Three-phase hierarchical model-based and hybrid inference

Svetlana Saarela [a,*], Petri Varvia [b], Lauri Korhonen [b], Zhiqiang Yang [c],
Paul L. Patterson [d], Terje Gobakken [a], Erik Næsset [a], Sean P. Healey [c], Göran Ståhl [e]

[a] Faculty of Environmental Sciences and Natural Resource Management, Norwegian University of Life Sciences, P.O. Box 5003, NO-1432, Ås, Norway
[b] School of Forest Sciences, University of Eastern Finland, P.O. Box 111, Joensuu FI-80101, Finland
[c] USDA Forest Service, Rocky Mountain Research Station, 507 25th St, Ogden, UT, USA
[d] USDA Forest Service, Rocky Mountain Research Station, 240 W Prospect, Fort Collins, CO 80526, USA
[e] Faculty of Forest Sciences, Swedish University of Agricultural Sciences, SLU Skogsmarksgränd 17, SE-90183, Umeå, Sweden

ARTICLE INFO

ABSTRACT

Global commitments to mitigating climate change and halting biodiversity loss require reliable information about Earth's ecosystems. Increasingly, such information is obtained from multiple sources of remotely sensed data combined with data acquired in the field. This new wealth of data poses challenges regarding the combination of different data sources to derive the required information and assess uncertainties. In this article, we show how predictors and their variances can be derived when hierarchically nested models are applied. Previous studies have developed methods for cases involving two modeling steps, such as biomass prediction relying on tree-level allometric models and models linking plot-level field data with remotely sensed data. This study extends the analysis to cases involving three modeling steps to cover new important applications. The additional step might involve an intermediate model, linking field and remotely sensed data available from a small sample, for making predictions that are subsequently used for training a final prediction model based on remotely sensed data:

- In cases where the data in the final step are available wall-to-wall, we denote the approach three-phase hierarchical model-based inference (3pHMB),
- In cases where the data in the final step are available as a probability sample, we denote the approach three-phase hierarchical hybrid inference (3pHHY).

Specifications table

Subject area:	Environmental Science
More specific subject area:	*Statistical Inference based on Earth Observation data*
Name of your method:	*Three-phase Hierarchical Model-based and Hybrid Inference*
Name and reference of original method:	*N.A.*
Resource availability:	*3pHMB_3pHHY_MC-based_valid.Rmd*

Method details

Background

Assessment of state and change of forest biomass has become important worldwide due to the large carbon dioxide fluxes incurred by deforestation and afforestation, as well as by tree growth and harvesting in managed forests (e.g., [1]). Intergovernmental Panel on Climate Change (IPCC) guidance specifies that parties must report changes in biomass carbon stocks as well as uncertainties of the reported figures. Since large parts of the world's forests are inaccessible, applying remotely sensed (RS) data is an important option for such assessments (ibid.). However, the accuracies of the assessments should be quantified using objective and consistent methods (e.g., [2]).

A standard model-based approach to forest biomass assessment is to specify and estimate a model that links biomass measured in the field with RS metrics. The estimated model is then applied across the forest area of interest, whereby biomass predictions and their uncertainties can be obtained utilizing standard methods of model-based inference (e.g., [3]). Uncertainties arise since a model cannot perfectly predict the actual biomass conditions. In some cases, several models are needed to make biomass assessments feasible. For example, models are often applied to predict the biomass of individual trees in the field from measurements of their diameters and heights. Subsequently, these predicted biomasses are used for developing models linking field biomass at the level of sample plots with RS metrics. In this case, the overall uncertainty arises from two modeling steps (e.g., [4]).

Remote sensing methods that utilize laser measurements have revolutionized forest resource assessments [5–7]. These methods provide information about the 3D structure of forests, which is closely linked to biomass. However, laser measurements are expensive, and thus in many cases they cannot be conducted wall-to-wall but only through sample coverage [8]. Important examples include the strip samples of laser measurements conducted in Alaska [9], and the samples of laser footprints from NASA's Global Ecosystem Dynamics Investigation (GEDI; [1,10] and the Ice, Cloud, and land Elevation Satellites 1 and 2 (ICESat-1 and −2; [11–17]. For many of these applications, existing methods for assessing uncertainties using frameworks such as model-assisted estimation (e.g., [9]), hybrid inference (e.g., [1,11]), and hierarchical model-based inference (e.g., [10]) have been sufficient.

Sometimes the available data structure requires additional levels in the hierarchy. New methods that utilize several RS-based models in hierarchical structures call for further development of methods to assess uncertainties. An example application involves modeling plot level biomass from field measurements, linking these with laser measurements through a second model, and utilizing laser-based predictions of biomass at sample locations as substitutes for field data in estimating a third model for biomass prediction based on RS data available wall-to-wall, such as data from the Landsat satellite [18]. In this case, three hierarchically nested modeling steps contribute to the overall uncertainty of the biomass prediction for the study area. Ideally, the different datasets involved should stem from about the same time point Hou et al. [19].

Note the difference between this type of hierarchically nested models and the hierarchy of data and models applied in standard multi-level modeling (e.g., [20]). In our case, the target response variable in the first modeling step is replaced by model predictions of the response variable in the subsequent modeling steps. In standard multi-level modeling, the response variable does not change, and multi-level modeling (also called mixed-effects modeling) refers to handling groups of observations, within which the observations are dependent.

This article aims to develop and demonstrate predictors and methods for uncertainty assessment for model-based inference involving three hierarchically nested modeling steps. If data in the final step are available wall-to-wall, we denote the method three-phase hierarchical model-based inference (3pHMB). If data in the last step are available from a probability sample, we denote the method three-phase hierarchical hybrid inference (3pHHY). The methods are similar in the first three steps, and differ only in the last step (see the Graphical Abstract).

Overview

In the following sections, we first derive and present the formulas needed for applying the 3pHMB and 3pHHY frameworks. Secondly, we provide further theoretical support for the approach through an analysis based on a multivariate superpopulation model; in this section, we also demonstrate important consequences of multi-stage hierarchical modeling. Lastly, we validate our theoretical results through Monte Carlo simulation.

Details of three-phase hierarchical model-based inference (3pHMB)

We base our description of 3pHMB inference on an example from forest inventory aiming at assessing the biomass for some large study area. We define our population as the grid-cells that tessellate the study area into N units. The size of the grid-cells corresponds to the size of plots utilized in field inventories and to the size of pixels for which RS data can be retrieved. For each grid-cell RS data are available, e.g., multispectral optical satellite data from the Landsat satellite (e.g., [21,22]). With the 3pHMB method, wall-to-wall RS data are used for predicting the biomass in each grid-cell and through averaging across the study area we predict the biomass density. All other data sources, described below, are utilized for providing data for estimating the parameters of this prediction model, which we denote the Landsat model.

When the 3pHMB method is applied, field data are not adequate for immediately estimating the parameters of the Landsat model, but "pseudo-field" biomass data are available from a sample of accurate predictions based on, e.g., airborne laser scanning (ALS) data. This sample dataset is used for estimating the Landsat model parameters. However, to make the ALS-based predictions, the

parameters of the ALS model must also be estimated. This model is estimated from a small sample of field biomass assessments with corresponding ALS data.

The biomasses of trees in the field also need to be predicted from models (often called allometric biomass models, e.g., [23]), since direct measurement is expensive and destroys the resource being monitored. Thus, biomass models are typically estimated from small samples of trees, which are cut down and carefully weighed during dedicated research studies (e.g., [24,25]). When applying the 3pHMB method we often aggregate model predictions of tree-level biomass to plot level, e.g., using the method described in Saarela et al. [[4], pp. 11–12, Section "Aggregation of tree-level AGB predictions to plot level"] and in Varvia et al. [[26], Appendix A.1]. In this article, we do not address the details of this aggregation but assume that plot level biomass is either available from aggregation of tree-level predictions (based on field data) or directly predicted from plot level measurements in the field of, e.g., basal area and mean height.

Thus, our study area of interest (AOI) is tessellated into N grid-cells that constitute the population elements. We assume that the objective is to predict the population mean, $\bar{y}_U$, of grid-cell level aboveground biomass density (AGBD), defined as

$$\bar{y}_U = \frac{1}{N} \sum_{i=1}^{N} y_i,$$

where U denotes the set of elements in the AOI. The RS data available for this set (e.g., Landsat data) are denoted $\boldsymbol{P}_U$, which is a matrix with N observations of t explanatory variables (plus a column of 1's for a model intercept). The dataset available for estimating the prediction model used in the final stage is denoted S_{III}. It has n_{III} sample elements and contains data of the kind described above (denoted $\boldsymbol{P}_{III}$ for this set) and intermediate RS data (e.g. from ALS) denoted $\boldsymbol{Z}_{III}$, which is a matrix with n_{III} observations of q explanatory variables (plus a column of 1's). Further, the dataset S_{II} has n_{II} elements and data in the datasets $\boldsymbol{Z}_{II}$ (as above, but with n_{II} observations) and $\boldsymbol{X}_{II}$, which is a matrix with n_{II} observations of h variables (plus a column of 1's). The latter data are field measurements (such as basal area, and plot averages of tree diameters and tree heights). Finally, the dataset S_I has n_I observations. It contains field measurements in $\boldsymbol{X}_I$ (as above, but with n_I observations) and data from actual measurements of AGBD, in the vector $\boldsymbol{y}_I$.

