Environmetrics

WILEY

RESEARCH ARTICLE OPEN ACCESS

A Separable Bootstrap Variance Estimation Algorithm for Hierarchical Model-Based Inference of Forest Aboveground Biomass Using Data From NASA's GEDI and Landsat Missions

Svetlana Saarela[1] | Sean P. Healey[2] | Zhiqiang Yang[2] | Bjørn-Eirik Roald[1] | Paul L. Patterson[2] | Terje Gobakken[1] | Erik Næsset[1] | Zhengyang Hou[3] | Ronald E. McRoberts[4,5] | Göran Ståhl[6]

[1]Faculty of Environmental Sciences and Natural Resource Management, Norwegian University of Life Sciences, NMBU, Ås, Norway | [2]USDA Forest Service, Rocky Mountain Research Station, Ogden, Utah, USA | [3]The Key Laboratory for Silviculture and Conservation of Ministry of Education, Beijing Forestry University, Beijing, China | [4]Raspberry Ridge Analytics, Hugo, Minnesota, USA | [5]Department of Forest Resources, University of Minnesota, Saint Paul, Minnesota, USA | [6]Faculty of Forest Sciences, Swedish University of Agricultural Sciences, Umeå, Sweden

Correspondence: Svetlana Saarela (svetlana.saarela@nmbu.no)

Received: 6 November 2023 | **Revised:** 4 July 2024 | **Accepted:** 5 October 2024

Funding: This work was supported by NASA's GEDI Science Team (80HQTR21T0013).

Keywords: aboveground biomass | GEDI | Landsat | random forest | regression analysis

ABSTRACT

The hierarchical model-based (HMB) statistical method is currently applied in connection with NASA's Global Ecosystem Dynamics Investigation (GEDI) mission for assessing forest aboveground biomass (AGB) in areas lacking a sufficiently large number of GEDI footprints for employing hybrid inference. This study focuses on variance estimation using a bootstrap procedure that separates the computations into parts, thus considerably reducing the computational time required and making bootstrapping a viable option in this context. The procedure we propose uses a theoretical decomposition of the HMB variance into two parts. Through this decomposition, each variance component can be estimated separately and simultaneously. For demonstrating the proposed procedure, we applied a square-root-transformed ordinary least squares (OLS) model, and parametric bootstrapping, in the first modeling step of HMB. In the second step, we applied a random forest model and pairwise bootstrapping. Monte Carlo simulations showed that the proposed variance estimator is approximately unbiased. The study was performed on an artificial copula-generated population that mimics forest conditions in Oregon, USA, using a dataset comprising AGB, GEDI, and Landsat variables.

1 | Introduction

Assessment of ecosystem state and change is becoming increasingly important in the context of mitigating climate change and biodiversity loss. Information from such assessments is crucial for monitoring trends and selecting relevant mitigation measures. Advances in remote sensing technology offer new possibilities for providing this type of essential information. For instance, in a recent study, Dubayah et al. (2022) applied space LiDAR data from NASA's Global Ecosystem Dynamics Investigation (GEDI) mission to create a forest aboveground biomass (AGB) map for tropical and temperate regions with 1-km resolution.

Coupled with the new wealth of possibilities to assess ecosystem state and change, assessing the reliability of information is becoming increasingly important. For example, wall-to-wall maps may convey a false impression of perfect information (e.g., Kangas, Myllymäki, and Mehtätalo 2023), hiding large uncertainties in the estimated state for each map unit. Thus, it is imperative to develop and apply appropriate methods for assessing information reliability. To address this, Saarela et al. (2020) demonstrated a method for assessing the mean square error of AGB predictions at the level of individual map units; Dubayah et al. (2022) utilized a hybrid inferential methodology to complement the GEDI map by providing estimates of uncertainty for 1 km map units, following methods devised by Patterson et al. (2019).

In some cases, the GEDI map units do not contain a sufficient number of LiDAR footprints to apply hybrid inference. For those map units, an alternative is to apply hierarchical model-based (HMB) inference (Saarela et al. 2016, 2018), which uses a combination of GEDI and Landsat data. With HMB, inference proceeds in two modeling steps. In the first step, the sample of GEDI data available in the neighborhood of a target grid cell is used for predicting pseudo-field AGB data. These data are subsequently applied for training a Landsat AGB model, which is then applied to all map units within the 1 km grid cell (Saarela et al. 2018). Thus, the variance of a predictor in HMB inference needs to account for the effects of two sources of modeling uncertainty.

Originally, HMB inference was developed for generalized linear and nonlinear parametric models (Saarela et al. 2018, 2020). However, due to the development of flexible and efficient machine-learning prediction methods, such as random forest (Breiman 2001), parametric models may not always be the preferred choice. With nonparametric models, the analytical methods for uncertainty assessment cannot be applied. Consequently, resampling methods, particularly based on bootstrapping (Efron 1979), are the main alternative.

Rubin (1987) proposed a method for assessing variances within design-based inference when the observed sample values are missing and instead imputed through an imputation model, which may be nonparametric. With this approach, during each imputation round a new set of imputed values for each sampled unit with nonresponse is generated and the population parameter estimate and the corresponding variance estimate are computed treating the imputed values as if they were observed. The target population parameter is estimated as the average of the estimates across all imputation rounds. The corresponding variance comprises two components: (i) the empirical population variance of the population parameter estimates across the imputations, plus (ii) the average of the estimated variances of the population parameter estimator computed across the imputation rounds. This method was adopted by McRoberts et al. (2016) in the context of developing bootstrapping-based uncertainty analysis for hybrid inference.

Some studies have raised concerns over bias in Rubin's (1987) variance estimator. Särndal (1992) analytically demonstrated that Rubin's estimator is, in fact, biased, and proposed an alternative variance estimator, which does not require multiple imputations. Kim et al. (2006) provided a comprehensive overview of the literature and suggested conditions under which Rubin's variance estimator is biased. In the context of hybrid inference, Fortin, Manso, and Schneider (2018) introduced a variance estimator that utilizes parametric bootstrapping as the basis for multiple imputation and also corrects for the bias inherent in Rubin's estimator. Recently, Fortin et al. (2024) applied a similar methodology for HMB, and parametric bootstrapping to address uncertainty due to the first-step model.

Rubin's (1987) and Fortin et al.'s (2018 and 2024) variance estimators involve two sources of uncertainty and a combination of bootstrap and analytical methods. The estimation procedure is computationally demanding but remains feasible because it requires only one bootstrap loop. McRoberts et al. (2016) used Rubin's (1987) estimator for hybrid inference and bootstrap methods to assess uncertainty from both the estimated model and the random sampling of covariates. This required a nested bootstrapping loop, which significantly increased the computational demands. With HMB, nested bootstrapping to account for uncertainty from both modeling steps involved is extremely computationally demanding. Thus, to facilitate the use of bootstrapping in connection with assessing uncertainties from combinations of nonparametric models in HMB, alternatives to nested bootstrapping are required. The need for such developments is even larger if additional modeling steps are introduced, such as when applying three-phase HMB (Saarela et al. 2023; Varvia et al. 2024).

1.1 | Objectives

The objective of this study was to develop and demonstrate a bootstrapping algorithm for HMB inference, which splits the computations into several independent steps, thus facilitating faster computations. The method is based on a formal separation of the variance of the HMB predictor into components. Each of these components can be estimated separately and in parallel, using bootstrapping algorithms such as pairwise bootstrap (e.g., Esteban et al. 2019) or the parametric bootstrap (Ene et al. 2018). The performance of the method was evaluated through Monte Carlo simulation.

2 | Material and Methods

2.1 | Simulated Data

For the numerical part of the study, we generated a superpopulation (e.g., Ene et al. 2012) of 3 million independent observations using copulas (e.g., Nelsen 2006). We applied regular vine (R-vine) copula models implemented in R by Nagler et al. (2023). From the superpopulation, random samples of 100,000 units were repeatedly selected to constitute the target population during the Monte Carlo analysis (see further down). Each population unit comprised AGB derived from field measurement, rh50 and rh98 from GEDI (corresponding to heights at which a certain percentile of returned energy is reached relative to the ground), and digital numbers from Landsat's red, near infrared (NIR) and shortwave infrared 1 (swir1) bands. Based on empirical data from Oregon, USA, and copula modeling, joint distributions of the variables were obtained for each unit in the superpopulation, mimicking conditions in Oregon. Field forest AGB data were obtained from plot measurements of the US Forest Inventory and Analysis (FIA) survey (e.g., Menlove and Healey 2020). The field data used to parameterize the copula population comprised 2313 FIA plots, measured between 2015 and 2019. We used only those plots with a single forested condition (defined on the basis of variables such as stand age and forest type) because it would not be possible to spatially differentiate multiple conditions using the remotely sensed covariates upon which the copula population is based. We processed Landsat time series through continuous change detection and classification (CCDC) and synthetic images (Zhu et al. 2015) corresponding to July 1 of the survey year for which GEDI L2A Geolocated Elevation Height Metrics Product were available (GEDI02_A: Dubayah et al. 2021). Table 1 provides an overview of the variables in the superpopulation.

