International Journal of Applied Earth Observation and Geoinformation 129 (2024) 103797

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag

Connecting spaceborne lidar with NFI networks: A method for improved estimation of forest structure and biomass

Paul B. May [a,b,*], Ralph O. Dubayah [b], Jamis M. Bruening [b], George C. Gaines III [c]

[a] Department of Mathematics, South Dakota School of Mines & Technology, Rapid City, SD, United States
[b] Department of Geographical Sciences, University of Maryland - College Park, College Park, MD, United States
[c] Rocky Mountain Research Station, U.S. Department of Agriculture (USDA) Forest Service, Missoula, MT, United States

ARTICLE INFO

Keywords:
Lidar
Forestry
Small area estimation
Regression
Spatial statistics

ABSTRACT

Spaceborne lidar provides a unique opportunity to supplement the field plot measurements of national forest inventories (NFIs) by providing dense measurements of vertical canopy structure. For full waveform instruments such as the Global Ecosystem Dynamics Investigation (GEDI), measurements take the form of reflected energy as a function of height within an observed footprint. Many forest attributes cannot be directly computed from the waveforms, and thus statistical models relating the measurements to target attributes must be trained using field plot data. Because the discrete footprint samples produced by spaceborne lidar instruments have little chance of collocation with pre-existing field plots, model calibration remains a primary challenge. We leveraged the density and spatial correlation of GEDI observations to make predictions of the cumulative waveform over 326,787 NFI plots distributed across the contiguous United States. The predictions yield probability distributions, giving not only a predicted waveform but quantification of prediction uncertainty. The product of this work is a data set comprised of the NFI plots paired with the predictive waveform distributions. The predicted waveforms and their uncertainties can be used to train models with "error-in-variables" techniques that account for and filter the uncertainty in the predicted waveforms. This data set and the corresponding techniques allow statistically rigorous model training without requirements of collocation, yielding downstream forest attribute predictions that are consistent with NFI measurements and estimates. We demonstrate the data set by training a model for forest above ground biomass density (AGBD) and then use the model to produce a spatially complete 1 km map of AGBD for the contiguous United States. We further compare our AGBD predictions to US NFI estimates at a 64,000 hectare scale, showing the increase in precision to be proportional to the collection of almost 500,000 additional NFI plots across forested regions of CONUS.

1. Introduction

Large scale monitoring of forest attributes is crucial for carbon accounting, forest management and promoting habitat suitability. For countries with developed national forest inventories (NFIs), this need has traditionally been addressed with random samples of *in situ* field plots on which forest attributes of interest are measured. Estimates of population averages/totals can then be constructed using these samples. The collection of field plots is laborious and expensive, limiting the spatial and temporal density of samples and therefore the resolution of the subsequent estimates. These limitations often preclude accurate information at resolutions that are actionable and policy relevant. For instance, landowners requiring audits of carbon stocks or timber supply typically control tracts of land far smaller than the sampling density of NFIs (Prisley et al., 2021), and regional species distribution models for species dependent on forest habitat have little to gain from population estimates representing hundreds of square kilometers (Burns et al., 2020). Such needs, faced with the limitations of plot-based NFI's, motivate the use of remote sensing to improve the resolution of estimates while maintaining a broad scope, both in terms of spatial extent and the attributes that can be estimated.

Light detection and ranging (lidar) has emerged as the premier remote sensing technology for measuring forest structure due to its sensitivity to the vertical and horizontal arrangement of canopy elements. A common implementation of lidar is through airborne laser scanning (ALS) platforms that provide spatially complete measurements at sub-meter resolutions. However, individual ALS campaigns are typically limited in spatial extent. Spaceborne lidar resolves this by providing measurements from orbiting satellites at near global extent. The Geoscience Laser Altimeter System (GLAS) onboard the

* Corresponding author at: Department of Mathematics, South Dakota School of Mines & Technology, Rapid City, SD, United States.
E-mail address: paul.may@sdsmt.edu (P.B. May).

https://doi.org/10.1016/j.jag.2024.103797
Received 10 November 2023; Received in revised form 7 March 2024; Accepted 25 March 2024
Available online 5 April 2024

Ice, Cloud, and Land Elevation Satellite (ICESat) was one of the first such instruments, collecting full waveform lidar measurements sampled at discrete 70 m diameter footprints throughout its operation from 2003–2010. While the primary focus of ICESat was measurement of ice sheets, clouds and aerosols, its data were frequently used for vegetation measurements (Boudreau et al., 2008; Saatchi et al., 2011; Baccini et al., 2012; Margolis et al., 2015). GLAS/ICESat was superseded by the Advanced Topographic Laser Altimeter System (ATLAS) onboard ICESat −2, currently in operation, which instead of full waveform measurements implemented a single photon counting technique (Abdalati et al., 2010). The most recent spaceborne lidar mission to launch is the Global Ecosystem Dynamics Investigation (GEDI), which is specifically designed to measure ecosystem structure (Dubayah et al., 2020a). Mounted on the International Space Station, GEDI samples full waveform lidar measurements at discrete 25 m diameter footprints.

While lidar is sensitive to forest structure, many attributes cannot be directly computed from the measurements and statistical models are required to relate the measurables to the target attribute. Models can be trained using field plot measurements which are considered 'ground truth'. Because preceding and existing spaceborne lidar implementations collect discrete footprint samples with significant gaps between, model training remains a primary challenge, as the probability of footprints overlapping existing field plots is small. A wide range of techniques have been employed for model training, most with the final goal of predicting forest biomass or carbon. Saatchi et al. (2011) calibrated GLAS measurements to field plots by taking a subset of the assembled plots that either intersected or shared a forest stand with a GLAS footprint. Baccini et al. (2012) instead undertook a post-hoc field campaign to collect plot data under GLAS footprints. Boudreau et al. (2008) and Margolis et al. (2015) used ALS data as a conduit, training models between the spatially complete ALS and the field plots, and then training models between GLAS and the ALS predictions. Duncanson et al. (2022) also used ALS data as a conduit, but instead used the dense ALS point returns to simulate GEDI waveforms over the plots (Hancock et al., 2019), yielding direct training pairs. These models drive the current GEDI biomass products (Dubayah et al., 2021, 2022a).

The above techniques force the omission of a large number of potential field plots. Requiring collocation of lidar footprints or ALS data with field plots by happenstance results in a convenience sample, rather than a randomly dispersed probability sample, and can leave entire ecoregions unrepresented in the training set. This can lead to biased predictions in these regions, as models extrapolate outside of the training domain Dubayah et al. (2022b) and Bruening et al. (2023). Further, for moderate-sized footprints such as from GEDI, the probability of intersecting footprints/plots is near zero, making mediating ALS data the only viable option within this category of techniques. Post-hoc field campaigns to collect data under footprints can produce representative training samples, but are expensive and arduous, somewhat foiling the original motivation of incorporating remote sensing data: to circumvent the cost of additional field samples.

It would be advantageous for several reasons if NFIs could be used as training databases for spaceborne lidar missions. Not only do NFIs constitute a massive pre-existing training set dispersed across a country, but public and private agencies often desire models and predictions consistent with existing and historical NFI measurements and estimates (Gillespie, 2000). While depending on ALS or footprint intersections would seriously limit the number of viable NFI plots, GEDI does collect dense footprint samples, and many NFIs plots have a concentration of nearby footprints. Like most geographic variables, the observed waveforms are spatially correlated, allowing inference at unobserved locations.

We leverage this spatial correlation with multivariate spatial models to make predictions of cumulative GEDI waveforms over 326,787 NFI plots evenly dispersed across the contiguous United States (CONUS). The spatial models yield not only predicted waveforms, but rich quantification of uncertainty through a full predictive distribution, giving a range of plausible waveforms for each NFI plot. The NFI plots paired with the predicted waveforms and their uncertainties form a powerful training data set, the primary product of this work. Using the data set and error-in-variables techniques to account for the waveform prediction uncertainty, GEDI models can be trained to the NFI measurements.

While some version of this technique could be employed with other spaceborne lidar instruments, GLAS/ICESat or ATLAS/ICESat −2, the GEDI instrument provides several advantages. Both ICESat iterations were optimized for ice sheets not vegetation. ATLAS employs photon-counting technology, where individual measurements must be aggregated along-track by hundreds of meters to adequately convey canopy structure; waveform lidar can depict canopy structure with a single measurement. GLAS employs waveform lidar technology, but the size of the footprints (60–120 m) creates topography related issues in the measurements. The GEDI instrument provides waveform lidar measurements within moderate-sized footprints at a greater density than either of the ICESat iterations.

Our advocated techniques allow rigorous models relating the lidar observables to ground-measured forest attributes without requirements of collocation. By utilizing NFI plot networks, we gain a large and geographically representative training data set and ensure the ability to produce predictions that are coherent and unbiased relative to the NFI measurements. Both the process of data set generation and model training are conducted in a probabilistic manner, accounting for uncertainties at each stage and yielding trustworthy final audits of prediction uncertainty. We illustrate the power of our method by training a model for forest above ground biomass density (AGBD) and subsequently producing a 1 km map of AGBD for CONUS. We further compare our model AGBD predictions to NFI estimates at a 64,000 ha scale, showing substantial improvements in precision. Biomass density is an ideal test variable due to its strong relationship to lidar measurements and its importance to NFIs from the standpoint of carbon accounting, climate treaties and general forest monitoring. We emphasize, however, that models can be trained between GEDI waveforms and any NFI-measured attribute with this method.

2. Data

2.1. GEDI

GEDI collects lidar samples within 25 m diameter footprints along 8 parallel ground tracks, with a 60 m center-to-center spacing between footprints along-track and 600 m between tracks. GEDI collected data from April 2019 to March 2023 and is planned to resume collection in late 2024. The relative height (RH) metrics are quantiles from the cumulative waveform, where RH50, for example, gives the height above ground below which 50% of the waveform energy was returned. Here we consider the RH metrics to be the core measurable of waveform lidar, as they are often used as predictors of forest attributes and the original intensity-versus-distance waveform can still be recovered. The GEDI L2A data product (Dubayah et al., 2020b) gives footprint observations of the RH metrics, discretized in 1% increments $(\mathrm{RH0}, \mathrm{RH1}, \dots, \mathrm{RH100})$. For brevity, we will simply refer to this vector of length 101 as the waveform. The L2A product also contains a quality flag, identifying the most useful waveform returns, and a leaf-off flag, identifying footprints collected over deciduous forests during leaf-off conditions. We filtered all GEDI data used in this study to meet the quality flag and to observe leaf-on conditions. Every waveform observation is associated with a point location, which we naturally consider to be the center of the circular footprint; however, current geolocation accuracy has a standard deviation of around 10 m, preventing exact geolocation of an observed footprint center (Beck et al., 2021).

2.2. FIA

The USDA Forest Service Forest Inventory and Analysis (FIA) Program administers the strategic NFI of the United States. At its core, this NFI is comprised of a nationwide network of over 300,000 sample locations, with a base sampling intensity of approximately 1 location per 2400 ha. The sampling intensity is sometimes regionally increased from the base intensity to satisfy specific inventory objectives. Ground measurements of forest attributes taken at these locations (every 5 or 7 years in the eastern U.S. and every 10 years in the western U.S.) facilitate statistical inference on a wide range of forest population parameters. The locations were selected according to an equal probability, pseudo-systematic sample design, which is assumed by the program to well approximate a simple random sample. At each location a suite of core forest attributes are measured.

Each plot consists of a cluster of four circular subplots of diameter 48 ft (∼14.6 m), with one central subplot surrounded by three plots at azimuths of 0, 120 and 240 degrees, with a center-to-center distance of 120 ft (∼36.6 m) to the central plot (Fig. 1). Each plot is associated with a point location, which is considered to be the center of the central subplot. We consider a GEDI footprint and an FIA plot collocated if their given point locations are identical; because of the differing geometry of a footprint and plot, this does not mean the two sources measure the exact same area (Fig. 1). For a thorough description of the FIA plot design, sampling design and estimation procedures, see Westfall et al. (2022).

FIA constructs estimates of forest attribute area averages/totals using direct, design-based estimates: Within a given area, only the contained plot measurements and the sampling design of the plot locations is used to drive statistical inference. An advantage of design-based estimation is that no distributional or model assumptions are imposed on the target forest attribute. The primary disadvantage is the estimate precision is proportional to the sample size within the area. This limits the scale at which precise estimates can be produced proportional to the sampling density of the plots.

We made GEDI waveform predictions over all base intensity plot locations contained within a L4 ecoregion: due to the random selection of plot locations within 2400 ha sampling hexagons, some plot locations are off the coast by random chance and therefore not contained in the land partition of the L4 ecoregions (but also not useful as forest attribute training plots). This gives a total of 326,787 field plot locations. The FIA plot locations are confidential to maintain land owner privacy and avoid tampering. FIA releases public coordinates, which are fuzzed by ∼1 km, and occasionally swapped within U.S. county boundaries. Under agreement and collaboration with the U.S. Forest Service, we utilized the true coordinates for this study. Like the GEDI footprints, many FIA plots are not geolocated perfectly. The most recent documentation from the agency estimates the average geolocation error to be ∼10 m (Hoppus and Lister, 2005), an error similar to that of the GEDI footprints.