For deriving a formula for the variance of the 3pHMB method, we start by defining a univariate linear model F, which links AGBD and field measurements at the grid-cell level, using data from S_I:

$$\text{F} : Y_i = \boldsymbol{x}_i \boldsymbol{\beta} + \epsilon_i, \ \epsilon_i \sim \mathcal{N}\left(0, \omega^2\right), \tag{1}$$

where Y_i is the variable of interest (AGBD), $\boldsymbol{x}_i$ is a $(h+1)$-length row vector of explanatory variables (field measurements) with a unit term for the intercept, $\boldsymbol{\beta}$ is a $(h+1)$-length column vector of model parameters to be estimated, and ϵ_i is a variability (a.k.a. error) term which we assume follows a normal distribution with zero mean and constant variance ω^2.[1]

The model parameters are estimated using the ordinary least squares (OLS) estimator (e.g., [27]):

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}_I^T \boldsymbol{X}_I\right)^{-1} \boldsymbol{X}_I^T \boldsymbol{y}_I$$

The estimated model is then used to predict the target variable (AGBD) for the elements in S_{II}:

$$\hat{\boldsymbol{y}}_F = \boldsymbol{X}_{II} \hat{\boldsymbol{\beta}} .$$

The index F is used to denote that the AGBD predictions in S_{II} are based on the model F (Eq. (1)). The values $\hat{\boldsymbol{y}}_F$ are predictions of AGBD based on field data, which are subsequently used for estimating the parameters of the model predicting AGBD from ALS data. The ALS model is thus the next model to be addressed in the model hierarchy.

To support the construction of the ALS model, we specify a multivariate model, G*, which links the $(h+1)$-multivariate response variable of field measurements $\boldsymbol{X}_i$ and a $(q+1)$-length row vector $\boldsymbol{z}_i$ of ALS explanatory variables:

$$\text{G}^* : \boldsymbol{X}_i = \boldsymbol{z}_i \mathbf{A}^* + \boldsymbol{v}_i^*, \ \boldsymbol{v}_i^* \sim \mathcal{N}\left(0, \boldsymbol{\Delta}^*\right),$$

where $\mathbf{A}^*$ is a $((q+1) \times (h+1))$ - matrix of model parameters to be estimated, $\boldsymbol{v}_i^*$ is a $(h+1)$-length multivariate vector of variability terms, which are assumed to be independent, normally distributed, have zero mean, and their variance-covariance matrix $\boldsymbol{\Delta}^*$ is of size $((h+1) \times (h+1))$.

By multiplying each model component in G* by $\boldsymbol{\beta}$ we obtain the univariate model ([28], Appendix A.3)

$$\boldsymbol{X}_i \boldsymbol{\beta} = \boldsymbol{z}_i \mathbf{A}^* \boldsymbol{\beta} + \boldsymbol{v}_i^* \boldsymbol{\beta}, \ \boldsymbol{v}_i^* \boldsymbol{\beta} \sim \mathcal{N}\left(0, \boldsymbol{\beta}^T \boldsymbol{\Delta}^* \boldsymbol{\beta}\right),$$

Denoting $\mathbf{A}^* \boldsymbol{\beta}$ as $\boldsymbol{\alpha}$, $\boldsymbol{v}_i^* \boldsymbol{\beta}$ as $\boldsymbol{v}_i$, and $\boldsymbol{\beta}^T \boldsymbol{\Delta}^* \boldsymbol{\beta}$ as δ^2, we obtain the following model, which we denote as G:

$$\text{G} : \boldsymbol{X}_i \boldsymbol{\beta} = \boldsymbol{z}_i \boldsymbol{\alpha} + \boldsymbol{v}_i, \ \boldsymbol{v}_i \sim \mathcal{N}\left(0, \delta^2\right) . \tag{2}$$

Here, $\boldsymbol{z}_i$ a $(q+1)$-length row vector of ALS explanatory variables with a unit term for the intercept and $\boldsymbol{\alpha}$ is a $(q+1)$-length column vector of model parameters to be estimated. This model links the expectation of AGBD, due to the model F, with ALS explanatory variables.

---

[1] In the case of heterogeneity and correlation among model error terms, the generalized least square estimator should be applied to estimate the model parameters. Varvia et al. [26] presents an example where heterogeneous variance was taken into account in the context of 3pHHY.

Table 1

A summary of datasets and estimation steps involved in 3pHMB prediction.

Dataset	Available information	Model training steps	Model application step
The field dataset S_I	AGBD and field measurements	A model linking AGBD and field measurements (model F)	–
The ALS dataset S_{II}	Field measurements and ALS variables	An ALS model, used to predict AGBD from ALS variables (model G)	The model F is applied on the S_{II} dataset
The Landsat dataset S_{III}	ALS and Landsat variables	A Landsat model, used to predict AGBD from Landsat variables (model Q)	The model G is applied on the S_{III} dataset
The AOI, U	Landsat data	–	The model Q is applied to the entire population

The predictions $\hat{\mathbf{y}}_F$ are used as the response variables in the model G to estimate its model parameters, $\boldsymbol{\alpha}$, based on the explanatory variables $\mathbf{Z}_{II}$ from S_{II} ([28], Appendix A.3):

$$\hat{\boldsymbol{\alpha}} = \left(\mathbf{Z}_{II}^T \mathbf{Z}_{II}\right)^{-1} \mathbf{Z}_{II}^T \hat{\mathbf{y}}_F = \left(\mathbf{Z}_{II}^T \mathbf{Z}_{II}\right)^{-1} \mathbf{Z}_{II}^T \mathbf{X}_{II} \hat{\boldsymbol{\beta}} = \widehat{\mathbf{A}^*} \widehat{\boldsymbol{\beta}}.$$

Here, $\widehat{\mathbf{A}^*} = (\mathbf{Z}_{II}^T \mathbf{Z}_{II})^{-1} \mathbf{Z}_{II}^T \mathbf{X}_{II}$ are estimated model parameters from the multivariate model G*.

At this point, we have utilized the actual measurements of AGBD from S_I to predict AGBD for all elements in S_{II} based on field measurement data. The AGBD predictions for S_{II} have subsequently been used for training an ALS model (the G model), through which we now can predict AGBD for S_{III}:

$$\hat{\mathbf{y}}_G = \mathbf{Z}_{III} \hat{\boldsymbol{\alpha}}.$$

The predicted values $\hat{\mathbf{y}}_G$ are used for training the model to be applied in the last stage, based on wall-to-wall RS data, i.e., the Landsat model.

In support of constructing the Landsat model, we introduce a second multivariate model, which links the $(q + 1)$-multivariate variable $\mathbf{Z}_i$ (ALS variables) as a multivariate response variable with a $(t + 1)$-length row vector of Landsat explanatory variables $\mathbf{p}_i$; the model is denoted Q*:

$$Q^* : \mathbf{Z}_i = \mathbf{p}_i \boldsymbol{\Gamma}^* + \mathbf{e}_i^*, \ \mathbf{e}_i^* \sim \mathcal{N}\left(0, \boldsymbol{\Theta}^*\right),$$

where $\boldsymbol{\Gamma}^*$ is a $((t + 1) \times (q + 1))$-matrix of model parameters, $\mathbf{e}_i^*$ is a $(q + 1)$-multivariate variable of variability terms which are independent, normally distributed, have zero mean and variance-covariance $\boldsymbol{\Theta}^*$ of size $((q + 1) \times (q + 1))$.

By multiplying model components in Q* with the model coefficients from G, we obtain the following univariate model

$$\mathbf{Z}_i \boldsymbol{\alpha} = \mathbf{p}_i \boldsymbol{\Gamma}^* \boldsymbol{\alpha} + \mathbf{e}_i^* \boldsymbol{\alpha}, \ \mathbf{e}_i^* \boldsymbol{\alpha} \sim \mathcal{N}\left(0, \boldsymbol{\alpha}^T \boldsymbol{\Theta}^* \boldsymbol{\alpha}\right).$$

Denoting $\boldsymbol{\Gamma}^* \boldsymbol{\alpha}$ as $\boldsymbol{\gamma}$, $\mathbf{e}_i^* \boldsymbol{\alpha}$ as e_i and $\boldsymbol{\alpha}^T \boldsymbol{\Theta}^* \boldsymbol{\alpha}$ as θ^2 we obtain the following model, which we denote as Q:

$$Q : \mathbf{Z}_i \boldsymbol{\alpha} = \mathbf{p}_i \boldsymbol{\gamma} + e_i, \ e_i \sim \mathcal{N}\left(0, \theta^2\right), \tag{3}$$

where, $\boldsymbol{\gamma}$ is an $(t + 1)$-length vector of model coefficients to be estimated.

The ALS-based predictions of AGBD, $\hat{\mathbf{y}}_G$, are used for estimating the model parameters, $\boldsymbol{\gamma}$, using information on explanatory Landsat variables $\mathbf{P}_{III}$ from S_{III}:

$$\hat{\boldsymbol{\gamma}} = \left(\mathbf{P}_{III}^T \mathbf{P}_{III}\right)^{-1} \mathbf{P}_{III}^T \hat{\mathbf{y}}_G = \left(\mathbf{P}_{III}^T \mathbf{P}_{III}\right)^{-1} \mathbf{P}_{III}^T \mathbf{Z}_{III} \hat{\boldsymbol{\alpha}} = \left(\mathbf{P}_{III}^T \mathbf{P}_{III}\right)^{-1} \mathbf{P}_{III}^T \mathbf{Z}_{III} \widehat{\mathbf{A}^*} \hat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\Gamma}^*} \widehat{\mathbf{A}^*} \hat{\boldsymbol{\beta}},$$

where $\widehat{\boldsymbol{\Gamma}^*} = (\mathbf{P}_{III}^T \mathbf{P}_{III})^{-1} \mathbf{P}_{III}^T \mathbf{Z}_{III}$ are estimated model parameters from the model Q*.

Finally, the estimated Q model is used to predict the AGBD across the entire AOI, i.e., for all the elements in U:

$$\hat{\mathbf{y}}_U = \mathbf{P}_U \hat{\boldsymbol{\gamma}}.$$

The 3pHMB population mean predictor is then:

$$\hat{\bar{y}}_{U_{3pHMB}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{p}_i \hat{\boldsymbol{\gamma}} = \bar{\mathbf{p}}_U \hat{\boldsymbol{\gamma}}, \tag{4}$$

where $\bar{\mathbf{p}}_U$ is the $(t + 1)$-length row vector of Landsat explanatory variable averages over the AOI.

