IOP Publishing Environ. Res. Lett. 14 (2019) 065007 https://doi.org/10.1088/1748-9326/ab18df

Environmental Research Letters

OPEN ACCESS

RECEIVED
9 October 2018

REVISED
28 March 2019

ACCEPTED FOR PUBLICATION
12 April 2019

PUBLISHED
13 June 2019

LETTER

Statistical properties of hybrid estimators proposed for GEDI—NASA's global ecosystem dynamics investigation

Paul L Patterson[1], **Sean P Healey**[2,9], **Göran Ståhl**[3], **Svetlana Saarela**[3], **Sören Holm**[3], **Hans-Erik Andersen**[4], **Ralph O Dubayah**[5,6], **Laura Duncanson**[5], **Steven Hancock**[5], **John Armston**[5], **James R Kellner**[6,7], **Warren B Cohen**[8] **and Zhiqiang Yang**[2]

1 USDA Forest Service, Rocky Mountain Research Station, 240W Prospect, Fort Collins, CO 80526, United States of America
2 USDA Forest Service, Rocky Mountain Research Station, 507 25th St, Ogden, UT, United States of America
3 Department of Forest Resource Management, Swedish University of Agricultural Sciences, Umeå, Sweden
4 USDA Forest Service, Pacific Northwest Research Station, Seattle, WA, United States of America
5 Department of Geographical Sciences, University of Maryland, College Park, MD, United States of America
6 Institute at Brown for Environment and Society, Brown University, Providence, RI, United States of America
7 Department of Ecology and Evolutionary Biology, Brown University, Providence, RI, United States of America
8 USDA Forest Service, Pacific Northwest Research Station, Corvallis, OR, United States of America
9 Author to whom any correspondence should be addressed.

E-mail: sean.healey@usda.gov

Keywords: carbon monitoring, lidar, forest biomass

Supplementary material for this article is available online

Abstract

NASA's Global Ecosystem Dynamics Investigation (GEDI) mission will collect waveform lidar data at a dense sample of ~25 m footprints along ground tracks paralleling the orbit of the International Space Station (ISS). GEDI's primary science deliverable will be a 1 km grid of estimated mean aboveground biomass density (Mg ha^{-1}), covering the latitudes overflown by ISS (51.6 °S to 51.6 °N). One option for using the sample of waveforms contained within an individual grid cell to produce an estimate for that cell is hybrid inference, which explicitly incorporates both sampling design and model parameter covariance into estimates of variance around the population mean. We explored statistical properties of hybrid estimators applied in the context of GEDI, using simulations calibrated with lidar and field data from six diverse sites across the United States. We found hybrid estimators of mean biomass to be unbiased and the corresponding estimators of variance appeared to be asymptotically unbiased, with under-estimation of variance by approximately 20% when data from only two clusters (footprint tracks) were available. In our study areas, sampling error contributed more to overall estimates of variance than variability due to the model, and it was the design-based component of the variance that was the source of the variance estimator bias at small sample sizes. These results highlight the importance of maximizing GEDI's sample size in making precise biomass estimates. Given a set of assumptions discussed here, hybrid inference provides a viable framework for estimating biomass at the scale of a 1 km grid cell while formally accounting for both variability due to the model and sampling error.

1. Introduction

The NASA Global Ecosystem Dynamics Investigation (GEDI) mission mounted a full-waveform lidar instrument on the International Space Station (ISS) in late 2018. GEDI is designed to measure forest structure, and one of its primary science deliverables will be a 1 km grid of mean aboveground biomass density (AGBD in Mg ha^{-1}) estimates over the forested areas between 51.6 ° N and S (the range of the ISS). The basic lidar metric supporting these estimates will be waveforms related to the canopy height density profile for 25 m (diameter) footprints, spaced 60 m apart along ground tracks paralleling the ISS orbit. After two years of operation, GEDI will have covered the majority of 1 km cells with two or more ground tracks. Performance of other lidar instruments suggests that canopy height and structure metrics derived from GEDI data

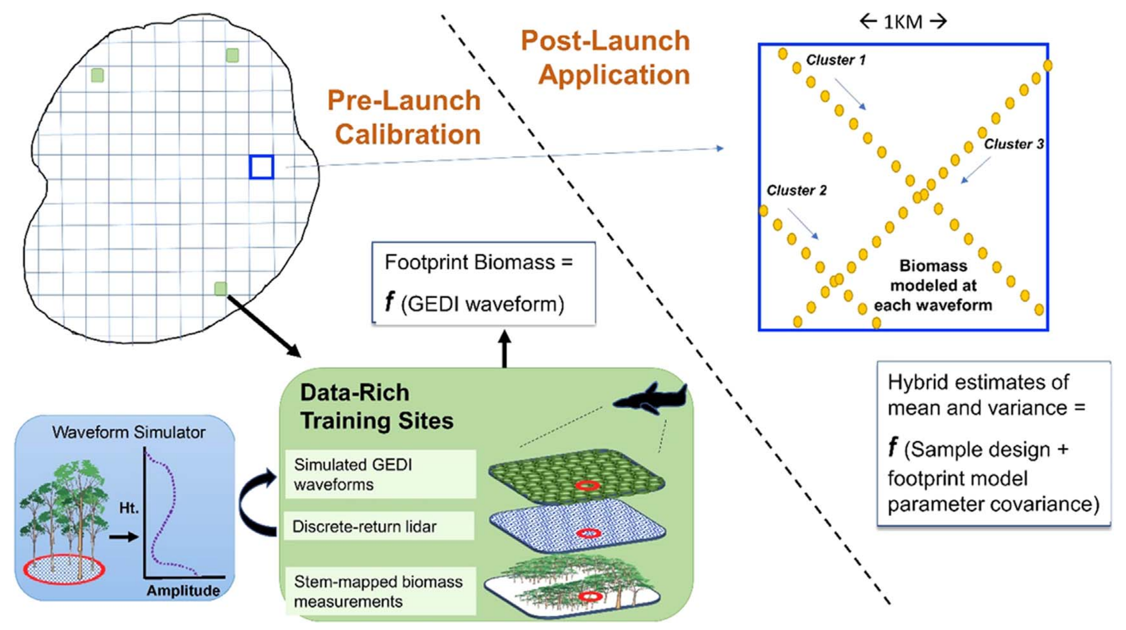

Figure 1. Proposed pre-launch calibration approach, using simulated GEDI waveforms collected over closely geo-registered field plots to develop waveform-to-AGBD relationships that may be applied to all footprints. Hybrid estimators access both properties of how GEDI samples the grid cell as well as the footprint model parameter covariance.

will be strongly correlated with AGBD (Zolkos *et al* 2013).

This paper concerns the challenge of using spatially discontinuous lidar footprint data acquired along tracks to estimate mean AGBD within each 1 km grid cell. One approach has been to 'scale up' from field data to spatially coincident lidar footprints using one level of models, apply those models to all lidar footprints, and then use the lidar-based AGBD predictions to calibrate another level of models that predict AGBD using coarse-resolution optical or radar data (e.g. Baccini *et al* 2017). However, while diagnostics such as Root Mean Square Error from the latter model can be used to indicate confidence for predictions at the scale of each coarse-resolution grid cell under such an approach, ignoring residual variance in the field-to-lidar model can hide substantial uncertainty (Saarela *et al* 2016), as can discounting the sampling uncertainty involved with associating fine-grain lidar measurements with coarser remote sensing data.

In this paper, we propose the use of hybrid inference (Fattorini 2012, Ståhl *et al* 2016) to estimate mean AGBD, with associated uncertainty, at the level of the GEDI grid cell while explicitly accounting for both field-to-lidar model error as well as sampling uncertainty. Hybrid methods have been used with both spaceborne (Healey *et al* 2012, Margolis *et al* 2015) and airborne (Ståhl *et al* 2011, Corona *et al* 2014) lidar instruments to estimate forest AGBD, though never for areas corresponding to grid cells. The calibration/application approach we propose (figure 1) relies upon lidar datasets contributed from around the world being converted to GEDI-like waveforms that realistically incorporate several types of uncertainty (Hancock *et al* 2019). Resulting simulated GEDI waveforms are being related to corresponding ground measurements to create parametric, footprint-level AGBD models (Kellner *et al* 2018).

Using hybrid estimation, these models will be applied to all footprints acquired by GEDI, and the resulting predictions will form the basis of a sample-based estimate of mean AGBD within each grid cell; no wall-to-wall imagery will be used. Our estimators of mean AGBD and the variance of that mean are similar to those used by Ståhl *et al* (2011), treating GEDI tracks intersecting a 1 km grid cell as a simple random sample of clustered observations. Variance is estimated as a function of: (1) the number of clustered predicted footprint-level AGBD values and the variability among them, as well as (2) the uncertainty of the parameter estimates used in the footprint-level models.

Key assumptions in this approach include the following: (1) the parameter covariance matrix generated during the footprint modeling process appropriately conveys the uncertainty of footprint-level AGBD predictions; (2) GEDI waveforms simulated from airborne data adequately represent signals received from the sensor, and; (3) the expected values of the proposed hybrid estimators are unbiased. This paper evaluates this third assumption.

The objectives of this study were to (1) propose a framework through which hybrid inference can be used with GEDI data to estimate mean AGBD and the variance of that mean; and (2) using GEDI waveforms simulated from airborne small-footprint lidar for 60 diverse grid cells, construct a simulation that tests the bias of the proposed estimators. Documenting estimator performance is an important step in evaluating figure 1's approach to turning GEDI's high-quality, spatially discontinuous observations into directly interpretable gridded estimates of mean AGBD across the globe.

2. Methods

This section describes the proposed GEDI sample framework and estimators, as well as a test of the proposed estimators. The first subsection describes the proposed sampling framework and our application of hybrid estimators to estimate both mean AGBD and variance of the mean. An empirical assessment of the properties of these estimators in the GEDI context is presented in the third subsection and in appendix S2 is available online at stacks.iop.org/ERL/14/065007/mmedia, while field and lidar data supporting this assessment are described in the second subsection and in appendix S1. Consideration of the role of spatial autocorrelation, specific to the case of the GEDI 1 km cell is given in appendix S3, while footprint-level modeling details are in appendix S4. We note that footprint models described in S4 are approximations of GEDI global models, which are being developed independently (Kellner *et al* 2018).

2.1. Estimators

The GEDI instrument will produce ~25 m (diameter) footprint observations along tracks paralleling the ISS obit. The intersection of a track with a 1 km cell produces a collection of GEDI footprints, to be denoted here as a cluster. We propose to view each 1 km grid cell as tessellated into equal-area non-overlapping population elements, each representing a potential GEDI footprint. GEDI footprints are population elements sampled by the GEDI instrument. Figure 2 illustrates this conception, adapted to the spatial dimensions of lidar and field data available for this study. Footprint (20 m) and grid cell dimensions (approximately 800 m) used here (illustrated in figure 2) were considered sufficiently similar to GEDI to support applied testing of the estimators described in the following. For simplicity, we consider tracks as having either a 45° or 135° degree inclination for descending or ascending orbits, respectively, although ISS inclination patterns will vary across latitudes. In this idealization, we assert that there are two potential disjoint clusters for the intersection of a track with a grid cell—one starting on the border footprint, and one starting with the subsequent along-track footprint. The sample will be the set of clusters from the tracks that intersect the 1 km grid cell; due to variability of the ISS orbit this sample can be considered a simple random sample of clusters of GEDI footprints. In this exercise, we do not consider missing data issues that will arise because of clouds.