The FIA plot parameters used in our study are the measurements of forest AGBD, the reported plot locations and the date of measurement. Forest AGBD is an important attribute, directly conveying the amount of tree matter and indirectly terrestrial carbon stocks. Particular to FIA, a field plot must be at least 10% stocked by tree species to be considered forested. Forest attributes such as AGBD on non-forested plots are recorded as zeros. We will refer to exact-zero AGBD plot values, but this is best understood as AGBD magnitudes below a detection threshold, as some marginal amount of forest biomass may still be present. The geographic coordinates of the plots are critical as they determine the relative position to nearby GEDI footprints, which drives the interpolation methods used to generate predictions. Finally, the time of measurement is useful during model training to filter plots deemed to possess measurements too old to relate to contemporary GEDI observations.

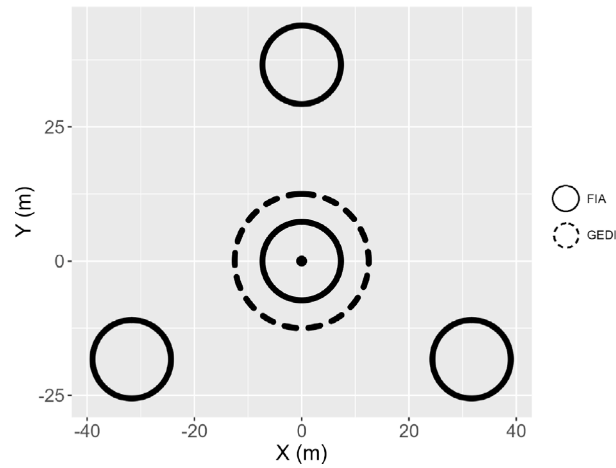


Fig. 1. A GEDI footprint and FIA field plot collocated, meaning they share the same point location given by center of the footprint for GEDI and the center of the central subplot for an FIA plot.

2.3. Auxiliary data

We used a forest classification map derived from the National Land Cover Database (NLCD) 2019 land cover product (Dewitz, 2021) to assist the waveform prediction. The product gives predicted land cover classification at a 30 m pixel resolution derived from Landsat imagery. We convert the raster into a binary map, delineating forest and non-forest. The binary value associated with a given footprint or NFI plot is the pixel containing the footprint/plot center.

The spatial models used to predict GEDI waveforms depend on estimated parameters that may be ecoregion specific. We used the EPA Level 4 (L4) ecoregions to partition CONUS into distinct model domains (https://www.epa.gov/eco-research/level-iii-and-iv-ecoregions-continental-united-states). The L4 partition consists of 967 ecoregions characterized by relative internal-homogeneity of environmental resources. Finally, we used the validation map of Menlove and Healey (2020), which provided a comparison to demonstrate the improvements in estimation precision from our method. The map consists of design-based estimates (and estimate uncertainties) of AGBD within 64,000 ha hexagons partitioning CONUS (12,591 hexagons). The hexagon estimates were produced with the FIA plots within each respective hexagon according to typical FIA procedures. On average, 27 plots per hexagon were used to construct the estimates, with a total of 338,606 FIA plots across all hexagons. The data set provides the average measurement year for the plots within each hexagon. The median of the average measurement year is 2015, with a minimum of 2009.

3. Methods

The methods of this work are decomposed into three stages (Fig. 2). In the first stage, the dense but spatially incomplete GEDI observations are interpolated to make predictions at NFI plot locations. The predicted waveforms and their uncertainties paired with the plot measurements form a training data set to serve the second stage: Regression models between GEDI RH metrics and NFI measurements are trained using error-in-variables techniques to account for the uncertainties in the waveform predictions. In the third stage, the regression model from the previous stage is applied to the densely observed GEDI footprints to yield predictions of the target forest attribute. We demonstrate the latter two stages using forest AGBD as the target attribute.

This section gives a general overview of the statistical methods and a more specific description of the logistics for our implementation over CONUS. More specific mathematical and computational details on the methods can be found in Appendix and the provided code.

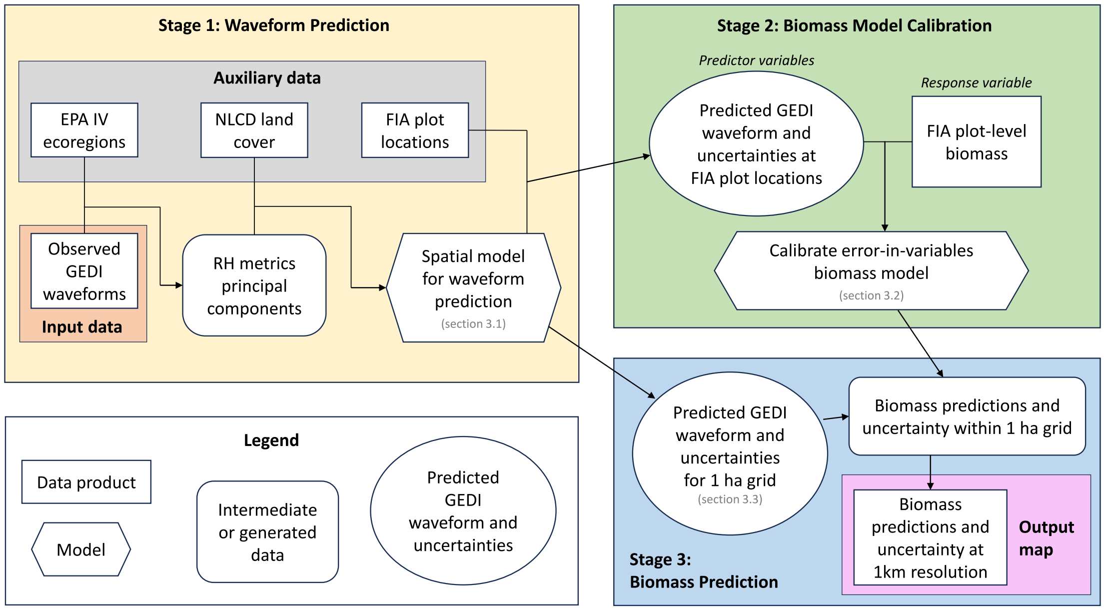


Fig. 2. A method flowchart depicting the three distinct stages. Stages 2 and 3 are executed using forest AGBD as the target attribute, but a different NFI-measured attribute could be used with no fundamental changes to the methodology.

3.1. Waveform prediction

The probability of perfect collocation between a GEDI footprint and an NFI plot is slight. However, there is substantial probability of a multitude of GEDI footprints near enough to an NFI plot to be informative. We used a multivariate spatial model similar to Taylor-Rodriguez et al. (2019) and May et al. (2022) to leverage the spatial correlation of the GEDI observations and generate predictions at NFI plot locations (Fig. 3).

Let $\boldsymbol{x}(s)$ be a GEDI waveform at location s, a vector of length 101 containing the RH metrics from 0 to 100. The observed waveforms are projected onto their principal components, ordered from largest to lowest magnitude of variation. The principal components are approximately uncorrelated, and are therefore modeled and interpolated separately onto the NFI plot locations. For each principal component, we used the NLCD forest/non-forest map as a binary regression covariate. This binary covariate had little overall explanatory power, but slightly improved the spatial coherence of the residual variation by accounting for abrupt breaks in the landscape. After the principal components have been predicted, the predictive distribution of the waveform can then be reconstructed. The predicted waveform is given by the expected value (the center of mass) of the predictive distribution and the prediction uncertainty is given by the dispersion of the distribution around the expected value.

As a computational expedient, we considered only the spatial correlation of the first k principal components, assuming the rest to be spatially independent, where k is the smallest integer such that the first k principal components account for over 99.99% of the total waveform variation. The rationale for this is two-fold: First, by definition the remaining $101-k$ principal components hold little additional information. Second, the principal components become increasingly spatially incoherent, making informative interpolation increasingly futile.

We partitioned the interpolation over CONUS by the EPA L4 ecoregions. The GEDI footprints were first filtered by the L2A quality flag and for leaf-on observations. For each ecoregion, a new set of principal components was computed and model parameters were estimated for each principal component. Using these estimates, predictions for the NFI plot locations within that ecoregion were computed. To limit the magnitude of GEDI data handled by the spatial model, we used the nearest 25 GEDI footprints for every NFI plot. omitting the rest. If this resulted in fewer than 1000 GEDI footprints for an ecoregion, a random sample of the omitted GEDI footprints was reclaimed to raise the sample size to 1000 and ensure confident model parameter estimates. The median/mean number of footprints used per ecoregion was 3654/6945.

To gauge the confidence of the predictions across the ecoregions, we examine the total prediction variance for each NFI plot location (sum of prediction variances across all RH metrics) relative to the total population variance (sum of the sample variance across all RH metrics):

$$\mathrm{RV}_{ij} = \frac{\mathrm{tr}(\boldsymbol{\Sigma}_{i,\mathrm{pred}})}{\mathrm{tr}(\boldsymbol{\Sigma}_{j,\mathrm{pop}})}, \tag{1}$$

where $\mathrm{tr}(\cdot)$ denotes the trace of a matrix, or sum of the elements on the main diagonal, $\boldsymbol{\Sigma}_{i,\mathrm{pred}}$ is the 101×101 prediction covariance matrix for the 101 RH metrics at the ith NFI plot, and $\boldsymbol{\Sigma}_{j,\mathrm{pop}}$ is the population sample covariance matrix, same dimensions, for the RH metrics within ecoregion j. A relative variance of 0 indicates the prediction is perfect, possible only through a direct collocation of a GEDI footprint and that plot (this never occurred). A relative variance of 1 indicates that the spatial model prediction is no more confident than simply using the sample mean of waveforms for that ecoregion, implying that there are no GEDI footprints within the spatial correlation range of the NFI plot: Indeed, if there are no GEDI footprints within the spatial correlation range of the NFI plot, the multivariate spatial model will predict the estimated mean waveform for that ecoregion. A relative variance near 1 does not mean that the prediction is completely uninformative, since an ecoregion sample mean is somewhat locally specific, but does imply much higher uncertainty and a less informative contribution to the training data set. For each ecoregion, we examine the first, second and third quartiles (25%, 50%, 75% quantiles) for the relative variance. For example, a second quartile of 0.6 indicates that 50% of the plot predictions within that ecoregion have a relative variance less than 0.6.

The primary factors driving the level of prediction certainty are the range of the spatial correlation (how close does a GEDI footprint need to be to an NFI plot to be informative?) and the density of quality GEDI observations. The range of the spatial correlation is dictated by model parameters that are estimated from the data for each ecoregion. Within the range of latitudes for CONUS, the distribution of GEDI observations is somewhat affected by latitude. GEDI tracks converge at the orbital limits of 51.6°N and 51.6°S, therefore footprint densities

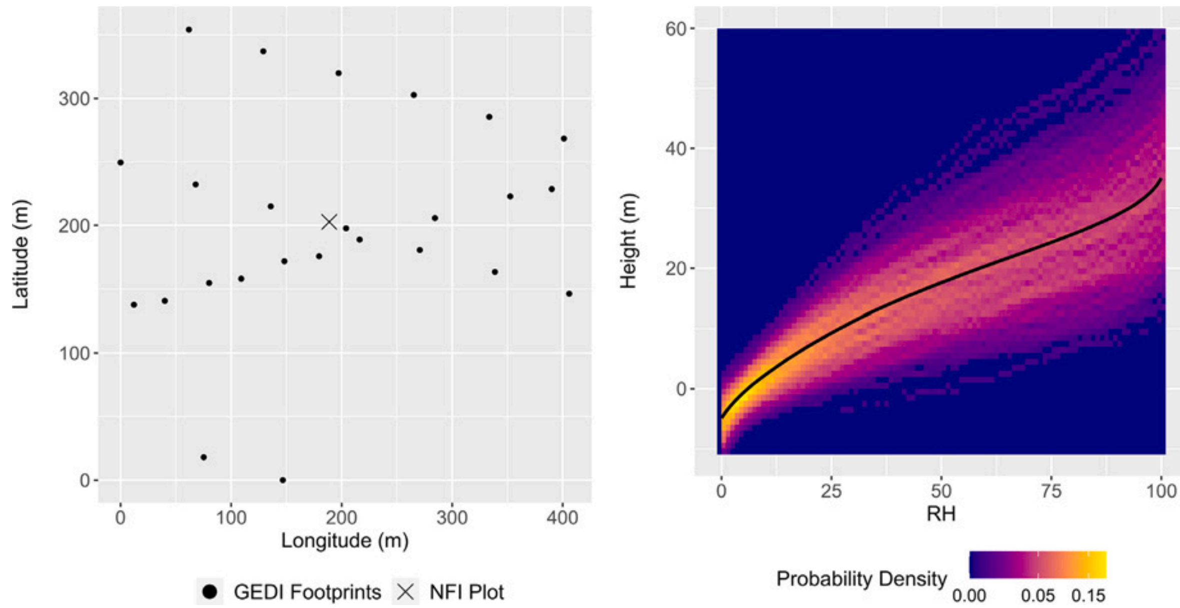


Fig. 3. An example of a waveform prediction over a randomly selected location posing as an NFI plot for the example. The footprint-observed waveforms and their spatial correlation are used to make a waveform prediction at the plot location. The predictions yield not only a predicted waveform (black line, right graphic) but a probability distribution of plausible waveforms, providing uncertainties as well.

increase further from the equator. However, the primary driver of footprint density in our study was presence of deciduous trees: we filter observations for leaf-on conditions, truncating a large amount of data for deciduous areas.