Table 1 summarizes the description of the datasets and the estimation steps. In the table, the datasets used in the analyses are listed with descriptions of available variables and whether they are used for model training or model application.

Fig. 1 gives a graphical overview of 3pHMB prediction.

The 3pHMB predictor variance and its estimator

Application of 3pHMB for making predictions, only, across some AOI is fairly straightforward, and does not require all the model formalism from the previous section. What is required is a procedure where the AGBD is first measured on the elements in S_I, then predicted in a stepwise manner for the elements in S_{II}, S_{III}, and U. However, the formalism introduced in the previous section
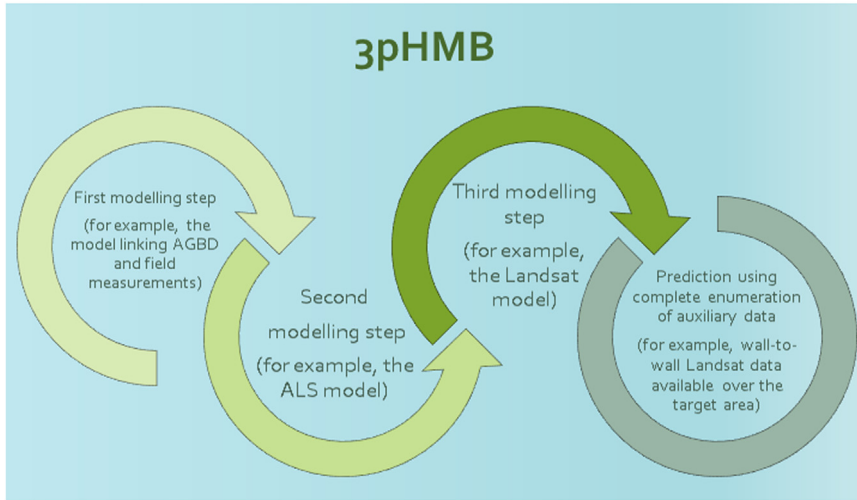
Fig. 1. A graphical overview of 3pHMB prediction based on wall-to-wall auxiliary (remotely sensed) data.

is needed for assessing the uncertainty associated with predicting the AGBD for the AOI. In this section, we build on the previous section and derive a formula for the variance of the predictor of AGBD, and its corresponding estimator.

For deriving the 3pHMB predictor variance, we begin with its expectation:

$$E\left[\widehat{\bar{y}}_{U_{3pHMB}}\right] = E\left[\bar{p}_U \hat{\gamma}\right] = \bar{p}_U E[\hat{\gamma}].$$

Since independent datasets (normally) are used for the three modeling steps, the estimated parameters in each step are independent and thus

$$E[\hat{\gamma}] = E\left[\widehat{\Gamma^*}\widehat{A^*}\hat{\beta}\right] = \Gamma^* A^* \beta = \gamma.$$

Thus, the expectation of the predictor of the population mean is

$$E\left[\widehat{\bar{y}}_{U_{3pHMB}}\right] = \bar{p}_U \gamma.$$

Following the definition, the variance of $\widehat{\bar{y}}_{U_{3pHMB}}$ is

$$
\begin{aligned}
V\left(\widehat{\bar{y}}_{U3pHMB}\right) &= E\left[\left(\widehat{\bar{y}}_{U3pHMB} - E\left[\widehat{\bar{y}}_{U3pHMB}\right]\right)^2\right] \\
&= E\left[\left(\bar{p}_U \hat{\gamma} - E\left[\bar{p}_U \hat{\gamma}\right]\right)^2\right] \\
&= \bar{p}_U E\left[(\hat{\gamma} - E[\hat{\gamma}])^2\right]\bar{p}_U^T \\
&= \bar{p}Cov(\hat{\gamma})\bar{p}_U^T.
\end{aligned}
\tag{5}
$$

It can be noted that the covariance of the estimated model parameters is at the core of this expression. Since several modeling steps have been involved in the estimation of the model parameters $\hat{\gamma}$, the covariance can be decomposed into the following terms using the law of total covariance [29]:

$$
\begin{aligned}
Cov(\hat{\gamma}) &= Cov\left(\widehat{\Gamma^*}\hat{\alpha}\right) \\
&= E_G\left[Cov_{Q*}\left(\widehat{\Gamma^*}\hat{\alpha}\right)\right] + Cov_G\left(E_{Q*}\left[\widehat{\Gamma^*}\hat{\alpha}\right]\right) \\
&= E_G\left[\hat{\alpha}^T Cov_{Q*}\left(\widehat{\Gamma^*}\right)\hat{\alpha}\right] + \Gamma^* Cov_G(\hat{\alpha})\Gamma^{*T} \\
&= \alpha^T Cov_{Q*}\left(\widehat{\Gamma^*}\right)\alpha + \Gamma^* Cov_G(\hat{\alpha})\Gamma^{*T} + Tr\left[Cov_{Q*}\left(\widehat{\Gamma^*}\right)Cov_G(\hat{\alpha})\right],
\end{aligned}
\tag{6}
$$

where the step from the second last to the last line is obtained from

$$
\begin{aligned}
E_G\left[\hat{\alpha}^T Cov_{Q*}\left(\widehat{\Gamma^*}\right)\hat{\alpha}\right] &= E_G\left[\sum_{i=1}^{(q+1)}\sum_{j=1}^{(q+1)} \hat{\alpha}_i Cov_{Q*}(\hat{\gamma}_i, \hat{\gamma}_j)\hat{\alpha}_j\right] \\
&= \sum_{i=1}^{(q+1)}\sum_{j=1}^{(q+1)} Cov_{Q*}(\hat{\gamma}_i, \hat{\gamma}_j) E_G[\hat{\alpha}_i \hat{\alpha}_j]
\end{aligned}
$$

$$= \sum_{i=1}^{(q+1)} \sum_{j=1}^{(q+1)} Cov_{Q*}(\hat{\gamma}_i, \hat{\gamma}_j)(\alpha_i \alpha_j + Cov(\hat{\alpha}_i, \hat{\alpha}_j))$$

$$= \sum_{i=1}^{(q+1)} \sum_{j=1}^{(q+1)} Cov_{Q*}(\hat{\gamma}_i, \hat{\gamma}_j)\alpha_i \alpha_j + \sum_{i=1}^{(q+1)} \sum_{j=1}^{(q+1)} Cov_{Q*}(\hat{\gamma}_i, \hat{\gamma}_j)Cov(\hat{\alpha}_i, \hat{\alpha}_j)$$

$$= \boldsymbol{\alpha}^T Cov_{Q*}\left(\widehat{\boldsymbol{\Gamma}*}\right)\boldsymbol{\alpha} + Tr\left[Cov_{Q*}\left(\widehat{\boldsymbol{\Gamma}*}\right)Cov_G(\hat{\boldsymbol{\alpha}})\right]$$

However, $Cov_G(\hat{\boldsymbol{\alpha}})$ in Eq. (6) emanates from several interacting models, and in order to present the variance formula in a format that would later on allow estimation of, $Cov_G(\hat{\boldsymbol{\alpha}})$ is further decomposed using the law of total covariance:

$$Cov_G(\hat{\boldsymbol{\alpha}}) = Cov_G\left(\widehat{\mathbf{A}*}\hat{\boldsymbol{\beta}}\right)$$

$$= E_F\left[Cov_{G*}\left(\widehat{\mathbf{A}*}\hat{\boldsymbol{\beta}}\right)\right] + Cov_F\left(E_{G*}\left[\widehat{\mathbf{A}*}\hat{\boldsymbol{\beta}}\right]\right)$$

$$= E_F\left[\hat{\boldsymbol{\beta}}^T Cov_{G*}\left(\widehat{\mathbf{A}*}\right)\hat{\boldsymbol{\beta}}\right] + \mathbf{A}*Cov_F(\hat{\boldsymbol{\beta}})\mathbf{A}*^T$$

$$= \boldsymbol{\beta}^T Cov_{G*}\left(\widehat{\mathbf{A}*}\right)\boldsymbol{\beta} + \mathbf{A}*Cov_F(\hat{\boldsymbol{\beta}})\mathbf{A}*^T + Tr\left[Cov_{G*}\left(\widehat{\mathbf{A}*}\right)Cov_F(\hat{\boldsymbol{\beta}})\right], \tag{7}$$

where the third term ($Tr[Cov_{G*}(\widehat{\mathbf{A}*})Cov_F(\hat{\boldsymbol{\beta}})]$) follows from $E_F[\hat{\boldsymbol{\beta}}^T \hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}^T \boldsymbol{\beta} + Cov(\hat{\boldsymbol{\beta}})$.

