is composed of two disjoint possible clusters: one starts on the edge of the 1-km x 1-km cell and one starts one shot in. If we combine over all the 45° lines, the population units are distributed along M non-overlapping clusters. For the GEDI sample, there will be with T_i shots in the ith cluster. We can express the population attribute, average of GEDI predictions over the N population elements, in terms of the clusters; that is:

$$\mu_Y = \frac{\sum_{i=1}^{M} \sum_{t=1}^{T_i} g(\boldsymbol{x}_{it}, \widehat{\boldsymbol{\alpha}})}{N} \quad [2]$$

Note that each population element is in one ascending line and one descending line. In the sampling design there are three possibilities, all ascending tracks, all descending tracks, or a combination of the two. In the first two cases M is the 2*number of lines and N is the number of 20mx20m cells in the population. In the second case both M and N are twice that number, but the average $\overline{Y}$ is the same. Therefore, in all of the three cases, we are estimating the same number.

2.3 Sufficiency of Proposed Estimators for 1-km square cells

The approach taken here, described in more detail by Patterson et al. (2019) is similar to the approach described by Ståhl et al. (2010) as appropriate for "large area" estimation problems. The size of the area is an issue because, while residuals of the footprint-level model are expected to sum to near zero over a large number of predictions, residual error may add uncertainty to the estimate in cases where fewer predictions are made, particularly if model residual errors show strong spatial autocorrelation. In such cases, it might be advisable to include a term in the variance estimator accounting for residual error.

Simulations described in Appendix S3 of Patterson et al. (2019) demonstrate that in most cases, a 1-km grid cell is large enough to assume negligible residual error. Therefore, no term has been added to our hybrid variance estimator.

2.4 Test of Proposed Estimators

It is often helpful to formally test the assertion that an estimator is unbiased for a particular application. A test of the algorithms described in Section 3 was conducted using airborne lidar and field data collected in six areas of the US (Figure 3) by a NASA CMS (Carbon Monitoring System) project: "An Historically Consistent and Broadly Applicable MRV System Based on Lidar Sampling and Landsat Time-series" (PI: Cohen). GEDI footprints were simulated from wall-to-wall discrete return lidar using the GEDI waveform simulator, which simulates large-footprint lidar waveforms using the

method presented in Blair and Hofton (1999), with noise added following Hancock et al. (2011) and Davidson and Sun (1988).

Sixty random areas (10 from each of the areas depicted in Figure 2) of approximately the same dimensions of a GEDI 1-km grid cell were selected from the CMS dataset. Alternative GEDI overpass patterns were simulated within each cell, with the inclination of each overpass at either 45° or 315°. Simulated shots were extracted in each overpass scenario, using realistic GEDI along-track spacing. Four thousand overpass patterns were generated for each case from 2 to 6 total overpasses in each test grid cell.

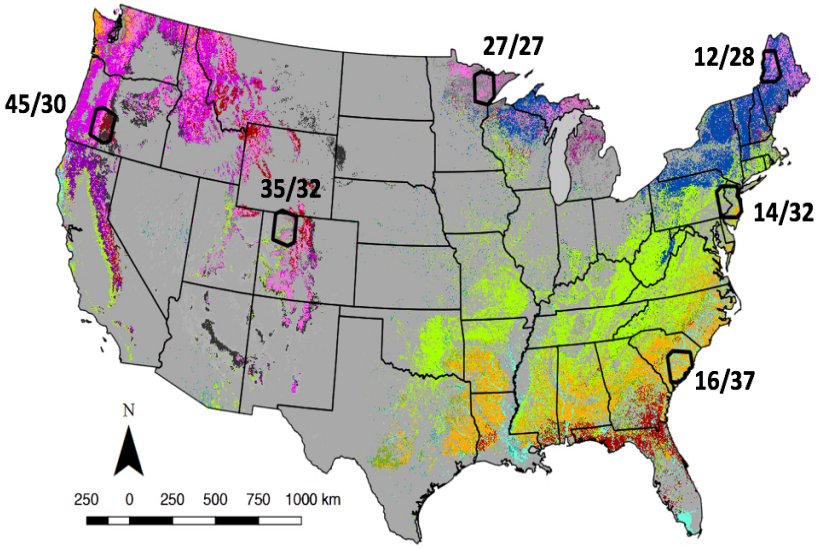


Figure 3. Study locations from which airborne lidar and field biomass data used to test GEDI's proposed estimator were acquired. North-South strips of discrete return lidar were collected at each site, along with field biomass data from 50 purposively distributed field plots. The numeric Landsat WRS-2 Path/Row for each site is also shown.

Regionally specific biomass models were created for the six study areas shown in Figure 3. Models for each region were calibrated at 50 plots (300 total supporting 6 regional models), placed strategically to cover the presumed range of biomass values. Simple polynomial models using simulated GEDI relative height metrics were fit to explain biomass for each plot. The covariance matrix of parameter values for each model is the primary vehicle for representing model uncertainty in hybrid variance estimators (see Section 3.4). In this test, that matrix was used to generate 4000 different biomass models for each of the 4000 overpass patterns generated above. Mean biomass and estimate variance were both estimated using the proposed hybrid estimators, and estimates were compared to empirical values derived from these simulations.

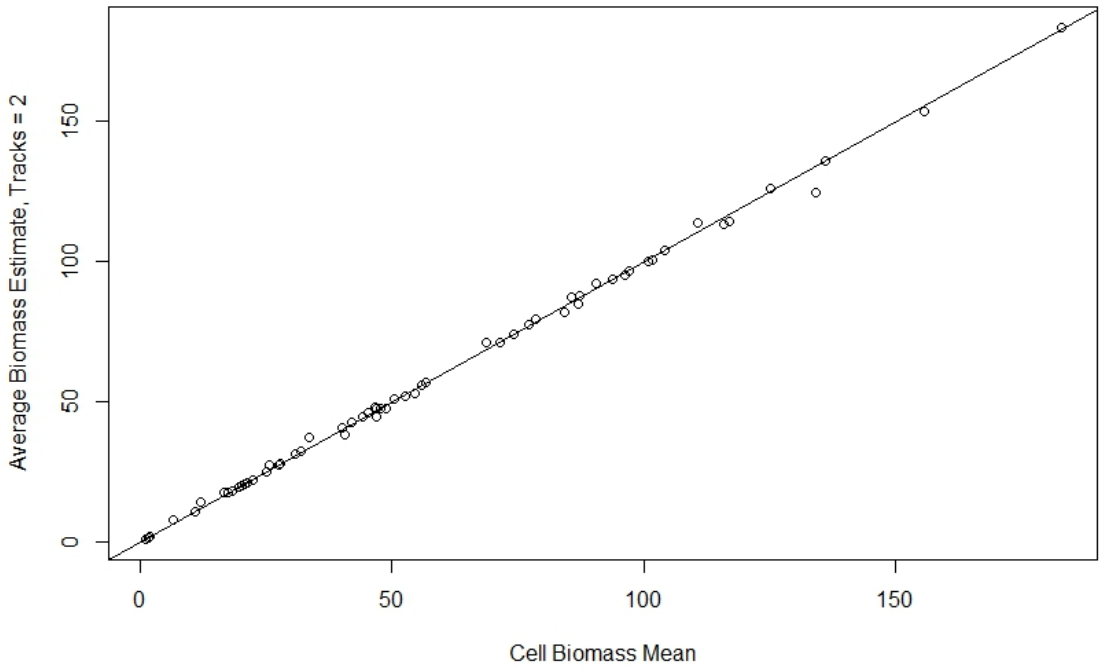


Figure 4. Empirical versus estimated biomass (Mg/ha) at 60 randomly located grid cells. This is the result for the 2-overpass case; more overpasses showed even less deviation from the 1:1 line.