3.2. Model training

The methods of Section 3.1 yield a data set comprised of GEDI waveform predictions paired with NFI plot measurements to serve the training of regression models between RH metrics and the plot attributes. Differing from the typical regression training set, the covariates (the RH metrics) are predicted with error as opposed to observed with certainty. This gives a Berkson error-in-variables problem (Berkson, 1950): the predicted waveforms are compressed towards their ecoregion mean, with more uncertain predictions having more severe compression. The variances/covariances of the covariate errors are provided by the spatial model that executed the waveform prediction, allowing the use of error-in-variables techniques to account for and filter the covariate error and rigorously estimate the regression model parameters (Dellaportas and Stephens, 1995). Given a fixed training sample size, the confidence in the model parameter estimates will be lower in the error-in-variables scenario, proportional to the magnitudes of the covariate uncertainties, than if the covariates had been observed perfectly. Indeed, the magnitude of uncertainty for a given predicted waveform is often quite large. However, this is counterbalanced by the massive training sample sizes that can be achieved by not requiring collocation.

We used the predicted waveform data set to train a regression model between square-root RH95 and cube-root NFI-measured AGBD. The transformations help induce a linear relationship between the covariate and AGBD, and predictive distributions are easily back-transformed to yield predictions of AGBD. For the remainder of the section we simply write RH95 and AGBD. To attenuate temporal decorrelation between the GEDI and NFI measurements, we used NFI plots with a measurement year of 2015 or later, leaving 209,816 plots for model training. This filter provided a compromise between contemporary measurements and maintaining a large training data set distributed across CONUS.

The FIA AGBD plot measurements have a large proportion (61%) of exact-zero measurements: due to random location selection, many plots will be located in non-forested locations. It is certainly desirable to be able to predict non-forest status, and correspondingly predict zero forest biomass, but this does require an additional modeling component. We used the zero-inflated spatial model of Finley et al. (2011) (see Fig. 4). The first modeling layer is a logistic regression, leveraging RH95 and a smooth spatial adjustment to predict whether a positive biomass is present. Given a positive biomass, the magnitude of biomass is predicted with a continuous regression model leveraging RH95 and a different smooth spatial adjustment. The spatial adjustments are given by Gaussian processes, accounting for geographic variations in the relationship between RH95 and AGBD (Gelfand and Schliep, 2016; May et al., 2023). The variation of the Gaussian processes are inferred from the data during model training.

We used a Bayesian approach to model training. In this context, accounting for the prediction error in the training waveforms is conceptually straight-forward, as the true unobserved waveforms are simply additional unknown model quantities, the prior distributions for which are given by the waveform predictive distributions (Dellaportas and Stephens, 1995). Bayesian inference takes prior beliefs on unknown quantities and updates them into posterior beliefs given new data. Here, the waveform predictions using the nearby GEDI footprints form the prior beliefs on the waveforms collocated with the NFI plots. Given a model for a particular forest attribute, the plot measurements of this attribute are the 'new data' that inform the posterior distribution for the collocated waveforms. During model training, plausible waveforms are repeatedly sampled from their posterior distribution. This creates a feedback loop, where the model influences the sampled waveforms and the sampled waveforms influence the model. If there is no detectable relationship between the waveforms and the plot attribute, the prior and posterior distribution for the waveforms will be identical.

We emphasize the correct interpretation of the model and resulting predictions. Because the response variables used in model training are NFI plot measurements, the model does not predict the target attribute within the confines of the 25 m diameter GEDI footprint. Rather the model predicts a hypothetical NFI measurement centered at that location within the particular geometry of the NFI plot design. When aggregating many predictions to make estimates of area averages the distinction means little, but may be pertinent when interpreting individual footprint predictions.

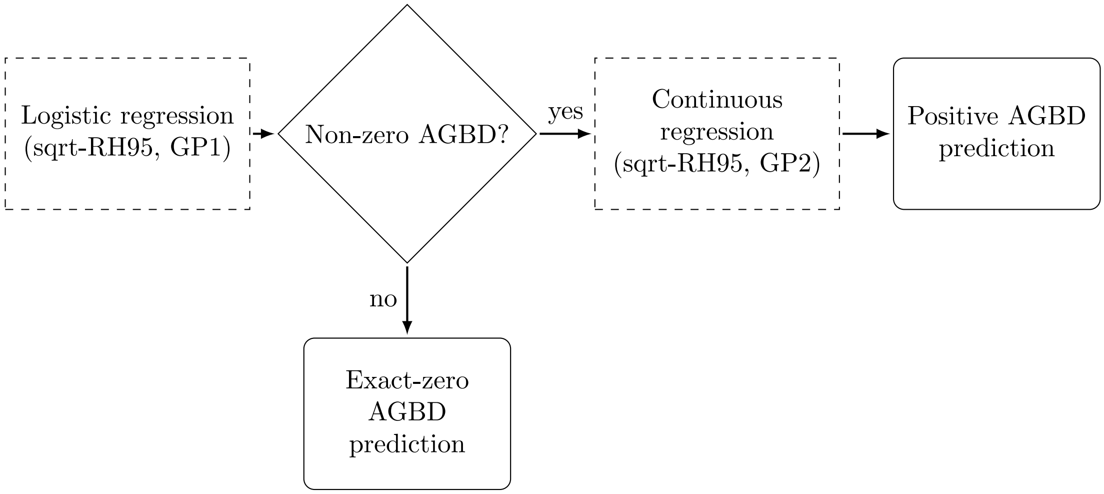


Fig. 4. Diagram of the zero-inflated spatial model.

3.3. Forest attribute prediction

The motivation to train regression models between GEDI waveforms (and remote sensing measurements in general) and field-measured attributes is to leverage the density of the remote sensing observations to provide precise estimates of population estimates at finer scales: Within CONUS, there is on average 105 leaf-on and quality-flagged GEDI footprints per square kilometer, in contrast to approximately one FIA plot per 24 km^2. Once a regression model has been trained, a number of options exist to convert GEDI footprint predictions to estimates of area averages.

The GEDI L4B product uses hybrid inference (Patterson et al., 2019) to produce AGBD estimates at 1 km. Hybrid inference takes a design-based approach to the observed RH metrics paired with a regression model to relate the RH metrics to AGBD. The method is both statistically defensible and computationally inexpensive. The disadvantage is that 1 km cells with less than two GEDI orbits cannot yield an estimate due to the design-based portion of the method.

We employ another spatial model similar to that of Section 3.1, except now only the RH metrics used in the regression model must be considered. First, the spatial model produces RH metric predictions (and uncertainties) on a one hectare grid, representing a hypothetical complete grid of FIA plots. The predicted RH metrics (and uncertainties) are fed through the regression model to yield a 1 ha grid of forest attribute predictions. The predicted grid can then be aggregated to arbitrary resolutions/areas, yielding predictive distributions at that resolution.

An advantage of the spatial model is the ability to make gap-free predictions. Areas with few or no GEDI observations can leverage neighboring observations to make predictions, though the resulting uncertainties may be large, depending on the strength of the area's spatial coherence. A disadvantage of the spatial model is the substantial computational expense. For the 1 km AGBD map presented in Section 4.2, we partitioned CONUS into ∼12,000 hexagons of area 64,000 ha each (Menlove and Healey, 2020). For each hexagon, all the GEDI observations within a 2 km buffer of the hexagon were used to make predictions for 1 km cells within the hexagon. A new spatial model was trained for each hexagon. This alleviated computational burden and also allowed local adaptations for the parameters of the spatial model. However, it does have the undesirable property that two 1 km cells immediately across a hexagon border from each other will have used potentially different parameter estimates to make their predictions. This induced some occasional edge effects, although our investigations have found these to be mild: at the greatest extreme, two hexagons east of the southern Sierra Nevada had a discontinuity from their neighboring hexagons of 15–20 Mg/ha. For all other hexagons, the edge effects were on the scale of 1 Mg/ha when present at all, which is negligible compared to the natural variation of biomass.

3.4. Validation

We conducted two separate cross-validation studies to validate the predicted waveforms and the example AGBD model respectively. For the first study, we selected three L4 ecoregions with distinct forest types (North Cascades, Middle Rockies, Central Appalachians) to investigate the validity of the predicted waveforms and their reported uncertainties (Fig. 5). The Cascades and Middle Rockies are primarily evergreen with high and moderate biomass levels respectively. The Appalachians are primarily deciduous with moderate to moderate-high biomass levels. Within each ecoregion, a random sample of 1000 GEDI footprints was withheld as a test set of hypothetical NFI plots. Using the remaining footprints and the methods in Section 3.1 we predicted waveform distributions over the test locations. Using the waveform distributions, we generated 95% prediction intervals for all RH metrics, i.e., the model claims there is a 95% probability the true RH value is within this interval. If the predicted distributions are accurate, the coverage rate of the prediction intervals will be approximately 95%. Because of the finite test sample, we would not expect even a perfect model to achieve exactly a 95% coverage rate. Rather, we construct a range of appropriate values as the 95% credible interval on success rates given 1000 independent tests, each with a 95% probability of success. This gives a range of appropriate coverage rates between 93.6% and 96.3%.

The second study validates the example AGBD model and its predictions to demonstrate the ability of the data set to generate models (and subsequent predictions) that are consistent with NFI measurements. The 209,816 field plots used for training were split into ten folds. Iteratively, each fold was withheld and the model trained with the remaining nine folds. Using the trained model, plot-level predictions were made for each plot within the withheld fold. Note that both the model training and the subsequent predictions utilize waveforms predicted over the plot, therefore this study simultaneously validates the example AGBD model and the RH95 values extracted from the predicted waveforms. For the forested plots (positive AGBD measurements), we examined the coverage of the 95% credible intervals. Again, the coverage rate should be approximately 95% for a well-performing model. Additionally, we examined the calibration curve (Pearce and Ferrier, 2000, Section 4) of the logistic regression model that discriminates between non-forest (zero AGBD) and forest (positive AGBD). The calibration curve compares the smoothed predicted probabilities

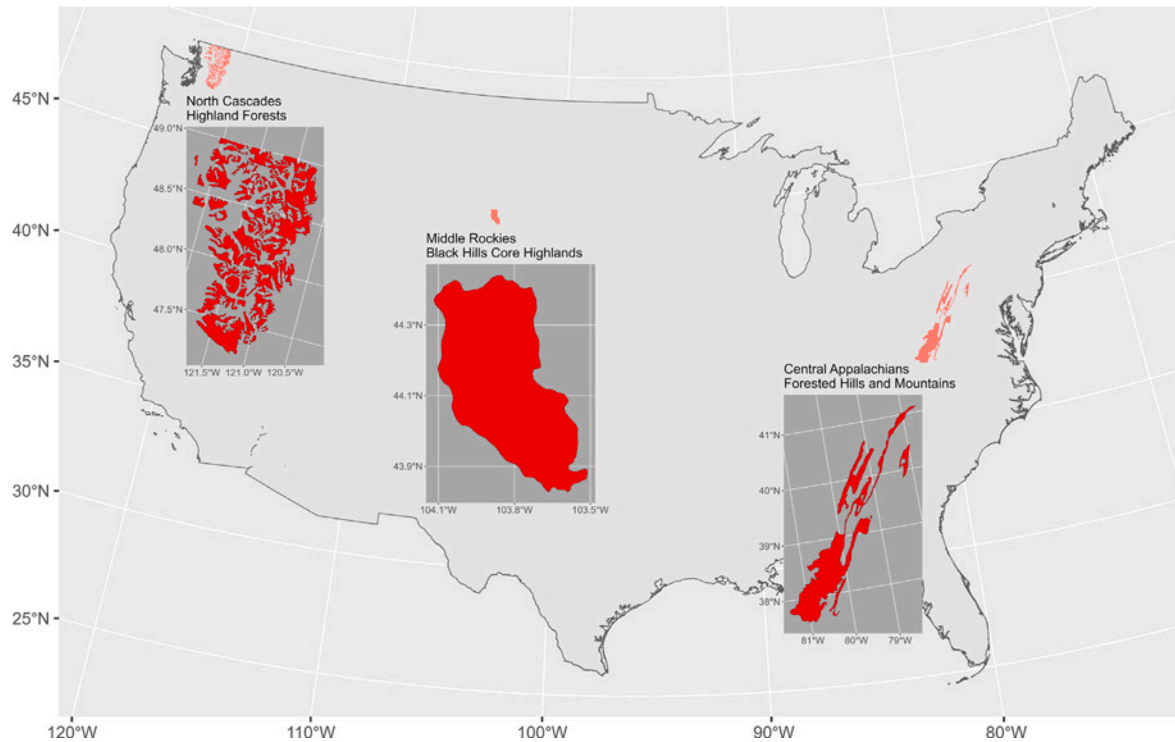


Fig. 5. The three EPA L4 ecoregions selected for the waveform cross-validation study, manually selected to be diverse test cases.

of forest-presence versus the smoothed actual proportions of forest-presence. For a well-performing model, the two will be near-identical and the calibration curve will follow the one-to-one line.