The population parameter of interest is the average of the true AGBD over the N population elements (potential GEDI footprints), where N is the total number of population elements in a grid cell. For a 1 km grid cell, N is equal to 2500 20 $\times$ 20 m population elements, where the footprints in this realization would have a 20 m diameter. There is a parametric model, $g(\boldsymbol{x}, \boldsymbol{\alpha})$, of AGBD in Mg ha^{-1}, where $\boldsymbol{x}$ are characteristics of the GEDI waveform for the GEDI footprint, and $\boldsymbol{\alpha}$ is a vector of parameters. The true AGBD for a footprint is assumed to deviate from the $g(\boldsymbol{x}, \boldsymbol{\alpha})$ by a random quantity, ε, with an expected sum of zero. Under the assumption that the model is an unbiased predictor of the true AGBD, and ignoring model misspecification error, the modeled value $g(\boldsymbol{x}, \boldsymbol{\alpha})$ will be used in place of the true AGBD. The mean can be expressed in terms of GEDI ground clusters. There are M non-overlapping clusters distributed across all possible tracks. There are T_i footprints in the ith cluster. The reader may have noted that each population element is in one ascending cluster and one descending cluster. We can express the population attribute, the average expected predicted AGBD over the N population elements, in terms of the clusters; that is:

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

where the modeled AGBD for each population element occurs in the numerator twice.

Equation (1) will be expressed in an equivalent form, for which there is an estimator and a method of calculating an approximate variance. Let $G_i = \sum_{t=1}^{T_i} g(\boldsymbol{x}_{it}, \boldsymbol{\alpha})$, equal the cluster total of the predicted AGBD per hectare for the footprints in the ith cluster. Equation (1) can be expressed as the ratio of the mean (or sum of) AGBD per cluster and the mean (or sum of) number of elements per cluster:

$$\mu_Y = \frac{\sum_i^M G_i}{\sum_i^M T_i} = \frac{M^{-1}\sum_i^M G_i}{M^{-1}\sum_{i=1}^M T_i}. \tag{2}$$

Parameter estimates, $\hat{\boldsymbol{\alpha}}$, for the model $g(\boldsymbol{x}, \boldsymbol{\alpha})$ are based on a separate sample, S_m, that is independent of the 1 km grid cell. In equations (1) and (2) the parameters, $\boldsymbol{\alpha}$, are replaced by the parameter estimates, $\hat{\boldsymbol{\alpha}}$, and the population number of clusters, M, by the sample number of clusters, m, to get the estimator $\hat{\mu}_Y$, i.e. G_i with $\hat{G}_i = \sum_{t=1}^{T_i} g(\boldsymbol{x}_{it}, \hat{\boldsymbol{\alpha}})$

$$\hat{\mu}_Y = \frac{\sum_i^m\sum_{t=1}^{T_i} g(\boldsymbol{x}_{it}, \hat{\boldsymbol{\alpha}})}{\sum_{i=1}^m T_i} = \frac{\sum_{i=1}^m \hat{G}_i}{\sum_{i=1}^m T_i}. \tag{3}$$

The second component of the equation expresses the estimator $\hat{\mu}_Y$, the sample mean of all predicted AGBD (per sample element) which is an estimator of the population average expected AGBD (per element). The population true average AGBD differs randomly from the expected value μ_Y by the population average $\bar{\varepsilon}$ of the N deviations ε_{it}. For large N the contribution of $\bar{\varepsilon}$ to the total random error can be assumed negligible compared to the other two sources of uncertainty (the sampling error and the variability due to the model, see below), even if spatial autocorrelation is

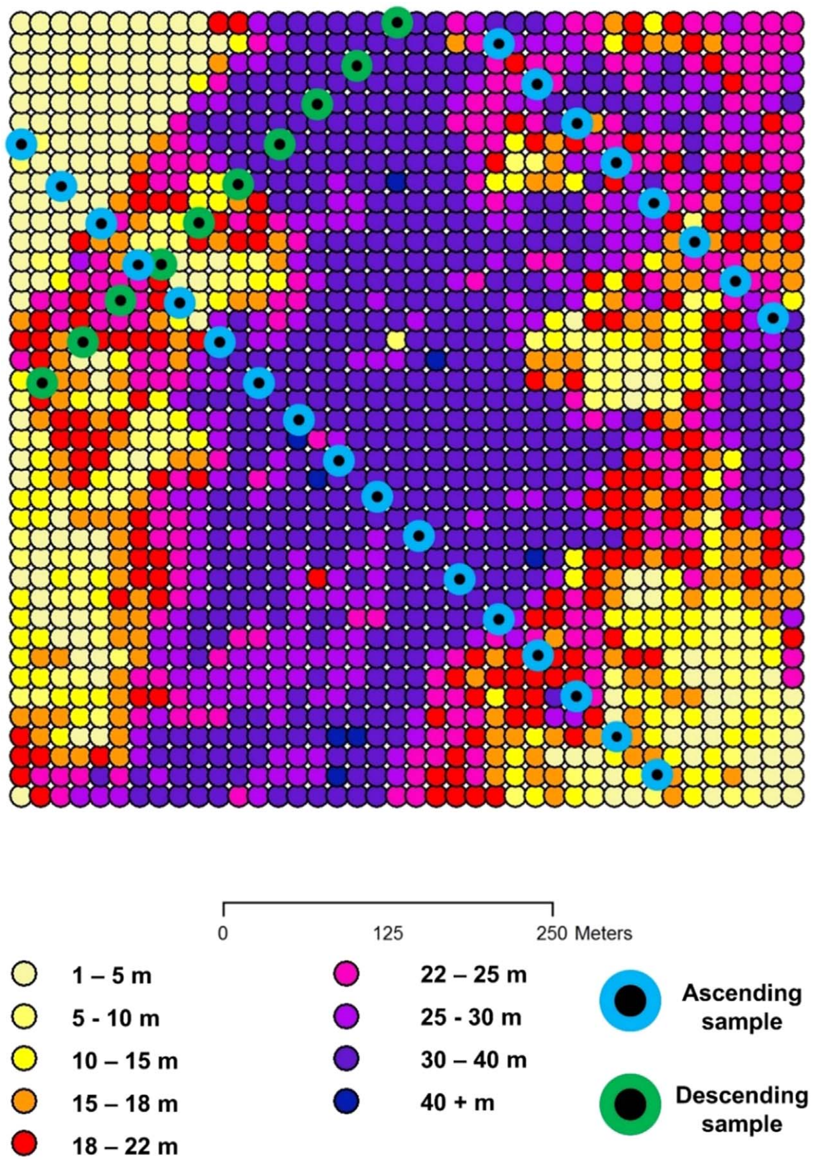

Figure 2. Near canopy top height (height from below which 96% of lidar energy returns) for GEDI footprints simulated from airborne data in one of six sites in the Pennsylvania/New Jersey study area. A possible sample of three clusters is shown.

present. Hence, it is sufficient to assume the size of the grid cell (and thus N) is large enough to imply that the average $\bar{\varepsilon}$ will be close to zero. This assumption, along with the assumptions related to model fit will be covered in the discussion section.

As noted in the Introduction, the GEDI sample will be treated as a random sample of clusters of GEDI footprints. The estimator in equation (3) can be viewed as estimating the mean of cluster totals and mean number of footprints separately, and then combining as a ratio estimator:

$$\hat{\mu}_Y = \frac{\sum_i^m \hat{G}_i}{\sum_{i=1}^m T_i} = \frac{m^{-1}\sum_i^m \hat{G}_i}{m^{-1}\sum_{i=1}^m T_i} = \frac{\bar{\hat{G}}}{\bar{T}}. \quad (4)$$

This estimation strategy combines a probability sample of auxiliary information (as opposed to wall-to-wall auxiliary information) with a prediction for the population element value, AGBD, instead of directly observing AGBD. Ståhl *et al* (2011) proposed an estimator of the form of equation (4) and derived an estimator of the approximate variance. Due to the small sample size used in that study in relation to the population size, Ståhl *et al* (2011) did not include a finite population correction. A finite population correction may need to be taken into account in GEDI's case. Expression (5) below is the variance estimator proposed in Ståhl *et al* (2011), with the addition of the finite population correction:

$$\hat{V}(\hat{\mu}_Y) = \frac{1}{\bar{T}^2}\left(1 - \frac{m}{M}\right)\frac{\sum_{i=1}^m (\hat{G}_i - \hat{\mu}_Y T_i)^2}{m(m-1)} + \frac{1}{\bar{T}^2}\sum_{j=1}^{p}\sum_{k=1}^{p}\widehat{\mathrm{Cov}}_{S_m}(\hat{\alpha}_j, \hat{\alpha}_k)\,\widehat{\bar{G}}_j^{\prime}\,\widehat{\bar{G}}_k^{\prime} \quad (5)$$

The first term is due to the sampling error and the second term is due to effects of uncertainty of the $\hat{\boldsymbol{\alpha}}$

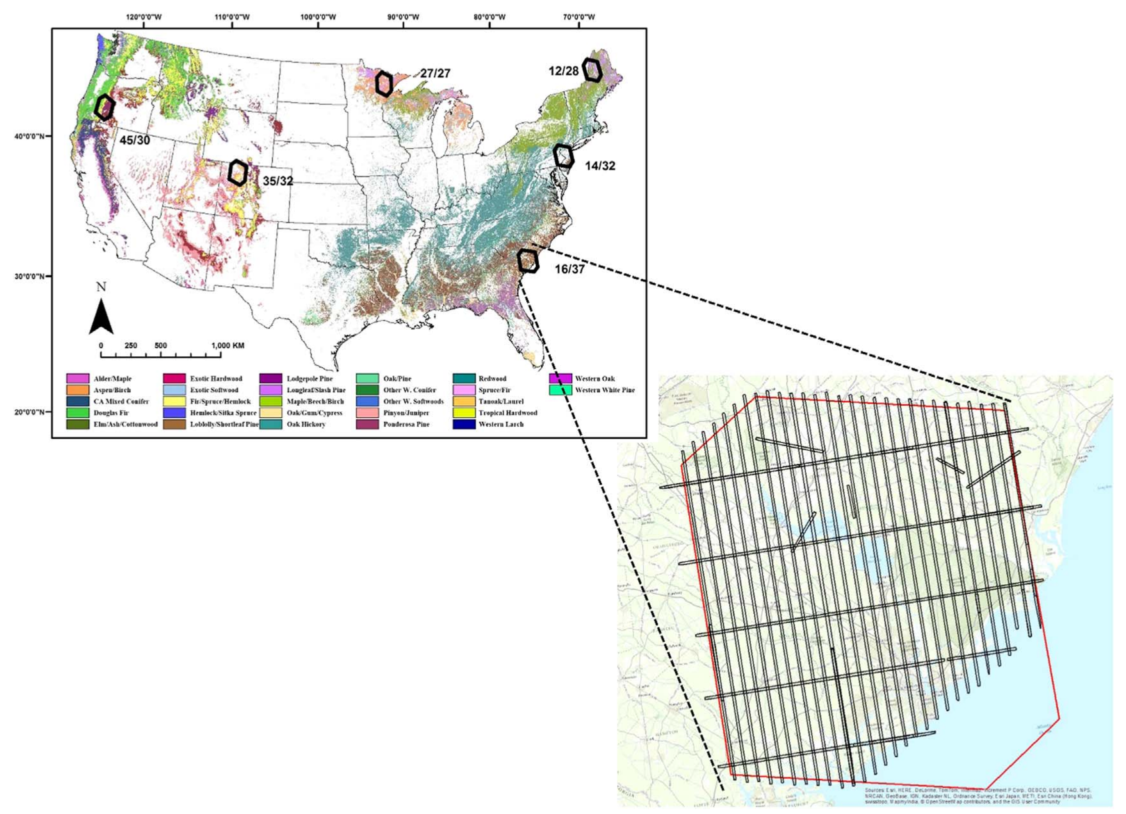

Figure 3. Location of the study's six focal areas, with detail for the SC site (24/34) showing airborne lidar flight lines. The width of each line was variable: ~300–900 m. Adapted from Healey *et al* (2018).