Results showed that estimates of mean biomass derived from the proposed hybrid estimators were unbiased across the sixty pseudo-GEDI grid cells (Figure 4). The same exercise showed proposed variance estimators to be asymptotically unbiased at higher numbers of overpasses (Figure 5). At least across the range of ecological variability captured in the six sampled study areas, the estimators proposed here for both mean 1-km biomass and variance of the estimate of the mean seem to be unbiased. It should be repeated that these estimators assume the footprint-level biomass models (described in the GEDI-L4A ATBD; Kellner et al., 2023) to be correctly specified and parameterized; to the degree that this assumption is met, the proposed estimators appear appropriate for GEDI.

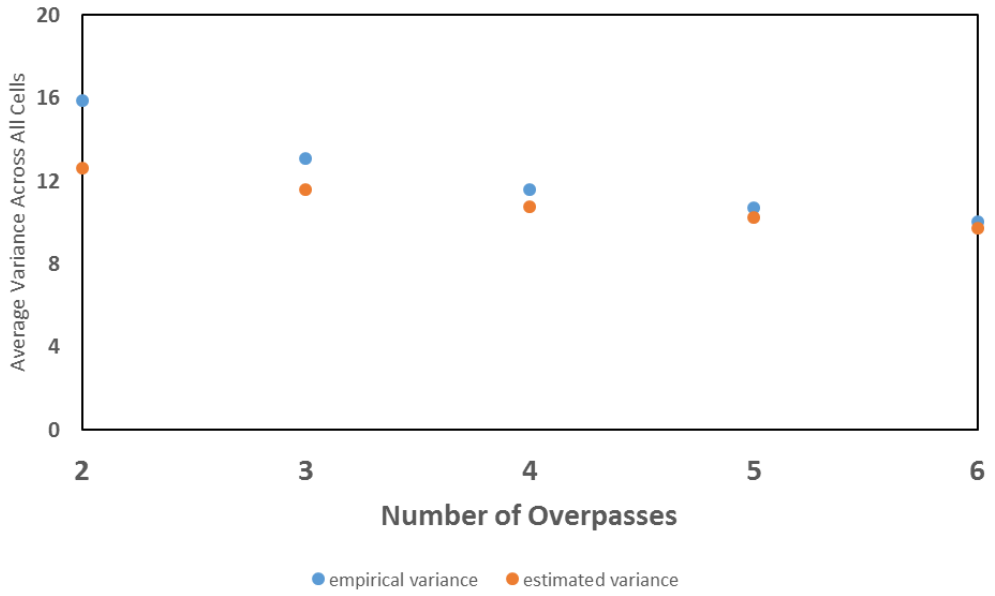


Figure 5. Average variance estimated for biomass estimates at the 60 simulated GEDI grid cells. The graph shows a comparison of empirical and analytical (estimated) mean biomass density estimate variance (square units for Mg/ha). Slight under-prediction of variance using hybrid estimators was noted when there are only two GEDI tracks intersecting a grid cell, but estimates appear to be asymptotically unbiased with increasing numbers.

2.5 Rules for footprint inclusion in the sample

The mission's precision goal does not specify just the forested parts of each cell – estimates must be stated in terms of mean ABGD over the entirety of the 1-km pixel. For this reason, every clear footprint – regardless of suspected land cover type – must be included in the estimation process. A significant bias would result from screening out (i.e. omitting from the sample) non-forest footprints, even in mixed 1-km cells that are partially covered by water features. Ancillary land cover information is nevertheless used to improve the biomass predicted for some waveforms.

The algorithm uses datasets available in the GEDI L4A product to exclude footprints that do not have sufficient geometric or radiometric quality for inclusion in the sample. There are flags in the L4A product that define whether each GEDI observation is representative of the conditions under which GEDI04_A models were developed (Kellner et al., 2023), however these are not used in L4B to avoid systematically omitting high biomass footprints for some prediction strata. L4A footprints are included in the sample used for the L4B algorithm based on the following criteria (pseudo-code in italics):

1. Shots flagged as high quality by the GEDI L2A Footprint Height and Elevation metric product (Hofton and Blair, 2020), which identifies surface waveforms with high fidelity
   l2_quality_flag == 1

2. Only shots with a beam sensitivity >0.98 for tropical Evergreen Broadleaf Tree prediction strata, and beam sensitivity >0.95 elsewhere, were included. These beam sensitivity thresholds were calculated using a 3-sigma signal threshold and selected to provide a sufficiently high signal-to-noise ratio to penetrate the highest canopy cover expected in these regions (Tang et al., 2019).
   if predict_stratum in (EBT_Af, EBT_SA, EBT_SAs):
      geolocation/sensitivity_a2 > 0.98
   else:
      geolocation/sensitivity_a2 > 0.95
3. Shots with high degradation of geolocation performance were excluded from the sample since these may fall outside the geographic extent of a 1 km cell.
   (*degrade_flag* // 10) is not in [5,7,8,9] &
   (*degrade_flag* % 10) is not in [1,2,4,5,6,7,9]

The L4A models are not applicable to waveforms collected under leaf-off conditions in deciduous forests and woodlands. The *land_cover_data/leaf_off_flag* dataset indicates whether the footprint was collected under leaf-off or leaf-on conditions and was derived for a 1 km EASE 2.0 grid using the VIIRS land surface phenology product VNP22Q2 (Zhang et al., 2016). *land_cover_data/leaf_off_flag* = 1 when the footprint was collected after the onset of maximum greenness and before the midpoint of the senescence phase for the given 1 km grid cell. Some L4A models use exclusively RH98 to predict AGBD. When this is the case, the L4B algorithm does not consider the leaf_off_flag because the impact of leaf-off conditions is assumed to be minimal for RH98.