Thus, the variance of the AGBD predictor can be expressed as:

$$V\left(\hat{\bar{y}}_{U_{3pHMB}}\right) = \bar{p}_U\left(\boldsymbol{\alpha}^T Cov_{Q*}\left(\widehat{\boldsymbol{\Gamma}*}\right)\boldsymbol{\alpha}\right)\bar{p}_U^T$$

$$+ \bar{p}_U\left(\boldsymbol{\Gamma}*\boldsymbol{\beta}^T Cov_{G*}\left(\widehat{\mathbf{A}*}\right)\boldsymbol{\beta}\,\boldsymbol{\Gamma}*^T\right)\bar{p}_U^T$$

$$+ \bar{p}_U\left(\boldsymbol{\Gamma}*\mathbf{A}*Cov_F(\hat{\boldsymbol{\beta}})\mathbf{A}*^T\,\boldsymbol{\Gamma}*^T\right)\bar{p}_U^T$$

$$+ \bar{p}_U\left(\boldsymbol{\Gamma}*Tr\left[Cov_{G*}\left(\widehat{\mathbf{A}*}\right)Cov_F(\hat{\boldsymbol{\beta}})\right]\boldsymbol{\Gamma}*^T\right)\bar{p}_U^T$$

$$+ \bar{p}_U\left(Tr\left[Cov_{Q*}\left(\widehat{\boldsymbol{\Gamma}*}\right)\left(\boldsymbol{\beta}^T Cov_{G*}\left(\widehat{\mathbf{A}*}\right)\boldsymbol{\beta} + \mathbf{A}*Cov_F(\hat{\boldsymbol{\beta}})\mathbf{A}*^T + Tr\left[Cov_{G*}\left(\widehat{\mathbf{A}*}\right)Cov_F(\hat{\boldsymbol{\beta}})\right]\right)\right]\right)\bar{p}_U^T. \tag{8a}$$

In the simulation analyses that we performed to validate our methodology (illustrated in Table 7), we found that all the terms involving traces were small, and could be neglected in the variance formula, for simplification. Thus, a good approximation of the variance of the AGBD predictor would be

$$V\left(\hat{\bar{y}}_{U3pHMB}\right) = \bar{p}_U\left(\boldsymbol{\alpha}^T Cov_{Q*}\left(\widehat{\boldsymbol{\Gamma}*}\right)\boldsymbol{\alpha}\right)\bar{p}_U^T + \bar{p}_U\left(\boldsymbol{\Gamma}*\boldsymbol{\beta}^T Cov_{G*}\left(\widehat{\mathbf{A}*}\right)\boldsymbol{\beta}\,\boldsymbol{\Gamma}*^T\right)\bar{p}_U^T + \bar{p}_U\left(\boldsymbol{\Gamma}*\mathbf{A}*Cov_F(\hat{\boldsymbol{\beta}})\mathbf{A}*^T\,\boldsymbol{\Gamma}*^T\right)\bar{p}_U^T. \tag{8b}$$

By replacing the covariances of estimated model parameters and the model parameters by their corresponding estimators, we obtain a variance estimator. In the simulation analysis that we performed to validate our methodology, it was found that the variance estimators were approximately unbiased.

If we would like to derive the MSE of the predictor rather than the variance, we may define the true AGBD over the AOI through the models involved. We first express the AGBD variable using the model Q as

$$\mathbf{Z}_U \boldsymbol{\alpha} = \mathbf{P}_U \boldsymbol{\gamma} + \mathbf{e}_U,$$

Using model G, we can express $\mathbf{Z}_U \boldsymbol{\alpha}$ as $\mathbf{X}_U \boldsymbol{\beta} - \boldsymbol{v}_U$ and thus $\mathbf{X}_U \boldsymbol{\beta} - \boldsymbol{v}_U = \mathbf{P}_U \boldsymbol{\gamma} + \mathbf{e}_U$, leading to

$$\mathbf{X}_U \boldsymbol{\beta} = \mathbf{P}_U \boldsymbol{\gamma} + \mathbf{u}_U + \boldsymbol{v}_U.$$

Using model F, we can express $\mathbf{X}_U \boldsymbol{\beta}$ as $\mathbf{y}_U - \boldsymbol{\epsilon}_U$ and thus $\mathbf{y}_U - \boldsymbol{\epsilon}_U = \mathbf{P}_U \boldsymbol{\gamma} + \mathbf{u}_U + \boldsymbol{v}_U$, which implies that the variable of interest for all elements in the AOI can be expressed using models F, G and Q as

$$\mathbf{y}_U = \mathbf{P}_U \boldsymbol{\gamma} + \mathbf{e}_U + \boldsymbol{v}_U + \boldsymbol{\epsilon}_U.$$

The AGBD mean for U can thus be expressed as

$$\bar{y}_U = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{p}_i \boldsymbol{\gamma} + e_i + v_i + \epsilon_i) = \bar{p}_U \boldsymbol{\gamma} + \bar{e}_U + \bar{v}_U + \bar{\epsilon}_U.$$

The MSE of the predictor $\hat{\bar{y}}_{U_{3pHMB}}$ is then, according to the definition of MSE,

$$MSE\left(\hat{\bar{y}}_{U_{3pHMB}}\right) = E\left[\left(\hat{\bar{y}}_{U_{3pHMB}} - \bar{y}_U\right)^2\right]$$

$$= E\left[\left(\bar{p}_U \hat{\boldsymbol{\gamma}} - \bar{p}_U \boldsymbol{\gamma} - \bar{e}_U - \bar{v}_U - \bar{\epsilon}_U\right)^2\right].$$

Under an assumption of independence between the variability terms, the MSE can be expressed as

$$MSE\left(\hat{\bar{y}}_{U3pHMB}\right) = \bar{p}_U Cov(\hat{\boldsymbol{\gamma}})\bar{p}_U^T + \frac{\theta^2}{N} + \frac{\delta^2}{N} + \frac{\omega^2}{N} = V\left(\hat{\bar{y}}_{U3pHMB}\right) + \frac{\theta^2}{N} + \frac{\delta^2}{N} + \frac{\omega^2}{N}. \tag{9}$$

However, note that it is not obvious that the three variability terms are independent, and typically there is also spatial autocorrelation among them. Thus, MSE expressions become complicated and require further investigation.
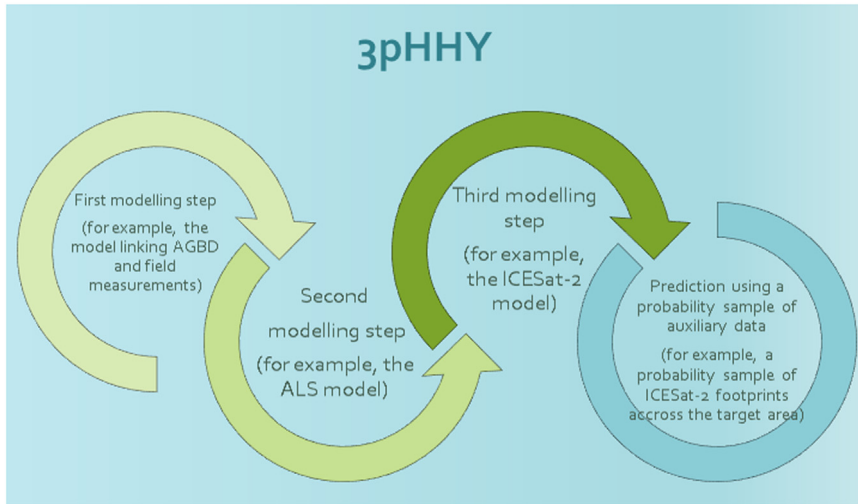
Fig. 2. A graphical overview on the 3pHHY prediction based on a probability sample of auxiliary (remotely sensed) data.

Details of three-phase hierarchical hybrid inference (3pHHY)

Next, we present our second case, i.e., when the last step auxiliary information is available only for a probability sample from the AOI, rather than wall-to-wall. This step is the only difference compared to 3pHMB inference. In our forest inventory example, the auxiliary information in this case could be retrieved from the ICESat-2 spaceborne LiDAR.

Thus, $\bar{p}_U$ in this case is estimated using the design-based estimator (e.g., [30], p. 42)

$$\hat{\bar{p}}_U = \frac{1}{N} \sum_{S_{IV}} \frac{p_k}{\pi_k},$$

where S_{IV} is a probability sample drawn from the AOI, p_k is a vector of ICESat-2 auxiliary values for the kth element in the sample, and π_k is the probability of including the kth element to the sample S_{IV}.

The design-based expectation of $\hat{\bar{p}}_U$ is

$$E_D\left[\hat{\bar{p}}_U\right] = \bar{p}_U.$$

The 3pHHY predictor of the AGBD mean over the AOI is then (cf., [31–33])

$$\hat{\bar{y}}_{U3pHHY} = \frac{1}{N} \sum_{S_{IV}} \frac{p_k}{\pi_k}\hat{\gamma}. \tag{10}$$

Fig. 2 gives a graphical overview of 3pHHY prediction.

The 3pHHY predictor variance and its estimator

To derive the variance of the 3pHHY predictor, we begin with the expectation of the predictor, which can be decomposed as

$$E\left[\hat{\bar{y}}_{U3pHHY}\right] = E\left[\frac{1}{N} \sum_{S_{IV}} \frac{p_k}{\pi_k}\hat{\gamma}\right] = E_D\left[E_Q\left[\frac{1}{N} \sum_{S_{IV}} \frac{p_k}{\pi_k}\hat{\gamma}\right]\right]$$

$$= E_D\left[\frac{1}{N} \sum_{S_{IV}} \frac{p_k}{\pi_k}E_Q[\hat{\gamma}]\right] = \bar{p}_U E_Q[\hat{\gamma}] = \bar{p}_U \mathbf{\Gamma}^*\mathbf{A}^*\boldsymbol{\beta} = \bar{p}_U \gamma,$$

where the subscript D denotes expectation due to the design.