estimates (Ståhl *et al* 2011). The $\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\alpha}})$ is the estimated covariance matrix of the p-parameter estimates, where the estimate is based on the separate set of data, S_m, that is independent of the 1 km grid cell. For simplicity the first term is referred as the sample component and second term the model component. If we assume a linear model, that is $g(\boldsymbol{x}_{it}, \hat{\boldsymbol{\alpha}}) = \hat{\boldsymbol{\alpha}}_1 + \sum_{j=2}^{p} x_{itj}\hat{\boldsymbol{\alpha}}_j$, where x_{itj} is the jth component of $\boldsymbol{x}_{it}$, then, $\widehat{\widehat{G}}_j' = \frac{1}{m}\sum_{i=1}^{m}\sum_{t=1}^{T_i} x_{itj}$. Let $\widehat{\widehat{\boldsymbol{x}}}$ be the vector of the means of the cluster totals of the predictor variables, i.e. $\widehat{\widehat{x}}_j = \frac{1}{m}\sum_{i=1}^{m}\sum_{t=1}^{T_i} x_{itj}$, for $j = 1, \ldots, p$. Then the second component on the right side of equation (5) can be expressed in matrix notation:

$$\sum_{j=1}^{p}\sum_{k=1}^{p}\widehat{\mathrm{Cov}}_{S_m}(\hat{\alpha}_j, \hat{\alpha}_k)\widehat{\widehat{G}}_j'\widehat{\widehat{G}}_k' = \widehat{\widehat{\boldsymbol{x}}}^T\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\alpha}})\widehat{\widehat{\boldsymbol{x}}}. \qquad (6)$$

Note that $\widehat{\widehat{x}}_1$ is equal to $\overline{T}$. In the construction of equation (5), a Taylor linearization process is used twice; first with regards to the model parameters of $g(\boldsymbol{x}, \hat{\boldsymbol{\alpha}})$ (Ståhl *et al* 2011) and second to estimate variance of a ratio estimator (Särndal *et al* 1992).

The empirical data described in section 2.2 were used to calibrate the simulation study described in section 2.3 to evaluate the statistical properties of these estimators in the GEDI setting.

2.2. Supporting datasets

Lidar and field data collected over six diverse areas in the United States supported this study. These areas were defined by the non-overlapping window (Thiessen Scene Area) of a local Landsat scene. The six scenes (given by numeric WRS-2 Path/Row) included one each in: northern Maine ('ME': 12/28, excluding the Canadian portion); eastern Pennsylvania and central New Jersey ('PA/NJ': 14/32); coastal South Carolina ('SC': 16/37); northern Minnesota ('MN': 27/27); northwestern Colorado ('CO': 35/32); and western Oregon ('OR': 45/30) (figure 3). The selected areas represented a wide range of forest ecosystems and disturbance processes, as described by Cohen *et al* (2017) and Healey *et al* (2018). Small-footprint lidar was collected at each site in the pattern displayed in figure 3, and a waveform simulator developed by the GEDI Science Definition Team was used to simulate GEDI waveforms from the airborne data. Data collection and waveform simulation is described in appendix S1.

2.3. Evaluating estimator performance

Simulations were conducted to examine the properties of the above estimators of mean AGBD and variance of the estimated mean. Modeled AGBD values treated as truth for each population element were developed for sixty square grid cells (ten randomly chosen grid cells at each of the six test areas in figure 3) of dimensions

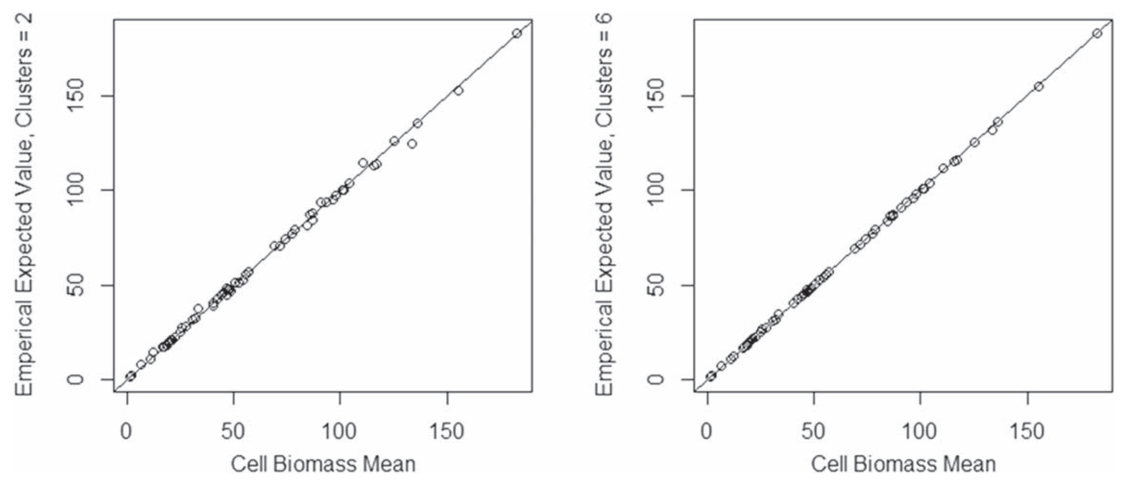

Figure 4. Expected estimate of AGBD (Mg ha^{-1}) versus true AGBD (on the X axis) for both two and six tracks. The line represents $y = x$.

approximately similar to GEDI's product specifications. These populations were created by applying a linear model to the GEDI metrics simulated from actual small-footprint lidar acquisitions (figure 2). Appendix S2 details the process of both establishing the true population and simulating different potential cluster patterns and model parameter combinations to test the proposed estimators under a range of conditions. The advantage of using real data to develop the 'true' population, as opposed to simulating arbitrary AGBD surfaces, was that it ensured realistic covariance among lidar metrics. To the degree that lidar metrics are correlated with AGBD (see appendix S3), this approach also created a realistic spatial representation of biomass and of residual prediction error when different realizations of model parameters were applied to the lidar metrics. The swath width of the available lidar data limited the size of these test cells to between 680 × 680 m and 800 × 800 m (instead of 1 km); as mentioned earlier, this was considered adequate for the tests described below.

3. Results

The mean true AGBD of the 60 simulated grid cell-scale populations ranged from 1 to 183 Mg ha^{-1}. There was little difference between the true and expected estimated values across simulated grid cells (figure 4). Divergence between these quantities centered around zero, and decreased with more clusters (figure 5).

The empirical percent bias of the variance estimator (equation (5)) as a function of number of clusters is shown in figure 6. The estimator appears to be only asymptotically unbiased, with mean underestimation of variance at approximately 21% with two clusters and 2% with six clusters. Outliers in figures 5 and 6, which measure bias as a percent of the mean AGBD, tended to be low-biomass locations.

The variance under-prediction at low numbers of clusters may be traced to the sample component of equation (5), not the model component (figure 7).

In most cases, the sampling component of the variance was larger than the model component (figure 8). With more tracks, overall variance of the estimate went down, and model variance accounted for a larger percentage of that variance.

4. Discussion

4.1. Advantages and assumptions

Earlier work with NASA's Geoscience Laser Altimeter System instrument pioneered the use of spaceborne lidar to make forest AGBD maps, often employing models that related field biomass to lidar and then, in a separate stage, related lidar-modeled biomass to synoptic data from sources like the MODIS satellites. While some work has attempted to impose a statistical framework on the use of such maps (e.g. Nelson *et al* 2017), most have employed ad hoc approaches to uncertainty. Mitchard *et al* (2014) pointed out two such ad hoc error budgeting approaches that yielded estimates across the Amazon that did not overlap with each other's relatively narrow confidence intervals, suggesting unacknowledged uncertainty in one or both estimates.

Saarela *et al* (2016) concluded that ignoring error occurring in the field-to-lidar step, as is often done, can lead to underestimation of variance by a factor of three. The spatial mismatch inherent in modeling AGBD in large pixels from constituent smaller-footprint measurements can also introduce uncertainties not clearly addressed in variance estimates (Réjou-Méchain *et al* 2014). The proposed hybrid approach does explicitly account for field-to-lidar model error, and it also accounts for sample error as footprints are combined to infer mean 1 km AGBD.

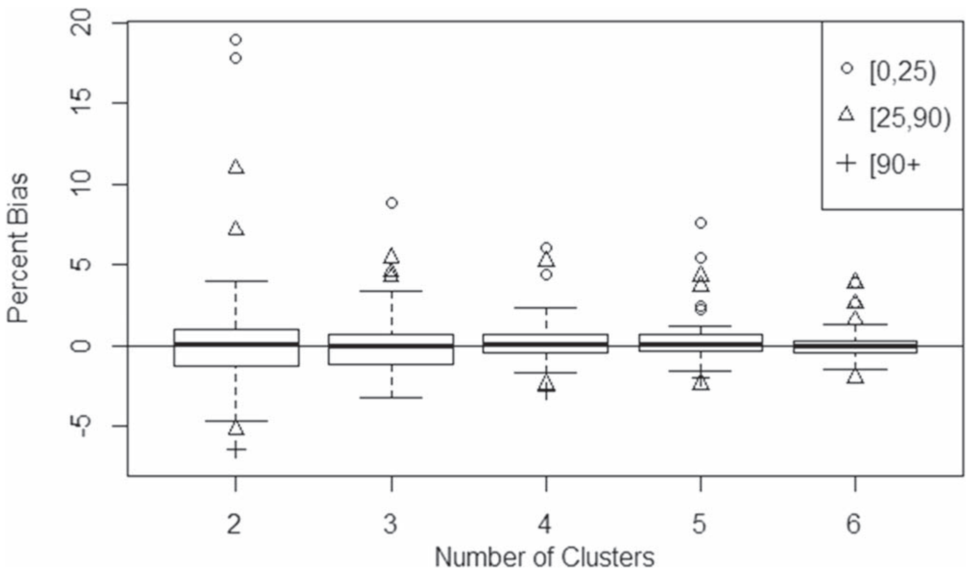

Figure 5. Boxplot of percent bias for ABGD estimates (equation (4)) at the grid cell level. Outliers are divided into categories defined by the mean AGBD for the simulated cell (0–25, 25–90, and 90 + Mg ha^{-1}).