Finally, we generated predictions of AGBD within the 64,000 ha hexagons of Menlove and Healey (2020) and compared our predictions and uncertainties to the design-based estimates and their uncertainties. In addition to comparing the uncertainties directly, a metric we used to quantify the improvements in precision is the effective sample increase (ESI), which represents the number of additional field plots required for the design-based estimate to achieve the same precision as the model prediction. Let $\hat{\sigma}_d^2$ be the estimated variance (squared standard error) of the design-based estimate, n be the number of plots used to construct the design-based estimate and σ_m^2 be the variance of the model prediction. The estimated ESI is

$$\hat{\text{ESI}} = n \cdot \left(\frac{\hat{\sigma}_d^2}{\sigma_m^2} - 1 \right), \tag{2}$$

which we computed for every hexagon with a model prediction greater than 50 Mg/ha and design-based standard error greater than 10 Mg/ha (2774 hexagons). We made such restrictions to avoid massive (but often not meaningful) estimated ESI values. If the design-based estimate is already precise in absolute terms, large relative improvements are not worth mentioning: a standard error of 2 Mg/ha reduced to 1 Mg/ha implies a three-fold increase in field plot samples, but these hypothetical additional samples are arguably wasted. For similar reasons, large relative improvements over low biomass areas will produce misleading ESI estimates.

Two caveats to the above comparison are warranted. First, the two measures of uncertainty have different theoretical interpretations, each appropriate to their respective statistical origins. The design-based variance represents the variability of the estimate given the repeated sampling of new plot locations (by how much might the estimate change if a different $n = 27$ plot locations within the hexagon were selected?). The Bayesian, model-based variance represents the variability in the posterior distribution, i.e., the dispersion in our beliefs on the unknown quantity given the observed data. However, in practice the two measures of uncertainty will be interpreted similarly. Second, the hexagon NFI estimates used many plot measurements older than the temporal cutoff we used for model training, 2015. Indeed the median of the hexagon-average measurement year is 2015. Given the same temporal cutoff, and consequently fewer field plots, it would be expected that the design-based variances (and our estimated ESI) would be higher. But enforcing our time frame would leave many hexagons without an equal-probability sample, making defensible design-based estimates impossible.

4. Results

Section 4.1 details the results of the waveform prediction over the NFI plots as well as the waveform cross-validation study. Section 4.2 provides the fitted AGBD model, the resulting 1 km AGBD map for CONUS, the results of the AGBD model cross-validation study, and the comparison to the hexagon design-based estimates.

4.1. Waveform prediction

We generated predictive waveform distributions over 326,787 NFI plots across CONUS using the methods in Section 3.1. The confidence in the predictions varies widely by region. The primary factors driving prediction uncertainty are the density of observed GEDI footprints and the strength of the spatial correlation. While the strength of the spatial correlation varies substantially by ecoregion, the more drastic factor is the density of observations, with the footprint density of the deciduous east far lower due to the filtering of leaf-off observations (Fig. 6). Some regions in the Appalachians have a filtered footprint density as low as 25 footprints per km^2, whereas some regions in the Pacific Northwest have densities nearing 250 footprints per km^2. This results in low prediction confidence for many NFI plots in the east, whereas the remainder of CONUS possesses an abundance of plots with informative predictions (Fig. 7). To conduct the predictions, we spatially interpolated the first k principal components, treating the rest as spatially independent, where k is the smallest integer such the remaining $101 - k$ components hold less than 0.01% of the total

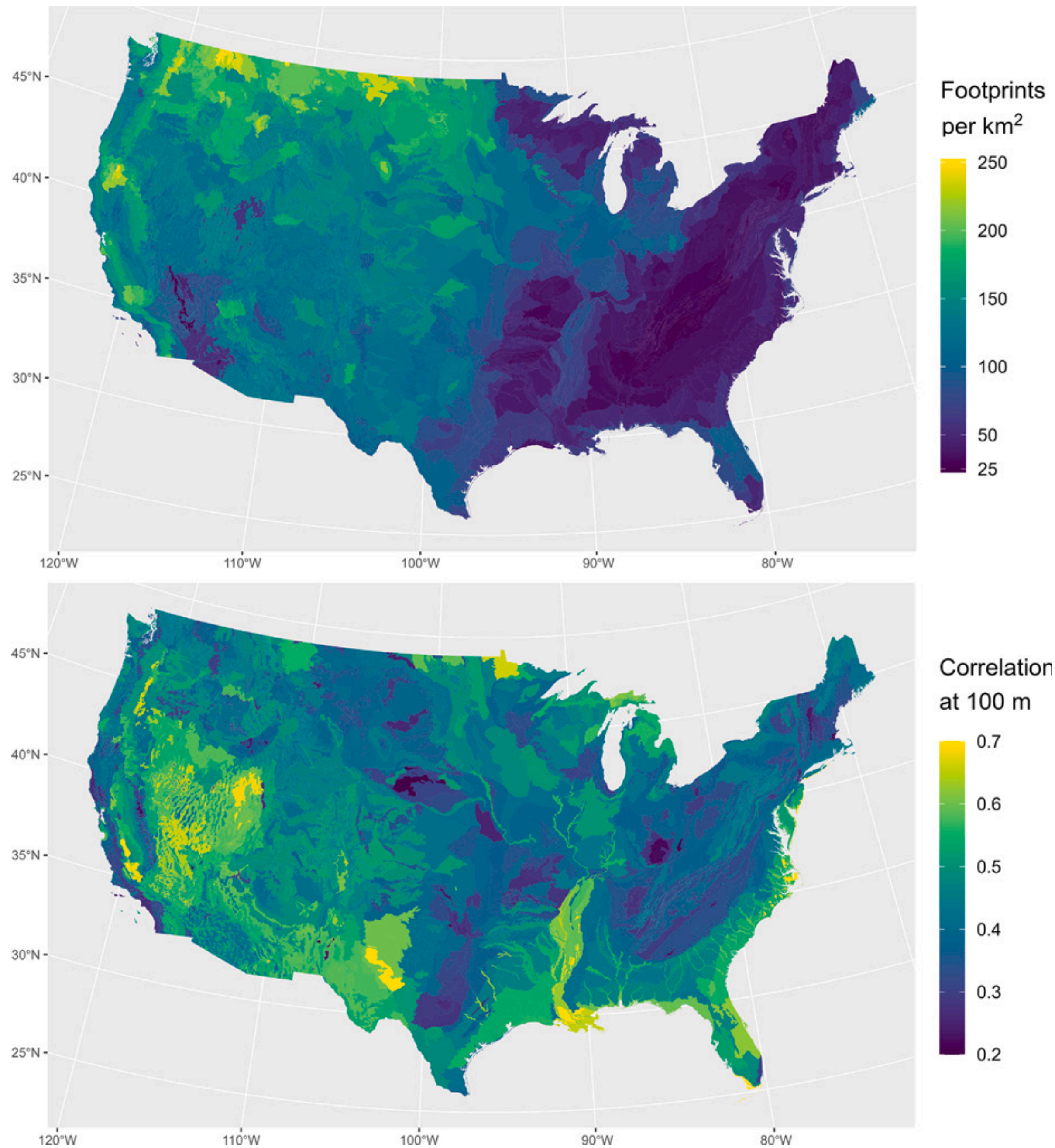


Fig. 6. Density of GEDI footprints (top) and the spatial correlation at 100 m for the dominant principal component of the waveform (bottom). The footprint density for the deciduous eastern CONUS is low due to the filtering of leaf-off observations.

waveform variation. The median k across all ecoregions was 9 and 90% of ecoregions had a k between 7 and 12.

We now turn to the waveform cross-validation study, where for each of the three selected ecoregions we predicted waveform distributions over 1000 randomly withheld test footprints. The coverage rates of the 95% credible intervals were near nominal across all RH metrics and ecoregions, with some RH metrics for the Middle Rockies and Central Appalachians deviating slightly from the range of appropriate values (Fig. 8). For example, the coverage rate of the 95% credible intervals for RH 61 – 94 within the Central Appalachians was below the adequate threshold of 93.6%, reaching as low as 90.0%. The interpretation of this result is that the reported theoretical uncertainties are slightly lower than the true error observed in the cross-validation study. We give an example taken from each ecoregion comparing the predictive distribution to the witheld waveform in Fig. 9.

4.2. Model for forest AGBD

We fit the example model between square-root RH95 and NFI measurements of AGBD measured 2015 or later (Section 3.2). Comparing the prior expected values of RH95 versus AGBD for the forested plots, we see compression along the covariate axis due to the Berkson errors, but an underlying positive polynomial relationship to AGBD is still evident (Fig. 10). Training the model with error-in-variable techniques accounts for the waveform prediction uncertainty and corrects the horizontal compression. Using this model, we produced a 1 km resolution map of forest AGBD for the entire CONUS (Fig. 11). The uncertainties for the eastern US are the largest, with relative standard deviations (standard deviation divided by expected value) often above 25%, whereas the western US is consistently below 25% (Fig. 12).

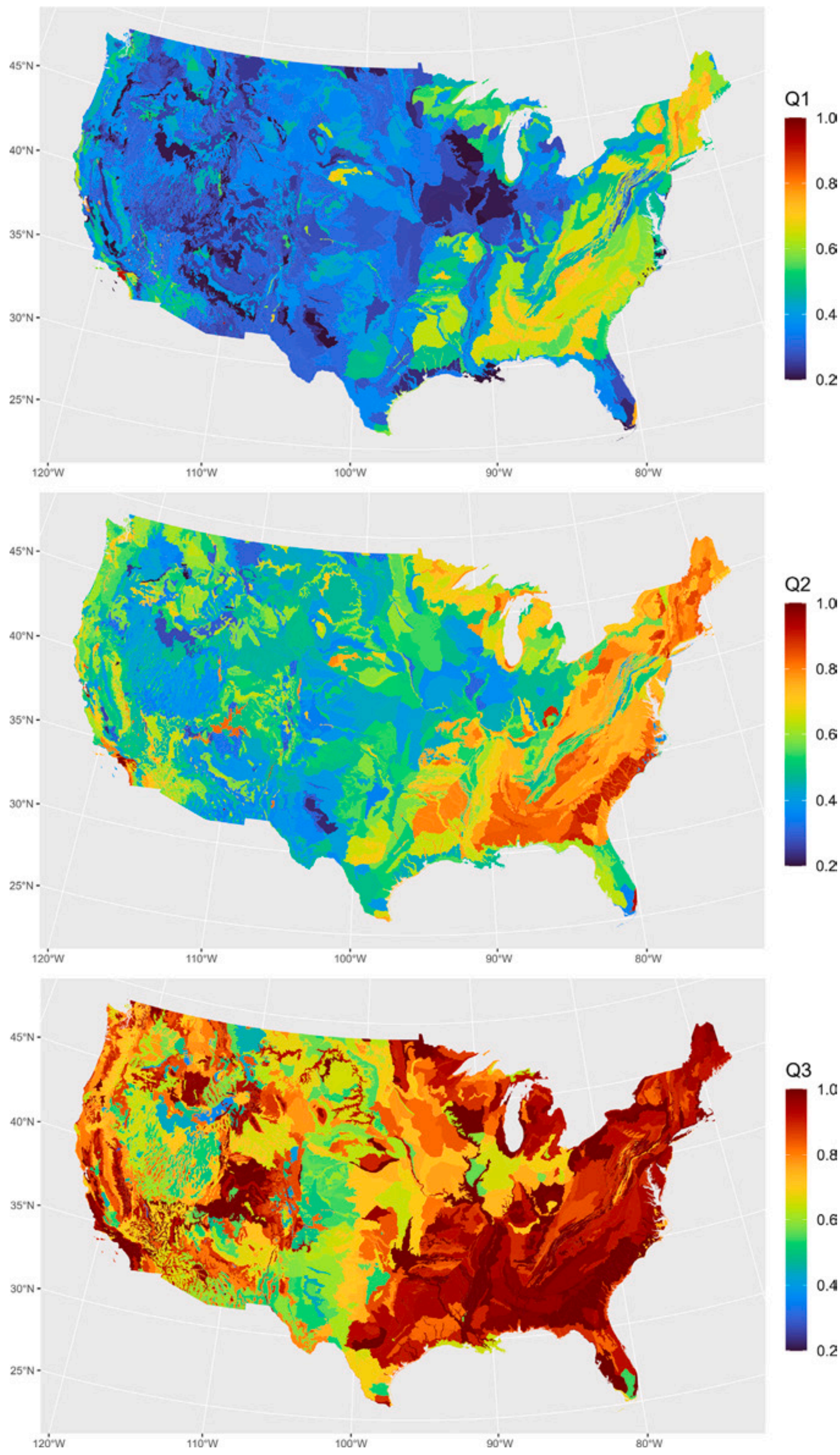
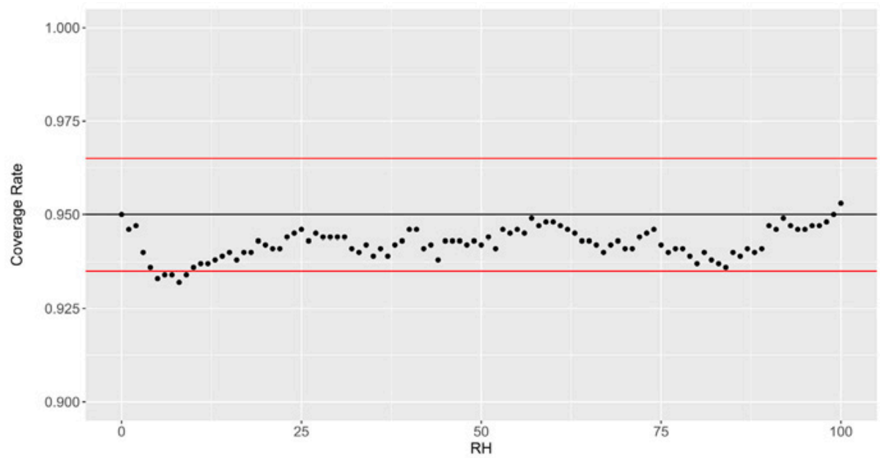