Figure 1 provides a visualization of the data. On the diagonal, the histogram of each variable is displayed. The lower corner of the off-diagonal squares shows the joint density plot of paired variables, while the upper corner of the off-diagonal squares provides the Pearson correlation coefficient (PCC) for the corresponding pairs of variables.

2.2 | HMB Inference

Let $U = \{1, \ldots, i, \ldots, N\}$ be a finite population. Each population unit has a multivariate variable associated with it, that is, the variables described in the previous section. The joint distribution of the variable defines the superpopulation model (Cassel et al. 1977). The multivariate variable comprises the target variable Y_i (the AGB) and the auxiliary variables $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ based on GEDI and Landsat data, respectively. We assume that three datasets S, Sa and U are independently selected from the superpopulation. The dataset S comprises AGB and GEDI data and is used to train the first-step model. The dataset Sa comprises GEDI and Landsat data and is used to train the second-step model. The target population U comprises wall-to-wall Landsat data, only. The objective pursued within HMB inference was to predict the target population mean, $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ for the given realization of U from the superpopulation, through HMB inference. (Here, the value y_i is a realization of Y_i for the target population,U.) We assume that our target population is large enough so that the target population mean is approximately equal to the superpopulation mean, which implies that the variance of the predictor would be approximately the same as the mean square error of the predictor (e.g., Ståhl et al. 2016; Saarela et al. 2022). We make this assumption because we know that the mean square error is the most relevant uncertainty measure to address. However, estimating the mean square errors is more demanding than estimating the variance because it involves assessing not only the variability of the predictor but also other components, such as the variability of the true value (e.g., McRoberts et al. 2018).

We begin by introducing two conditional models that follow from the superpopulation model. The first model is denoted 'Model I' and describes the relationship between the target variable Y_i and the auxiliary GEDI information $\boldsymbol{X}_i$:

$$\text{Model I} : Y_i = E_{\text{I}}[Y_i | \boldsymbol{X}_i = \boldsymbol{x}_i] + \epsilon_i, \tag{1}$$

where ϵ_i is the model error term, and $E_{\text{I}}[Y_i | \boldsymbol{X}_i = \boldsymbol{x}_i]$ is a conditional expectation of Y_i for a given realization of $\boldsymbol{X}_i = \boldsymbol{x}_i$. The second model is denoted 'Model II':

$$\text{Model II} : E_{\text{I}}[Y_i | \boldsymbol{X}_i] = E_{\text{II}}[E_{\text{I}}[Y_i | \boldsymbol{X}_i] | \boldsymbol{Z}_i = \boldsymbol{z}_i] + \upsilon_i, \tag{2}$$

where $E_{\text{I}}[Y_i | \boldsymbol{X}_i]$ is the conditional expectation of Y_i given $\boldsymbol{X}_i$, $E_{\text{II}}[E_{\text{I}}[Y_i | \boldsymbol{X}_i] | \boldsymbol{Z}_i = \boldsymbol{z}_i]$ is the conditional expectation of $E_{\text{I}}[Y_i | \boldsymbol{X}_i]$ for a given realization of Landsat data $\boldsymbol{Z}_i = \boldsymbol{z}_i$, and υ_i is the model error term. Since the expectation $E_{\text{I}}[Y_i | \boldsymbol{X}_i]$ is conditional on a variable, it is a variable in itself; properties of the response variable $E_{\text{I}}[Y_i | \boldsymbol{X}_i]$ under normality assumptions were presented and discussed in Saarela et al. (2023).

Within HMB inference the dataset S, comprising AGB and GEDI, data are used to train Model I. Subsequently, the estimated Model I is applied to the dataset Sa to predict the variable of interest; these predictions, that is, "pseudo-field" AGB, are denoted $\hat{\boldsymbol{y}}_{Sa}$, using explanatory variables $\boldsymbol{X}_{Sa}$ from GEDI. Then, the predicted variable $\hat{\boldsymbol{y}}_{Sa}$ is used to train Model II, utilizing explanatory variables $\boldsymbol{Z}_{Sa}$ from Landsat. The estimated Model II is then applied to the target population U to predict the variable of interest, the AGB, using the Landsat wall-to-wall auxiliary variables $\boldsymbol{Z}_U$.

The target population mean predictor under HMB inference is

$$\hat{\bar{y}}_{U_{\text{HMB}}} = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_i, \tag{3}$$

that is, the average of predicted values using estimated Model II based on wall-to-wall Landsat auxiliary data across the target area.

To derive the variance of the predictor $\hat{\bar{y}}_{U_{\text{HMB}}}$,we decompose its deviation from its expectation as:

$$\begin{aligned} \hat{\bar{y}}_{U_{\text{HMB}}} - E_{\text{I}}E_{\text{II}}\left[\hat{\bar{y}}_{U_{\text{HMB}}}\right] &= \hat{\bar{y}}_{U_{\text{HMB}}} - E_{\text{II}}\left[\hat{\bar{y}}_{U_{\text{HMB}}} | \text{I}\right] \\ &\quad + E_{\text{II}}\left[\hat{\bar{y}}_{U_{\text{HMB}}} | \text{I}\right] - E_{\text{I}}E_{\text{II}}\left[\hat{\bar{y}}_{U_{\text{HMB}}}\right] \\ &= \left(E_{\text{II}}\left[\hat{\bar{y}}_{U_{\text{HMB}}} | I\right] - E_{I}E_{II}\left[\hat{\bar{y}}_{U_{\text{HMB}}}\right]\right) + \left(\hat{\bar{y}}_{U_{\text{HMB}}} - E_{II}\left[\hat{\bar{y}}_{U_{\text{HMB}}} | I\right]\right) \\ &= \text{model error term due to Model I} \\ &\quad + \text{model error term due to Model II, conditional on Model I,} \end{aligned} \tag{4}$$

TABLE 1 | Overview of the variables in the superpopulation.

	AGB, [Mg·ha^{-1}]	rh50	rh98	Red	Nir	Swir1
Min	0.00	−1.87	2.52	0.01	0.10	0.03
Mean	295.55	12.72	28.95	0.03	0.23	0.10
Max	1906.44	44.21	68.72	0.13	0.48	0.28
SD	232.41	8.27	11.49	0.01	0.06	0.04

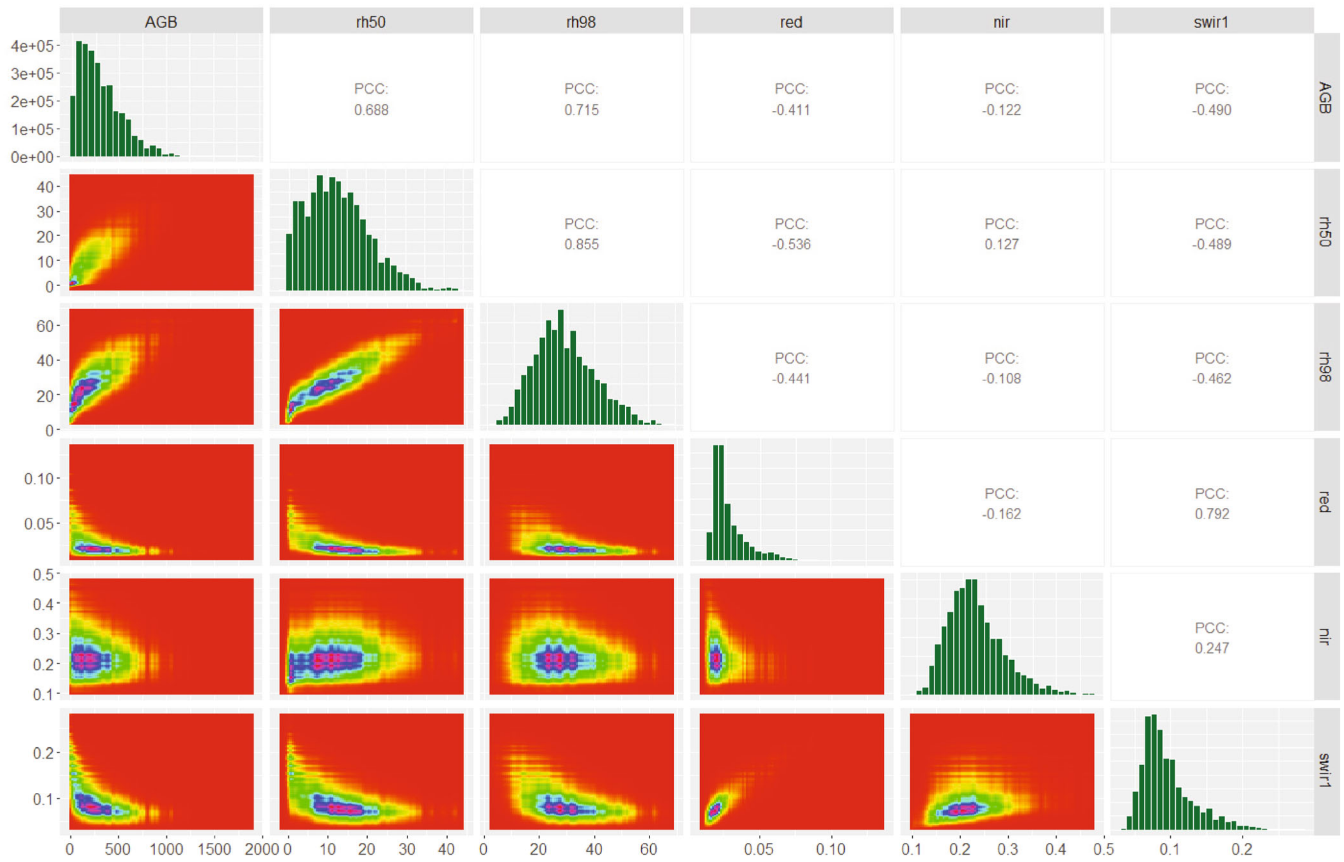

FIGURE 1 | A graphical visualization of the correlation matrix for the variables included in the superpopulation. The lower corner of the off-diagonal squares shows the joint density plot of paired variables, while the upper corner of the off-diagonal squares provides the Pearson correlation coefficient (PCC) for the corresponding pairs of variables. On the diagonal, the histogram of each variable is displayed.