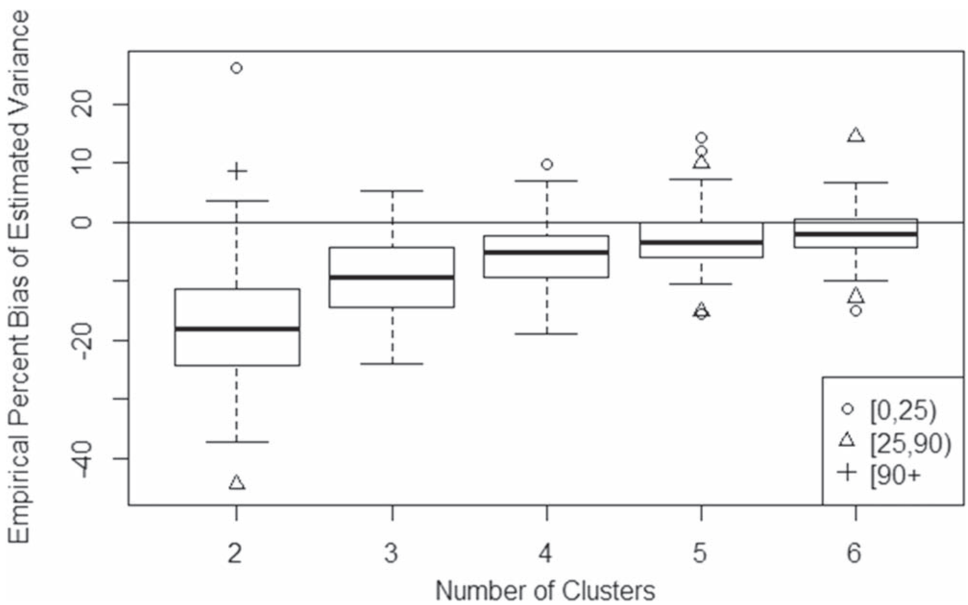

Figure 6. Boxplot of the empirical percent bias in the variance estimator for each grid cell (equation (5)). Outliers are divided into categories defined by the mean AGBD for the simulated cell (0–25, 25–90, and 90 + Mg ha^{-1}).

However, the hybrid approach summarized in figure 1 is subject to at least three important assumptions. The first is that the GEDI waveform simulator (Hancock *et al* 2019; described in appendix S2) correctly translates pulse counts to waveform data while representing the important sources of error. The benefit of using simulated GEDI data at precisely known locations to build footprint-level models is that it obviates potential error related to the positional uncertainty of GEDI footprints (currently estimated to be on the order of 10 m, post-processed). The risk is that any systematic simulator error may be propagated through footprint-level models to bias the 1 km mean AGBD estimates. However, the simulator is based on well-validated work by Hofton and Blair and Hofton (1999) and it is unlikely that violations of this assumption are substantial.

A second assumption is that one square kilometer is a large enough domain, given spatial autocorrelation of the population, that residual model error may be considered negligible. Mean residual error in the footprint AGBD model is presumed to tend toward zero over a large number of predictions. In an area that is small, particularly if spatial autocorrelation of residual error is high, residual errors may not sum to zero, introducing additional error. While GEDI's 1 km spatial domain is smaller than other applications of

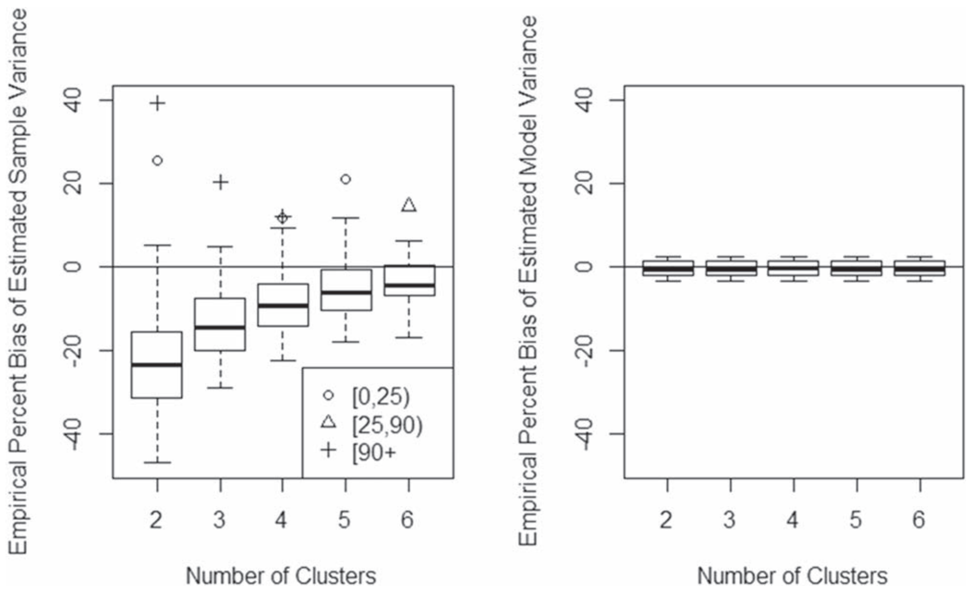

Figure 7. Decomposition of bias in overall grid cell-level variance estimates (figure 6) into sample and model contributions.

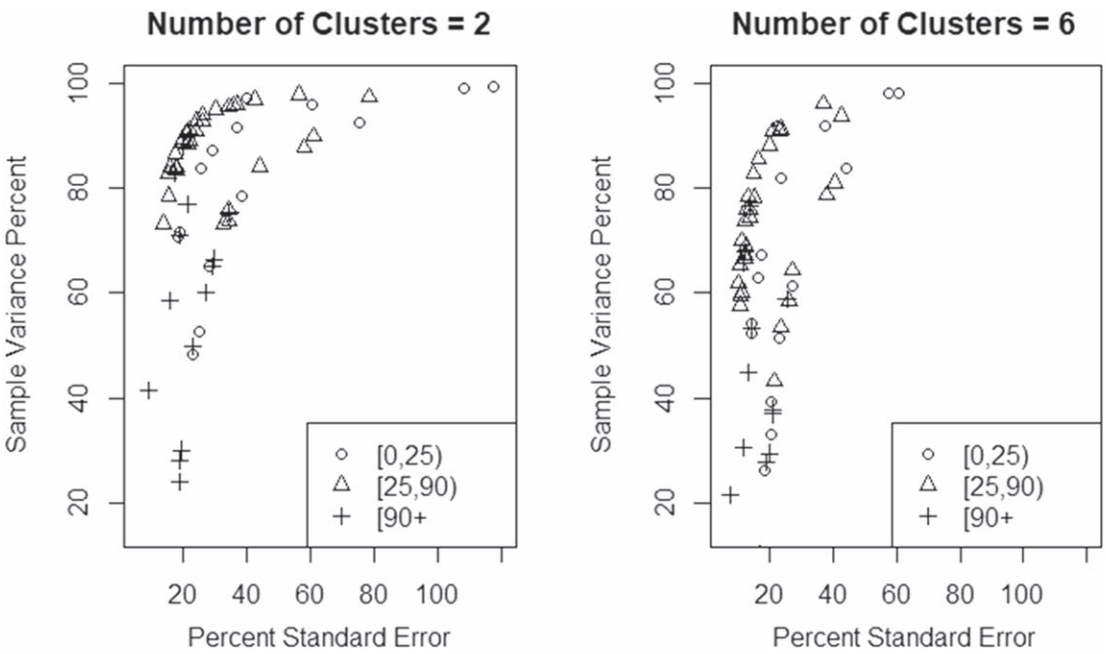

Figure 8. Sample variance as a percentage of the total variance versus the percent standard error (square root of the variance) for 2 clusters and 6 clusters. More clusters reduced both overall variance and the importance of the sample component of the variance.

hybrid inference (e.g. Ståhl *et al* 2011), initial exploration using the data from this study suggests that spatial autocorrelation at GEDI's 25 m grain size is relatively low, as is the effect of omitting residual error (also called individual error in appendix 3) from the proposed hybrid variance estimator (appendix 3). In some cases, however, the impact of residual error may reach 15% or more of the variance, so further work is needed to develop ways to include it in the hybrid approach.

Finally, it is assumed that the footprint-level AGBD model is correctly specified and 'applies' to the population in the cell. This notion implies that the number and order of terms are correctly specified, and that the parameter covariance matrix used in the variance calculation (equations (5) and (6)) adequately reflects uncertainty in the relationship between lidar and AGBD within the cell. Since even small model difference can generate strongly divergent population estimates when multiplied over large areas, this is a critical assumption. Although the GEDI team is collating a comprehensive global database of training data, the reality is that ground data are sparse in some areas. In such areas, model misspecification is an important risk. In this respect, hybrid inference is no different from any other remote sensing approach that relies upon field data to calibrate remotely sensed observations.

4.2. Performance of hybrid estimators in the context of GEDI's sample

Properties of the hybrid estimators proposed for the GEDI mission were evaluated here using simulations in which thousands of potential GEDI cluster patterns were tested in the context of model covariance across forests in 60 diverse grid cells. Bearing in mind the above assumptions associated with the approach outlined in figures 1, 4 and 5 suggest that the proposed estimator for mean AGBD exhibits negligible bias across cells. Our tests and the results in figures 4 and 5 presumed proper model specification, an assumption of model-based inference, although residual plots (figure S4(1)) suggested that trends in model residual error may exist in at least one study region. In such cases, corrective measures such as variable transformations may be required.

The variance estimator, on the other hand, appeared to be only asymptotically unbiased (figure 6); that is, bias approached zero as larger numbers of cluster samples were acquired. The Taylor linearization used to compute the sample component of the variance (as opposed to the model component) likely contributed to this underestimation. While Taylor linearization simplifies calculation of the variance, it is known to lead to variance underestimation for small sample sizes (Särndal *et al* 1992, p 176). One area of future work may center around methods proposed to adjust for under-estimation (Cochran 1977, p 156). Linearization was also used in deriving the model component of the variance estimator (equation (5)), but the linear model used in this study for footprint-level prediction likely eliminated the impact of linearization in the model component.

Negative bias in the variance estimates at low numbers of clusters should be considered when hybrid methods are used with GEDI data, and information about the number of clusters will accompany GEDI hybrid estimates for each cell. Low sample numbers will occur both at the beginning of the mission and in equatorial regions where the ISS orbit provides fewer 'looks' and where persistent cloud cover may frequently obscure the surface. An advantage of hybrid estimation is that the same methodology applied at the 1 km scale may also be applied to much larger, irregularly shaped areas such as entire countries; instead of having to combine grid cell estimates, all clusters intersecting a given country may be used in the context of a single hybrid estimate. GEDI country-scale sample size will be very large, leading to greatly diminished design-based contribution to variance estimates. The apparent dominance of the sample component in the variance estimator (figure 8) may be relevant to GEDI operational planning. There are signal-to-noise parameters that may allow GEDI to consider a larger sample of noisier footprints (and models) or a smaller sample of footprints collected under ideal atmospheric conditions. It should be recognized that the relative contribution of the sampling and model components of the variance estimator may differ in different settings as properties of the model change due, for example, to the size of S_m or complexity of the underlying lidar/forest structure relationship. Across the diverse ecosystems we studied, though, results suggest that including a larger sample will have a bigger impact on the magnitude and unbiasedness of estimated variance than a marginal improvement in model fit.

5. Conclusions

The GEDI mission was conceived to support estimation of forest biomass at greater accuracy and resolution than has previously been possible by greatly increasing the number of available observations of forest structure. The sampling pattern employed by GEDI is largely constrained by its technology: its deployment platform on the ISS, the number and strength of its lasers, and its mission duration. Central to the GEDI approach has been the creation of a framework within these observational constraints for properly estimating both mean AGBD and the variance around that estimated mean.