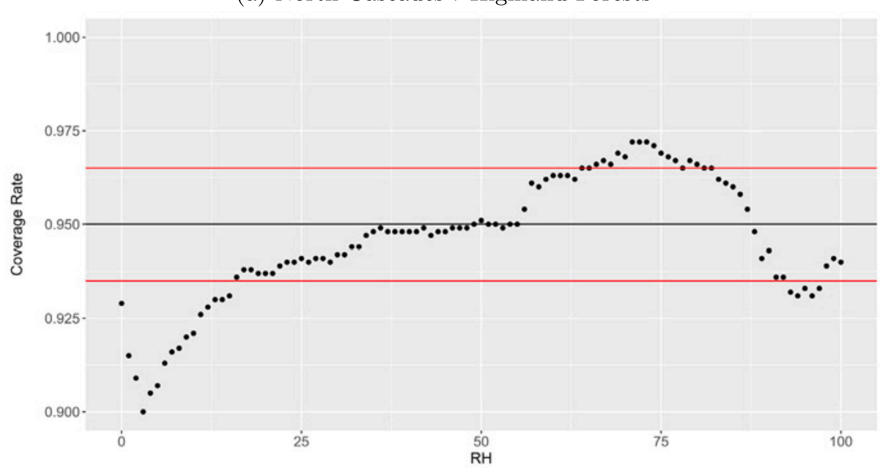
Fig. 7. Quartiles for the relative variance (see Eq. (1) and accompanying text) for each EPA L4 ecoregion.

We next present the ten-fold cross-validation study testing the plot-level AGBD predictions. For the 82,185 forested plots, the coverage rate of the 95% credible intervals was 96.5%. Further, the calibration curve for the logistic regression was near-ideal, never deviating from the one-to-one line by more than a few percentage points (Fig. 13(b)). Biomass is highly variable at the plot scale, and the model covariate (RH95) needed to be obtained from predicted waveforms, introducing further uncertainty. Even still, the accuracy of the predictions was reasonable, with a predictive R^2 of 0.60 across all plots and 0.45 restricting to forested plots.

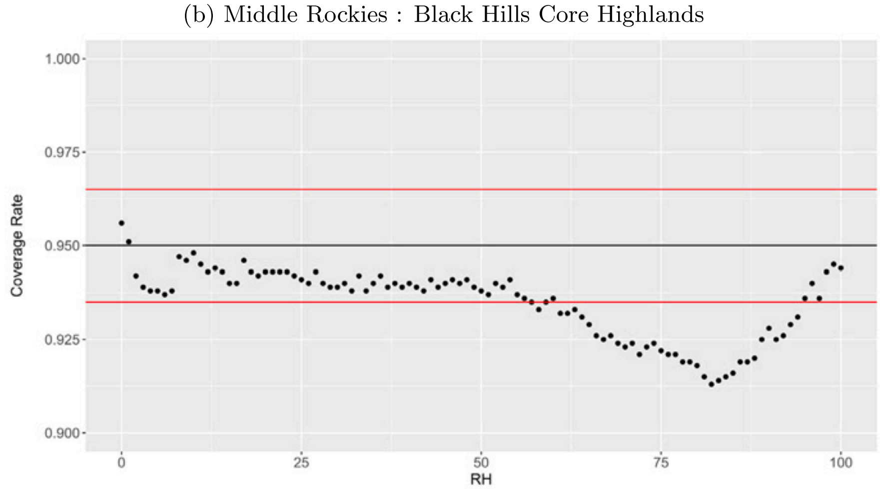
Finally, we compared the model AGBD predictions to NFI estimates within 64,000 ha hexagons. The model predictions concord well with the NFI estimates, with the only striking deviations from the one-to-one line occurring for NFI estimates that used less than 10 field plots, and therefore were highly uncertain (Fig. 14(a)). Further, the uncertainties of the model predictions were typically far lower than those of



(a) North Cascades : Highland Forests



(b) Middle Rockies : Black Hills Core Highlands

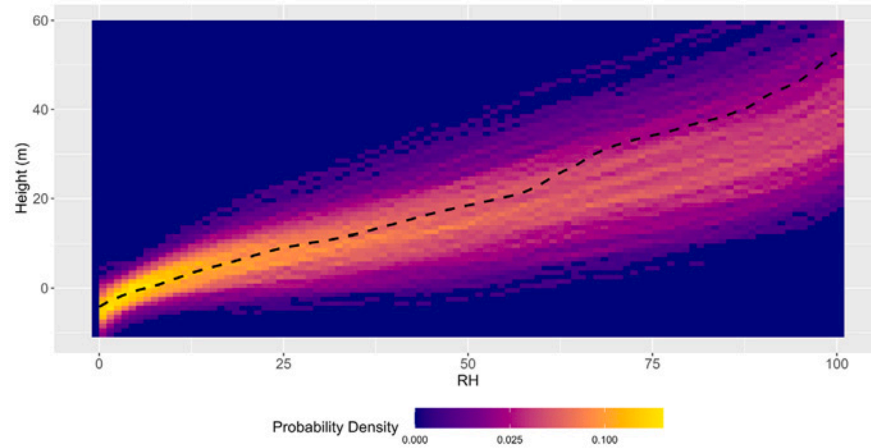


(c) Central Appalachians : Forested Hills and Mountains

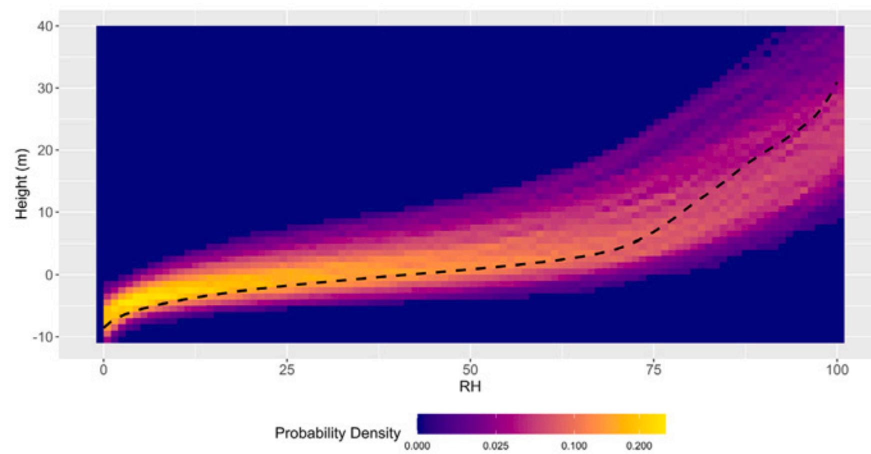
Fig. 8. Coverage rates of predictive 95% credible intervals across all RH metrics. Red lines indicate a 95% credible interval for the success rate given 1000 independent tests each with a 95% success probability. Coverage rates for RH 0 – 15 for the Middle Rockies and RH 61 – 94 for the Central Appalachians deviate below the range of appropriate rates, suggesting slight prediction overconfidence for these RH metrics within the respective regions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the NFI estimates. The model standard deviations never exceeded 25 Mg/ha, whereas the NFI standard errors were as large as 164 Mg/ha (the corresponding model standard deviation for this hexagon was 7 Mg/ha).

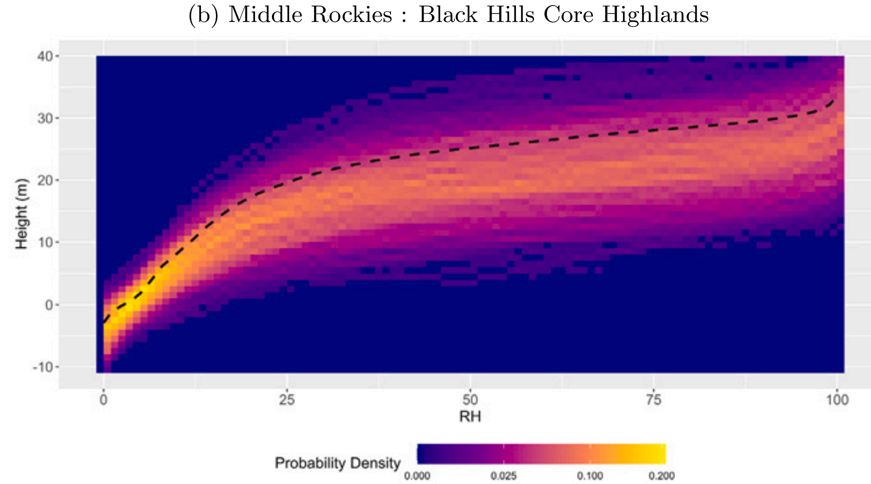
The increase in precision from the model can be stated in terms of the number of additional NFI plots required to achieve the same precision using the NFI design-based estimates (ESI). Across the 2774 hexagons considered (model prediction greater than 50 Mg/ha, NFI



(a) North Cascades : Highland Forests



(b) Middle Rockies : Black Hills Core Highlands



(c) Central Appalachians : Forested Hills and Mountains

Fig. 9. An example waveform predictive distribution for each of the cross-validation regions with a dotted line representing the true test waveform.

standard error greater than 10 Mg/ha), the median ESI was 135 plots. Only 50 hexagons had an ESI less than 27 (an effective doubling of the average hexagon sample size) and 7 hexagons had an ESI less than 10. The total ESI across the considered hexagons was 464,804 field plots, creating an effective sample size seven times the original.

5. Discussion

By generating probabilistic GEDI waveform predictions over NFI plots and accounting for the prediction uncertainties during model training with error-in-variables techniques, we alleviated the need for collocation or mediating ALS data to produce a massive regression

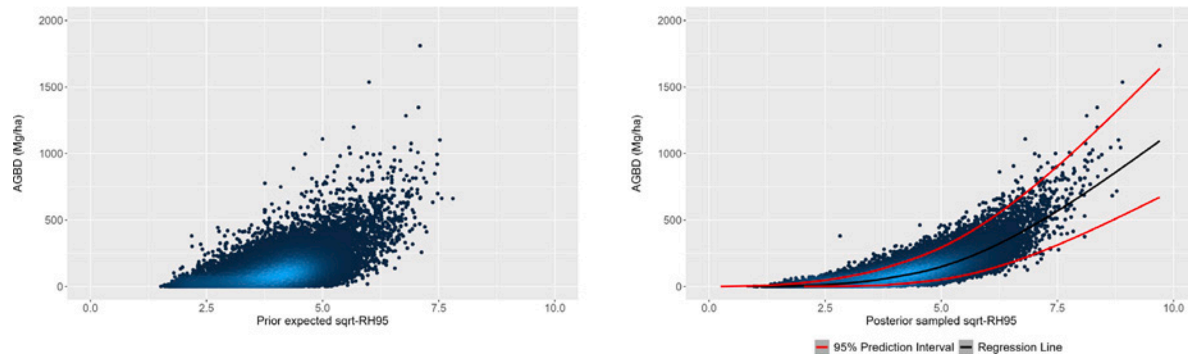


Fig. 10. The left panel gives the predicted sqrt-RH95 values versus NFI plot-measured AGBD. Note the compression along the horizontal axis due to the Berkson errors, where predicted sqrt-RH95 values are compressed towards their ecoregion mean. Such errors are filtered during model training to yield the right panel, where the horizontal compression is no longer evident. Within the Bayesian interpretation, the predicted covariates on the left panel are a prior expectation of sqrt-RH95. The covariates on the right panel are a random sample from the posterior distribution, i.e. a plausible configuration given the additional information from the NFI plot data. The right panel also gives the regression line and 95% prediction intervals for the final trained model.

training data set dispersed across CONUS. The data set can be used to train regression models between GEDI RH metrics and NFI-measured forest attributes, and subsequently leverage the density of GEDI observations to make population estimates at finer scales and at higher precision than possible with the field data alone. The data set produced in this work is for the US, but the method could be expanded to other nations to enhance their forest inventory.

The example AGBD model performed well. The coverage rate of the 95% credible intervals was 96.5% over the forested plots and the calibration curve showed the logistic regression correctly assessed the probability of forest presence. These positive results demonstrate the ability of the data set to produce models that make NFI-consistent predictions at the plot scale, which can be aggregated into NFI-consistent population estimates. In particular, we aggregated predictions to the 1 km scale, a resolution for which estimates cannot be produced using typical FIA design-based procedures due to the sampling density of the field plots. We further aggregated to a 64,000 ha scale to allow comparison to the design-based estimates. At this resolution, the model predictions remained congruent with the design-based estimates, but achieved far greater precision, often equivalent to sampling hundreds of additional NFI plots (Fig. 15).