The variance of the 3pHHY predictor can be decomposed using the law of total variance as

$$V\left(\hat{\bar{y}}_{U3pHHY}\right) = V\left(\frac{1}{N} \sum_{S_{IV}} \frac{p_k}{\pi_k}\hat{\gamma}\right)$$

$$= V_D\left(E_Q\left[\frac{1}{N} \sum_{S_{IV}} \frac{p_k}{\pi_k}\hat{\gamma}\right]\right) + E_D\left[V_Q\left(\frac{1}{N} \sum_{S_{IV}} \frac{p_k}{\pi_k}\hat{\gamma}\right)\right]$$

$$= V_D\left(\frac{1}{N}\sum_{S_{IV}}\frac{p_k}{\pi_k}\gamma\right) + E_D\left[\frac{1}{N}\sum_{S_{IV}}\frac{p_k}{\pi_k}Cov(\gamma)\left(\frac{1}{N}\sum_{S_{IV}}\frac{p_k}{\pi_k}\right)^T\right]$$

$$= \gamma^T Cov\left(\frac{1}{N}\sum_{S_{IV}}\frac{p_k}{\pi_k}\right)\gamma + \bar{p}_U Cov(\hat{\gamma})\bar{p}_U^T + Tr\left[Cov\left(\frac{1}{N}\sum_{S_{IV}}\frac{p_k}{\pi_k}\right)Cov(\hat{\gamma})\right]. \tag{11}$$

The third term in the final expression of Eq. (11) is due to

$$E_D\left[\frac{1}{N}\sum_{S_{IV}}\frac{p_k}{\pi_k}\left(\frac{1}{N}\sum_{S_{IV}}\frac{p_k}{\pi_k}\right)^T\right] = \bar{p}_U\bar{p}_U^T + Cov\left(\frac{1}{N}\sum_{S_{IV}}\frac{p_k}{\pi_k}\right).$$

Our simulation analysis showed that the third term, $Tr[Cov(\frac{1}{N}\sum_{S_{IV}}\frac{p_k}{\pi_k})Cov(\hat{\gamma})]$, in Eq. (11) is small and may be neglected. Thus, for any given sampling design the 3pHHY variance is

$$V\left(\hat{\bar{y}}_{U3pHHY}\right) = \gamma^T Cov\left(\frac{1}{N}\sum_{S_{IV}}\frac{p_k}{\pi_k}\right)\gamma + \bar{p}_U Cov(\hat{\gamma})\bar{p}_U^T. \tag{12}$$

Under a simple random sampling without replacement design, the variance is

$$V\left(\hat{\bar{y}}_{U3pHHY}\right) = \frac{1}{n_{IV}}\left(1 - \frac{n_{IV}}{N}\right)\gamma^T Cov(\boldsymbol{P}_U)\gamma + \bar{p}_U Cov(\hat{\gamma})\bar{p}_U^T, \tag{13}$$

where n_{IV} is the sample size, $(1 - \frac{n_{IV}}{N})$ is the finite population correction factor, and $Cov(\boldsymbol{P}_U)$ is the population covariance of the ICESat-2 explanatory variables. In the simulations performed to validate the results, Eq. (13) was used, since simple random sampling was assumed.

By replacing the covariance of estimated model parameters, $Cov(\hat{\gamma})$, the model parameters, γ, and the design-based covariance $Cov(\frac{1}{N}\sum_{S_{IV}}\frac{p_k}{\pi_k})$ with their corresponding estimators we obtain a variance estimator. Our simulation analysis showed that the variance estimator is approximately unbiased.

If we would like to derive the MSE of the 3pHHY predictor, we may proceed along the route previously outlined for 3pHMB, i.e.

$$MSE\left(\hat{\bar{y}}_{U3pHHY}\right) = E\left[\left(\hat{\bar{y}}_{U3pHHY} - \bar{y}_U\right)^2\right]$$

$$= E\left[\left(\frac{1}{N}\sum_{S_{IV}}\frac{p_k}{\pi_k}\hat{\gamma} - \bar{p}_U\gamma - \bar{e}_U - \bar{v}_U - \bar{\in}_U\right)^2\right]$$

$$= E\left[\left(\frac{1}{N}\sum_{S_{IV}}\frac{p_k}{\pi_k}\hat{\gamma} - \bar{p}_U\hat{\gamma} + \bar{p}_U\hat{\gamma} - \bar{p}_U\gamma - \bar{e}_U - \bar{v}_U - \bar{\in}_U\right)^2\right]$$

$$= E\left[\left(\left(\frac{1}{N}\sum_{S_{IV}}\frac{p_k}{\pi_k} - \bar{p}_U\right)\hat{\gamma} + \bar{p}_U(\hat{\gamma} - \gamma) - \bar{e}_U - \bar{v}_U - \bar{\in}_U\right)^2\right]$$

$$= \gamma^T Cov\left(\frac{1}{N}\sum_{S_{IV}}\frac{p_k}{\pi_k}\right)\gamma + \bar{p}_U Cov(\hat{\gamma})\bar{p}_U^T + Tr\left[Cov\left(\frac{1}{N}\sum_{S_{IV}}\frac{p_k}{\pi_k}\right)Cov(\hat{\gamma})\right] + \frac{\theta^2}{N} + \frac{\delta^2}{N} + \frac{\omega^2}{N}$$

$$= V\left(\hat{\bar{y}}_{U3pHHY}\right) + \frac{\theta^2}{N} + \frac{\delta^2}{N} + \frac{\omega^2}{N}. \tag{14}$$

Like in the case of 3pHMB, this MSE formula assumes independence between the variability terms and absence of spatial auto-correlation. Thus, deriving a useful MSE expression is complicated and requires further investigation.

Study of hierarchical modeling in the context of a superpopulation model

The previous sections of this article have provided the formulas needed for applying 3pHMB and 3pHHY for prediction, and for assessing uncertainties. In this section, we study properties of hierarchical modeling that cannot be straightforwardly deduced from the formulas presented in the previous sections. We do this by specifying a superpopulation model with normally distributed random variables. Some of the non-standard models from the previous sections, e.g., Eqs. (2) and (3), will now be derived through conditional expectation, thus validating the previously presented models. This section will also provide insights into effects of hierarchical modeling that are important to understand, such as the reduced variability of the response variable the more hierarchically nested models are added. Examples of the magnitude of such variability reduction are provided.

We assume that with each population element a multivariate random variable is associated, i.e., for the ith observation, the values (y_i, x_i, z_i, p_i) are outcomes from a multivariate random variable (Y_i, X_i, Z_i, P_i). The joint distribution of the random variables for the elements in the population defines a superpopulation model denoted ξ ([34], p. 80).

The superpopulation model ξ shapes the relationship between the variables in the population, and the observations linked to them. We assume the same type of population elements, variables and observations as previously, i.e. y_i is the AGBD measured in the field, x_i is some field plot measurements, such as average tree diameter (DBH), z_i is some ALS variable, and p_i some Landsat (in the case of 3pHMB) or ICESat-2 variable (in the case of 3pHHY). The values (y_i, x_i, z_i, p_i) are realizations from the corresponding multivariate random variable (Y_i, X_i, Z_i, P_i).

We assume a multivariate normal distribution, i.e.

$$\begin{pmatrix} Y_i \\ X_i \\ Z_i \\ P_i \end{pmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\mu} = (\mu_Y, \mu_X, \mu_Z, \mu_P)$ is a vector of superpopulation means, and $\boldsymbol{\Sigma}$ is the superpopulation variance-covariance matrix with variances $\sigma^2_{(\cdot)}$ for each variable on the diagonal and covariances $\sigma_{(\cdot)(\cdot)}$ between the variables on the off-diagonal elements. We assume that our multivariate variables are exchangeable, i.e., every permutation of the order of the population elements follows the same joint distribution ξ ([34], p. 85).

We now derive the models previously presented Eqs. (1), ((2), and (3)) through the superpopulation model. We begin with describing the model linking AGBD and DBH, i.e., the model F. Following the joint distribution ξ, the conditional probability density function of Y_i, given that $X_i = x_i$ for the ith unit, is

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)},$$

where $f_{X,Y}(x,y)$ is the joint probability density function of Y and X, and $f_X(x)$ is the marginal density function for X. Under the normality assumption the conditional expectation of Y_i given $X_i = x_i$ can be expressed as ([35], p. 72)

$$E[Y_i|X_i = x_i] = E[Y_i] + \frac{Cov(Y_i, X_i)}{V(X_i)}(x_i - E[X_i]) = \mu_Y + r_{YX}\frac{\sigma_Y}{\sigma_X}(x_i - \mu_X) = \mu_{Y|X} \tag{15}$$

where r_{YX} is the correlation between Y_i and X_i, and $\sigma_{(\cdot)}$ is the superpopulation standard deviation for the given variable. The variance of the conditional distribution is ([35], p. 72)

$$Var(Y_i|X_i = x_i) = V(Y_i) - \frac{Cov^2(Y_i, X_i)}{V(X_i)} = \sigma_Y^2(1 - r_{YX}^2) = \sigma_{Y|X}^2. \tag{16}$$

The conditional distribution of Y_i, given that $X_i = x_i$ for the ith observation is then

$$Y_i|X_i = x_i \sim \mathcal{N}\left(\mu_{Y|X}, \ \sigma_{Y|X}^2\right).$$

Thus, in our example, the AGBD random variable associated with the ith population element has the expectation μ_Y and variance σ_Y^2; X_i is the DBH random variable with expectation μ_X and variance σ_X^2. However, for a fixed outcome x_i from X_i, there is a random subset of Y_i conditional on the fixed x_i; this subset is the variable $Y_i|X_i = x_i$ with expectation $\mu_{Y|X} = \mu_Y + r_{YX}\frac{\sigma_Y}{\sigma_X}(x_i - \mu_X)$ and variance $\sigma_{Y|X}^2 = \sigma_Y^2(1 - r_{YX}^2)$. This corresponds to regressing Y_i on X_i, which reveals how much information about Y_i is contained in an observation of X_i ([35], p. 73).

Therefore, using the conditional distribution, we can define our first model (F), which links AGBD and DBH as

$$F: \ Y_i = \mu_{Y|X} + \epsilon_i, \ \epsilon_i \sim \mathcal{N}\left(0, \sigma_{Y|X}^2\right), \tag{17}$$

where ϵ_i is the variability term. By denoting $\omega^2 = \sigma_{Y|X}^2$, $\beta_0 = (\mu_Y - r_{YX}\frac{\sigma_Y}{\sigma_X}\mu_X)$ and $\beta_1 = r_{YX}\frac{\sigma_Y}{\sigma_X}$ we can rewrite F in the form previously introduced, i.e.

$$F: \ Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \ \epsilon_i \sim \mathcal{N}\left(0, \omega^2\right),$$

and in linear algebra notation

$$F: \ Y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i, \ \epsilon_i \sim \mathcal{N}\left(0, \omega^2\right),$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ and $\boldsymbol{x}_i = (1, x_i)$. By substituting the variances and correlations in the parameter formulas with their corresponding estimators, we obtain the well-known formulas for parameter estimation from OLS regression.