A biomass of zero (with zero model parameter covariance) is assigned to footprints over permanent open water bodies, urban infrastructure, bare ground and permanent snow/ice. In the case of permanent open water bodies, this is assigned using *landsat_water_persistence < 10*, which is the annual water percentage (Pickens et al., 2020). In the urban infrastructure case, this is assigned using urban_percentage < 50, where buildings may create waveforms indicative of substantial tree cover. While urban trees do store biomass, the risk of substantial biomass overprediction is of greater magnitude than the risk of missing biomass in street trees. The variable *urban_percentage* is from a 25 m global urban mask developed by the GEDI Science Team using the TerraSAR-X and TanDEM-X global urban footprint (GUF) data product (Esch et al., 2013). In the bare ground and permanent snow/ice cases, this is assigned using the European Space Agency (ESA) WorldCover 10m 2021 V200 product land-use "bare/sparse" and "snow/ice" classes, mode aggregated to 30m spatial resolution, for which class-specific accuracies are above 92% (Zanaga et al., 2022).

Some users will require an estimate of biomass density associated only with the forested part of 1-km cells. Such users may determine what fraction of the cell using the fine-resolution map of their choice and divide the L4B AGBD estimate by that fraction to arrive at a mean for the forested area. By avoiding pre-defining a sub-kilometer forest/non-forest mask, GEDI allows users to choose whichever mask best suits their application.

3.0 RETRIEVAL ALGORITHM

3.1 Requirement

GEDI's L1A Science Requirement states that 80% of 1-km cells must be estimated to within a standard error of either 20 Mg/ha or 20% of the estimate, whichever is greater.

3.2 Product Specification

The 1 km^2 resolution global EASE-Grid 2.0 is used to partition the GEDI L4A dataset (by footprint center point) into grid cells. This grid features equal-area cells and compatibility with many existing biosphere data sets. More information on this grid can be found in Brodzic et al. (2012), Brodzic et al. (2016) and from NSIDC at https://nsidc.org/data/ease. The L4A footprint biomass predictions are described in that product's User Guide (see Dubayah et al., 2021). The elevation and height metric algorithms used by GEDI are described in Hofton et al. (2020).

GEDI's primary L4B algorithm is hybrid inference (Ståhl et al., 2010; Patterson et al., 2019), in which mean biomass is estimated from an incomplete sample of modeled biomass values (available via the L4A product). Conventionally, the kind of linear observations collected by GEDI, or other air- or space-borne platforms, is treated as a cluster sample. Since at least two clusters are required to calculate a variance under hybrid estimation, GEDI's primary algorithm may only be applied to 1-km cells that have been intersected by at least two ground tracks. At the end of the mission, mean biomass in cells without sufficient clusters will be estimated instead with Generalized Hierarchical Model-Based inference (GHMB; Saarela et al., 2018; Saarela et al., 2022). This approach makes use of a second-level model that extends biomass to a surface predicted from wall-to-wall imagery. The *"Mode of Inference"* L4B variable describes which method was used for a particular cell. Until mission completion, only those cells where hybrid inference is possible will be populated with a mean biomass value (others will have a value of 0).

The distribution of no-data cells in the *"Mode of Inference"* grid is not uniform, with higher non-response found: 1) earlier in the mission life; 2) closer to the equator where the ISS overpass pattern is sparser; 3) in cloudy areas; and 4) in areas where reference ground tracks were not sampled because of the Mission's Year 2 orbital resonance problem. The latter problem involved repeated coverage of some ground tracks at the expense of others because of an unscheduled change in ISS altitude. See Sections 3.4 and 3.5 for details for estimated uncertainty.

Both the hybrid and GHMB variance estimators have two components, the first of which for each is model covariance due to the L4A field-to-GEDI AGBD model (Variance Component 1 variable). For hybrid estimation, the second variance component relates to GEDI's sample design, while for GHMB the second component relates to the fit of an additional model that links L4A predictions to wall-to-wall imagery. The Variance Component 1 and 2 variables allow the user to decompose uncertainty expressed in the

Standard Error variable into its primary components, whether the Mode of Inference is hybrid or GHMB estimation.

The Quality Flag variable allows querying of product progress toward the mission's aforementioned precision target. Table 3.1 lists the outputs described above. In addition to estimated variance components, the user is also given the number of overpasses (sample clusters) and total observations for the grid cell. The minimum number of samples per cluster for GEDI is one; sample numbers are disclosed to provide context for uncertainty estimates. Sårndal et al. (1992) suggest that means and variances may become unstable if there are fewer than 10 samples per cluster (in this case, per overpass) and Thompson (2004) provides similar guidance but identifies the limit as 5 samples per cluster.

Table 3-1 Summary of Gridded Biomass Outputs. Output granules are < 1 Gb each.

Output GEDI Product	**Units of measurement**	**Description**	**Data type**	**No data value**
Mean	Mg ha^{-1}	Estimated mean AGBD for the 1-km grid cell, including forest and non-forest	Float32	-9999
Variance Component 1	N/A	Uncertainty in the estimate of mean biomass due to the field-to-GEDI model used in L4A, expressed as variance.	Float32	-9999
Variance Component 2	N/A	If Mode of Inference = 1, this is the uncertainty due to GEDI's sampling of the 1-km cell, expressed as variance. If Mode of Inference = 2, this is uncertainty due to the model predicting biomass using wall-to-wall data, calibrated with the L4A footprint product	Float32	-9999
Standard Error	Mg ha^{-1}	Standard Error of the mean estimate, combining sampling and modeling uncertainty	Float32	-9999
Percentage Standard Error	Percentage	Standard error as a fraction of the estimated mean AGBD	Uint8	255

Number of Clusters	Count	Number of unique GEDI ground tracks with at least 1 high-quality waveform intersecting the grid cell	UInt16	0
Number of Observations	Count	Total number of high-quality waveforms across all ground tracks within the grid cell	UInt32	0
Quality Flag	None	0 = Outside the GEDI domain 1 = Land surface 2 = Land surface and meets GEDI mission L1 requirement (Percent standard error <20% or Standard Error < 20 Mg ha^{-1})	UInt8	0
L4A Prediction Stratum	None	Determined by Plant Functional Type and continent. Associated with a L4A model parameter covariance matrix that contributes to the Model Error Variance	UInt16	0
Mode of Inference	None	0 = None applied 1 = Hybrid Model-Based 2 = Generalized Hierarchical Model-Based	UInt8	0

3.3 Algorithm Flow

The algorithm flow of the L4B Gridded Aboveground Biomass (GEDI04_B) algorithm is presented in Figure 6.

3.3.1 Inputs

The primary inputs for the biomass gridding process are the footprint-level predictions of aboveground biomass and an indexing of clear shots by cell and by overpass (Table 3-2). Section 1.3 describes how those predictions (which compose the GEDI-L4A product) are made. Using a global dataset of paired simulated waveforms and field measurements of biomass, two parametric models have been created: one using realistically "noised" lidar data; and one using "unnoised" data. While ephemeral noise factors like background illumination and atmospheric attenuation can introduce random error into the footprint-level biomass model, the larger concern from a model-building point of view is footprints where the ground-finding algorithm fails, potentially causing height errors of tens of meters. Such footprints will appear as outliers in height-to-biomass space and can exert strong leverage on the parameters that will ultimately be used to predict biomass at all sites. Accommodating these outliers in the model-building process may cause over- or under-prediction of biomass for the majority of non-problematic footprints.