To train the example AGBD model, we restricted the FIA plots to those measured in 2015 or later. The observed GEDI waveforms used to generate the predicted waveforms were observed from April 2019 to March 2023. This leaves a maximum possible temporal discrepancy between a predicted waveform and plot measurement of eight years. This could lead to additional noise and uncertainty in the model, especially in the case of acute disturbances between the two measurements times. Any analyst using the data set could make their own operational decisions pertaining to time frames, potentially using disturbance maps to inform their filtering. As FIA continues its data release, there will be more plot measurements temporally coincident with the GEDI measurement period: at the time of this writing, the most recent available measurements are from 2022 and the latest year for which measurements are available in all 48 contiguous states is 2019.

The confidence in the waveform predictions over the NFI plots is driven primarily by the spatial coherence of waveforms within an ecoregion and the density of GEDI observations. The Appalachians in the eastern US suffer from several issues, with a below-average estimated spatial correlation range and, more importantly, a low density of GEDI observations due to leaf-off filtering. The low density of GEDI observations has ill effects for both model training and subsequent predictions, giving lower quality local training data and therefore increased uncertainty for local relationships between RH metrics and forest attributes, and fewer GEDI footprint-level predictions to supplement the NFI plots. This is reflected in the absolute and relative standard deviations of the 1 km AGBD map. The spatial coherence of the ecoregions is somewhat of a fixed reality, but it would be beneficial if some information from the leaf-off observations could be reclaimed, rather than omitting them entirely.

We used a multivariate spatial model to make GEDI cumulative waveform predictions: the observed waveforms were projected onto their principal components, modeled and predicted separately, and then the full waveform predictive distribution was reconstructed. A cross-validation study showed this technique to perform reasonably well. Some RH metrics, varying across the three test ecoregions, had coverage rates for the 95% credible intervals outside the range of appropriate values. We generated this range of appropriate values by assuming 1000 independent tests each with a 95% success probability. Our tests were not completely independent, as two randomly selected test locations nearby may leverage the same observed footprints to make a prediction, and all predictions use the same estimated parameters. While this likely makes the given range too narrow, it is still safe to interpret a coverage rate of 90.0% for RH4 within the Middle Rockies and 91.3% for RH82 within the Central Appalachians as a model inadequacy. In our experience, mild shortcomings in the modeling of a given principal component can inflate to larger problems when reconstructing the distribution.

Both the GEDI footprints and FIA plots are believed to have geolocation error with approximately 10 m standard deviation (Hoppus and Lister, 2005; Beck et al., 2021). The waveform predictions in this work are essentially executed using kriging equations on the principal components of the waveforms (Section 3.1, Appendix). To make a prediction at an unobserved location, the kriging equations leverage the waveform values and reported locations of the nearby footprints. Correlation between the observed locations and prediction location is assessed through an exponential correlation function and random noise component. It is reasonable to suppose that unaccounted error in the reported locations would have ill effects on the resulting predictions. The scenario of kriging prediction with uncertain locations was studied in Gabrosek and Cressie (2002). They found the effect on predictions to be negligible if the location standard deviation was small relative to the spatial range parameter of the exponential correlation function. Across all 967 ecoregions, the minimum spatial range parameter estimated for the dominant waveform component was 22 m, with 99% of the ecoregions having an estimated spatial range of over 150 m. Most estimated range parameters dwarfed the geolocation uncertainty, suggesting the geolocation uncertainty had little effect on the predicted waveforms.

6. Conclusion

We produced a training data set pairing 326,787 NFI plots across CONUS with predicted GEDI cumulative waveforms. This data set is ready to serve a multitude of objectives across different US regions

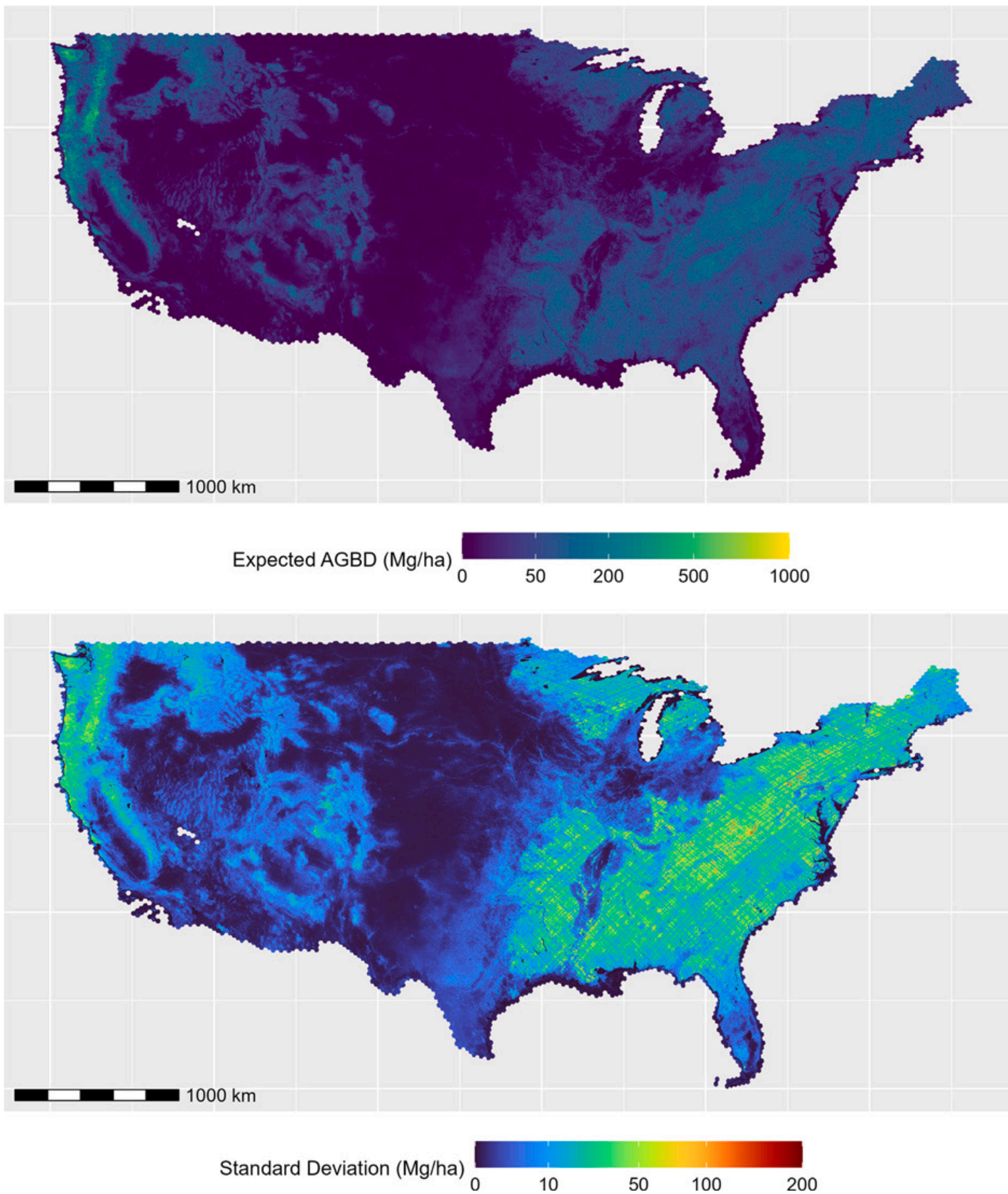


Fig. 11. Expected values and standard deviations for 1 km estimates of forest AGBD across CONUS. Note the diagonal patterns in the standard deviations from the orbital track pattern of GEDI. Uncertainties are generally lowest along the tracks. The filtering of leaf-off observations left fewer tracks in the deciduous east and therefore more obvious gaps between them.

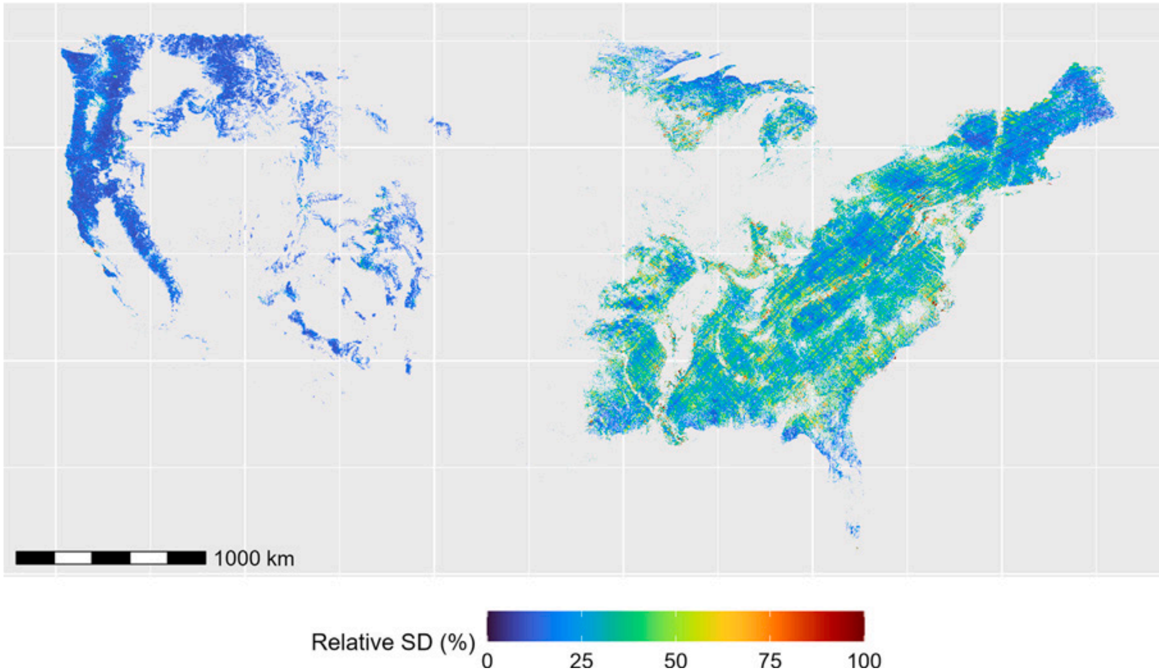


Fig. 12. Relative standard deviations (standard deviation divided by expected value) for 1 km cells with an expected value greater than 50 Mg/ha.

and forest attributes. We envision that the methods could be expanded to other NFIs outside of the FIA program of the US, allowing NFI-consistent GEDI models without requirements of collocation. We intend future work developing improved multivariate spatial models that can more accurately model and predict lidar waveforms and investigating methods to utilize leaf-off GEDI observations for deciduous areas.

CRediT authorship contribution statement

Paul B. May: Writing – original draft, Software, Methodology, Formal analysis, Conceptualization. **Ralph O. Dubayah:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization. **Jamis M. Bruening:** Writing – review & editing, Software, Data curation. **George C. Gaines III:** Writing – review & editing, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The FIA-GEDI data set will be made available at the LPDAAC, along with R code to query predictions over FIA plots by EPA L4 ecoregion. Further R code with C++ subroutines are provided to demonstrate waveform predictions and reproduce the waveform cross-validation study. We also provide R code for Bayesian inference on a linear regression model, accounting for error-in-variables. The 1 km AGBD maps for CONUS will be a made available at ORNL DAAC.

Acknowledgments

This research was supported by NASA Contract [#NNL15AA03C] for the development and execution of the GEDI mission (May, Dubayah, Bruening), a NASA FINESST Fellowship [#80NSSC21K1626] (Dubayah, Bruening), and a NASA Applications Supplement to a grant from NASA's Carbon Monitoring System [#80NSSC20K0023] (Dubayah, May).