where, $E_{\mathrm{I}}E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{\mathrm{HMB}}}\right]$ is the total expectation of the HMB predictor, and $E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{\mathrm{HMB}}}|\mathrm{I}\right]$ is its expectation due to Model II, conditional on Model I.

The variance of the HMB predictor is then

$$\begin{aligned} V\left(\hat{\bar{y}}_{U_{\mathrm{HMB}}}\right) &= E_{\mathrm{I}}E_{\mathrm{II}}\left[\left(\hat{\bar{y}}_{U_{\mathrm{HMB}}} - E_{\mathrm{I}}E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{\mathrm{HMB}}}\right]\right)^2\right] \\ &= E_{\mathrm{I}}E_{\mathrm{II}}\left[\left(\left(E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{\mathrm{HMB}}}|\mathrm{I}\right] - E_{\mathrm{I}}E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{\mathrm{HMB}}}\right]\right)\right.\right. \\ &\quad \left.\left. + \left(\hat{\bar{y}}_{U_{\mathrm{HMB}}} - E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{\mathrm{HMB}}}|\mathrm{I}\right]\right)\right)^2\right], \end{aligned} \tag{5}$$

which can be decomposed into three parts:

$$\begin{aligned} V\left(\hat{\bar{y}}_{U_{\mathrm{HMB}}}\right) &= E_{\mathrm{I}}E_{\mathrm{II}}\left[\left(E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{\mathrm{HMB}}}|\mathrm{I}\right] - E_{\mathrm{I}}E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{\mathrm{HMB}}}\right]\right)^2\right] \\ &\quad + E_{\mathrm{I}}E_{\mathrm{II}}\left[\left(\hat{\bar{y}}_{U_{\mathrm{HMB}}} - E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{\mathrm{HMB}}}|\mathrm{I}\right]\right)^2\right] \\ &\quad + 2E_{\mathrm{I}}E_{\mathrm{II}}\left[\left(E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{\mathrm{HMB}}}|I\right] - E_{\mathrm{I}}E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{\mathrm{HMB}}}\right]\right)\right. \\ &\quad \left.\left(\hat{\bar{y}}_{U_{\mathrm{HMB}}} - E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{\mathrm{HMB}}}|\mathrm{I}\right]\right)\right] \end{aligned} \tag{6}$$

The first component on the right side of (6) is the propagated uncertainty due to the estimation of Model I through Model II, that is,

$$V_{\mathrm{I}}\left(E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{\mathrm{HMB}}}|\mathrm{I}\right]\right). \tag{7}$$

The second component is the expectation of the variance due to the estimation of Model II conditionally on Model I, that is,

$$E_{\mathrm{I}}\left[V_{\mathrm{II}}\left(\hat{\bar{y}}_{U_{\mathrm{HMB}}}|\mathrm{I}\right)\right]. \tag{8}$$

The third component of the right side of Equation (6), involves the covariance:

$$\begin{aligned} &E_{\mathrm{I}}E_{\mathrm{II}}\left[\left(E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{HMB}}|\mathrm{I}\right] - E_{\mathrm{I}}E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{HMB}}\right]\right)\left(\hat{\bar{y}}_{U_{HMB}} - E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{HMB}}|\mathrm{I}\right]\right)\right] = \\ &\quad = E_{\mathrm{I}}E_{\mathrm{II}}\left[E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{HMB}}|\mathrm{I}\right]\hat{\bar{y}}_{U_{HMB}} - E_{\mathrm{I}}E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{HMB}}\right]\hat{\bar{y}}_{U_{HMB}}\right. \\ &\quad \left. - E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{HMB}}|\mathrm{I}\right]^2 + E_{\mathrm{I}}E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{HMB}}\right]E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{HMB}}|\mathrm{I}\right]\right]. \end{aligned} \tag{9a}$$

The first term on the right side of Equation (9a) can be elaborated as

$$\begin{aligned} E_{\mathrm{I}}E_{\mathrm{II}}\left[E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{\mathrm{HMB}}}|\mathrm{I}\right]\hat{\bar{y}}_{U_{\mathrm{HMB}}}\right] &= E_{\mathrm{I}}\left[E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{\mathrm{HMB}}}|\mathrm{I}\right]E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{\mathrm{HMB}}}|\mathrm{I}\right]\right] \\ &= E_{\mathrm{I}}\left[E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{\mathrm{HMB}}}|\mathrm{I}\right]^2\right]. \end{aligned}$$

The second and fourth terms in Equation (9a) are simplified to $E_{\mathrm{I}}E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{\mathrm{HMB}}}\right]^2$, and the third term is $E_{\mathrm{I}}\left[E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{\mathrm{HMB}}}|\mathrm{I}\right]^2\right]$. Thus, the covariance Equation (9a) equals zero, since the first and third terms, and the second and the fourth terms cancel each other, that is,

$$E_{\mathrm{I}}\left[E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{\mathrm{HMB}}}|\mathrm{I}\right]^2\right] - E_{\mathrm{I}}E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{\mathrm{HMB}}}\right]^2 - E_{\mathrm{I}}\left[E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{\mathrm{HMB}}}|\mathrm{I}\right]^2\right] + E_{\mathrm{I}}E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{\mathrm{HMB}}}\right]^2 = 0. \tag{9b}$$

Thus, the variance of the HMB predictor can be presented as a decomposition based on the law of total variance (Saarela et al. 2022, 2023) as

$$V\left(\hat{\bar{y}}_{U_{\mathrm{HMB}}}\right) = V_{\mathrm{I}}\left(E_{\mathrm{II}}\left[\hat{\bar{y}}_{U_{\mathrm{HMB}}}|\mathrm{I}\right]\right) + E_{\mathrm{I}}\left[V_{\mathrm{II}}\left(\hat{\bar{y}}_{U_{\mathrm{HMB}}}|\mathrm{I}\right)\right]. \tag{10}$$

This decomposition is the basis for the separable bootstrap algorithm, presented further down.

2.3 | HMB With Parametric Models

If Model I and Model II are estimated using ordinary least squares (OLS) regression analysis with a linear model, the HMB predictor is

$$\hat{\bar{y}}_{U_{\mathrm{HMB_{OLS}}}} = \bar{\boldsymbol{z}}_U\hat{\boldsymbol{\alpha}}_{\mathrm{OLS}}, \tag{11}$$

and then the variance has the form

$$V_{\mathrm{OLS}}\left(\hat{\bar{y}}_{U_{\mathrm{HMB_{OLS}}}}\right) = \bar{\boldsymbol{z}}_U\left(\boldsymbol{Z}_{Sa}^T\boldsymbol{Z}_{Sa}\right)^{-1}\boldsymbol{Z}_{Sa}^T\boldsymbol{X}_{Sa}\mathrm{Cov}\left(\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right) \boldsymbol{X}_{Sa}^T\boldsymbol{Z}_{Sa}\left(\boldsymbol{Z}_{Sa}^T\boldsymbol{Z}_{Sa}\right)^{-1}\bar{\boldsymbol{z}}_U^T + \bar{\boldsymbol{z}}_U\mathrm{Cov}\left(\hat{\boldsymbol{\alpha}}_{\mathrm{OLS}}|\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}\right)\bar{\boldsymbol{z}}_U^T, \tag{12}$$

where $\bar{\boldsymbol{z}}_U$ is a vector of the means of the Landsat variable values over the target population U (with unit terms for the intercept), $\boldsymbol{Z}_{Sa}$ is the matrix of Landsat variables (with unit terms for the intercept) for the dataset Sa, and $\boldsymbol{X}_{Sa}$ is the corresponding matrix of GEDI explanatory variables; $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ and $\hat{\boldsymbol{\alpha}}_{\mathrm{OLS}}$ are estimated model parameters for Model I and Model II respectively (Saarela et al. 2016). The first component on the right-hand side of (12) is the propagated uncertainty due to the estimated Model I. The second component is the variance due to the estimated Model II, conditionally on Model I. An estimator of the variance is obtained by replacing the covariances of the estimated model parameters with their corresponding estimators. HMB inference with generalized least squares and generalized nonlinear least squares models was presented in Saarela et al. (2018), (2020), respectively.