Similarly, to derive the model G (Eq. (2)), we introduce $E[Y_i|X_i]$, which is a random variable since we now condition on the random variable X_i rather than some fixed outcome $X_i = x_i$ ([27], p. 131–132). The variance of $E[Y_i|X_i]$ is

$$V\left(E[Y_i|X_i]\right) = V\left(\mu_Y + r_{YX}\frac{\sigma_Y}{\sigma_X}(X_i - \mu_X)\right) = V\left(r_{YX}\frac{\sigma_Y}{\sigma_X}X_i\right) = r_{YX}^2\frac{\sigma_Y^2}{\sigma_X^2}V(X_i) = r_{YX}^2\frac{\sigma_Y^2}{\sigma_X^2}\sigma_X^2 = r_{YX}^2\sigma_Y^2. \tag{18}$$

The expectation of $E[Y_i|X_i]$ can be derived through the law of iterated expectations ([27], p. 132)

$$E\left[E[Y_i|X_i]\right] = E[Y_i] = \mu_Y. \tag{19}$$

The next step is to introduce the ALS variable, Z_i. Based on our superpopulation model, the conditional distribution of $E[Y_i|X_i]$ given that $Z_i = z_i$ has the expected value (cf. Eq. (15))

$$E\big[E[Y_i|X_i]|Z_i = z_i\big] = E\big[E[Y_i|X_i]\big] + \frac{Cov\big(E[Y_i|X_i], Z_i\big)}{V(Z_i)}\big(z_i - E[Z_i]\big)$$

$$= \mu_Y + r_{XZ}r_{YX}\frac{\sigma_Y}{\sigma_Z}\big(z_i - \mu_z\big) = \mu_{Y|X|Z}, \tag{20}$$

where,

$$Cov\big(E[Y_i|X_i], Z_i\big) = Cov\left(\left(\mu_Y + r_{YX}\frac{\sigma_Y}{\sigma_X}\big(X_i - \mu_X\big)\right), Z_i\right)$$

$$= Cov\left(r_{YX}\frac{\sigma_Y}{\sigma_X}X_i, \ Z_i\right) = r_{YX}\frac{\sigma_Y}{\sigma_X}Cov\big(X_i, Z_i\big)$$

$$= r_{YX}\frac{\sigma_Y}{\sigma_X}r_{XZ}\sqrt{V(X_i)V(Z_i)} = r_{XZ}r_{YX}\sigma_Y\sigma_Z.$$

The variance is (cf. Eq. (16))

$$Var\big(E[Y_i|X_i]|Z_i = z_i\big) = V\big(E[Y_i|X_i]\big) - \frac{Cov^2\big(E[Y_i|X_i], Z_i\big)}{V(Z_i)}$$

$$= r_{YX}^2\sigma_Y^2 - \frac{\big(r_{XZ}r_{YX}\sigma_Y\sigma_Z\big)^2}{\sigma_Z^2} = \sigma_Y^2 r_{YX}^2\big(1 - r_{XZ}^2\big) = \sigma_{Y|X|Z}^2 \tag{21}$$

The conditional distribution of $E[Y_i|X_i]$, given that $Z_i = z_i$ is then

$$E\big[Y_i|X_i\big]|Z_i = z_i \sim \mathcal{N}\big(\mu_{Y|X|Z}, \ \sigma_{Y|X|Z}^2\big).$$

Thus, we can define the model G:

$$\text{G}: \ E\big[Y_i|X_i\big] = \mu_{Y|X|Z} + v_i, \ v_i \sim \mathcal{N}\big(0, \sigma_{Y|X|Z}^2\big). \tag{22}$$

Denoting model parameters $\alpha_0 = (\mu_Y - r_{XZ}r_{YX}\frac{\sigma_Y}{\sigma_Z}\mu_Z)$, $\alpha_1 = r_{XZ}r_{YX}\frac{\sigma_Y}{\sigma_Z}$, and the variance of the variability term v_i as $\delta^2 = \sigma_{Y|X|Z}^2$, we can rewrite G as

$$\text{G}: \ E\big[Y_i|X_i\big] = \alpha_0 + \alpha_1 z_i + v_i, \ v_i \sim \mathcal{N}\big(0, \delta^2\big)$$

and in linear algebra notation

$$\text{G}: \ \boldsymbol{X}_i\boldsymbol{\beta} = \boldsymbol{z}_i\boldsymbol{\alpha} + v_i, \ v_i \sim \mathcal{N}\big(0, \delta^2\big),$$

where $\boldsymbol{X}_i = (1, \ X_i)$, $\boldsymbol{\alpha} = (\alpha_0, \ \alpha_1)^T$ and $\boldsymbol{z}_i = (1, \ z_i)$. This model coincides with the model presented in Eq. (2) in the previous section.

It remains to derive the model Q, (Eq. (3)) in a similar fashion. Thus, we define a new random variable, $E[E[Y_i|X_i]|Z_i]$, which has variance

$$V\big(E[E[Y_i|X_i]|Z_i]\big) = V\left(\mu_Y - r_{XZ}r_{YX}\frac{\sigma_Y}{\sigma_Z}\big(Z_i - \mu_Z\big)\right) = V\left(r_{XZ}r_{YX}\frac{\sigma_Y}{\sigma_Z}Z_i\right) = r_{XZ}^2 r_{YX}^2\sigma_Y^2, \tag{23}$$

and expectation (following the law of iterated expectations)

$$E\big[E[E[Y_i|X_i]|Z_i]\big] = E[Y_i] = \mu_Y. \tag{24}$$

The conditional distribution of $E[E[Y_i|X_i]|Z_i]$ given $P_i = p_i$ is

$$E\big[E[Y_i|X_i]|Z_i\big]|P_i = p_i \sim \mathcal{N}\big(\mu_{Y|X|Z|P}, \ \sigma_{Y|X|Z|P}^2\big),$$

where

$$E\big[E[E[Y_i|X_i]|Z_i]|P_i = p_i\big] = \mu_Y + r_{ZP}r_{XZ}r_{YX}\frac{\sigma_Y}{\sigma_P}\big(p_i - \mu_P\big) = \mu_{Y|X|Z|P}, \tag{25}$$

given that the covariance

$$Cov\big((E[E[Y_i|X_i]|Z_i]), P_i\big) = Cov\left(\left(\mu_Y - r_{XZ}r_{YX}\frac{\sigma_Y}{\sigma_Z}\big(Z_i - \mu_Z\big)\right), P_i\right) = r_{XZ}r_{YX}\frac{\sigma_Y}{\sigma_Z}Cov\big(Z_i, P_i\big) = r_{XZ}r_{YX}\sigma_Y\sigma_P,$$

and the variance is

$$V\big(E[E[Y_i|X_i]|Z_i]|P_i = p_i\big) = \sigma_Y^2 r_{YX}^2 r_{XZ}^2\big(1 - r_{ZP}^2\big) = \sigma_{Y|X|Z|P}^2. \tag{26}$$

Table 2

The expectation and variance of the response variable used in models F, G and Q.

Model	Response variable	Expectation of the response variable	Variance of the response variable
F	Y_i	μ_Y	σ_Y^2
G	$E[Y_i\|X_i]$	μ_Y	$r_{YX}^2 \sigma_Y^2$
Q	$E[E[Y_i\|X_i]\|Z_i]$	μ_Y	$r_{XZ}^2 r_{YX}^2 \sigma_Y^2$

Table 3

The decrease of the variance of the response variable in model G ($r_{YX}^2 \sigma_Y^2$) for different correlations (r_{YX}) between the explanatory and response variables in model F.

$= 0.9$	$r_{YX} = 0.8$	$r_{YX} = 0.7$	$r_{YX} = 0.6$
$0.81 \times \sigma_Y^2$	$0.64 \times \sigma_Y^2$	$0.49 \times \sigma_Y^2$	$0.36 \times \sigma_Y^2$

Thus, the model Q is

$$Q: \quad E\big[E[Y_i|X_i]|Z_i\big] = \mu_{Y|X|Z|P} + e_i, \ e_i \sim \mathcal{N}\big(0, \sigma_{Y|X|Z|P}^2\big). \tag{27}$$

Denoting the model parameters as $\gamma_0 = (\mu_Y - r_{ZP} r_{XZ} r_{YX} \frac{\sigma_Y}{\sigma_P} \mu_P)$ and $\gamma_1 = r_{ZP} r_{XZ} r_{YX} \frac{\sigma_Y}{\sigma_P}$, the variance of the variability term e_i as $\theta^2 = \sigma_{Y|X|Z|P}^2$, we can rewrite Q as

$$Q: \quad E\big[E[Y_i|X_i]|Z_i\big] = \gamma_0 + \gamma_1 p_i + e_i, \ e_i \sim \mathcal{N}\big(0, \theta^2\big),$$

and in linear algebra notation as

$$Q: \quad \boldsymbol{Z}_i \boldsymbol{\alpha} = \boldsymbol{p}_i \boldsymbol{\gamma} + e_i, \ e_i \sim \mathcal{N}\big(0, \theta^2\big),$$

where $\boldsymbol{Z}_i = (1, \ Z_i)$, $\boldsymbol{\gamma} = (\gamma_0, \ \gamma_1)^T$ and $\boldsymbol{p}_i = (1, \ p_i)$, which coincides with Eq. (3) in the previous section.