Our research suggests that a hybrid approach that accounts for uncertainty due both to the model and the sampling design is appropriate and effective. Analysis here focused on 1 km cells, but hybrid inference using GEDI data could likewise be applied over any political or ecological units of interest, subject to minimum size and model applicability assumptions cited above. Our results confirm the importance of having abundant field plot estimates of biomass and associated airborne lidar, from which representative models that relate lidar metrics to AGBD may be created. The creation of a global database of such field and lidar data has been a priority of the GEDI mission before launch, and its continued expansion should both reduce the model uncertainty carried into 1 km variance estimates and support the local applicability assumptions that underlie model-based inference.

Lastly, our research also suggests that maximizing the number of clusters within a cell is key towards providing approximately unbiased variance estimates. While observational strategies to increase cluster density are limited, options such as purposefully targeting important areas, lengthening mission duration, or increasing the size of grid cells to encompass more clusters are worth exploring.

Acknowledgments

This work was supported by grants NNH13AW62I and 80HQTR18T0016 from NASA's Carbon

Monitoring System (solicitation NNH16ZDA001N) as well as support from the NASA GEDI Science Definition Team. The authors appreciatively acknowledge constructive input from three anonymous reviewers.

ORCID iDs

Sean P Healey https://orcid.org/0000-0003-3498-4266
Svetlana Saarela https://orcid.org/0000-0002-9044-7249
James R Kellner https://orcid.org/0000-0002-9861-4857

References

Baccini A, Walker W, Carvalho L, Farina M, Sulla-Menashe D and Houghton R A 2017 Tropical forests are a net carbon source based on aboveground measurements of gain and loss *Science* **358** 230–4

Blair J B and Hofton M A 1999 Modeling laser altimeter return waveforms over complex vegetation using high-resolution elevation data *Geophys. Res. Lett.* **26** 2509–12

Cochran W G 1977 *Sampling Techniques* 3rd ed (New York: Wiley)

Cohen W *et al* 2017 How similar are forest disturbance maps derived from different landsat time series algorithms? *Forests* **8** 98

Corona P, Fattorini L, Franceschi S, Scrinzi G and 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** 1303–11

Fattorini L 2012 Design-based or model-based inference? The role of hybrid approaches in environmental surveys *Studies in Honor of Claudio Scala* ed L Fattorini vol 2012 (Siena: Department of Economics and Statistics, University of Siena) pp 173–214

Hancock S, Armston J, Hofton M, Sun X, Tang H, Duncanson L I, Kellner J R and Dubayah R 2019 The GEDI simulator: a large-footprint waveform lidar simulator for calibration and validation of spaceborne missions *Earth Space Sci.* **6** 294–310

Healey S P *et al* 2018 Mapping forest change using stacked generalization: an ensemble approach *Remote Sens. Environ.* **204** 717–28

Healey S P, Patterson P L, Saatchi S S, Lefsky M A, Lister A J and Freeman E A 2012 A sample design for globally consistent biomass estimation using lidar data from the geoscience laser altimeter system (GLAS) *Carbon Balance Manage.* **7**

Kellner J, Duncanson L and Armston J 2018 Algorithm Theoretical Basis Document (ATBD) for GEDI L4 Footprint Aboveground Biomass Density (Greenbelt, MD: NASA Goddard Space Flight Center)

Margolis H A, Nelson R F, Montesano P M, Beaudoin A, Sun G, Andersen H-E and 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** 838–55

Mitchard E T A *et al* 2014 Markedly divergent estimates of Amazon forest carbon density from ground plots and satellites *Glob. Ecol. Biogeography* **23** 935–46

Nelson R *et al* 2017 Lidar-based estimates of aboveground biomass in the continental US and Mexico using ground, airborne, and satellite observations *Remote Sens. Environ.* **188** 127–40

Réjou-Méchain M *et al* 2014 Local spatial structure of forest biomass and its consequences for remote sensing of carbon stocks *Biogeosciences* **11** 5711–42

Saarela S, Holm S, Grafström A, Schnell S, Næsset E, Gregoire T G, Nelson R F and Ståhl G 2016 Hierarchical model-based inference for forest inventory utilizing three sources of information *Ann.Forest Sci.* **73** 895–910

Särndal C-E, Swensson B and Wretman J 1992 *Model Assisted Survey Sampling* (New York: Springer) p 694

Ståhl G, Holm S, Gregoire T G, Gobakken T, Naesset E and Nelson R 2011 Model-based inference for biomass estimation in a lidar sample survey in Hedmark County, Norway *Can. J. Forest Resour.* **41** 96–107

Ståhl G *et al* 2016 Use of models in large-area forest surveys: comparing model-assisted, model-based and hybrid estimation *Forest Ecosyst.* **3** 5

Zolkos S G, Goetz S J and Dubayah R O 2013 A meta-analysis of terrestrial aboveground biomass estimation using lidar remote sensing *Remote Sens. Environ.* **128** 289–98

Supplementary Materials

Appendix S1. Field and Lidar Data Collection Methods

Lidar data were collected using a Riegl LMS-Q680i system at five sites (SC, NJ/PA, ME, MN, CO) in the summer of 2014 and a sixth site (OR) in the summer of 2015. All data were collected during snow-free time periods when vegetation was fully leaf-on and non-senescent. The data was collected in north-south oriented flight lines spaced 5 km apart (Figure 3): some crossing flights were flown for calibration purposes. Technical specifications for the lidar acquisition included: point density of 4 pulses per square meter; altitude of 732 m; field of view of 60 degrees; pulse rate of 330000 Hz; nominal swath width of 812 m; horizontal and vertical accuracy (root mean square error in z) of 50 cm. The lidar data were processed in Fusion (McGaughey 2015) to produce rasters of various metrics with a 30 meter cell size to support the plot selection process described below.

Approximately fifty field inventory plots were established at each area in the summer of 2015 (Legner et al, in prep). To ensure that field plots represented the full range of biomass levels within forested areas of each scene, the lidar data was used to inform the selection of field plot locations via a stratified sampling procedure. After masking out non-forest areas within the lidar coverage using a lidar-based percent cover threshold (10%) and the National Land Cover Dataset (Homer et al. 2007), fifteen strata were delineated based upon three vegetation cover and five height classes. A minimum of 23-24 candidate plot locations were generated for each of the 15 identified strata. Final field plots were selected from these candidate plots on the basis of accessibility and valid forest land cover status. Once in the field, crews aimed to visit 50 sample plots at each study area, with 3-4 plots in each stratum.

At each plot location, live and dead trees with dbh > 12.7 cm were measured on a 16.2-meter radius plot, and trees with 2.54 cm < dbh < 12.7 cm were measured on a 4.57-m radius circular plot. Field protocols consistent with those used by the US Forest Service were used in measuring trees (CMS Field Guide, 2015). In addition, survey-grade GPS coordinates (< 1 meter positional error) were acquired for each plot center.

Large-footprint lidar waveforms similar to those expected from GEDI were simulated from the acquired lidar data using the method presented in Blair and Hofton (1999), in which waveforms are modeled as the sum of individual returns from surfaces at different heights, accounting for instrument-specific properties. Realistic noise was added to the simulated waveforms following Hancock et al. (2011) and Davidson and Sun (1988). The expected signal to noise ratio (SNR) of GEDI signals has been predicted through link margin analysis. For a given SNR, the probability of the ground elevation being correctly identified can be calculated and used to quantify the expected measurement accuracy. The simulator has been validated against real LVIS data (airborne large-footprint, full-waveform lidar similar to GEDI; (Blair et al. 1999)) in terms of waveform metrics and metric accuracy in the presence of noise (Hancock et al., in review).

Appendix S2. Simulations and Assessment Details

For each region, we fit the linear model $Y_{ri} = x_{ri}^2\alpha_r + \epsilon_r$, where, r indicates the region, x_{ri} is the simulated rh90 value (the height above the ground below which 90% of energy returned) for the ith plot in the rth region, and $\varepsilon_r \sim N(0, \sigma_r)$. For context, Figure 1 displays the rh96 value of a cell in the PANJ area. The above model was fit in each study region from the field data described in Appendix S1. Related parameter covariance matrices used by the proposed variance estimator (Eq. 5) therefore reflected realistic levels of model uncertainty. Because the "true" population was built with the same models used in the hybrid inference process (Equations 1 and 4), those models were assured to be both correctly specified and fit, important assumptions of model-based and hybrid inference. It must be noted that models generated for this dataset are not exactly the same as those fit with GEDI's global dataset (Kellner et al., 2018).

Each simulation varied both the parameters of the model used in the hybrid inference process and the spatial orientation of the sample (Figure 2 gives one permutation), comparing results of the proposed hybrid estimators to properties of the established "true" population. The covariance matrix of parameter values from the footprint-level AGBD model is the primary vehicle for representing model uncertainty in hybrid variance estimators (Equation. 4 and 5), so to evaluate the effect of the model, 3999 random vectors were generated from the multinomial normal distribution $MVN\left(\widehat{\boldsymbol{\alpha}}, \widehat{Cov}(\widehat{\boldsymbol{\alpha}})\right)$. These, along with the parameters, $\widehat{\boldsymbol{\alpha}}$, from the initial model, created a set of 4000 realizations of the model parameters.

Alternative spatial samples, representing different potential cluster patterns, were drawn for each grid cell to evaluate properties of the sample design. Four thousand cluster patterns were randomly drawn for scenarios of sample sizes from 2 to 6 clusters. As described in Section 2.1, clusters were considered simply as lines of alternating footprints with inclinations either of 45° (descending) or 135° (ascending). Each sample can be represented as a pair (m_a, m_d), where m_a is the number of ascending tracks and m_d is the number of descending tracks. Pilot testing in the Oregon study area suggested that the mix of ascending and descending clusters, such as $(0,3), (1,2), (2,1)$ and $(3,0)$, did not affect simulation results; only the sum, $m_a + m_d$, was important. Therefore, for simulations representing each number of clusters [2,...,6], the following cluster configurations were used: $(1,1), (1,2), (2,2), (3,2)$ and $(3,3)$.

For each sample/model combination (i, j), with $1 \leq i \leq 4000$ and $1 \leq j \leq 4000$, the ith sample i and the jth realization of the model parameters was used to calculate values for the estimator $\hat{\mu}_{Yij}$ (Equation 4) and the variance estimator $\hat{V}(\hat{\mu}_Y)_{ij}$ (Equations 5 and 6). This set of 16 million mean biomass and variance estimates was stored for each of 60 grid cells across simulations of from 2 to 6 clusters.

Assessment of the bias is based on comparing the expected value of the estimator to the "true" mean of AGBD for the population (the grid cell). The true mean is denoted μ_g, where g is the index of the grid cell. Each grid cell and cluster combination defines an estimator, $\hat{\mu}_{go}$, where o is the index for the cluster. For the expected value of the estimator, $E\left(\hat{\mu}_{go}\right)$, we used the mean of the 16 million entries in the estimate matrix. This is an estimate of the expected value of the estimator with respect to the both the sample and the model.

We assessed the bias in two ways. First we graphed true means at the grid cell level versus the expected (average) value of estimator. Second, we graphed the percent bias for each grid cell cluster combination

$$PcntBias_{go} = \frac{E\left(\hat{\mu}_{go}\right) - \mu_g}{\mu_g} * 100 \qquad [7]$$

versus the number of clusters.