For this reason, unnoised data are used to create the footprint-level biomass predictions. Waveforms derived with ground-finding errors will be incorrectly predicted, but we assume those cases will be rare, and biomass at most footprints will be predicted with the best possible models. However, the covariance matrix of model parameters used to propagate model error in the variance calculations (below) is generated from the same models updated with predictor variables (*xvar*) that are derived from waveform simulations with realistic estimates of noise. The magnitude of noise in these waveform simulations is determined by randomly sampling from the distribution of beam sensitivity and noise standard deviation observed on orbit, following the quality filtering outlined in Section 2.5, for input to the simulation approach developed in Hancock et al. (2019) and applied in Duncanson et al. (2022). This approach aims to account for a conservatively high amount of model error while ensuring that the majority of footprint-level predictions are as good as possible. In Table 3-2, model parameters come from the unnoised models, which are available in the L4A product. The parameter covariance matrices come from the noised models, which are generated by the L4B algorithm.

Figure 6 highlights the process of aggregating footprint measurements and predictions first at the track (cluster) level and at the grid cell level. Filtering rules based on waveform quality are described above (Section 2.5). It is notable that filtering does not occur based on variables such as slope or canopy cover that may be correlated with biomass; such systematic omissions may result in misleading inferences about mean biomass.

Table 3-2 Summary of GEDI and External Inputs

Input GEDI parameter	**Information needed**	**GEDI Product Source**
Footprint-level geolocation	Which shots are in each 1 km cell on the EASE 2.0 grid	GEDI Level 4A Datasets: *geolocation/lat_lowestmode* *geolocation/lon_lowestmode*
Footprint Identifier	Which shots are in each track (cluster) within each cell	GEDI Level 4A Datasets: *shot_number*
Footprint-level estimated biomass	The level of aboveground biomass (Mg/ha) is estimated for each footprint	GEDI Level 4A Datasets: *agbd* *agbd_t*
Footprint-level predictor variables	The predictor variables (transformed relative height metrics) used to estimate aboveground biomass (Mg/ha) is estimated for each footprint	GEDI Level 4A Datasets: *xvar*
Footprint-level land cover data	Land cover data used to determine which shots to apply a zero mean / zero covariance model, and if the shot was acquired in leaf-off conditions in deciduous forests and woodlands.	GEDI Level 4A Datasets: *land_cover_data/landsat_water_persistence* *land_cover_data/urban_proportion* *land_cover_data/leaf_off_flag* External: ESA WorldCover v002 (Zanaga et al., 2022)
Footprint-level model information	The parameters used to predict biomass at a footprint in each prediction stratum (includes variance/covariance matrix of model parameters)	GEDI Level 4A Datasets: */ANCILLARY/model_data* (compound) *predict_stratum*

Shot quality information	Factors used to filter land surface shots based on measurement quality: waveform fidelity, signal-to-noise ratio (beam sensitivity); and degradation of geolocation.	GEDI Level 4A datasets: *l2_quality_flag* *geolocation/sensitivity_a2* *degrade_flag*
Granule quality information	Factors used to statistically filter sub-orbit granules within individual tiles that have outlier predictor variables and spurious aboveground biomass estimates because of atmosphere (e.g., low flog and cloud).	GEDI Level 2A datasets: *digital_elevation_model* *elev_highest_return*
Global equal-area 1 x 1 km grid	Spatial extent of each cell defined by the EASE 2.0 grid	NA, external (Brodzic et al., 2012)

3.3.2 Processing

The primary processing tasks are described here and are designed to ensure the computing burden in each 1-km cell is light (Figure 6). Means must be calculated for biomass predictions within each overpass, and then a mean of those means must be calculated to produce the gridded estimate (see Equations 3-5 in Section 3.4). Variance estimates are slightly more complex. Two straightforward linearizations (not explicitly described here) limit the complexity of calculations.

The GEDI Level 4A product is partitioned into 72x72 km tiles on the EASE 2.0 grid using footprint ground elevation coordinates (geolocation/[lat|lon]_lowestmode) for global processing. Prior to application of the baseline algorithm (Section 3.4), orbit granules affected by atmosphere, including low cloud and fog, in each 72x72 km tile were identified and removed using an iterative local outlier detection algorithm. This algorithm involved the following steps for each 72x72 km tile:

1. Read all quality shots in a 108x108 km window centered over the 72x72 km tile and calculate the difference between GEDI derived canopy top elevation (*elev_highestreturn*) and the TanDEM-X DEM elevation (*digital_elevation_model*) from the GEDI Level 2A product.
2. At X spatial resolution and using the differences from step (1), calculate:
    a. The Z-score for each cluster with $Z_i = (x_i - \mu)/\sigma$,
    where x_i is the mean difference for cluster i, μ is the mean of all cluster means, and σ is the standard deviation of all cluster means
    b. $\max(x)$, the maximum value of cluster mean differences.
3. Select all clusters that have a least Y cells with a Z-score greater than the 0.99 quantile of Z-scores and a mean cluster difference > 30 m. Add their orbit number to the sub-orbit granule exclude list for that tile
4. Iterate steps 1-3 until there is no further change in the sub-orbit granule by tile exclusion list
5. A small number of sub-orbit granules may be manually identified and added to the sub-orbit granule exclude list in step (4) through visual comparison of the GEDI Level 4B 1 km mean and standard error grids with the output of step (2)

For the GEDI Level-4B PGE Version 2, *X* was 2-km in step (2), *Y* was 5 in step (4), and steps 1 to 3 were applied 5 times. Entire granules instead of tracks per tile were removed per tile to minimize the number of iterations required (all or multiple tracks from an orbit were usually affected).