Decreased variability of the response variable

Deriving the models through conditional expectation reveals properties of 3pHMB (and 3pHHY) that cannot be immediately observed from the first section of this article. An important property to note is that as more modeling steps are included, the variability of the response variable decreases. This decreased variability (of proxy AGBD in our example) may have important implications in applications. One obvious case is if the objective is to map the AGBD distribution across a landscape. In case several hierarchically nested models have been utilized, the mapped AGBD variability in the landscape would be substantially smaller than the real variability. A potential solution to avoiding such decreased variability could be to apply calibration (e.g., [36]).

The decreased variability may also have negative implications if the objective is to predict the AGBD, not least for domains. As pointed out by Chambers and Clark [37], a means to minimize the uncertainty of model-based predictors is to ensure first-order balanced samples, meaning that the mean values of explanatory variables in the dataset used for model estimation coincide with the mean values of the explanatory variables in the target population. Although not specifically studied in this article, we hypothesize that first-order balanced sampling becomes increasingly important the more modeling steps are included. Further, whereas the most relevant uncertainty measure in connection with model-based inference would be the MSE, several studies suggest that for large areas, the relative difference between MSE and variance would be small (e.g., [33,38]) in case model-unbiased predictors are applied. However, in case the variability of the response variable is substantially reduced, it remains to be evaluated if this assertion holds.

Thus, model-based inference should be applied with caution in case the variability of the (proxy) response variable is substantially reduced compared to the real-world variability of the response variable. Below, we demonstrate how the variance of the response variable is affected by multi-step modeling (the models F, G and Q).

Table 2 gives an overview of the models and the corresponding expectation and variance of the response variable, under the previously introduced superpopulation model.

From Table 2, we can see that whereas the response variables have the same expectation, their variability around the expectation is not the same. With the increased hierarchy of conditions, the variability of the proxy target variable is decreasing in comparison to the variability of the actual target variable.

In Tables 3 and 4 we show numerical examples of the reduction of the variance of the response variable. In Table 3, this is shown for model G, i.e., for the case of two modeling steps.

From Table 3 we can see that, e.g., if the correlation between the response and explanatory variables is 0.8, the variance of the response variable in model G is 36% smaller than the original variance of the response variable; if the correlation is 0.7, about half of the variability is lost.

Table 4 gives numerical examples for the case of three modeling steps. Now, the decreased variance of the response variable is further accentuated. For example, when both correlations are 0.8, 59% of the original variance is lost; if both are 0.7, the corresponding figure is 76%.

Table 4

The variance of the response variable for model Q ($r_{XZ}^2 r_{YX}^2 \sigma_Y^2$) for different correlations between the explanatory and response variables in model F and model G (r_{YX} and r_{XZ}).

	$r_{YX} = 0.9$	$r_{YX} = 0.8$	$r_{YX} = 0.7$	$r_{YX} = 0.6$
$r_{XZ} = 0.9$	$0.66 \times \sigma_Y^2$	$0.52 \times \sigma_Y^2$	$0.40 \times \sigma_Y^2$	$0.29 \times \sigma_Y^2$
$r_{XZ} = 0.8$	$0.52 \times \sigma_Y^2$	$0.41 \times \sigma_Y^2$	$0.31 \times \sigma_Y^2$	$0.23 \times \sigma_Y^2$
$r_{XZ} = 0.7$	$0.40 \times \sigma_Y^2$	$0.31 \times \sigma_Y^2$	$0.24 \times \sigma_Y^2$	$0.18 \times \sigma_Y^2$
$r_{XZ} = 0.6$	$0.29 \times \sigma_Y^2$	$0.23 \times \sigma_Y^2$	$0.18 \times \sigma_Y^2$	$0.13 \times \sigma_Y^2$

Table 5

Input information.[2]

Description	Input values				
Superpopulation means	μ_Y	μ_X	μ_Z	μ_P	
	62.62	10.94	7.16	8.28	
Superpopulation standard deviations	sd_Y	sd_X	sd_Z	sd_P	
	49.53	6.30	4.34	5.69	
Superpopulation correlations between variables	r_{YX}	r_{XZ}	r_{ZP}		
	0.77	0.75	0.76		
Sizes of datasets involved in the analysis	n_I	n_{II}	n_{III}	n_{IV}	N
	102	943	1721	5760	38,400

[2] The input data mimics the conditions of the empirical case study presented in Varvia et al. [26]. Note that, although $n_I < n_{II} < n_{III} < n_{IV} < N$ in this case, the presented methodological framework does not require any specific dataset size relations.

Table 4 suggests that even if the correlation between explanatory variables and the response variable is fairly strong in the first two modeling steps, the last model is estimated using a response variable with substantially reduced variability compared to the response variable it is mimicking.

Methods validation

We validated the correctness of the presented variance formulas for 3pHMB and 3pHHY prediction through Monte Carlo (MC) simulation. A real-world application of 3pHHY is presented by Varvia et al. [26].

An R Markdown file is available as supplementary material to this article. It provides an R code for the simulation with a step-by-step description of how the simulations were performed. The input information mimics boreal forest conditions in the northern part of Finland. The target variable is the AGBD (Y_i). The explanatory variable for model F is DBH (X_i); in model G it is an ALS variable related to vegetation height (Z_i), and in model Q, it is an ICESat-2 variable related to vegetation height (P_i). The superpopulation means ($\mu_{(.)}$) and standard deviations ($sd_{(.)}$) of the variables and the correlations between them are presented in Table 5. The sizes of the datasets are also given in the table.

The evaluation of the performance of the proposed predictors and variance formulas was conducted following the steps described below.

- *Step 1a:* Using the input information on the superpopulation means and standard deviations, the "true" model parameters and variances of variability terms in models F, G*, G, Q* and Q were obtained following the theoretical outline presented in the section "Study of hierarchical modeling in the context of a superpopulation model".
- *Step 1b:* The explanatory variables ($\boldsymbol{X}_I$ for model F over the dataset n_I, $\boldsymbol{Z}_{II}$ for model G over the dataset n_{II}, $\boldsymbol{P}_{III}$ for model Q over the dataset n_{III}, and $\boldsymbol{P}_U$ over U) were generated based on their mean values and standard deviations following normal distributions.
- *Step 1c:* The variances of the 3pHMB and 3pHHY predictors were computed according to Eqs. (8a) and (11) based on the results from steps 1a and 1b.

Steps 1a – 1c are preparations for the MC simulation, but located outside the MC loop, since they are repeated only once. The MC iterations included the following steps:

- *Step 2a:* The variability terms for the models F, G*, G, Q* and Q were generated randomly based on Step 1a. As a consequence, the response variables for models were also generated for all elements in the datasets.
- *Step 2b:* The model parameters of the models F, G*, G, Q* and Q were estimated based on the simulated data, and applied for predicting the AOI population mean following 3pHMB inference. The predicted value was recorded for each MC iteration.
- *Step 2c:* A simple random probability sample (without replacement) was drawn from $\boldsymbol{P}_U$ and the AOI population mean was predicted following 3pHHY inference. Note that the predicted value was recorded for each MC iteration.
- *Step 2d:* Variance estimators were applied to estimate the variance of the 3pHMB and 3pHHY predictors based on the simulated outcomes from each MC iteration. The estimates were recorded for each MC iteration.

Table 6

The variance of the 3pHMB and 3pHHY predictors and their corresponding estimated variances; $V(\hat{y}_{U(\cdot)})$ is the variance of the population mean predictor based on Eqs. (8a) and (11), $V_{MC}(\hat{y}_{U(\cdot)})$ is the empirical variance of the population mean predictor from the MC iterations, and $\hat{V}(\hat{y}_{U(\cdot)})$ is the average of estimated variances across the MC iterations.

3pHMB			3pHHY		
$V(\hat{y}_{U_{3pHMB}})$	$V_{MC}(\hat{y}_{U_{3pHMB}})$	$\hat{V}(\hat{y}_{U_{3pHMB}})$	$V(\hat{y}_{U_{3pHHY}})$	$V_{MC}(\hat{y}_{U_{3pHHY}})$	$\hat{V}(\hat{y}_{U_{3pHHY}})$
10.94	10.96	10.96	11.03	11.04	11.04

Table 7

The contribution of the terms involving traces to the overall variance of 3pHMB and 3pHHY predictors.

Predictor	Predictor variance based on Eq. (8a) and (11)	Terms involving traces	Relative contribution of terms involving traces
3pHMB	10.94	2.0×10^{-3}	$1.8 \times 10^{-2}\%$
3pHHY	11.03	0.8×10^{-3}	$0.7 \times 10^{-2}\%$

Steps 2a – 2d were repeated one million times. Based on the MC simulations, the empirical variances of the 3pHMB and 3pHHY predictors were obtained and could be compared with the variances according to the results from Eqs. (8a) and (11) as well as the corresponding variance estimators.

The results of the simulations are presented in Tables 6 and 7. In addition, it was observed that the average of the predicted AOI population means over the MC iterations based on the 3pHMB and 3pHHY predictors corresponded well with the AGBD superpopulation mean, which shows that the predictors are approximately unbiased.

Table 6 shows the results from comparing the analytically derived variances of the predictors with the corresponding empirical variances from the MC iterations, and the average values of estimated variances. The Table shows that Eqs. (8a) and (11) are valid as variance formulas and that the corresponding variance estimators are approximately unbiased.

Referring to the Eqs. (8a) and (8b), it was suggested that terms with traces in (8a) could be neglected due to their small contribution to the overall variance of the 3pHMB and 3pHHY predictors. The empirical results in Table 7 validate this assertion.

Conclusions

The main purpose of this study was to present predictors, variances, and variance estimators for hierarchical model-based inference and hierarchical hybrid inference involving three modeling steps. Previous studies have proposed solutions for methods involving two modeling steps (e.g., Saarela et al. [39]). The need for methods of this kind emanates from forest resources assessment utilizing multiple sources of remotely sensed data in combination with field data. In Monte Carlo simulations, the correctness of the proposed formulas was validated. In addition, it was shown that caution should be exercised when applying several modeling steps based on models with weak correlation between the explanatory variable(s) and the response variable, since multiple modeling steps dramatically reduce the variability of the predicted target variable, which may lead to problems in some applications. For example, a map produced on the basis of multiple modeling steps would display substantially reduced variability for the target variable compared to the real variability.