Appendix. Multivariate spatial model for waveform prediction

Let $\boldsymbol{x}(s)$ be a vector of length 101 representing the RH metrics (cumulative waveform) at location s. We assume the spatial model

$$\boldsymbol{x}(s) = \boldsymbol{\Gamma} \boldsymbol{z}(s) \tag{3}$$

where $\boldsymbol{\Gamma}$ is a 101×101 orthonormal matrix ($\boldsymbol{\Gamma}^T \boldsymbol{\Gamma} = \boldsymbol{I}$) and $\boldsymbol{z}(s)$ is a vector of independent spatial processes, $z_j(s)$; $j = 1, \dots, 101$. This is similar to the full linear coregionalization model of May et al. (2022), and much of the following is synthesized from that work. Note that

$$\boldsymbol{\Sigma}_x \equiv \mathrm{Var}[\boldsymbol{x}(s)] = \boldsymbol{\Gamma} \boldsymbol{\Lambda}_z \boldsymbol{\Gamma}^T, \tag{4}$$

where $\boldsymbol{\Lambda}_z$ is a diagonal matrix with entries $\mathrm{Var}[z_j(s)]$. Therefore, $\boldsymbol{\Gamma}$ gives the eigenvectors of $\boldsymbol{\Sigma}_x$, the covariance matrix of $\boldsymbol{x}(s)$ and

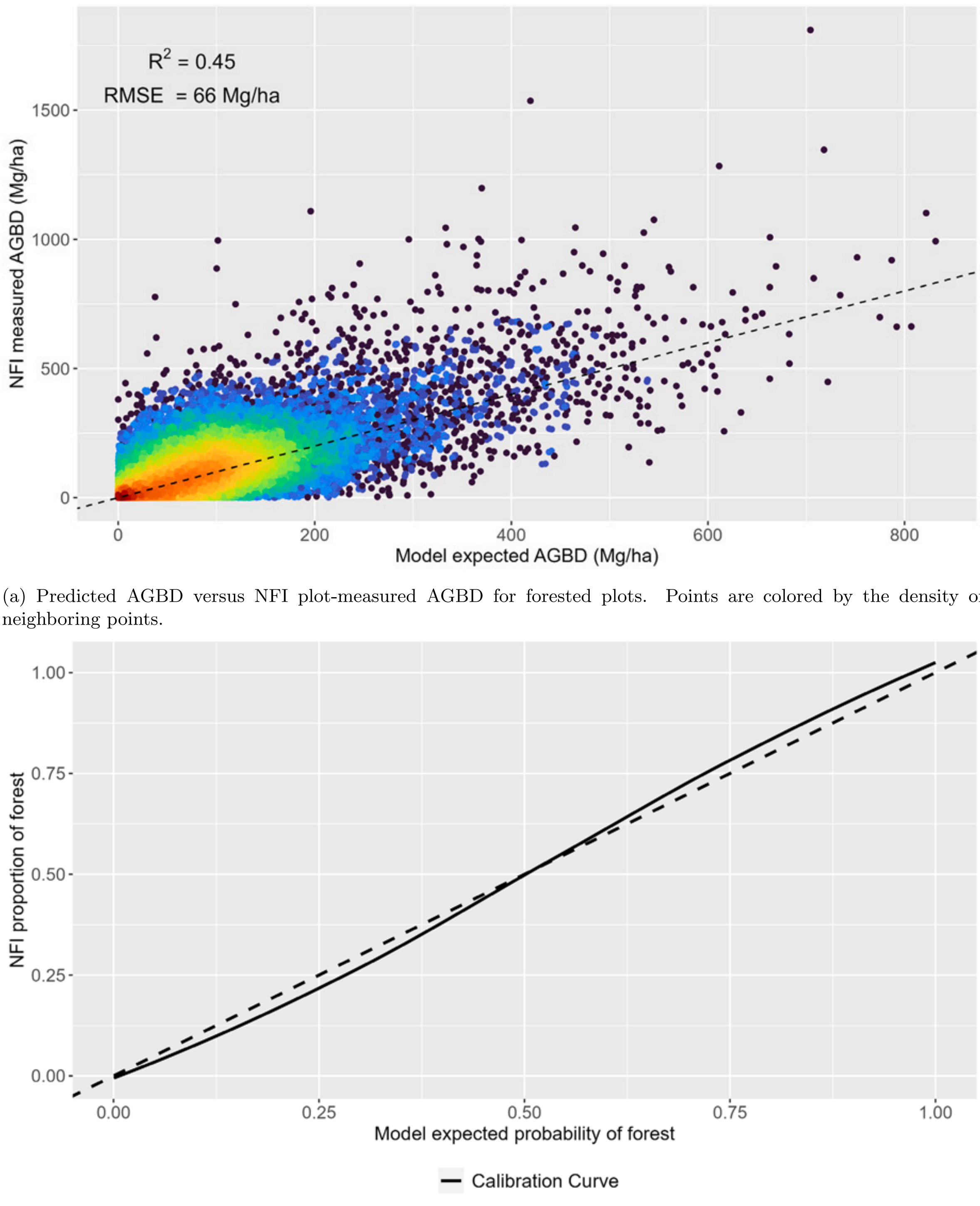
$$\boldsymbol{z}(s) = \boldsymbol{\Gamma}^T \boldsymbol{x}(s), \tag{5}$$

implying $\boldsymbol{z}(s)$ is a vector of principal components for $\boldsymbol{x}(s)$. Let $z_j(s)$ be ordered according to decreasing variance. We assume

$$g_j\left(z_j(s)\right) = \begin{cases} \alpha_j + b(s)\beta_j + w_j(s) + \epsilon_j(s); & \text{for } j \le k \\ \alpha_j + b(s)\beta_j + \epsilon_j(s); & \text{for } j > k \end{cases} \tag{6}$$

where α_j, β_j are regression coefficients, $b(s) = 0, 1$, is the NLCD forest classification value, $w_j(s)$ is a spatially correlated Gaussian process with exponential covariance function and $\epsilon_j(s)$ is a spatially independent, normally distributed error. Index k is the point where $w_j(s)$ is removed and the spatial processes are assumed to be spatially independent. Finally, functions $g_j(\cdot)$ are Box–Cox transformations (Box and Cox, 1964; Sakia, 1992) to promote assumptions of normally distributed random effects (without transformation, the leading principal components are often highly skewed, but become increasingly "normal-like" in distribution as j increases).

Taking a fully Bayesian approach to the above model would require intensive Markov chain Monte Carlo (MCMC) chains that would be difficult to monitor across all 967 ecoregions. However, all of the difficult parameters are assumed constant across the ecoregion, for which there are effectively limitless training samples (GEDI footprints) to estimate with. Therefore, we fix the difficult parameters at point estimates and assume the uncertainty to be negligible. For each ecoregion, we fix



(a) Predicted AGBD versus NFI plot-measured AGBD for forested plots. Points are colored by the density of neighboring points.

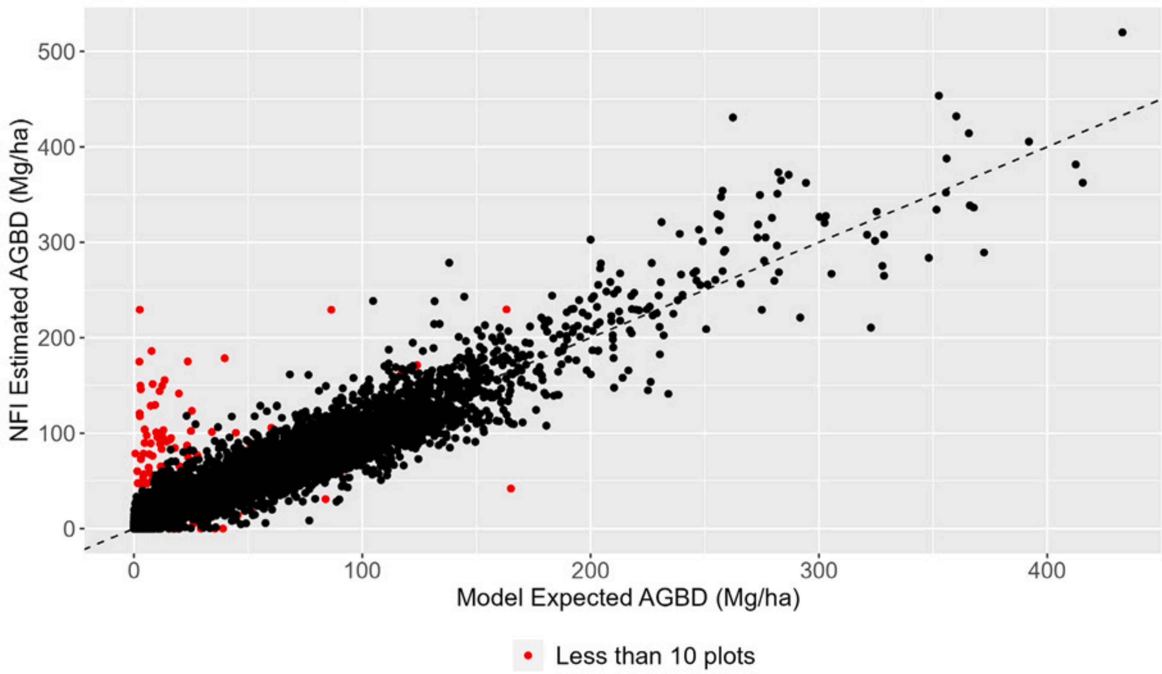
(b) Calibration curve for the logistic regression discriminating between forest and non-forest.

Fig. 13. Results of the plot-level cross-validation study. Dotted lines give the one-to-one line.

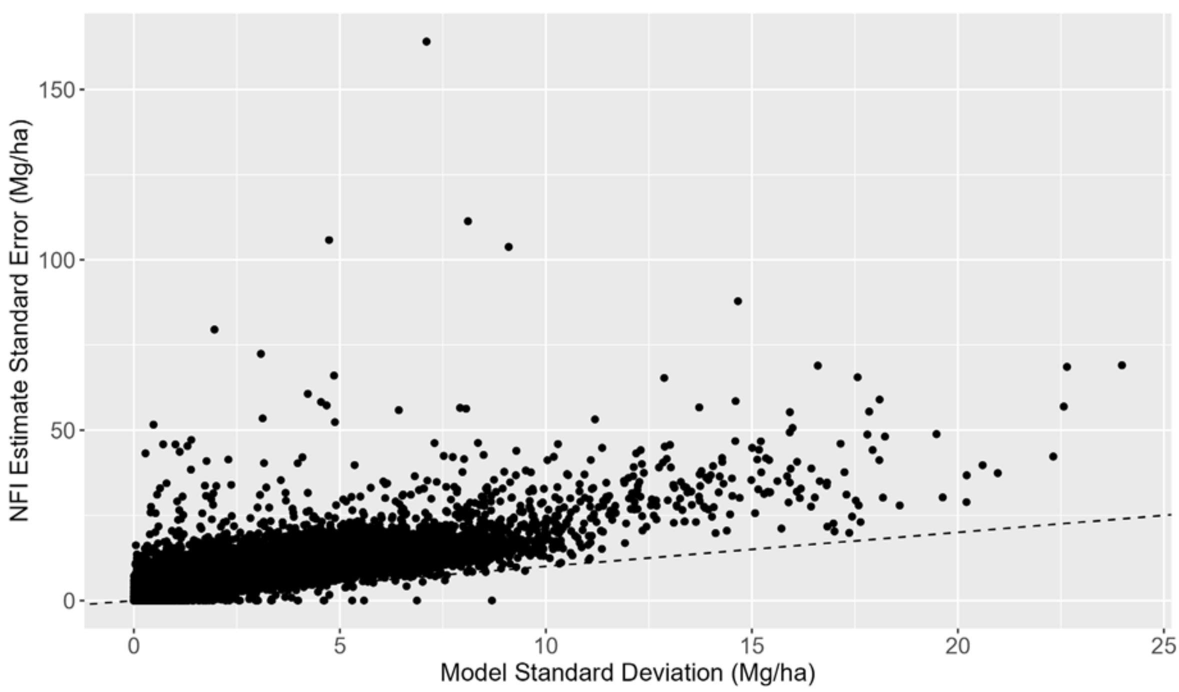
$\boldsymbol{\Gamma}$ at the eigendecomposition of the sample covariance matrix for $\boldsymbol{x}(s)$ and fix the parameters for the Box–Cox transformations, $g_j(\cdot)$, and the parameters for the random effects, $w_j(s)$, $\epsilon(s)$ at their maximum likelihood estimates. Posterior distributions for the regression coefficients and predictions $g_j\left(\boldsymbol{z}_j(s^*)\right)$ at an NFI plot location s^* are then available in closed form.

To select k, the point at which the Gaussian process $w_j(s)$ drops out, an operational decision was made. We selected k as the minimum index at which the first k sample variances account for over 99.99% of the total variance. The rationale for this is two-fold: First, by definition the remaining $101 - k$ principal components hold little additional information. Second, the principal components become increasingly spatially incoherent, making informative interpolation increasingly futile. Indeed, for every ecoregion the estimated spatial correlation of $w_j(s)$ decayed below practical significance before index j reached the enforced cut-off, k.

The closed-form posterior predictive distributions for $g_j\left(\boldsymbol{z}_j(s^*)\right)$ are saved at every NFI plot location. Monte Carlo methods are required



(a) Model predicted AGBD versus NFI estimated AGBD. The two sets are in broad agreement, with major deviations occurring only for NFI estimates that used than 10 field plots.



(b) Model standard deviations versus the NFI estimate standard errors. The model uncertainties are typically far lower than the NFI counterparts.

Fig. 14. Comparison of the model predictions and uncertainties to the NFI estimates and uncertainties across all 12,591 hexagons. Dotted lines give the one-to-one line.

to reclaim the cumulative waveform distribution and the marginal RH metric distributions. A multitude of samples (thousands, depending on the accuracy required) are drawn from the posterior distributions of $g_j\left(\boldsymbol{z}_j(s^*)\right)$; $j = 1, \dots, 101$. These samples are back-transformed with $g_j^{-1}(\cdot)$ to yield posterior samples of $\boldsymbol{z}_j(s^*)$; $j = 1, \dots, 101$ and then matrix-multiplied by $\boldsymbol{\Gamma}$ to achieve multivariate posterior samples of $\boldsymbol{x}(s_*)$. If only certain RH metrics are desired, $\boldsymbol{\Gamma}$ can be row-subsetted to yield only samples of those RH metrics. Further, if posterior distributions of a transformation of the RH metrics are desired (e.g., square-root RH95 in our example AGBD model), this transformation is simply applied the posterior samples of the RH metrics. All of this is handled internally by the `getRHmetrics()` function of the associated R code, allowing users to query expected values and (co)variances for any set RH metrics, under any valid transformation, for any of the predicted FIA plots.