The assessment of the bias in variance estimator was based on an empirical estimate of the percent bias for each grid cell:

$$PcntBiasEV_{go} = \frac{E\left(\hat{V}(\hat{\mu}_Y)_{go}\right) - V(\hat{\mu}_Y)_{go}}{V(\hat{\mu}_Y)_{go}} * 100 \qquad [8]$$

where the subscript g is the index for the grid cells and o is the index for the cluster, $V(\hat{\mu}_Y)_{go}$ is the variance of the estimator for the gth grid cell and the oth cluster, and $E\left(\hat{V}(\hat{\mu}_Y)_{go}\right)$ is the expected value of the variance estimator for the gth grid cell and the oth cluster. The variance of the 16 million mean biomass estimates is used as an empirical estimate of the variance of the estimator, $V(\hat{\mu}_Y)_{go}$. The mean of the 16 million estimated variances is used as the expected value of variance estimator, $E\left(\hat{V}(\hat{\mu}_Y)_{go}\right)$.

To test the sufficiency of the 16 million simulated values we graphed the variance of the mean biomass estimates versus the number of simulation repetitions. Figure S2-1 shows that estimator stabilizes somewhere the between 10,000 and 15,000 repetitions. The line is mean of variance for number of repetitions between 10,000 and 16,000.

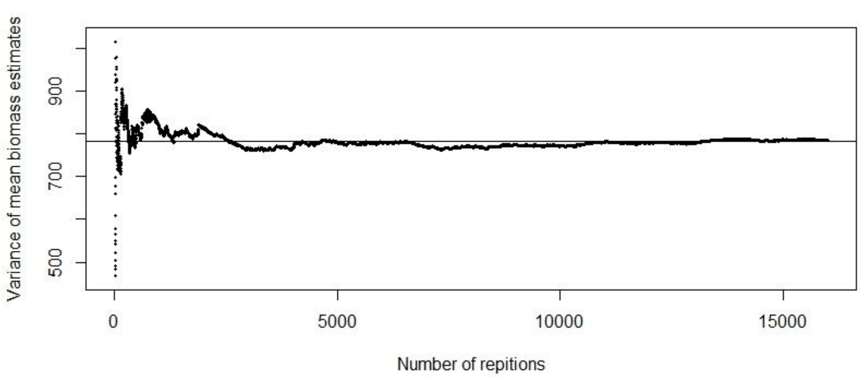

Figure S2-1: The variance of the mean biomass estimates versus number of repetitions. The line is the mean of variance of the mean biomass estimates for repetitions between 10,000 and 16,000.

As noted in section 2.1, the variance estimator, Equation 5, has two components; the first is due to the sampling design and the second to uncertainty in the model. The relative contribution of each component can be evaluated using the simulation results. For each of 16 million (i, j) combinations, we decomposed Equation 5 into sample and model variance, and then took the mean of each as the expected analytical value. Empirical sample variance was calculated by referencing the original 16 million prediction of mean AGBD, allowing that one could take the variance of all estimates for any given sample (i) across all sets of model parameters (j), and one could likewise take the variance of all estimates for any set of parameters (j) across all samples (i). The mean of estimate variances across model parameters was taken as an empirical estimate of the sample contribution to the variance, and the mean of estimate variances across samples was taken as an empirical estimates of the model contribution. Biases for the respective components were then evaluated using Equation 8.

Appendix S3

In this article we focus on the variance as a measure of uncertainty for the hybrid estimator, i.e. we assess the expected squared deviation between the estimator and its expected value. However, especially for small populations (see, e.g., McRoberts *et al.* 2018) it might be more relevant to use the mean square error (MSE) as an uncertainty measure, since the MSE is the expected squared deviation between the estimator and the true population parameter value. However, while the variance can typically be straightforwardly estimated based on empirical survey data this is normally not the case for the MSE. Thus, if we use the variance as a measure of uncertainty we should ideally make sure that the difference in numerical value between the variance and the MSE is small for the type of applications addressed.

In this appendix we first clarify the difference between the variance and the MSE for the hybrid estimator, and then assess numerical differences between the variance and the MSE under assumptions that mimic GEDI mission core applications.

For a given realization from a superpopulation model (e.g., Cassel et al., 1977) that we assume to have generated our data, the true biomass, μ, per area unit (a random variable) within the target study area can be defined as

$$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N} \tilde{g}\,(x_i, \alpha, \varepsilon_i) = \frac{1}{N}\sum_{i=1}^{N} g\,(x_i, \alpha) + \frac{1}{N}\sum_{i=1}^{N} \varepsilon_i, \qquad [9]$$

where N is the population size (number of potential GEDI footprints in the study area), y_i is the true biomass (per area unit) for population unit i (equivalent with $\tilde{g}$(.)), *g(.)* is the fixed part of the superpopulation model, x_i is the value of the auxiliary GEDI variable (usually a vector) for unit *i*, α is (the vector of) parameters in the model, and ε_i is an individual deviation for unit *i*. The individual deviations may follow certain probability distributions and may be spatially correlated.

Based on a sample, the model parameters in g(.) are estimated, and we obtain the estimated model $g(x_i, \hat{\alpha})$, which is assumed to be unbiased (at least with good approximation), i.e. $\mathrm{E}[g(x_i, \hat{\alpha})] = g(x_i, \alpha)$.

A generic Horvitz-Thompson-type (Ståhl et al., 2016) hybrid estimator of the biomass per area unit in the study area is

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{n}\frac{1}{\pi_i}g(x_i,\hat{\alpha}), \qquad [10]$$

where π_i is the inclusion probability of population unit *i* and *n* is the number of sampled units.

The MSE of the estimator is defined as $E[(\hat{\mu} - \mu)^2]$ whereas the variance of the estimator is defined as $E[(\hat{\mu} - E[\mu])^2]$. The MSE comprises three different terms, as shown below, since the deviation, D, between the estimator $\hat{\mu}$ and the true (random) value μ can be partitioned into three components:

$$\begin{aligned} D &= \hat{\mu} - \mu \\ &= \frac{1}{N}\sum_{i=1}^{n}\frac{1}{\pi_i}g(x_i,\hat{\alpha}) - \frac{1}{N}\sum_{i=1}^{N}\tilde{g}\,(x_i,\alpha,\varepsilon_i) \\ &= \frac{1}{N}\sum_{i=1}^{n}\frac{1}{\pi_i}g(x_i,\hat{\alpha}) - \frac{1}{N}\sum_{i=1}^{n}\frac{1}{\pi_i}g(x_i,\alpha) \\ &\quad + \frac{1}{N}\sum_{i=1}^{n}\frac{1}{\pi_i}g(x_i,\alpha) - \frac{1}{N}\sum_{i=1}^{N}g\,(x_i,\alpha) \\ &\quad + \frac{1}{N}\sum_{i=1}^{N}g\,(x_i,\alpha) - \frac{1}{N}\sum_{i=1}^{N}\tilde{g}\,(x_i,\alpha,\varepsilon_i) \end{aligned} \qquad [11]$$

Here,

$$D_M = \frac{1}{N}\sum_{i=1}^{n}\frac{1}{\pi_i}g(x_i,\hat{\alpha}) - \frac{1}{N}\sum_{i=1}^{n}\frac{1}{\pi_i}g(x_i,\alpha) \qquad [12]$$

is defined as model error.

$$D_S = \frac{1}{N}\sum_{i=1}^{n}\frac{1}{\pi_i}g(x_i,\alpha) - \frac{1}{N}\sum_{i=1}^{N} g\,(x_i,\alpha) \qquad [13]$$

is sampling error and

$$D_I = \frac{1}{N}\sum_{i=1}^{N} g\,(x_i,\alpha) - \frac{1}{N}\sum_{i=1}^{N} \tilde{g}\,(x_i,\alpha,\varepsilon_i) \qquad [14]$$

is the individual error. Since the three terms are uncorrelated (proof is omitted here), the MSE will be the sum of expected squared deviations for each of the terms. However, the variance will only include the expected squared deviation of the first two terms.

Thus, the difference between the MSE and the variance is the expected squared deviation of (Eq 14). Note that this deviation is taken across all units in the population and that it is invariant to what sampling and modeling strategies were applied. The expected squared deviation of this term can be written as:

$$E\left[\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\varepsilon_i\,\varepsilon_j\right] \qquad [15]$$

The size of Eq. 15 will depend on the size of the individual deviations and their spatial auto-correlation. Modeling based on data that are strongly correlated with the target variable, such as GEDI data with biomass, will tend to have smaller individual deviations compared to models based on data that are weakly correlated with the target variable.

We evaluated the difference between the MSE and the variance of our hybrid estimator for a number of cases based on data from our study areas (see section 2.2). In our case a ratio hybrid estimator was applied (see Eq 4 for notation):

$$\hat{\mu}_Y = \frac{\bar{\bar{\hat{G}}}}{\bar{\bar{T}}} = \frac{m^{-1}\sum_i^m \hat{G}_i}{m^{-1}\sum_{i=1}^m T_i}$$

Its variance (see Eq.[5]) is:

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

However, as pointed out above, the difference between the MSE and the variance is still given by Eq. 15.

To perform the evaluation we employed the Monte Carlo simulation approach and estimated a ratio between empirical hybrid estimator standard error (SE) and empirical root mean square error (RMSE), i.e.

$$ratio = \frac{SE_{emp}(\hat{\mu}_Y)}{RMSE_{emp}(\hat{\mu}_Y)}\ \frac{\sqrt{V_{emp}(\hat{\mu}_Y)}}{\sqrt{MSE_{emp}(\hat{\mu}_Y)}} \qquad [16]$$

where empirical variance is defined as (cf. Cassel et al. 1977, p. 94)

$$V_{emp}(\hat{\mu}_Y) = E[(\hat{\mu}_Y - E[\hat{\mu}_Y])^2] = \frac{1}{\Omega}\sum_{i=1}^{\Omega}\left(\hat{\mu}_{Y_i} - E[\hat{\mu}_Y]\right)^2 \qquad [17]$$

with $E[\hat{\mu}_Y] = \frac{1}{N}\sum_{i=1}^{N} g\,(x_i, \alpha)$ and Ω is a number of iterations, and empirical MSE (cf. Cassel et al. 1977, p. 94, Eq. 2.4) give that $\hat{\mu}_Y$ is design-unbiased (Ståhl et al., 2011)

$$MSE_{emp}(\hat{\mu}_Y) = E[(\hat{\mu}_Y - E[\hat{\mu}_Y])^2] + V_{emp}(\mu) + \left(B_{emp}[\hat{\mu}_Y]\right)^2$$
$$= \frac{1}{\Omega}\sum_{i=1}^{\Omega}\left(\hat{\mu}_{Y_i} - E[\hat{\mu}_Y]\right)^2 + \frac{1}{\Omega}\sum_{i=1}^{\Omega}(\mu_i - E[\mu])^2 + \left(\frac{1}{\Omega}\sum_{i=1}^{\Omega}\left(\hat{\mu}_{Y_i} - \mu_i\right)\right)^2 \qquad [18]$$

with $E[\mu] = \frac{1}{\Omega}\sum_{i=1}^{\Omega}\mu_i$ and $B_{emp}[\hat{\mu}_Y]$ an empirical bias due to the model (cf. Cassel et al. 1977, p. 94).

In Table S3-1 a summary of the case study examples we evaluated are given.

Table S3-1. Summary of the case study examples

Number of overpassing GEDI tracks	{2; 6}
Number of units used for the model training (“field plots”)	50
Coefficient of determination, R^2	{10^{-3}; 0.2; 0.5; 0.7; 0.9}
Spatial autocorrelation between two neighboring population elements (isotropic correlation)	{10^{-3}; 0.2; 0.5; 0.7; 0.9}
Number of Monte Carlo simulation iterations, Ω	1.5×10^4

In Table S3-2, results from the case study examples are given.

Table S3-2. The ratio between the hybrid estimator empirical SE and RMSE for different study cases.