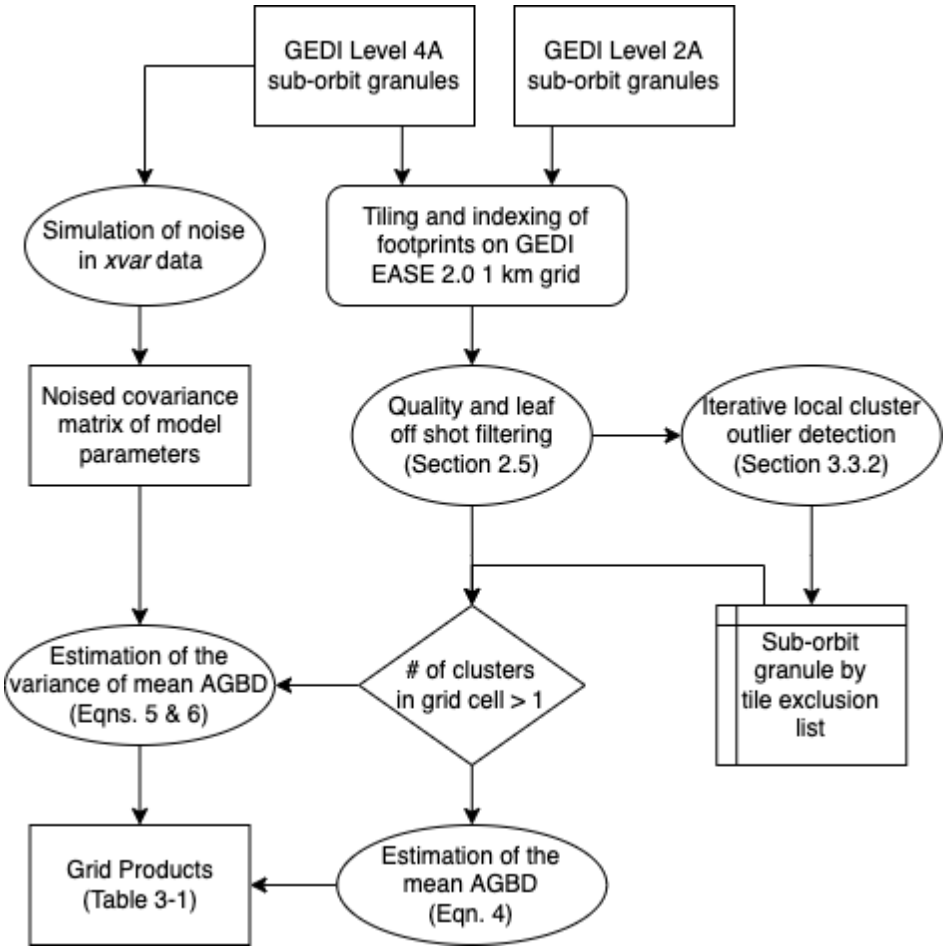


Figure 6. Algorithm Flow Diagram

3.4 Baseline Algorithm

Figure 7 summarizes the calculations and data needed to produce output estimates of mean biomass in each 1-km cell as well as estimates of the standard error around those mean estimates. These processes, also discussed here in detail, will operate at the grid cell level, considering all footprints occurring within the cell's borders. A contingency for cells which do not have the two or more overpasses needed to supply a cross-cluster variance is presented in Section 3.5.

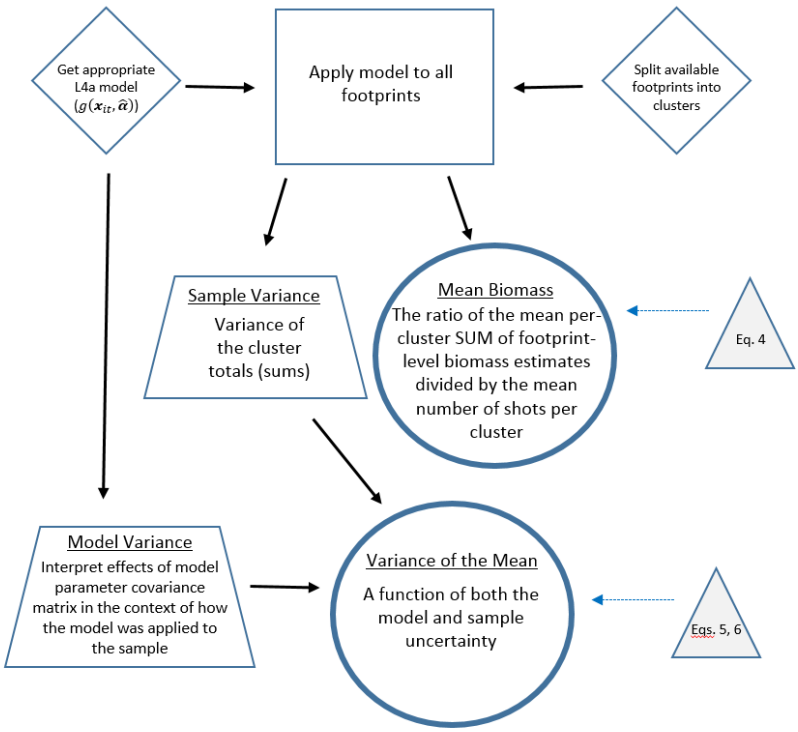


Figure 7. Interaction of equations in derivation of mean and standard error estimates.

Equation [2] will be expressed in an equivalent form, for which there is an estimator allowing calculation of variance. Let $\hat{G}_i = \sum_{t=1}^{T_i} g(\boldsymbol{x}_{it}, \hat{\boldsymbol{\alpha}})$, the cluster total of the predicted biomass per hectare for the pixels in the *i*th cluster. Then equation [2] can be expressed in terms of the ratio of the sum of the cluster totals of the biomass per hectare and sum of the number of pixels per cluster, which then can be expressed as the ratio of mean over the clusters of the cluster totals of the predicted biomass per hectare for each pixel in the cluster and the mean over clusters of the number of pixels per cluster,

$$\mu_Y = \frac{\sum_i^M \hat{G}_i}{\sum_{i=1}^M T_i} = \frac{M^{-1} \sum_i^M \hat{G}_i}{M^{-1} \sum_{i=1}^M T_i} \qquad [3]$$

We propose to use a single random sample of the clusters (i.e., the GEDI tracks). We will estimate the mean of cluster totals and average length separately and combine as a ratio estimator.

$$\hat{\mu}_Y = \frac{\bar{\hat{G}}}{\bar{T}} = \frac{m^{-1} \sum_i^m \hat{G}_i}{m^{-1} \sum_{i=1}^m T_i} = \frac{\sum_i^m \hat{G}_i}{\sum_{i=1}^m T_i} = \frac{\sum_i^m \sum_{t=1}^{T_i} g(\boldsymbol{x}_{it}, \hat{\boldsymbol{\alpha}})}{n} \qquad [4]$$

where $n = \sum_{i=1}^m T_i$ is the number of GEDI shots in the sample. This estimator combines design-based sampling with model-based estimation, i.e., there is a sample of auxiliary information (as opposed to wall-to-wall auxiliary information, such as landsat) and the population element value biomass is estimated using a model. These types of estimators

have been denoted hybrid estimation (Ståhl et al, 2016). In setting of "Stratification and post-stratification", Ståhl et al (2011), proposed an estimator of the form of equation [4] and derived an approximate variance and proposed an estimated variance. In expression [5] below is estimated variance proposed in Ståhl et al (2011), with the addition of the finite correction factor.