2.4 | HMB With Nonparametric Models

In case nonparametric models are applied in either or both modeling steps, the variance formula presented above cannot be applied. Fortin et al. (2024) presented a variance estimator involving bootstrapping for cases where nonparametric (or complicated parametric) modeling was applied in the first modeling step. (However, with nonparametric models in both steps, the approach by Fortin et al. (ibid.) cannot be applied.) The multiple imputation framework proposed by Rubin (1987) was used as the basis for Fortin's variance estimator. The uncertainty due to Model I was assessed through parametric bootstrapping. Following Fortin et al. (2024), the population mean predictor is

$$\hat{\bar{y}}_{U_{\mathrm{HMB}_F}} = \frac{1}{B}\sum_{i=1}^{B}\hat{\bar{y}}_{U_{\mathrm{HMB}_i}} = \frac{1}{B}\sum_{i=1}^{B}\bar{\boldsymbol{z}}_U\hat{\boldsymbol{\alpha}}_i, \tag{13}$$

where B is the number of bootstrap iterations, and $\hat{\boldsymbol{\alpha}}_i$ is a vector of estimated model parameters for the ith bootstrap iteration. The total variance of the predictor $\hat{\bar{y}}_{U_{\mathrm{HMB}_F}}$ comprises two components: (i) the empirical population variance of the target population estimates across the bootstrap rounds $\left(\bar{\boldsymbol{z}}_U\frac{\sum_{i=1}^{B}\left(\hat{\boldsymbol{\alpha}}_i - \frac{1}{B}\sum_{i=1}^{B}\hat{\boldsymbol{\alpha}}_i\right)^2}{B}\bar{\boldsymbol{z}}_U^T\right)$, and (ii) a component related to the analytically-based variance estimator for the second-step model, that is, $\left(2\bar{\boldsymbol{z}}_U\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\alpha}}_{\mathrm{bts}})\bar{\boldsymbol{z}}_U^T - \bar{\boldsymbol{z}}_U\frac{\sum_{i=1}^{B}\widehat{\mathrm{Cov}}_i(\hat{\boldsymbol{\alpha}}_i)}{B}\bar{\boldsymbol{z}}_U^T\right)$. Thus, the variance is estimated as

$$\hat{V}_F\left(\hat{\bar{y}}_{U_{\mathrm{HMB}_F}}\right) = \bar{\boldsymbol{z}}_U\frac{\sum_{i=1}^{B}\left(\hat{\boldsymbol{\alpha}}_i - \frac{1}{B}\sum_{i=1}^{B}\hat{\boldsymbol{\alpha}}_i\right)^2}{B}\bar{\boldsymbol{z}}_U^T + 2\bar{\boldsymbol{z}}_U\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\alpha}}_{\mathrm{bts}})\bar{\boldsymbol{z}}_U^T - \bar{\boldsymbol{z}}_U\frac{\sum_{i=1}^{B}\widehat{\mathrm{Cov}}_i(\hat{\boldsymbol{\alpha}}_i)}{B}\bar{\boldsymbol{z}}_U^T \tag{14}$$

In (14), $\hat{\boldsymbol{\alpha}}_{\mathrm{bts}}$ is estimated using the average of $\hat{y}_{Sa}$ predictions across the bootstrap iterations. The first and the third components of the variance estimator Equation (14) (Fortin et al. 2024, Equation 9) coincide with Rubin's variance estimator, if added. By introducing the difference term $\left(2\bar{\boldsymbol{z}}_U\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\alpha}}_{\mathrm{bts}})\bar{\boldsymbol{z}}_U^T - \bar{\boldsymbol{z}}_U\frac{\sum_{i=1}^{B}\widehat{\mathrm{Cov}}_i(\hat{\boldsymbol{\alpha}}_i)}{B}\bar{\boldsymbol{z}}_U^T\right)$, the bias of Rubin's (1987) variance estimator is removed (Fortin et al. 2024).

The first component of (14) is the propagated uncertainty due to the estimation of Model I, since the bootstrapping procedure is conducted for Model I, and Model II is re-estimated within each bootstrap iteration. No bootstrapping procedure is employed for Model II at this stage.

The component $\left(2\bar{\boldsymbol{z}}_U\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\alpha}}_{\mathrm{bts}})\bar{\boldsymbol{z}}_U^T - \bar{\boldsymbol{z}}_U\frac{\sum_{i=1}^{B}\widehat{\mathrm{Cov}}_i(\hat{\boldsymbol{\alpha}}_i)}{B}\bar{\boldsymbol{z}}_U^T\right)$ is the estimated uncertainty due to Model II. The estimation is based on an analytical expression of the model-based variance. However, if the second model is nonparametric a bootstrapping approach could be applied instead. If so, a nested bootstrap procedure would be necessary. In other words, during each outer bootstrap loop for Model I, an inner bootstrap loop should be executed for Model II. Such an approach would significantly increase the computational load (see Section 2.7).

In the following, we propose a separable bootstrapping solution for situations where resampling is applied to assess the uncertainty in both modeling steps. This solution does not require nested resampling.

2.5 | A Separable Bootstrap HMB Variance Estimation Algorithm (HMB.Bts)

In the following, we propose a bootstrapping variance estimation framework for HMB inference that utilizes nonparametric models in both modeling steps. The proposed framework does not require nested bootstrapping; instead, the computation process can be performed in single, and potentially parallel, bootstrap loops. In this framework, the population mean predictor is

$$\hat{\hat{y}}_{U_{\mathrm{HMB.bts}}} = \frac{\sum_{i=1}^{B} \hat{\hat{y}}_{U_{\mathrm{HMB.bts}_{\mathrm{I}_i}}} + \sum_{i=1}^{B} \hat{\hat{y}}_{U_{\mathrm{HMB.bts}_{\mathrm{II}_i}}}}{2B} \tag{15}$$

and its variance is

$$\hat{V}_{\mathrm{HMB.bts}}\left(\hat{\hat{y}}_{U_{\mathrm{HMB.bts}}}\right) = \frac{\sum_{i=1}^{B}\left(\hat{\hat{y}}_{U_{\mathrm{HMB.bts}_{\mathrm{I}_i}}} - \frac{1}{B}\sum_{i=1}^{B}\hat{\hat{y}}_{U_{\mathrm{HMB.bts}_{\mathrm{I}_i}}}\right)^2}{B-1} + \frac{\sum_{i=1}^{B}\left(\hat{\hat{y}}_{U_{\mathrm{HMB.bts}_{\mathrm{II}_i}}} - \frac{1}{B}\sum_{i=1}^{B}\hat{\hat{y}}_{U_{\mathrm{HMB.bts}_{\mathrm{II}_i}}}\right)^2}{B-1} \tag{16}$$

In (15), $\hat{\hat{y}}_{U_{\mathrm{HMB.bts}_{\mathrm{I}_i}}}$ represents the predicted population mean from the i^{th} bootstrap iteration when Model I is bootstrapped. As a consequence of bootstrapping the first model, the second model is re-fitted and then reapplied to the target population U, resulting in a new predicted value for the target population mean. It should be noted that when the first model is bootstrapped, the second one changes as a consequence, but not because it is being bootstrapped. At each bootstrap iteration, another population mean prediction, $\hat{\hat{y}}_{U_{\mathrm{HMB.bts}_{\mathrm{II}_i}}}$, is obtained by bootstrapping the second model in parallel to the first bootstrapping procedure. As a result of the proposed bootstrap procedures, two second-step models are trained simultaneously and independently during each (parallel) bootstrap iteration. One propagates the errors due to bootstrapping the first model, and the other is due to bootstrapping the second model. Figure 2 provides a graphical overview of the HMB.bts bootstrap procedure.

Table 2 describes the bootstrapping procedure proposed by Fortin et al. (2024) and the HMB.bts bootstrapping proposed in this study.

2.6 | Motivating the HMB.Bts Procedure

From Equation (10) we know that a generic breakdown of the total variance due to uncertainty from two modeling steps can be expressed as $V\left(\hat{\hat{y}}_{U_{\mathrm{HMB}}}\right) = V_{\mathrm{I}}\left(E_{\mathrm{II}}\left[\hat{\hat{y}}_{U_{\mathrm{HMB}}}|\mathrm{I}\right]\right) + E_{\mathrm{I}}\left[V_{\mathrm{II}}\left(\hat{\hat{y}}_{U_{\mathrm{HMB}}}|\mathrm{I}\right)\right]$.To motivate HMB.bts, we will now show how the proposed bootstrap procedure estimates each of the two components.