Spatial autocorrelation	R^2	Ratio between empirical SE and RMSE, [%]	
		2 GEDI tracks	6 GEDI tracks
10^{-3}	10^{-3}	99.0	99.0
	0.2	96.7	99.0
	0.5	91.7	98.4
	0.7	88.8	97.6
	0.9	84.4	95.2
0.2	10^{-3}	97.5	97.6
	0.2	95.4	97.4
	0.5	91.1	97.0
	0.7	88.5	96.9
	0.9	84.3	94.5
0.5	10^{-3}	91.2	91.8
	0.2	90.1	91.7
	0.5	88.2	91.2
	0.7	86.5	91.6
	0.9	83.9	91.8
0.7	10^{-3}	81.6	82.1
	0.2	82.0	82.4
	0.5	82.2	82.2
	0.7	82.4	83.1

	0.9	82.1	85.0
0.9	10^{-3}	72.4	72.2
	0.2	72.2	72.7
	0.5	73.2	72.1
	0.7	73.5	73.3
	0.9	76.2	73.1

It can be seen that with increased spatial autocorrelation between neighboring population elements the difference between the estimator variance and MSE is increasing. However, the difference does not exceed the 27%. Table S3-3 presents components of the empirical MSE, i.e. the empirical variance of hybrid estimator of the population mean

$$V_{emp}(\hat{\mu}_Y) = \frac{1}{\Omega}\sum_{i=1}^{\Omega}\left(\hat{\mu}_{Y_i} - E[\hat{\mu}_Y]\right)^2 \quad [19]$$

the empirical variance of the population mean

$$V_{emp}(\mu) = \frac{1}{\Omega}\sum_{i=1}^{\Omega}(\mu_i - E[\mu])^2 \quad [20]$$

and the squared empirical bias due to the model

$$(B[\hat{\mu}_Y])^2 = \left(\frac{1}{\Omega}\sum_{i=1}^{\Omega}\left(\hat{\mu}_{Y_i} - \mu_i\right)\right)^2 \quad [21]$$

Table S3-3. The three components of the empirical MSE.

Spatial autocorrelation	R^2	2 GEDI tracks			6 GEDI tracks		
		$V_{emp}(\hat{\mu}_Y)$	$V_{emp}(\mu)$	$\left(B_{emp}[\hat{\mu}_Y]\right)^2$	$V_{emp}(\hat{\mu}_Y)$	$V_{emp}(\mu)$	$\left(B_{emp}[\hat{\mu}_Y]\right)^2$
10^{-3}	10^{-3}	2271.2	44.0	2.7	2266.2	44.0	5.1×10^{-1}
	0.2	10.7	1.7×10^{-1}	5.8×10^{-1}	9.8	1.7×10^{-1}	3.2×10^{-2}
	0.5	3.5	4.1×10^{-2}	6.1×10^{-1}	2.5	4.1×10^{-2}	4.1×10^{-2}
	0.7	2.2	2.3×10^{-3}	5.8×10^{-1}	1.2	2.3×10^{-3}	4.1×10^{-2}
	0.9	1.5	4.2×10^{-4}	6.0×10^{-1}	0.5	4.2×10^{-4}	3.9×10^{-2}
0.2	10^{-3}	2240.3	114.8	9.1×10^{-2}	2267.6	114.8	3.0×10^{-1}
	0.2	11.0	4.9×10^{-1}	6.1×10^{-1}	9.7	4.9×10^{-1}	3.1×10^{-2}
	0.5	3.5	1.2×10^{-1}	5.9×10^{-1}	2.6	1.2×10^{-1}	5.0×10^{-2}
	0.7	2.2	5.1×10^{-2}	5.7×10^{-1}	1.2	5.1×10^{-2}	3.2×10^{-2}
	0.9	1.5	1.1×10^{-2}	6.0×10^{-1}	0.5	1.1×10^{-2}	4.2×10^{-2}
0.5	10^{-3}	2663.0	510.5	9.2×10^{-1}	2724.2	510.5	3.2×10^{-1}
	0.2	11.5	2.1	5.7×10^{-1}	11.2	2.1	3.1×10^{-2}
	0.5	3.8	5.5×10^{-1}	5.7×10^{-1}	2.9	5.5×10^{-1}	2.9×10^{-2}
	0.7	2.4	2.2×10^{-1}	5.6×10^{-1}	1.4	2.2×10^{-1}	4.8×10^{-2}
	0.9	1.6	6.1×10^{-2}	6.0×10^{-1}	0.5	6.1×10^{-2}	4.1×10^{-2}
0.7	10^{-3}	3673.2	1809.8	6.2×10^{-2}	3749.9	1809.8	1.3×10^{-2}
	0.2	15.8	7.0	5.2×10^{-1}	15.0	7.0	4.7×10^{-2}
	0.5	4.9	1.8	5.7×10^{-1}	3.8	1.8	4.3×10^{-2}
	0.7	2.8	7.9×10^{-1}	5.5×10^{-1}	1.8	7.9×10^{-1}	4.1×10^{-2}
	0.9	1.7	2.0×10^{-1}	6.0×10^{-1}	0.6	2.0×10^{-1}	4.1×10^{-2}

0.9	10^{-3}	13654.7	11749.4	2.8	12795.9	11749.4	3.5×10^{-1}
	0.2	56.4	49.2	4.6×10^{-1}	55.2	49.2	2.9×10^{-2}
	0.5	14.6	12.8	5.6×10^{-1}	13.9	12.8	4.8×10^{-2}
	0.7	6.7	5.3	5.5×10^{-1}	6.1	5.3	4.1×10^{-2}
	0.9	2.8	1.4	5.9×10^{-1}	1.7	1.4	4.9×10^{-2}

Table S3-3 shows that the hybrid estimator squared bias due to the model is in the range 2.7 to 6.2×10^{-2} depending on the spatial autocorrelation, R^2 and number of GEDI tracks. Although, it can be considered as rather small, its portion in MSE is increasing with an increase of the coefficient of determination due to the decrease of the hybrid estimator empirical variance and of the population mean empirical variance. This explains the decrease of variance and MSE ratio over an increase of R^2 for population with low spatial correlation values in Table S3-2.

Since individual error is not addressed by the hybrid estimators proposed in this paper, it is of interest to search Table S3-2 for levels of spatial autocorrelation and R^2 typical of the GEDI mission to understand the impact of this omission. While it is outside of our scope to define a global distribution of these parameters for the mission, the data collected for this study offer some initial, if limited, insight.

Because our spatial data reflect lidar metrics instead of the actual biomass, we must use spatial autocorrelation of a metric such as RH 90 as a proxy for autocorrelation of AGBD, the target variable (and the variable represented in Table S3-2). We assessed the similarity of the spatial structure of a GEDI height metric (RH 90) with AGBD using one of the few publically available joint datasets combining simulated GEDI footprints over a large area with spatially correspondent field measurements of biomass (Bradford et al., 2014; Hancock et al., 2019). At this study site (the Robson Creek SuperSite, a tropical rainforest), the spatial structure of biomass and a corresponding height metric were similar at the GEDI spatial grain. The height metric, in fact, appeared more spatially autocorrelated than AGBD (Figure S3-1), although we emphasize that this is a result from a single site.

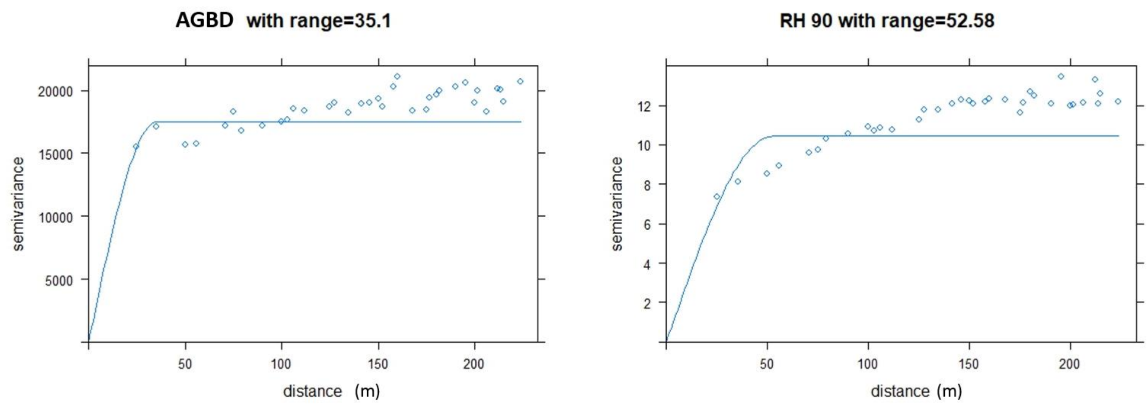

Figure S3-1. Semivariance as function of distance (spatial autocorrelation) of AGBD and simulated GEDI RH 90 at the Robson Creek SuperSite.

Table S3-4 shows the R^2 and spatial autocorrelation (Moran's i coefficient, averaged across 10 cells) for each of the six study sites. AGBD at our sites exhibited a strong relationship with lidar-based height metrics, which showed low spatial autocorrelation at the 25-m grain size of GEDI observations.

Table S3-4. Summary of model fit and average spatial autocorrelation for the 6 study areas

Site	R^2	Average Spatial Autocorrelation (Moran's i)
CO	0.76	0.14
ME	0.90	0.12
MN	0.88	0.15

OR	0.83	0.1
PANJ	0.61	0.15
SC	0.85	0.18

Combining results from Tables S3-2 and S3-4, the impact of omitting a term for individual error from the hybrid variance estimator proposed here is on the order of 1-15%, with the impact decreasing when a larger sample is available.

Appendix S4. Footprint Level Models

For each of the six regions, we fit the linear model $Y_{ri} = x_{ri}^2\alpha_r + \epsilon_r$, where, r indicates the region, x_{ri} is the simulated rh90 value (the height above the ground below which 90% of energy returned) for the ith plot in the rth region, and $\varepsilon_r \sim N(0, \sigma_r)$. The model parameter estimates (Table S4-1) were fit using OLS based on approximately fifty field inventory plots, which cover the full range of biomass levels within forested areas of each region (for details see Appendix S1).

Table S4-1 . Model parameter estimates for each of the size regions. Also included are the standard error of the parameter (StdErr), the p-value for the null hypothesis of alpha = 0, and the coefficient of determination (R^2).

Region	alpha	StdErr	p-value	R^2
co	0.4742	0.0394	1.20E-15	0.76
me	0.5017	0.0246	5.23E-25	0.90
mn	0.2966	0.0154	1.55E-24	0.88
or	0.3093	0.0202	2.61E-20	0.83
panj	0.2439	0.0282	2.31E-11	0.61
sc	0.2747	0.0167	1.26E-21	0.85

For each region, the residuals were plotted separately against rh90 and fitted AGBD values (Figures S4-1 to S4-6). Residual plots do exhibit some heteroscedasticity, and there are no negative residuals at zero because of the form of the linear model (above). However, heteroscedasticity was not considered strong enough to warrant special accommodation in the proposed approach to hybrid inference. Alternatives for this kind of accommodation would include either a transformation of the data or the use of a heteroscedasticity-consistent covariance matrix estimator (White, 1980).