$$\hat{V}(\hat{\mu}_Y) = \frac{1}{\bar{T}^2}\left(1 - \frac{m}{M}\right)\frac{\sum_{i=1}^{m}\left(\hat{G}_i - \hat{\mu}_Y T_i\right)^2}{m(m-1)} + \frac{1}{\bar{T}^2}\sum_{j=1}^{p}\sum_{k=1}^{p}\widehat{Cov}_{S2}\left(\hat{\alpha}_j, \hat{\alpha}_k\right)\hat{\bar{G}}_j'\hat{\bar{G}}_k' \quad [5]$$

The first term is due to the sample design and the second term is due to model uncertainty. If we assume a linear model, that is $g(\boldsymbol{x}_{it}, \hat{\boldsymbol{\alpha}}) = \alpha_1 + \sum_{j=2}^{p} x_{itj}\alpha_j$, where x_{itj} is the *j*th component of $\boldsymbol{x}_{it}$, then,

$$\hat{\bar{G}}_j' = \frac{1}{m}\sum_{i=1}^{m}\sum_{t=1}^{T_i} x_{itj}$$

Using the linear model assumption, we can rewrite the second term of equation [5]. Let $\hat{\bar{\boldsymbol{x}}}$ be the vector of the means of the cluster totals of the predictor variables, i.e., $\hat{\bar{\boldsymbol{x}}}_j = \frac{1}{m}\sum_{i=1}^{m}\sum_{t=1}^{T_i} x_{itj}$, for $j = 1, \ldots, p$. Then the second component of right-hand side of equation [7] can be expressed in matrix notation.

$$\sum_{j=1}^{p}\sum_{k=1}^{p}\widehat{Cov}_{S2}\left(\hat{\alpha}_j, \hat{\alpha}_k\right)\hat{\bar{G}}_j'\hat{\bar{G}}_k' = \hat{\bar{\boldsymbol{x}}}^T\widehat{Cov}(\hat{\boldsymbol{\alpha}})\hat{\bar{\boldsymbol{x}}} \quad [6]$$

where $\widehat{Cov}(\hat{\boldsymbol{\alpha}})$ is the predicted covariance matrix of the fitted parameters. Note that $\hat{\bar{\boldsymbol{x}}}_1$ is equal to $\bar{T}$.

3.5 Contingency Algorithm

Cells not meeting precision requirements (Quality Flag = 1) with hybrid inference at the end of the mission will follow a different biomass estimation process. Generalized Hierarchical Model-Based inference (GHMB), as described and tested in the GEDI context by Saarela et al. (2018), relies upon both a sample of high-quality data and a wall-to-wall modeled surface of biomass to produce inferences about the forest population of an area such as a 1-km cell. AGBD is measured at a set of field plots presumed to be representative of the population that includes the 1-km cell; these plots do not need to fall within the cell. AGBD is then modeled from the plots to a set of coincident real or simulated GEDI waveforms, allowing AGBD prediction at all GEDI shots (this process is equivalent to the L4A process). Predicted biomass at all shots is then compared to coincident satellite imagery (likely from Landsat, Sentinel, or TanDEM-X). In GHMB, uncertainties from both models (field to lidar, lidar to Landsat)

are combined appropriately, such that estimate of the mean and standard error may be derived for areas of interest. Saarela et al. (2018) demonstrate that this approach is likely to be appropriate for many 1-km cells. Research by the GEDI Science Team is demonstrating that in most realistic cases hybrid and GHMB estimators yield approximately equivalent estimates of mean biomass and estimated variance (Saarela et al., 2018; Qi et al., 2019; Saarela et al., 2022). GHMB estimates of mean AGBD and the standard error of those estimates will fill in estimates for the zero- and one-pass cells across the grid.

4.0 CALIBRATION AND VALIDATION

The Forest Inventory and Analysis Program of the US Forest Service (FIA; Bechtold and Patterson, 2005) has a mission similar to GEDI's: to report biomass and other forest characteristics over areas so large that they would be infeasible to exhaustively measure. A range of corporate and municipal forest managers rely upon FIA's estimates for decisions related to timber supply, finding sites for bioenergy projects, old-growth and endangered species management, and carbon accounting. Assessments in these fields require unbiased estimates for which uncertainty can be clearly stated. Sample theory allows FIA, under a few simple and clearly stated assumptions, to calculate the standard error of its estimates on the basis of sample number and variance. While GEDI's approach must also consider footprint-level model error, sample theory likewise performs a similar function: both inferences and uncertainties may be formally derived for an area of interest (i.e., each 1-km grid cell) under clear assumptions.

The GEDI mission, like FIA's national forest inventory, relies upon its own theoretically calculated estimates of precision for purposes of determining mission performance with respect to its science goals. The primary risk associated with this strategy is incorrect biomass model specification at the footprint level. Model-based inference (including the hybrid approach used here) assumes that the parameter selection process during model-building includes variance that is representative of conditions in the domain in which the model is being applied. Significant model misspecification at this stage, due possibly to selection or measurement bias, can result in unacknowledged estimation errors.

Post-launch calibration and validation efforts are therefore primarily focused at the scale of individuals footprints. Efforts include expanding the acquisition of co-located simulated GEDI and field data in ecosystems under-represented in the GEDI FSBD used for model-building. Such data will provide insight into previous effects of generalization error on cell-level estimates and supply a better basis for calibrating the footprint-level models. Updated footprint-level models will also support future reprocessing of the grid-based biomass estimates. Details of the footprint-level calibration and validation plan can be found in the theoretical basis document for that product (L4A; Kellner et al., 2023).

Independent validation is also important to assess unacknowledged estimation errors in the GEDI Level 4 strategy for biomass inference. Activities include comparison against design-based estimates from national forest inventories (e.g., Menlove & Healey, 2020) at multiple scales (e.g., Dubayah et al., 2022), which are not subject to the same modeling errors that the GEDI estimates are. Large area acquisition of a sample of GEDI grid cells by NASA's Land Vegetation and Ice Sensor (LVIS) and other ALS platforms may also be used to support validation of the GEDI Level 4B product. However only a very small fraction of GEDI's global grid may be sampled with this approach. Ignoring ground-to-airborne model error can also introduce significant unappreciated uncertainty (Saarela et al., 2016). Accounting for ground-to-airborne model error is possible (Nelson et al., 2017), therefore good practice protocols need to be followed (see Duncanson et al., 2021) to ensure expected uncertainties are substantially lower than GEDI's stated levels of error. Generally, GEDI's internal estimates of uncertainty, based upon sampling theory

and properties of GEDI's sample and models, should be the most globally consistent and comprehensive metric of the precision of the L4B gridded biomass product.

5.0 REFERENCES

Bechtold WA, Patterson PL. 2005. The enhanced Forest Inventory and Analysis program – national sampling design and estimation procedures. Gen Tech Rep SRS-80. Asheville, NC. USDA Forest Service Southern Research Station. 85 p.

Blair JB and Hofton MA. 1999. Modeling laser altimeter return waveforms over complex vegetation using high-resolution elevation data. Geophysical Research Letters, 26: 2509-2512.