The first component is the variance due to the first modeling step when applying expected values from the second model to predict the population means, conditional on the outcome of Model I, that is, the estimated parameter values in Model I for a given bootstrap iteration. The estimated random forest model (Model II) approximately provides the expected values for a given

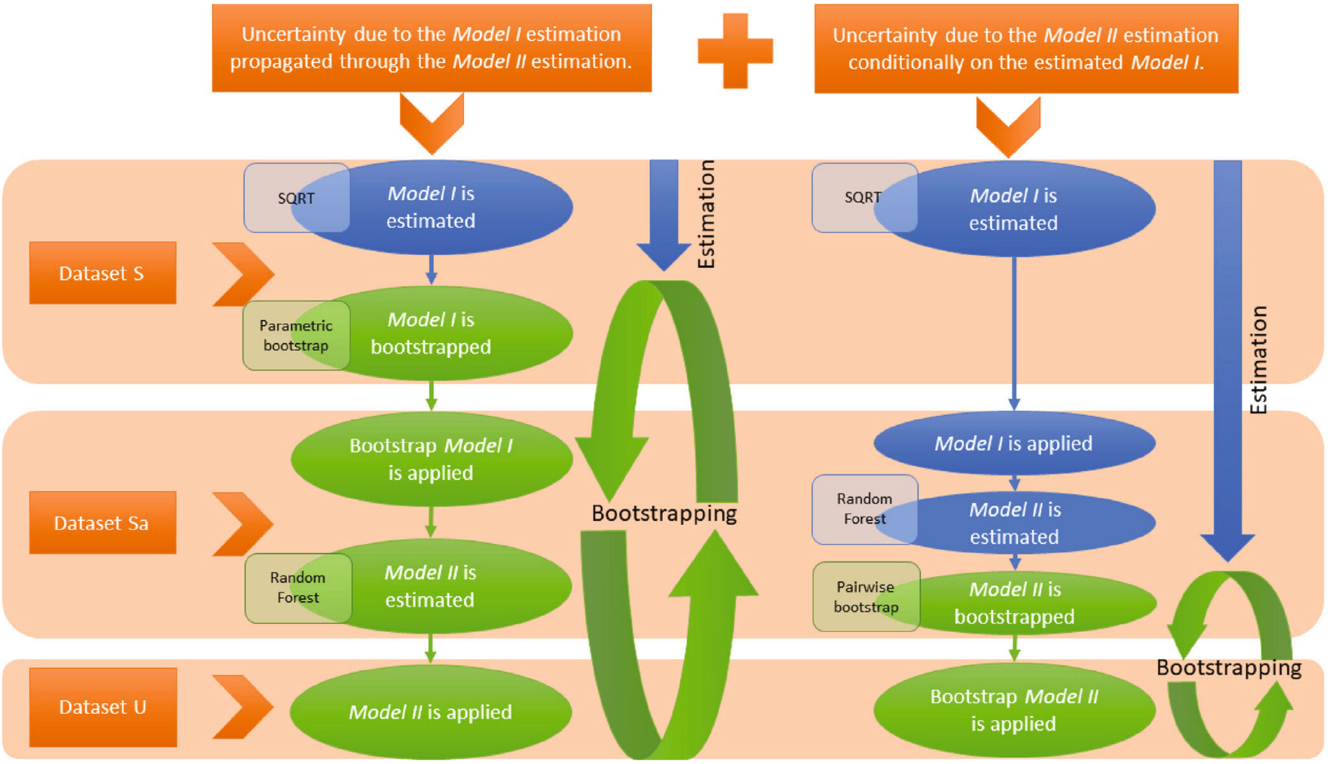
Uncertainty due to the *Model I* estimation propagated through the *Model II* estimation.
Uncertainty due to the *Model II* estimation conditionally on the estimated *Model I*.
Dataset S
Dataset Sa
Dataset U
SQRT
Model I is estimated
Parametric bootstrap
Model I is bootstrapped
Bootstrap *Model I* is applied
Random Forest
Model II is estimated
Model II is applied
Estimation
Bootstrapping
SQRT
Model I is estimated
Model I is applied
Random Forest
Model II is estimated
Pairwise bootstrap
Model II is bootstrapped
Bootstrap *Model II* is applied
Estimation
Bootstrapping

FIGURE 2 | Graphical overview of the HMB.bts estimation procedure is provided. Estimation steps color-coded in green are part of the bootstrapping procedure, while estimation steps color-coded in blue are estimated only once. Dataset S is used to train Model I, the dataset Sa is used to train Model II and the dataset U is the target population (see Section 2.2).

TABLE 2 | Comparison of Fortin's (2024) bootstrap procedure with the procedure proposed in this study.

Steps	Fortin et al. (2024)	HMB.bts
Step (i)	Model I is fitted using the information on AGB (y_S) and GEDI data ($\boldsymbol{X}_S$) from the dataset S.	Same as in Fortin et al.
Step (ii)	A set of model parameters is generated from a multivariate normal distribution using estimated model parameters in Model I, these model parameters are applied for predicting AGB ($\widehat{y}_{Sa}$) over the dataset Sa.	Same as in Fortin et al.
Step (iii)	The AGB predictions $\widehat{y}_{Sa}$ are used for training Model II using Landsat data ($\boldsymbol{Z}_{Sa}$). Model II is parametric.	Same as in Fortin et al., but Model II is nonparametric (which means that Fortin's method cannot be applied).
Step (iv)	Model II is applied across the target area U, utilizing available wall-to-wall Landsat data ($\boldsymbol{Z}_U$) to predict the population mean value. The predicted mean values over the bootstrap iterations are used to estimate the first variance component of the variance estimator Equation (14): $\bar{\boldsymbol{z}}_U \frac{\sum_{i=1}^{B}\left(\widehat{\boldsymbol{\alpha}}_i - \frac{1}{B}\sum_{i=1}^{B}\widehat{\boldsymbol{\alpha}}_i\right)^2}{B} \bar{\boldsymbol{z}}_U^T$	Model II is applied across the target area U, utilizing available wall-to-wall Landsat data ($\boldsymbol{Z}_U$) to predict the population mean value. The predicted mean values over the bootstrap iterations are used to estimate the first variance component of the variance estimator Equation (16): $\frac{\sum_{i=1}^{B}\left(\hat{\bar{y}}_{U_{\mathrm{HMB.bts}_{\mathrm{I}i}}} - \frac{1}{B}\sum_{i=1}^{B}\hat{\bar{y}}_{U_{\mathrm{HMB.bts}_{\mathrm{I}i}}}\right)^2}{B-1}$
Step (v)	An analytical expression based on Model II is employed to estimate the variance associated with predicting the population mean, that is, $\bar{\boldsymbol{z}}_U \widehat{\mathrm{Cov}}_i(\widehat{\boldsymbol{\alpha}}_i) \bar{\boldsymbol{z}}_U^T$ for all bootstrap iterations. Throughout this process, the AGB pseudo-field data ($\widehat{y}_{Sa}$**)** are treated as observed data, that is, the variance is conditional on the predictions based on Model I. The average of the estimated variances across the bootstrap iterations is then estimated as $\bar{\boldsymbol{z}}_U \frac{\sum_{i=1}^{B}\widehat{Cov}_i(\widehat{\boldsymbol{\alpha}}_i)}{B} \bar{\boldsymbol{z}}_U^T$	The pseudo-field AGB field data predictions ($\widehat{y}_{Sa}$) obtained by applying Model I across the dataset Sa, are used to train Model II using random forest. Model II is then bootstrapped using pairwise bootstrapping.
Step (vi)	After the bootstrap procedure is completed, the average of the $\widehat{y}_{Sa}$ predictions across the bootstrap iterations is used to estimate another set of model parameters corresponding Model II, that is, the $\widehat{\boldsymbol{\alpha}}_{\mathrm{bts}}$ values. The model parameters are then used to estimate the variance of prediction following the analytical variance expression: $\bar{\boldsymbol{z}}_U \widehat{Cov}(\widehat{\boldsymbol{\alpha}}_{\mathrm{bts}}) \bar{\boldsymbol{z}}_U^T$ To obtain the second component on the right side of the Equation (14). Thus, the uncertainty due to the estimated Model II is $\left(2\bar{\boldsymbol{z}}_U \widehat{\mathrm{Cov}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{bts}}) \bar{\boldsymbol{z}}_U^T - \bar{\boldsymbol{z}}_U \frac{\sum_{i=1}^{B}\widehat{\mathrm{Cov}}_i(\widehat{\boldsymbol{\alpha}}_i)}{B} \bar{\boldsymbol{z}}_U^T\right)$	Model II is applied across the target population U to predict the population mean. The predicted population mean values over the bootstrap iterations are used to estimate the second component of the variance estimator, Equation (16), that is, the uncertainty due to the estimation of Model II as: $\frac{\sum_{i=1}^{B}\left(\hat{\bar{y}}_{U_{\mathrm{HMB.bts}_{\mathrm{II}_i}}} - \frac{1}{B}\sum_{i=1}^{B}\hat{\bar{y}}_{U_{\mathrm{HMB.bts}_{\mathrm{II}i}}}\right)^2}{B-1}$

set of explanatory Landsat variables, conditional on the outcome of Model I at each bootstrap iteration. Thus, by applying the proposed algorithm, where Model I is estimated differently in each iteration, and the consequential uncertainty is propagated through Model II, the first variance component of (10) is estimated. The estimation procedure is at least approximately unbiased, and it coincides with the first term of the estimation procedure proposed by Fortin et al. (2024).