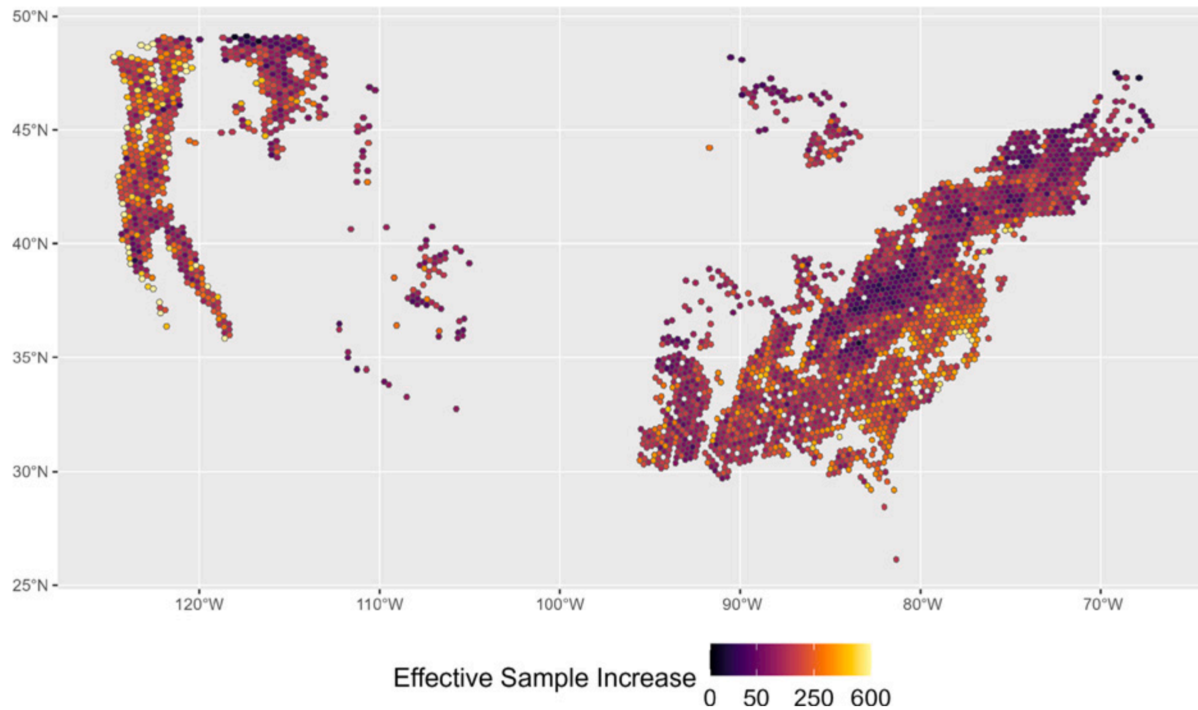


Fig. 15. Estimated ESI across the 2774 hexagons with a model prediction greater than 50 Mg/ha and an NFI standard error greater than 10 Mg/ha. A total of 29 hexagons had an ESI greater than 600, but were truncated in the color scale to improve visualization. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

References

Abdalati, W., Zwally, H.J., Bindschadler, R., Csatho, B., Farrell, S.L., Fricker, H.A., Harding, D., Kwok, R., Lefsky, M., Markus, T., et al., 2010. The ICESat-2 laser altimetry mission. Proc. IEEE 98 (5), 735–751.

Baccini, A., Goetz, S., Walker, W., Laporte, N., Sun, M., Sulla-Menashe, D., Hackler, J., Beck, P., Dubayah, R., Friedl, M., et al., 2012. Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps. Nature Clim. Change 2 (3), 182–185.

Beck, J., Wirt, B., Luthcke, S., Hofton, M., Armston, J., 2021. Global Ecosystem Dynamics Investigation (GEDI) Level 02 User Guide Version 2.0. US Geological Survey, Earth Resources Observation and Science Center.

Berkson, J., 1950. Are there two regressions? J. Amer. Statist. Assoc. 45 (250), 164–180.

Boudreau, J., Nelson, R.F., Margolis, H.A., Beaudoin, A., Guindon, L., Kimes, D.S., 2008. Regional aboveground forest biomass using airborne and spaceborne LiDAR in Québec. Remote Sens. Environ. 112 (10), 3876–3890.

Box, G.E., Cox, D.R., 1964. An analysis of transformations. J. R. Stat. Soc. Ser. B Stat. Methodol. 26 (2), 211–243.

Bruening, J.M., May, P.B., Armston, J.D., Dubayah, R.O., 2023. Precise and unbiased biomass estimation from GEDI data and the US forest inventory. Front. For. Glob. Change 6, 1149153.

Burns, P., Clark, M., Salas, L., Hancock, S., Leland, D., Jantz, P., Dubayah, R., Goetz, S.J., 2020. Incorporating canopy structure from simulated GEDI lidar into bird species distribution models. Environ. Res. Lett. 15 (9), 095002.

Dellaportas, P., Stephens, D.A., 1995. Bayesian analysis of errors-in-variables regression models. Biometrics 1085–1095.

Dewitz, J., 2021. National Land Cover Database (NLCD) 2019 Products [Data set]. Technical Report, U.S. Geological Survey.

Dubayah, R., Armston, J., Healey, S.P., Bruening, J.M., Patterson, P.L., Kellner, J.R., Duncanson, L., Saarela, S., Ståhl, G., Yang, Z., et al., 2022b. GEDI launches a new era of biomass inference from space. Environ. Res. Lett. 17 (9), 095001.

Dubayah, R., Armston, J., Healey, S., Yang, Z., Patterson, P., Saarela, S., Stahl, G., Duncanson, L., Kellner, J., 2022a. GEDI L4B Gridded Aboveground Biomass Density, Version 2 [Data set]. ORNL DAAC, Oak Ridge, Tennessee, USA.

Dubayah, R., Armston, J., Kellner, J., Duncanson, L., Healey, S., Patterson, P., Hancock, S., Tang, H., Hofton, M., Blair, J., Luthcke, S., 2021. GEDI L4A footprint level aboveground biomass density, version 1. http://dx.doi.org/10.3334/ORNLDAAC/1907, URL https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1907.

Dubayah, R., Blair, J.B., Goetz, S., Fatoyinbo, L., Hansen, M., Healey, S., Hofton, M., Hurtt, G., Kellner, J., Luthcke, S., Armston, J., Tang, H., Duncanson, L., Hancock, S., Jantz, P., Marselis, S., Patterson, P.L., Qi, W., Silva, C., 2020a. The global ecosystem dynamics investigation: High-resolution laser ranging of the Earth's forests and topography. Sci. Remote Sens. 1, 100002.

Dubayah, R., Hofton, M., Blair, J., Armston, J., Tang, H., Luthcke, S., 2020b. GEDI L2A Elevation and Height Metrics Data Global Footprint Level V001 [Data set]. NASA EOSDIS Land Processes DAAC.

Duncanson, L., Kellner, J.R., Armston, J., Dubayah, R., Minor, D.M., Hancock, S., Healey, S.P., Patterson, P.L., Saarela, S., Marselis, S., Silva, C.E., Bruening, J., Goetz, S.J., Tang, H., Hofton, M., Blair, B., Luthcke, S., Fatoyinbo, L., Abernethy, K., Alonso, A., Andersen, H.-E., Aplin, P., Baker, T.R., Barbier, N., Bastin, J.F., Biber, P., Boeckx, P., Bogaert, J., Boschetti, L., Boucher, P.B., Boyd, D.S., Burslem, D.F., Calvo-Rodriguez, S., Chave, J., Chazdon, R.L., Clark, D.B., Clark, D.A., Cohen, W.B., Coomes, D.A., Corona, P., Cushman, K., Cutler, M.E., Dalling, J.W., Dalponte, M., Dash, J., de Miguel, S., Deng, S., Ellis, P.W., Erasmus, B., Fekety, P.A., Fernandez-Landa, A., Ferraz, A., Fischer, R., Fisher, A.G., García-Abril, A., Gobakken, T., Hacker, J.M., Heurich, M., Hill, R.A., Hopkinson, C., Huang, H., Hubbell, S.P., Hudak, A.T., Huth, A., Imbach, B., Jeffery, K.J., Katoh, M., Kearsley, E., Kenfack, D., Kljun, N., Knapp, N., Král, K., Krůček, M., Labrière, N., Lewis, S.L., Longo, M., Lucas, R.M., Main, R., Manzanera, J.A., Martínez, R.V., Mathieu, R., Memiaghe, H., Meyer, V., Mendoza, A.M., Monerris, A., Montesano, P., Morsdorf, F., Næsset, E., Naidoo, L., Nilus, R., O'Brien, M., Orwig, D.A., Papathanassiou, K., Parker, G., Philipson, C., Phillips, O.L., Pisek, J., Poulsen, J.R., Pretzsch, H., Rüdiger, C., Saatchi, S., Sanchez-Azofeifa, A., Sanchez-Lopez, N., Scholes, R., Silva, C.A., Simard, M., Skidmore, A., Stereńczak, K., Tanase, M., Torresan, C., Valbuena, R., Verbeeck, H., Vrska, T., Wessels, K., White, J.C., White, L.J., Zahabu, E., Zgraggen, C., 2022. Aboveground biomass density models for NASA's global ecosystem dynamics investigation (GEDI) lidar mission. Remote Sens. Environ. 270, 112845.

Finley, A.O., Banerjee, S., MacFarlane, D.W., 2011. A hierarchical model for quantifying forest variables over large heterogeneous landscapes with uncertain forest areas. J. Amer. Statist. Assoc. 106 (493), 31–48.

Gabrosek, J., Cressie, N., 2002. The effect on attribute prediction of location uncertainty in spatial data. Geogr. Anal. 34 (3), 262–285.

Gelfand, A.E., Schliep, E.M., 2016. Spatial statistics and Gaussian processes: A beautiful marriage. Spatial Stat. 18, 86–104.

Gillespie, A.J., 2000. Pros and Cons of Continuous Forest Inventory: Customer Perspectives. United States Department of Agriculture Forest Service General Technical Report NC, pp. 60–66.

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 Space Sci. 6 (2), 294–310.

Hoppus, M., Lister, A., 2005. The status of accurately locating forest inventory and analysis plots using the global positioning system. In: Proceedings of the Seventh Annual Forest Inventory and Analysis Symposium, Portland, OR, USA. Vol. 36, 179184.

Margolis, H.A., Nelson, R.F., Montesano, P.M., Beaudoin, A., Sun, G., Andersen, H.-E., Wulder, M.A., 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.

May, P., Biesecker, M., Rekabdarkolaee, H.M., 2022. Response envelopes for linear coregionalization models. J. Multivariate Anal. 192, 105015.

May, P., McConville, K.S., Moisen, G.G., Bruening, J., Dubayah, R., 2023. A spatially varying model for small area estimates of biomass density across the contiguous United States. Remote Sens. Environ. 286, 113420.

Menlove, J., Healey, S.P., 2020. A comprehensive forest biomass dataset for the USA allows customized validation of remotely sensed biomass estimates. Remote Sens. 12 (24), 4141.

Patterson, P.L., Healey, S.P., Ståhl, G., Saarela, S., Holm, S., Andersen, H.-E., Dubayah, R.O., Duncanson, L., Hancock, S., Armston, J., et al., 2019. Statistical properties of hybrid estimators proposed for GEDI—NASA's global ecosystem dynamics investigation. Environ. Res. Lett. 14 (6), 065007.

Pearce, J., Ferrier, S., 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecol. Model. 133 (3), 225–245.

Prisley, S., Bradley, J., Clutter, M., Friedman, S., Kempka, D., Rakestraw, J., Sonne Hall, E., 2021. Needs for small area estimation: Perspectives from the US private forest sector. Front. For. Glob. Change 4, 746439.

Saatchi, S.S., Harris, N.L., Brown, S., Lefsky, M., Mitchard, E.T., Salas, W., Zutta, B.R., Buermann, W., Lewis, S.L., Hagen, S., et al., 2011. Benchmark map of forest carbon stocks in tropical regions across three continents. Proc. Natl. Acad. Sci. 108 (24), 9899–9904.

Sakia, R.M., 1992. The Box–Cox transformation technique: A review. J. R. Stat. Soc.: Ser. D (Stat.) 41 (2), 169–178.

Taylor-Rodriguez, D., Finley, A.O., Datta, A., Babcock, C., Andersen, H.-E., Cook, B.D., Morton, D.C., Banerjee, S., 2019. Spatial factor models for high-dimensional and large spatial data: An application in forest variable mapping. Statist. Sinica 29, 1155.

Westfall, J.A., Coulston, J.W., Moisen, G.G., Andersen, H.-E., 2022. Sampling and Estimation Documentation for the Enhanced Forest Inventory and Analysis Program. Technical Report, U.S. Department of Agriculture, Forest Service, Northern Research Station, Madison, WI.