Brodzik, M. J., B. Billingsley, T. Haran, B. Raup, M. H. Savoie. 2012. EASE-Grid 2.0: Incremental but Significant Improvements for Earth-Gridded Data Sets. *ISPRS* International Journal of Geo-Information, 1(1):32-45, doi:10.3390/ijgi1010032. http://www.mdpi.com/2220-9964/1/1/32.

Brodzik, M. J., B. Billingsley, T. Haran, B. Raup, M. H. Savoie. 2014. Correction: Brodzik, M. J. et al. EASE-Grid 2.0: Incremental but Significant Improvements for Earth-Gridded Data Sets. ISPRS International Journal of Geo-Information 2012, 1, 32-45. ISPRS International Journal of Geo-Information, 3(3):1154-1156, doi:10.3390/ijgi3031154. http://www.mdpi.com/2220-9964/3/3/1154

Corona P, Fattorini L, Franceschi S, Scrinzi G, Torresan C (2014) Estimation of standing wood volume in forest compartments by exploiting airborne laser scanning information: model-based, design-based, and hybrid perspectives. Can J Forest Res 44:1303–1311, http://dx.doi.org/10.1139/cjfr-2014-0203

Davidson, FM and Sun X. 1988, Gaussian approximation versus nearly exact performance analysis of optical communication systems with PPM signaling and APD receivers. Communications, IEEE Transactions on, 36, 1185–1192.

Dubayah, R.O., J. Armston, J.R. Kellner, L. Duncanson, S.P. Healey, P.L. Patterson, S. Hancock, H. Tang, J. Bruening, M.A. Hofton, J.B. Blair, and S.B. Luthcke. 2021. GEDI L4A Footprint Level Aboveground Biomass Density, Version 2. ORNL DAAC, Oak Ridge, Tennessee, USA. https://doi.org/10.3334/ORNLDAAC/1986

Dubayah, R., Armston, J., Healey, S.P., Bruening, J.M., Patterson, P.L., Kellner, J.R., Duncanson, L., Saarela, S., Ståhl, G., Yang, Z., Tang, H., Blair, J.B., Fatoyinbo, L., Goetz, S., Hancock, S., Hansen, M., Hofton, M., Hurtt, G., Luthcke, S., 2022. GEDI launches a new era of biomass inference from space. Environ. Res. Lett. 17, 095001. https://doi.org/10.1088/1748-9326/ac8694

Duncanson, L., Armston, J., Disney, M., Avitabile, V., Barbier, N., Calders, K., Carter, S., Chave, J., Herold, M., MacBean, N., McRoberts, R., Minor, D., Paul, K., Réjou-Méchain, M., Roxburgh, S., Williams, M., Albinet, C., Baker, T., Bartholomeus, H., Bastin, J.F., Coomes, D., Crowther, T., Davies, S., de Bruin, S., De Kauwe, M., Domke, G., Dubayah, R., Falkowski, M., Fatoyinbo, L., Goetz, S., Jantz, P., Jonckheere, I.,

Jucker, T., Kay, H., Kellner, J., Labriere, N., Lucas, R., Mitchard, E., Morsdorf, F., Næsset, E., Park, T., Phillips, O.L., Ploton, P., Puliti, S., Quegan, S., Saatchi, S., Schaaf, C., Schepaschenko, D., Scipal, K., Stovall, A., Thiel, C., Wulder, M.A., Camacho, F., Nickeson, J., Román, M., Margolis, H. (2021). Aboveground Woody Biomass Product Validation Good Practices Protocol. Version 1.0. In L. Duncanson, M. Disney, J. Armston, J. Nickeson, D. Minor, and F. Camacho (Eds.), Good Practices for Satellite Derived Land Product Validation, (p. 236): Land Product Validation Subgroup (WGCV/CEOS), doi:10.5067/doc/ceoswgcv/lpv/agb.001

Duncanson L. *et al.* 2022. Aboveground biomass density models for NASA's Global Ecosystem Dynamics Investigation (GEDI) lidar mission. Remote Sensing of Environment, 270, 112845.

Esch, T., Marconcini, M., Felbier, A., Roth, A., Heldens, W., Huber, M., Schwinger, M., Taubenböck, H., Müller, A., Dech, S., 2013. Urban Footprint Processor—Fully Automated Processing Chain Generating Settlement Masks From Global Data of the TanDEM-X Mission. IEEE Geoscience and Remote Sensing Letters, 10, 1617–1621. https://doi.org/10.1109/LGRS.2013.2272953

Gregoire, T.G., G. Ståhl, E. Næsset, T. Gobakken, R. Nelson, & S. Holm. 2011. Model-assisted estimation of biomass in a LiDAR sample survey in Hedmark County, Norway. *Canadian Journal of Forest Research* 41: 83-95. doi: 10.1139/X10-195.

Gregoire TG, Næsset E, McRoberts RE, Ståhl G, Andersen H-E, Gobakken T, Ene L, Nelson R. 2016. Statistical rigor in LiDAR-assisted estimation of aboveground forest biomass. Remote Sensing of Environment, 173: 98-108.

Hancock, S., Armston, J., Hofton, M., Sun, X., Tang, H., Duncanson, L.I., Kellner, J.R., Dubayah, R., 2019. The GEDI Simulator: A Large-Footprint Waveform Lidar Simulator for Calibration and Validation of Spaceborne Missions. Earth and Space Science 6, 294–310. https://doi.org/10.1029/2018EA000506

Hancock, S., Disney, M., Muller, J.-P., Lewis, P., and Foster, M., 2011, A threshold insensitive method for locating the forest canopy top with waveform lidar. Remote Sensing of Environment, 115, 3286–3297.

Healey SP, Patterson PL, Saatchi SS, Lefsky MA, Lister AJ, Freeman EA. 2012. A sample design for globally consistent biomass estimation using lidar data from the Geoscience Laser Altimeter System (GLAS). Carbon Balance and Management, 7:1-16.

Hofton, M.A., Blair, J.B., 2020. Algorithm theoretical basis document (ATBD) for GEDI transmit and receive waveform processing for L1 and L2 products.

Houghton RA, Hall F, Goetz SJ. 2009. Importance of biomass in the global carbon cycle. Journal of Geophysical Research, 114: 1-13.

Kellner, J. R., Armston, J., and Duncanson, L. 2023. Algorithm theoretical basis document for GEDI footprint aboveground biomass density. Earth and Space Science, 10, e2022EA002516. https://doi.org/10.1029/2022EA002516

Magnussen S. 2015. Arguments for a model-dependent inference? Forestry, 88(3): 317-325.