The second variance component is the expectation of the variance among population mean predictions, for different outcomes of Model II, conditional on the outcome of Model I. That is, for a given estimate of Model I, bootstrap iterations are run for Model II, to assess the variability due to estimating the second model conditional on the first model. Then, the expectation of these variances is computed across different realizations of Model I. In our standard procedure for HMB.bts, however, we argue that the main variability in this case is due to variability following different realizations of Model II, in which case we avoid nested bootstrapping by selecting only a single outcome of Model I rather than computing the mean value from several outcomes of Model I (which would require nested bootstrapping). A motivation for this is that, in normal applications of HMB, the first model should be accurate enough to provide pseudo-field variable values, in which case variability due to the outcome of Model I should be relatively small. Further, although only a single outcome of Model I is used, the proposed procedure remains approximately unbiased. The unbiasedness was tested using a Monte Carlo sampling-based simulation, described in Section 2.8. In this simulation, the HMB.bts framework was applied to different samples, independently selected from the same superpopulation in each iteration. This approach ensures that the second variance component, estimated through bootstrapping Model II based on a single outcome of Model I, is assessed on different outcomes of Model I in each Monte Carlo iteration.

2.7 | Time Complexity Comparison of Nested Bootstrap With Parallel Separable Bootstrap Procedures

In this section, we perform an analytical comparison of nested bootstrap with the proposed separable bootstrap procedure, which is based on two bootstrap loops running in parallel, by means of time complexity, also called "growth rate" in some literature (e.g., Cormen 2009, p. 23–29). In the analysis, we assume that the average time required to train Model I (t_1) and the average time for training Model II (t_2) are the same for both approaches over a large number of experiments. Assuming an average case with an average number of threads and an average level of parallelization, we obtain the following expression for the overall time required to perform the nested bootstrap procedure:

$$t_{\text{nested}} = \left(\frac{B}{C_1}\right)t_1 + \left(\frac{B}{C_1}\right)\left(\frac{B}{C_2}\right)t_2 = \left(\frac{B}{C_1}\right)t_1 + \left(\frac{B^2}{C_1C_2}\right)t_2$$

where, C_1 and C_2 are numbers of parallel threads used for training Models I and II respectively, and B is the number of bootstrap iterations.

Thus, the amount of computer time needed to perform the nested bootstrap procedure is order of B squared, that is,

$$t_{\text{nested}}(B) = O\left(B^2\right). \tag{17}$$

In the case of the parallel bootstrapping procedure, the time required for running the algorithm can be expressed as:

$$t_{\text{parallel}} = \left(\frac{B}{C_1}\right)t_1 + \left(\frac{2B}{C_1C_2}\right)t_2$$

and thus

$$t_{\text{parallel}}(B) = O(B). \tag{18}$$

Equations (17) and (18) show that with nested bootstrap, the amount of computer time increases quadratically with the number of bootstrap iterations. However, with the proposed separable algorithm based on parallel bootstrap loops, the amount of computer time increases linearly with the number of bootstrap iterations.

In the section below, we describe how we validated the HMB.bts estimation framework through Monte Carlo simulation, using an example related to NASA's GEDI mission.

2.8 | Monte Carlo Simulation

We applied Monte Carlo simulation to evaluate the performance of the separable bootstrap variance estimation algorithm. In each Monte Carlo iteration, the datasets U, Sa, and S were selected from the superpopulation described in section 2.2 using simple random sampling without replacement. These datasets were chosen independently and simultaneously, representing one realization of each dataset from the same superpopulation. The design-independent selection of datasets from the superpopulation mimics the data structure of the GEDI mission, where the dataset S used to train Model I is an independently selected set of data that may not be a subsample of neither the target population U nor the GEDI sample Sa. The same principle is applied regarding the design-independence of the dataset Sa, that is, the set of GEDI footprints may not be a subsample of the target population U, since the HMB framework is applied in areas where there are no GEDI footprints or the subsample is sparse, and thus GEDI footprints from outside of the target population are used to train Model II (Saarela et al. 2018).

A square-root-transformed OLS (SQRT) model was employed as Model I to mimic GEDI's L4A models (Duncanson et al. 2022). Then, the model was applied to the GEDI sample of footprints, that is, the dataset Sa to predict AGB using GEDI data; the obtained predictions were then used to train Model II by applying the random forest nonparametric model (Breiman 2001), based on Landsat data. Model II was then applied across the target population U. The target population mean and the corresponding variance were estimated as follows Equations (15) and (16) according to the procedure described in Table 2 employing the parametric bootstrap to assess uncertainty due to estimating Model I, and the pairwise bootstrap to assess the uncertainty due to estimating Model II. After completing the Monte Carlo iterations, we computed the average of the population mean predictions as follows:

$$\hat{\bar{y}}_{U\,\text{MC}} = \frac{\sum_{i=1}^{\text{MC}} \hat{\bar{y}}_{U_{\text{HMB.bts}_i}}}{\text{MC}}, \tag{19}$$

where MC is the number of Monte Carlo iterations and $\hat{\bar{y}}_{U_{\text{HMB.bts}_i}}$ is obtain by applying Equation (15) at i^{th} Monte Carlo iteration. The empirical variance was computed as

$$V_{\text{MC}}\left(\hat{\bar{y}}_{U_{\text{HMB.bts}}}\right) = \frac{\sum_{i=1}^{\text{MC}}\left(\hat{\bar{y}}_{U_{\text{HMB.bts}\,i}} - \hat{\bar{y}}_{U\,\text{MC}}\right)^2}{\text{MC}-1}. \tag{20}$$

The variance $V_{\text{MC}}\left(\hat{\bar{y}}_{U_{\text{HMB.bts}}}\right)$ was used to validate the HMB.bts variance estimator by taking the averages of the variance estimates over the Monte Carlo simulations

$$\hat{V}_{\text{MC}}\left(\hat{\bar{y}}_{U_{\text{HMB.bts}}}\right) = \frac{\sum_{i=1}^{\text{MC}} \hat{V}_{\text{HMB.bts}_i}\left(\hat{\bar{y}}_{U_{\text{HMB.bts}_i}}\right)}{\text{MC}} \tag{21}$$

and comparing the average with the Monte Carlo-based empirical variance.

The empirical coverage of estimated confidence intervals across Monte Carlo iterations is calculated as follows

$$\text{CI}_{\text{emp}} = 100 \times \frac{\sum_{i=1}^{\text{MC}} \widehat{\text{CI}}_{\text{cover}}}{\sum_{i=1}^{\text{MC}} \widehat{\text{CI}}_{\text{total}}}, \tag{22}$$

where, $\sum_{i=1}^{\mathrm{MC}}\widehat{\mathrm{CI}}_{\mathrm{cover}}$ is the number of estimated 95% confidence intervals that cover the superpopulation mean value of AGB (295.55 Mg·ha^{-1}: Table 1), and $\sum_{i=1}^{\mathrm{MC}}\widehat{\mathrm{CI}}_{\mathrm{cover}}$ is the total number of estimated 95% confidence intervals.

The empirical mean squared error (MSE) of the HMB.bts predictor is estimated using the Monte Carlo simulation outcome as

$$\mathrm{MSE}_{\mathrm{MC}}\left(\hat{\bar{y}}_{U_{\mathrm{HMB.bts}}}\right) = \frac{\sum_{i=1}^{\mathrm{MC}}\left(\hat{\bar{y}}_{U_{\mathrm{HMB.bts}\,i}} - \bar{y}_i\right)^2}{\mathrm{MC}}, \tag{23}$$

where, $\bar{y}_i$ is the true population mean of the target population U for the given i^{th} realization through the Monte Carlo simulation.

3 | Results

The following results are based on 6′000 Monte Carlo iterations, each consisting of 750 bootstrap iterations to estimate the variance through the proposed bootstrapping algorithm. Within each bootstrap iteration, two random forest models were trained with 750 trees (the two random forest models are a consequence of the two bootstrap loops running in parallel, see Figure 2). Table 2 presents the results for the average of the predicted target population mean following the HMB.bts procedure, the empirical variance of the predicted population mean, the average of HMB.bts estimated variances, and the averages of the two estimated variance components over the Monte Carlo iterations.

From Table 3, it is evident that the propagated uncertainty resulting from estimating Model I accounts for approximately 90% of the overall estimated variance in our demonstration example. The uncertainty attributed to the second model is about 14%. Table 2 also demonstrates that the average of the estimated variances is nearly identical to the empirical variance of the HMB.bts predictor. This finding is further demonstrated in Figure 3.

In Figure 3, the Monte Carlo simulation-based empirical variance (cumulative variance) of the HMB.bts predictor for the target population mean predictor converges to the average of the estimated variances, that is, the cumulative mean of the estimated variance across the Monte Carlo iterations. Figure 4 presents histogram plots and boxplots of the estimated variance distribution and predicted population mean over the Monte Carlo iterations.