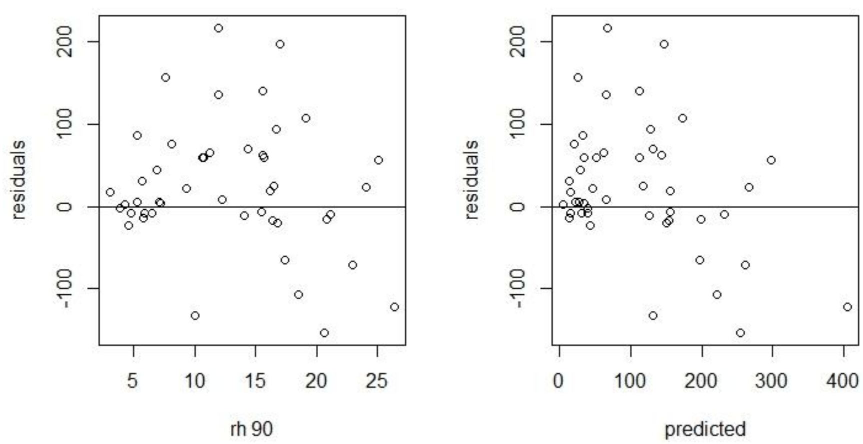

Figure S4-1. Plots of model residuals vs rh90 and model residuals vs predicted values for CO region.

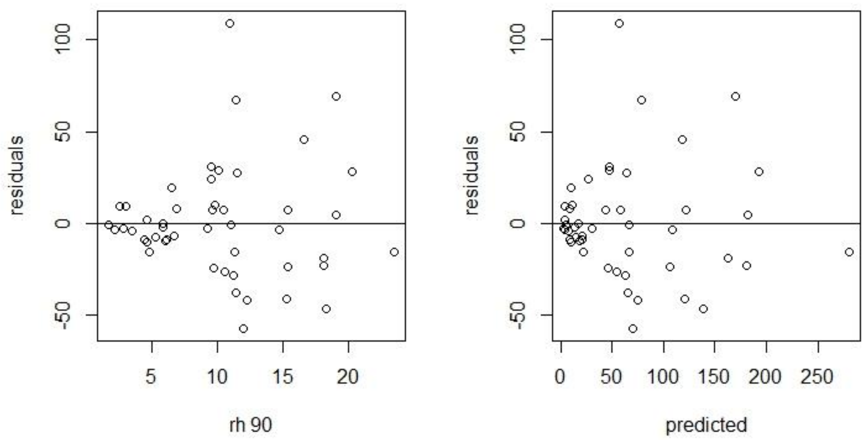

Figure S4-2. Plots of model residuals vs rh90 and model residuals vs predicted values for ME region.

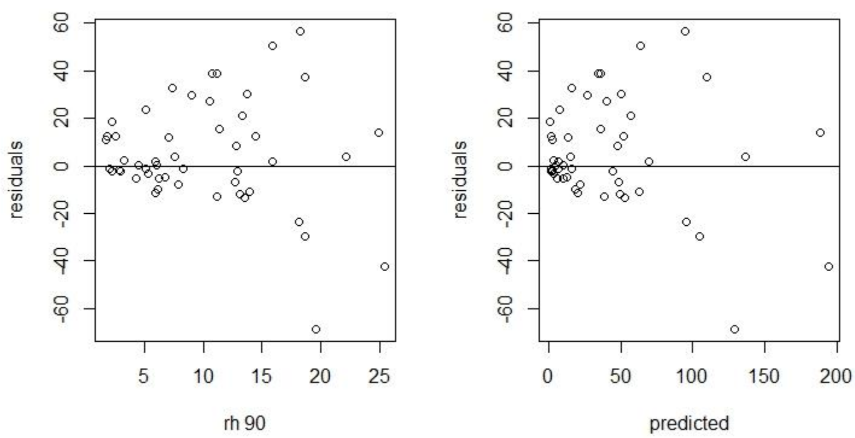

Figure S4-3. Plots of model residuals vs rh90 and model residuals vs predicted values for MN region.

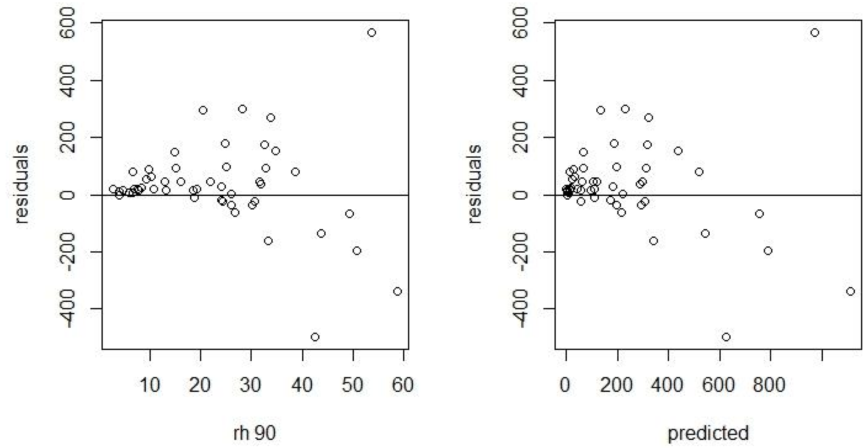

Figure S4-4. Plots of model residuals vs rh90 and model residuals vs predicted values for OR region.

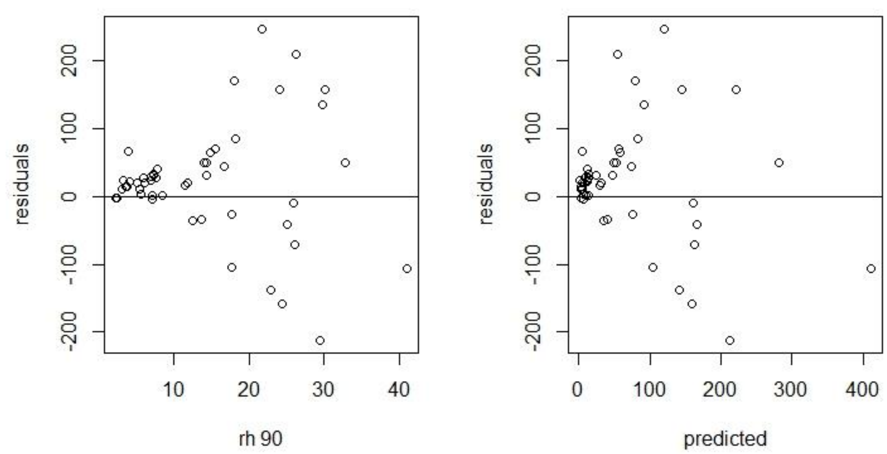

Figure S4-5. Plots of model residuals vs rh90 and model residuals vs predicted values for PANJ region.

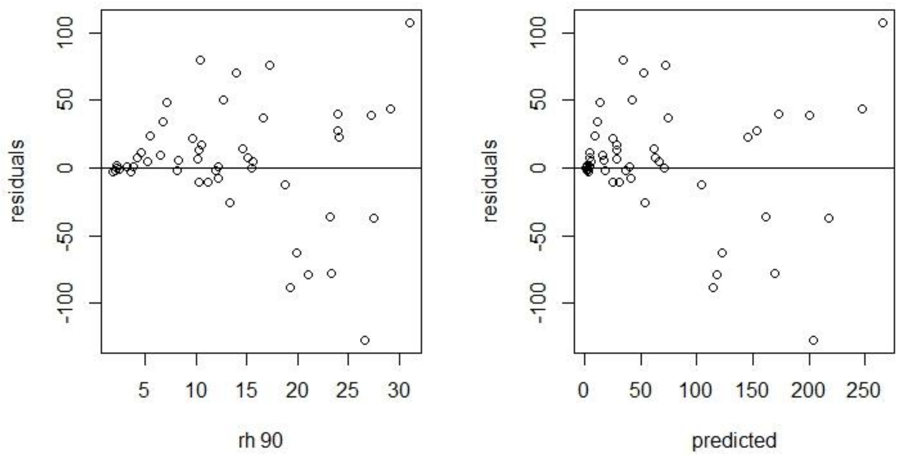

Figure S4-6. Plots of model residuals vs rh90 and model residuals vs predicted values for SC region.

Appendix References

Baccini, A., Walker, W., Carvalho, L., Farina, M., Sulla-Menashe, D., & Houghton, R.A. (2017). Tropical forests are a net carbon source based on aboveground measurements of gain and loss. Science, 358, 230-234

Blair, J.B., & Hofton, M.A. (1999). Modeling laser altimeter return waveforms over complex vegetation using high-resolution elevation data. Geophysical Research Letters, 26, 2509-2512

Blair, J.B., Rabine, D.L., & Hofton, M.A. (1999). The Laser Vegetation Imaging Sensor: a medium-altitude, digitisation-only, airborne laser altimeter for mapping vegetation and topography. ISPRS Journal of Photogrammetry and Remote Sensing, 54, 115-122

Bradford, M.G., Metcalfe, D.J. Ford, A.J. Liddell, M.J. & McKeown, A.T. (2014) Floristics, stand structure and above ground biomass of a 25 ha rainforest plot in the Wet Tropics of Australia. *Journal of Tropical Forest Science* 26(4):543-553.

Cassel, C. M., Sarndal, C. E., & Wretman, J. H. (1977). *Foundations of inference in survey sampling.* New York: Wiley.

Cochran, W.G. (1977). Sampling Techniques, 3rd Edition. New York: John Wiley

Davidson, F.M., & Xiaoli, S. (1988). Gaussian approximation versus nearly exact performance analysis of optical communication systems with PPM signaling and APD receivers. IEEE Transactions on Communications, 36, 1185-1192

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

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

Homer, C., Dewitz, J., Fry, J., Coan, M., Hossain, N., Larson, C., Herold, N., McKerrow, A., VanDriel, J.N., & Wickham, J. (2007). Completion of the 2001 National Land Cover Database for the Conterminous United States. Photogrammetric Engineering & Remote Sensing, 73, 337-341

McGaughey, R.J. (2015). Fusion Lidar Processing Software. http://forsys.cfr.washington.edu/fusion/fusion_overview.html

McRoberts, R.E., Næsset, E., Gobakken, T., Chirici, G., Condés, S., Hou, Z., Saarela, S., Chen, Q., Ståhl, G., & Walters, B.F. (2018). Assessing components of the model-based mean square error estimator for remote sensing assisted forest applications. *Canadian Journal of Forest Research, 48*, 642-649

Saarela, S., Holm, S., Grafström, A., Schnell, S., Næsset, E., Gregoire, T.G., Nelson, R.F., & Ståhl, G. (2016). Hierarchical model-based inference for forest inventory utilizing three sources of information. *Annals of Forest Science, 73*, 895-910

Särndal, C.-E., Swensson, B., & Wretman, J. (1992). Model Assisted Survey Sampling. In (p. 694). New York: Springer-Verlag Inc.

Ståhl, G., Holm, S., Gregoire, T.G., 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 Research, 41(1),* pp.96-107.

Ståhl, G., Saarela, S., Schnell, S., Holm, S., Breidenbach, J., Healey, S.P., Patterson, P.L., Magnussen, S., Næsset, E., McRoberts, R.E. & Gregoire, T.G. (2016). Use of models in large-area forest surveys: comparing model-assisted, model-based and hybrid estimation. *Forest Ecosystems, 3(1),* p.5.

White, Halbert. (1980). "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity." Econometrica, 48, 817-838.

Zolkos, S.G., Goetz, S.J., & Dubayah, R.O. (2013). A meta-analysis of terrestrial aboveground biomass estimation using lidar remote sensing. *Remote Sensing of Environment, 128*, 289-298