Margolis HA, Nelson RF, Montesano PM, Beaudoin A, Sun G, Andersen HE, Wulder M. 2015. Combining satellite lidar, airborne lidar and ground plots to estimate the amount and distribution of aboveground biomass in the Boreal forest of North America. Can J Forest Res 45(7):838–855, http://dx.doi.org/10.1139/cjfr-2015-0006

McRoberts RE, Næsset E, Gobakken T. 2014. Estimation for inaccessible and non-sampled forest areas using model-based inference and remotely sensed auxiliary information. Remote Sensing of Environment, 154, 226-233

Menlove J, Healey SP. 2020. A Comprehensive Forest Biomass Dataset for the USA Allows Customized Validation of Remotely Sensed Biomass Estimates. Remote Sensing, 12(24):4141. https://doi.org/10.3390/rs12244141

Mitchard TA et al. 2014. Markedly divergent estimates of Amazon forest carbon density from ground plots and satellites. Global Ecology and Biogeography, 23(8): 935-946. 10.1111/geb.12168

Neigh, C.S.R., Nelson, R.F., Ranson, K.J., Margolis, H.A., Montesano, P.M., Sun, G., et al. 2013. Taking stock of circumpolar forest carbon with ground measurements, airborne, and space LiDAR. Remote Sensing of Environment, 137 (2013), 274-287. doi: 10.1016/j.rse.2013.06.019.

Nelson R, Boudreau J, Gregoire TG, Margolis H, Næsset E, Gobakken T et al., 2009. Estimating Quebec provincial forest resources using ICESat/GLAS. Canadian Journal of Forest Research, 39, 862-881.

Nelson R, Gobakken T, Næsset E, Gregoire TG, Ståhl G, Holm S, Flewelling J. 2012. Lidar sampling - using an airborne profiler to estimate forest biomass in Hedmark County, Norway. Remote Sensing Environment, 123:563–578, http://dx.doi.org/10.1016/j.rse.2011.10.036

Nelson R, Margolis H, Montesano P, Sun G, Cook B, Corp L, Andersen H-E, deJong B, Pellat FP, Fickel T, Kauffman J, Prisley S. 2017. Lidar-based estimates of aboveground biomass in the continental US and Mexico using ground, airborne, and satellite observations. Remote Sensing of Environment, 188, 127-140.

Patterson, P.L., Healey, S.P., Ståhl, G., Saarela, S., Holm, S., Andersen, H.-E., Dubayah, R.O., Duncanson, L., Hancock, S., Armston, J., Kellner, J.R., Cohen, W.B., Yang, Z.,

2019. Statistical properties of hybrid estimators proposed for GEDI—NASA's global ecosystem dynamics investigation. Environ. Res. Lett. 14, 065007. https://doi.org/10.1088/1748-9326/ab18df

Pickens, A. H., Hansen, M. C., Hancher, M., Stehman, S. V., Tyukavina, A., Potapov, P., et al., 2020. Mapping and sampling to characterize global inland water dynamics from 1999 to 2018 with full Landsat time-series. Remote Sensing of Environment, 243, 111792. https://doi.org/10.1016/j.rse.2020.111792

Qi, W., Saarela, S., Armston, J., Ståhl, G. & Dubayah, R. Forest biomass estimation over three distinct forest types using TanDEM-X InSAR data and simulated GEDI lidar data. Remote Sensing of Environment 232, 111283 (2019).

Saarela S, Schnell S, Grafström A, Tuominen S, Nordkvist K, Hyyppä J, Kangas A, Ståhl G. 2015 Effects of sample size and model form on the accuracy of model-based estimators of growing stock volume in Kuortane, Finland. Can J Forest Re 45:1524–1534. http://dx.doi.org/10.1139/cjfr-2015-0077

Saarela S, Holm S, Grafström A, Schnell S, Næsset E, Gregoire TG, Nelson RF, Ståhl G. 2016. Hierarchical model-based inference for forest inventory utilizing three sources of information. Annals of Forest Science, DOI 10,1007/s13595-016-0590-1. 16p.

Saarela S, Holm S, Healey SP, Andersen H-E, Petersson H, Prentius W, Patterson PL, Næsset E, Gregoire TG, Ståhl G. 2018. Generalized Hierarchical Model-Based Estimation for Aboveground Biomass Assessment Using GEDI and Landsat Data. *Remote Sensing*, 10(11):1832. https://doi.org/10.3390/rs10111832

Saarela, S., Holm, S., Healey, S.P., Patterson, P.L., Yang, Z., Andersen, H.E., Dubayah, R.O., Qi, W., Duncanson, L.I., Armston, J.D. and Gobakken, T., 2022. Comparing frameworks for biomass prediction for the Global Ecosystem Dynamics Investigation. Remote Sensing of Environment, 278, p.113074.

Särndal C, Swensson B, Wretman J. 1992. Model assisted survey sampling. Springer-Verlag, New York. 694 p.

Ståhl G, Holm S, Gregoire TG, Gobakken T, Næsset E, Nelson R 2011. Model-based inference for biomass estimation in a LiDAR sample survey in Hedmark County, Norway. Canadian Journal of Forest Resources, 41, 96-107

Ståhl G, Saarela S, Schnell S, Holm S, Breidenbach J, Healey SP, Patterson PL Magnussen S, Næsset E, McRoberts R, Gregoire T. 2016. Use of models in large-area forest surveys: comparing model-assisted, model-based and hybrid estimation. Forest Ecosystems, 3, 5

Thompson, S. K. 2002. Sampling (2nd ed.). New York: John Wiley & Sons, Inc. 367 pp.

Woodall CW, Coulston JW, Domke GM et al., 2015. The US forest carbon accounting framework: stocks and stock change. Gen Tech Rep NRS-154. Newtown Square, PA: USDA Forest Service Northern Research Station. 49p.

Wulder MA, White JC, Nelson RF, Næsset E, Ørka HO, Coops NC, Hilker T, Bater CW, Gobakken T. 2012. Lidar sampling for large-area forest characterization: A review. Remote Sensing of Environment, 121, 196-209.

Zanaga, D., Van De Kerchove, R., Daems, D., De Keersmaecker, W., Brockmann, C., Kirches, G., Wevers, J., Cartus, O., Santoro, M., Fritz, S., Lesiv, M., Herold, M., Tsendbazar, N.E., Xu, P., Ramoino, F., and Arino, O., 2022. ESA WorldCover 10 m 2021 v200, https://doi.org/10.5281/zenodo.7254221

Zhang, X., Friedl, M., A., Henebry, G., M., 2016. Algorithm theoretical basis document: VIIRS land surface phenology product.

Zolkos SG, Goetz SJ, Dubayah R. 2013. A meta-analysis of terrestrial aboveground biomass estimation using lidar remote sensing. Remote Sensing of Environment: 128: 289-298.

GLOSSARY/ACRONYMS

CAL/VAL	Calibration and Validation
FIA	Forest Inventory and Analysis
GEDI	Global Ecosystem Dynamics Investigation
L4A	Level 4A
L4B	Level 4B
AGBD	Aboveground Biomass Density
GHMB	Generalized Hierarchical Model-Based inference
RH	Relative Height metrics