To obtain results for the cumulative variance in Figure 3 and for the predicted population mean in Figure 4, we performed an additional Monte Carlo simulation with 100,000 iterations that

TABLE 3 | Monte Carlo simulation results.

Average of predicted population means per Equation (19), [Mg·ha^{-1}]	296.90
Monte Carlo-based mean squared error (MSE) per Equation (23)	34.18
Monte Carlo-based empirical variance per Equation (20)	32.17
Average of estimated variances per Equation (21)	32.17
Average of estimated first variance components, that is, the propagated uncertainty due to the Model I estimation	27.81
Average of estimated second variance components, that is, the uncertainty due to the Model II estimation	4.36
Empirical coverage of formal 95% confidence intervals per Equation (22), [%]	94.67

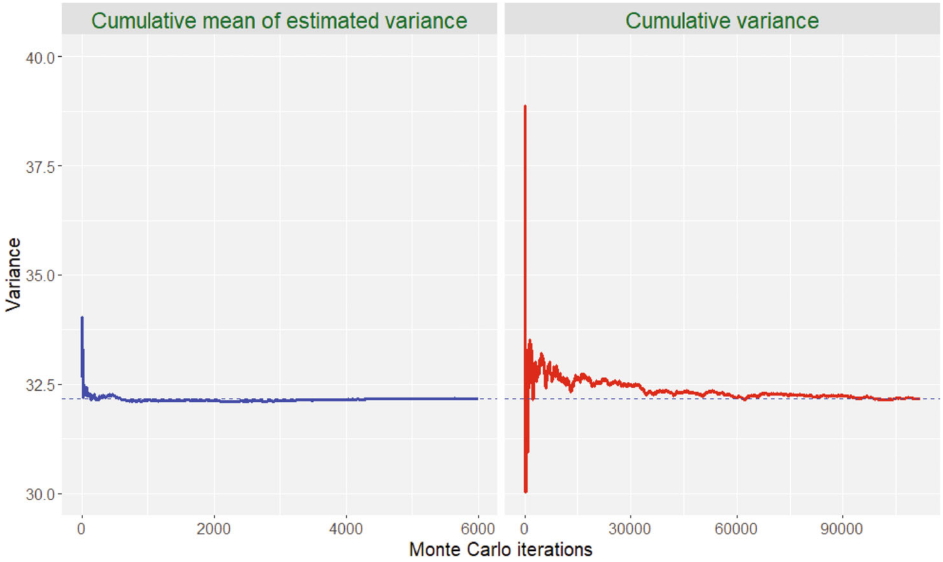

FIGURE 3 | Cumulative variance of the predicted population mean (red line) and cumulative mean of the estimated variance (blue line) following HMB.bts estimation algorithm across Monte Carlo iterations.

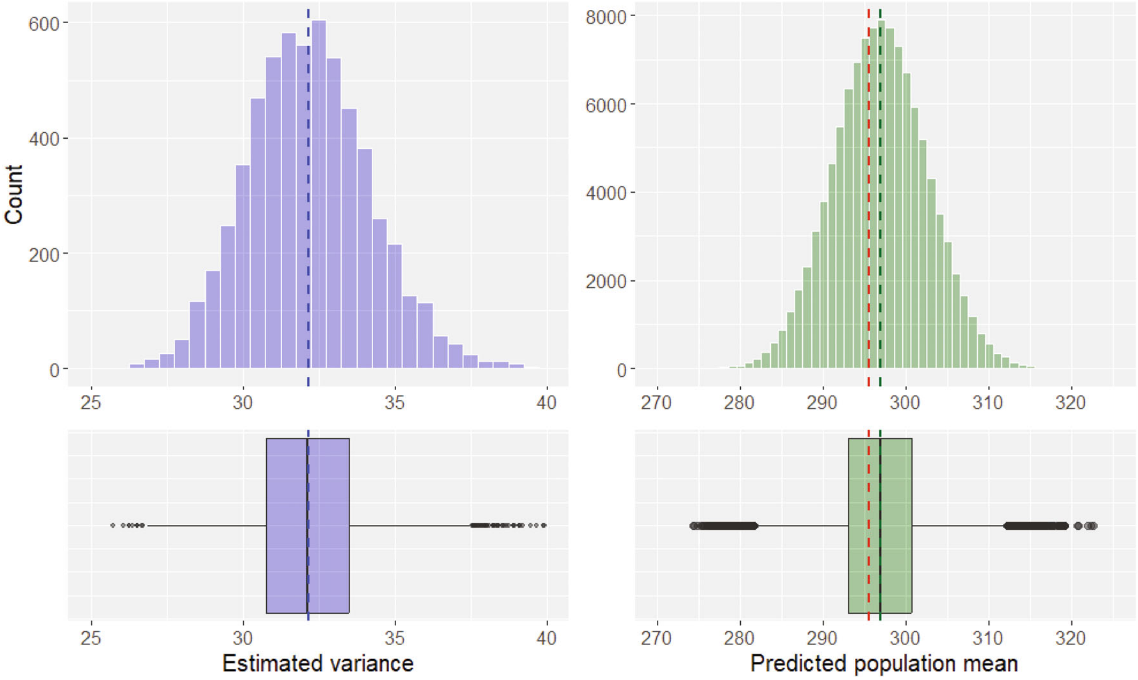

FIGURE 4 | Histogram and boxplot of estimated variances and predicted population mean values obtained through the HMB.bts estimation framework across Monte Carlo iterations. In the left histogram and boxplot, the vertical red dashed line represents the Monte Carlo-based empirical variance, while the vertical blue dashed line represents the average of the estimated variances. Since the empirical and average variances are numerically equal to 32.17 (see Table 3), the red dashed line is hidden beneath the blue dashed line. In the right histogram and boxplot, the vertical red dashed line represents the AGB superpopulation mean, which is 295.55 Mg·ha^{-1} (see Table 2), and the vertical blue dashed line represents the average of the predicted population means, which is 296.90 Mg·ha^{-1} (see Table 3).

did not include variance estimation but only contained a prediction of the target population mean. Thus, the right panel results in Figures 3 and 4 are based on 106,000 iterations, while the left panel results are based on 6000 iterations. This was done to save computational time, as running the full-scale Monte Carlo simulation, including variance estimation was very computationally demanding (several months of computations were required to obtain results for 6000 iterations). To obtain reliable endpoint estimates of the cumulative mean of the estimated variance and the cumulative variance, we used a very large number of iterations (cf. McRoberts et al. 2023).

4 | Discussion

Combining different data sources for efficient large-scale forest surveys is becoming increasingly important for providing information for policies and decision-making, at national, regional, and global scales. The wealth of remotely sensed data offers many possibilities in this regard, but also substantial challenges related to specifying estimators/predictors and assessing uncertainties. In this study, we address challenges involved in assessing uncertainties in HMB inference (e.g., Saarela et al. 2016), and suggest a new bootstrapping procedure for this purpose. The novelty of the proposed method is that it separates the computations into independent steps, thereby making it possible to run the bootstrapping in parallel and substantially reducing the computational time required. The method has similarities to previous procedures for variance estimation following multiple imputation described by Rubin (1987) and for HMB by Fortin et al. (2024). However, the procedures described by Fortin et al. require a hierarchically nested procedure, whereas the method proposed in this study makes use of the law of total variance (e.g., Feller 1977) to decompose the variance into separate terms.

The Monte Carlo simulation analysis demonstrated that the proposed bootstrapping estimation algorithm provides variance estimates, which are at least approximately unbiased. This outcome aligns with our expectations, as the bootstrap procedure theoretically should provide approximately unbiased estimates of each of the terms emanating from decomposing the total variance into components. However, in practice, it is likely that the proposed procedure is not fully unbiased, for example, because it is based on an assumption that model predictions of expected values, for all levels of the explanatory variables involved, can be used as a basis for estimating the expectation of the population mean, and subsequently the variance among those expectations. However, similar approximations are made in other bootstrapping procedures as well (e.g., Fortin et al. 2024).

To avoid the need for nested bootstrapping, we recommend estimating the variance attributable to the second model, conditional on the outcome of the first model, from a single outcome of the first model. However, such an estimator might slightly underestimate the variance. Särndal (1992) proposed an additional term to address this underestimation, which would necessitate the use of nested bootstrapping. Nevertheless, as our simulation analysis demonstrated, the underestimation appears to be negligible.

The separable bootstrap estimation algorithm reduces the computational burden by running two bootstrap procedures in parallel. As shown through the analytical expressions in Section 2.7, with the increase in bootstrap iterations, the amount of computer time increases quadratically in the case of a nested bootstrap procedure, whereas our proposed algorithm requires an amount of time that is linearly proportional to the number of bootstrap iterations.

The proposed procedure can be applied in any survey where proxy values are used in place of observed data, and these proxy values are obtained through imputation modeling methods utilizing auxiliary information. Furthermore, the framework permits an increase in the number of modeling steps, such as the three-phase HMB estimation, without a significant increase in the computational burden.

5 | Conclusions

We have demonstrated a separable bootstrap estimation algorithm for predicting population parameters and estimating their corresponding variances within the HMB inferential framework. This estimation algorithm provides approximately unbiased predictors for population parameters and their corresponding variances. We validated the method through Monte Carlo simulations, using data that simulate forest conditions in the state of Oregon, USA. The simulation results showed a close correspondence between the empirical Monte Carlo variance, and the average variance estimated through the proposed bootstrapping algorithm.

Acknowledgments

Funding for the study was provided by a grant from NASA's GEDI Science Team (80HQTR21T0013). The authors acknowledge the Orion High-Performance Computing Center (OHPCC) at the Norwegian University of Life Sciences (NMBU) for providing computational resources that have contributed to the research results reported within this paper. URL (internal): https://orion.nmbu.no. We are thankful to Lennart Noordermeer for his comments on early versions of the manuscript. We are also thankful to the Associate Editor and the two anonymous Reviewers for their constructive comments, which helped us improve the clarity of our manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

Breiman, L. 2001. "Random Forests." *Machine Learning* 45: 5–32. https://doi.org/10.1023/A:1010933404324.

Cassel, C. M., C. E. Särndal, and J. H. Wretman. 1977. *Foundation of inference in statistical survey*. New York, NY: Wiley, New York.

Cormen, T. H., ed. 2009. *Introduction to Algorithms*. 3rd ed. Cambridge, Mass: MIT Press.

Dubayah, R., J. Armston, S. P. Healey, et al. 2022. "GEDI Launches a New Era of Biomass Inference From Space." *Environmental Research Letters* 17: 095001. https://doi.org/10.1088/1748-9326/ac8694.

Dubayah, R., M. Hofton, J. Blair, J. Armston, H. Tang, and S. Luthcke. 2021. GEDI L2A Elevation and Height Metrics Data Global Footprint Level V002 https://doi.org/10.5067/GEDI/GEDI02_A.002.

Duncanson, L., J. R. Kellner, J. Armston, et al. 2022. "Aboveground Biomass Density Models for NASA's Global Ecosystem Dynamics Investigation (GEDI) Lidar Mission." *Remote Sensing of Environment* 270: 112845. https://doi.org/10.1016/j.rse.2021.112845.

Efron, B. 1979. "Computers and the Theory of Statistics: Thinking the Unthinkable." *Society for Industrial and Applied Mathematics Review* 21: 460–480. https://doi.org/10.1137/1021092.

Ene, L. T., T. Gobakken, H.-E. Andersen, et al. 2018. "Large-Area Hybrid Estimation of Aboveground Biomass in Interior Alaska Using Airborne Laser Scanning Data." *Remote Sensing of Environment* 204: 741–755. https://doi.org/10.1016/j.rse.2017.09.027.

Ene, L. T., E. Næsset, T. Gobakken, T. G. Gregoire, G. Ståhl, and R. Nelson. 2012. "Assessing the Accuracy of Regional LiDAR-Based Biomass Estimation Using a Simulation Approach." *Remote Sensing of Environment* 123: 579–592. https://doi.org/10.1016/j.rse.2012.04.017.

Esteban, J., R. McRoberts, A. Fernández-Landa, J. Tomé, and E. Næsset. 2019. "Estimating Forest Volume and Biomass and Their Changes Using Random Forests and Remotely Sensed Data." *Remote Sensing* 11: 1944. https://doi.org/10.3390/rs11161944.

Feller, W. 1977. *An Introduction to the Probability Theory and Its Application*. London: John Willey and Sons INC.

Fortin, M., R. Manso, and R. Schneider. 2018. "Parametric Bootstrap Estimators for Hybrid Inference in Forest Inventories." *Forestry: An International Journal of Forest Research* 91: 354–365. https://doi.org/10.1093/forestry/cpx048.

Fortin, M., O. Van Lier, J.-F. Côté, H. Erdle, and J. White. 2024. "A Bootstrap-Based Approach to Combine Individual-Based Forest Growth Models and Remotely Sensed Data." *Forestry: An International Journal of Forest Research* cpae003: 649–661. https://doi.org/10.1093/forestry/cpae003.

Kangas, A., M. Myllymäki, and L. Mehtätalo. 2023. "Understanding Uncertainty in Forest Resources Maps." *Silva Fennica* 52: 22026. https://doi.org/10.14214/sf.22026.

Kim, J. K., J. Michael Brick, W. A. Fuller, and G. Kalton. 2006. "On the Bias of the Multiple-Imputation Variance Estimator in Survey Sampling." *Journal of the Royal Statistical Society, Series B: Statistical Methodology* 68: 509–521. https://doi.org/10.1111/j.1467-9868.2006.00546.x.

McRoberts, R. E., Q. Chen, G. M. Domke, G. Ståhl, S. Saarela, and J. A. Westfall. 2016. "Hybrid Estimators for Mean Aboveground Carbon Per Unit Area." *Forest Ecology and Management* 378: 44–56. https://doi.org/10.1016/j.foreco.2016.07.007.

McRoberts, R. E., E. Næsset, T. Gobakken, et al. 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. https://doi.org/10.1139/cjfr-2017-0396.

McRoberts, R. E., E. Næsset, Z. Hou, et al. 2023. "How Many Bootstrap Replications Are Necessary for Estimating Remote Sensing-Assisted, Model-Based Standard Errors?" *Remote Sensing of Environment* 288: 113455. https://doi.org/10.1016/j.rse.2023.113455.

Menlove, J., and S. P. Healey. 2020. "A Comprehensive Forest Biomass Dataset for the USA Allows Customized Validation of Remotely Sensed Biomass Estimates." *Remote Sensing* 12: 4141. https://doi.org/10.3390/rs12244141.

Nagler, T., U. Schepsmeier, J. Stoeber, E. C. Brechmann, B. Graeler, and T. Erhardt. 2023. "VineCopula: Statistical Inference of Vine Copulas." *R Package Version 2.5.0* https://CRAN.R-project.org/package=VineCopula.

Nelsen, R. B. 2006. *An Introduction to Copulas, Springer Series in Statistics*. New York, NY: Springer New York. https://doi.org/10.1007/0-387-28678-0.

Patterson, P. L., S. P. Healey, G. Ståhl, et al. 2019. "Statistical Properties of Hybrid Estimators Proposed for GEDI—NASA's Global Ecosystem Dynamics Investigation." *Environmental Research Letters* 14: 065007. https://doi.org/10.1088/1748-9326/ab18df.

Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys, 1st Ed, Wiley Series in Probability and Statistics*. Hoboken, NJ: Wiley. https://doi.org/10.1002/9780470316696.

Saarela, S., S. Holm, A. Grafström, et al. 2016. "Hierarchical Model-Based Inference for Forest Inventory Utilizing Three Sources of Information." *Annals of Forest Science* 73: 895–910. https://doi.org/10.1007/s13595-016-0590-1.

Saarela, S., S. Holm, S. Healey, et al. 2018. "Generalized Hierarchical Model-Based Estimation for Aboveground Biomass Assessment Using GEDI and Landsat Data." *Remote Sensing* 10: 1832. https://doi.org/10.3390/rs10111832.

Saarela, S., S. Holm, S. P. Healey, et al. 2022. "Comparing Frameworks for Biomass Prediction for the Global Ecosystem Dynamics Investigation." *Remote Sensing of Environment* 278: 113074. https://doi.org/10.1016/j.rse.2022.113074.

Saarela, S., P. Varvia, L. Korhonen, et al. 2023. "Three-Phase Hierarchical Model-Based and Hybrid Inference." *MethodsX* 11: 102321. https://doi.org/10.1016/j.mex.2023.102321.

Saarela, S., A. Wästlund, E. Holmström, et al. 2020. "Mapping Aboveground Biomass and Its Prediction Uncertainty Using LiDAR and Field Data, Accounting for Tree-Level Allometric and LiDAR Model Errors." *Forest Ecosystems* 7: 43. https://doi.org/10.1186/s40663-020-00245-0.

Särndal, C.-E. 1992. "Methods for Estimating the Precision of Survey Estimates When Imputation Has Been Used." *Survey Methodology* 18: 241–252.

Ståhl, G., S. Saarela, S. Schnell, et al. 2016. "Use of Models in Large-Area Forest Surveys: Comparing Model-Assisted, Model-Based and Hybrid Estimation." *Forest Ecosystems* 3: 5. https://doi.org/10.1186/s40663-016-0064-9.

Varvia, P., S. Saarela, M. Maltamo, et al. 2024. "Estimation of Boreal Forest Biomass From ICESat-2 Data Using Hierarchical Hybrid Inference." *Remote Sensing of Environment* 311: 114249. https://doi.org/10.1016/j.rse.2024.114249.

Zhu, Z., C. E. Woodcock, C. Holden, and Z. Yang. 2015. "Generating Synthetic Landsat Images Based on all Available Landsat Data: Predicting Landsat Surface Reflectance at any Given Time." *Remote Sensing of Environment* 162: 67–83. https://doi.org/10.1016/j.rse.2015.02.009.