Although this article has focused on applications based on remotely sensed data, the methods are general and can be applied in any discipline where empirical data are lacking but proxies can be obtained through hierarchically nested models.

Declaration of Competing Interest

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

CRediT authorship contribution statement

Svetlana Saarela: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Writing – review & editing. **Petri Varvia:** Data curation, Methodology, Writing – review & editing. **Lauri Korhonen:** Data curation, Resources, Funding acquisition, Writing – review & editing. **Zhiqiang Yang:** Software, Writing – review & editing. **Paul L. Patterson:** Methodology, Writing – review & editing. **Terje Gobakken:** Resources, Writing – review & editing. **Erik Næsset:** Resources, Methodology, Writing – review & editing. **Sean P. Healey:** Resources, Funding acquisition, Writing – review & editing. **Göran Ståhl:** Methodology, Writing – original draft, Writing – review & editing.

Data availability

The data were simulated inside the validation code provided by the authors. An R Markdown file is attached as supplementary material.

Acknowledgments

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.mex.2023.102321.

References

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

[2] A. Araza, S. de Bruin, M. Herold, S. Quegan, N. Labriere, P. Rodriguez-Veiga, V. Avitabile, M. Santoro, E.T.A. Mitchard, C.M. Ryan, O.L. Phillips, S. Willcock, H. Verbeeck, J. Carreiras, L. Hein, M.-J. Schelhaas, A.M. Pacheco-Pascagaza, P. da Conceição Bispo, G.V. Laurin, G. Vieilledent, F. Slik, A. Wijaya, S.L. Lewis, A. Morel, J. Liang, H. Sukhdeo, D. Schepaschenko, J. Cavlovic, H. Gilani, R. Lucas, A comprehensive framework for assessing the accuracy and uncertainty of global above-ground biomass maps, Remote Sens. Environ. 272 (2022) 112917, doi:10.1016/j.rse.2022.112917.

[3] R.E. McRoberts, Probability- and model-based approaches to inference for proportion forest using satellite imagery as ancillary data, Remote Sens. Environ. 114 (2010) 1017–1025, doi:10.1016/j.rse.2009.12.013.

[4] S. Saarela, A. Wästlund, E. Holmström, A.A. Mensah, S. Holm, M. Nilsson, J. Fridman, G. Ståhl, Mapping aboveground biomass and its prediction uncertainty using LiDAR and field data, accounting for tree-level allometric and LiDAR model errors, For. Ecosyst. 7 (2020) 43, doi:10.1186/s40663-020-00245-0.

[5] J. Hyyppä, H. Hyyppä, D. Leckie, F. Gougeon, X. Yu, M. Maltamo, Review of methods of small-footprint airborne laser scanning for extracting forest inventory data in boreal forests, Int. J. Remote Sens. 29 (2008) 1339–1366, doi:10.1080/01431160701736489.

[6] E. Næsset, Predicting forest stand characteristics with airborne scanning laser using a practical two-stage procedure and field data, Remote Sens. Environ. 80 (2002) 88–99, doi:10.1016/S0034-4257(01)00290-5.

[7] M.A. Wulder, J.C. White, R.F. Nelson, E. Næsset, H.O. Ørka, N.C. Coops, T. Hilker, C.W. Bater, T. Gobakken, Lidar sampling for large-area forest characterization: a review, Remote Sens. Environ. 121 (2012) 196–209, doi:10.1016/j.rse.2012.02.001.

[8] T. Gobakken, E. Næsset, R. Nelson, O.M. Bollandsås, T.G. Gregoire, G. Ståhl, S. Holm, H.O. Ørka, R. Astrup, Estimating biomass in Hedmark County, Norway using national forest inventory field plots and airborne laser scanning, Remote Sens. Environ. 123 (2012) 443–456, doi:10.1016/j.rse.2012.01.025.

[9] H.-E. Andersen, J. Strunk, H. Temesgen, Using airborne light detection and ranging as a sampling tool for estimating forest biomass resources in the Upper Tanana Valley of Interior Alaska, West. J. Appl. For. 26 (2011) 157–164, doi:10.1093/wjaf/26.4.157.

[10] S. Saarela, S. Holm, S. Healey, H.-E. Andersen, H. Petersson, W. Prentius, P. Patterson, E. Næsset, T. Gregoire, G. Ståhl, Generalized hierarchical model-based estimation for aboveground biomass assessment using GEDI and landsat data, Remote Sens. 10 (2018) 1832, doi:10.3390/rs10111832.

[11] S. Holm, R. Nelson, G. Ståhl, Hybrid three-phase estimators for large-area forest inventory using ground plots, airborne lidar, and space lidar, Remote Sens. Environ. 197 (2017) 85–97, doi:10.1016/j.rse.2017.04.004.

[12] M.A. Lefsky, D.J. Harding, M. Keller, W.B. Cohen, C.C. Carabajal, F. Del Bom Espirito-Santo, M.O. Hunter, R. de Oliveira, Estimates of forest canopy height and aboveground biomass using ICESat, Geophys. Res. Lett. 32 (2005) n/a-n/a, doi:10.1029/2005GL023971.

[13] H.A. Margolis, R.F. Nelson, P.M. Montesano, A. Beaudoin, G. Sun, H.-E. Andersen, M.A. Wulder, 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. For. Res. 45 (2015) 838–855, doi:10.1139/cjfr-2015-0006.

[14] T. Markus, T. Neumann, A. Martino, W. Abdalati, K. Brunt, B. Csatho, S. Farrell, H. Fricker, A. Gardner, D. Harding, M. Jasinski, R. Kwok, L. Magruder, D. Lubin, S. Luthcke, J. Morison, R. Nelson, A. Neuenschwander, S. Palm, S. Popescu, C. Shum, B.E. Schutz, B. Smith, Y. Yang, J. Zwally, The Ice, Cloud, and land Elevation Satellite-2 (ICESat-2): science requirements, concept, and implementation, Remote Sens. Environ. 190 (2017) 260–273, doi:10.1016/j.rse.2016.12.029.

[15] L. Narine, L. Malambo, S. Popescu, Characterizing canopy cover with ICESat-2: a case study of southern forests in Texas and Alabama, USA, Remote Sens. Environ. 281 (2022) 113242, doi:10.1016/j.rse.2022.113242.

[16] S.C. Popescu, T. Zhou, R. Nelson, A. Neuenschwander, R. Sheridan, L. Narine, K.M. Walsh, Photon counting LiDAR: an adaptive ground and canopy height retrieval algorithm for ICESat-2 data, Remote Sens. Environ. 208 (2018) 154–170, doi:10.1016/j.rse.2018.02.019.

[17] P. Varvia, L. Korhonen, A. Bruguière, J. Toivonen, P. Packalen, M. Maltamo, S. Saarela, S.C. Popescu, How to consider the effects of time of day, beam strength, and snow cover in ICESat-2 based estimation of forest biomass? Remote Sens. Environ. 280 (2022) 113174, doi:10.1016/j.rse.2022.113174.

[18] M.A. Wulder, J.C. White, C.W. Bater, N.C. Coops, C. Hopkinson, G. Chen, Lidar plots — A new large-area data collection option: context, concepts, and case study, Can. J. Remote Sens. 38 (2012) 600–618, doi:10.5589/m12-049.

[19] Z. Hou, Q. Xu, R.E. McRoberts, J.A. Greenberg, J. Liu, J. Heiskanen, S. Pitkänen, P. Packalen, Effects of temporally external auxiliary data on model-based inference, Remote Sens. Environ. 198 (2017) 150–159, doi:10.1016/j.rse.2017.06.013.

[20] T.A.B. Snijders, R.J. Bosker, Multilevel analysis: an Introduction to Basic and Advanced Multilevel Modeling, 2. ed., SAGE, Los Angeles, Calif, 2012 ed.

[21] D.P. Roy, M.A. Wulder, T.R. Loveland, C.E W, R.G. Allen, M.C. Anderson, D. Helder, J.R. Irons, D.M. Johnson, R. Kennedy, T.A. Scambos, C.B. Schaaf, J.R. Schott, Y. Sheng, E.F. Vermote, A.S. Belward, R. Bindschadler, W.B. Cohen, F. Gao, J.D. Hipple, P. Hostert, J. Huntington, C.O. Justice, A. Kilic, V. Kovalskyy, Z.P. Lee, L. Lymburner, J.G. Masek, J. McCorkel, Y. Shuai, R. Trezza, J. Vogelmann, R.H. Wynne, Z. Zhu, Landsat-8: science and product vision for terrestrial global change research, Remote Sens. Environ. 145 (2014) 154–172, doi:10.1016/j.rse.2014.02.001.

[22] M.A. Wulder, J.C. White, T.R. Loveland, C.E. Woodcock, A.S. Belward, W.B. Cohen, E.A. Fosnight, J. Shaw, J.G. Masek, D.P. Roy, The global Landsat archive: status, consolidation, and direction, Remote Sens. Environ. 185 (2016) 271–283, doi:10.1016/j.rse.2015.11.032.

[23] R.E. McRoberts, Q. Chen, G.M. Domke, G. Ståhl, S. Saarela, J.A. Westfall, Hybrid estimators for mean aboveground carbon per unit area, For. Ecol. Manag. 378 (2016) 44–56, doi:10.1016/j.foreco.2016.07.007.

[24] J. Repola, Biomass equations for Scots pine and Norway spruce in Finland, Silva Fenn 43 (2009), doi:10.14214/sf.184.

[25] J. Repola, Biomass equations for birch in Finland, Silva Fenn 42 (2008), doi:10.14214/sf.236.

[26] Varvia, P., Saarela, S., Maltamo, M., Packalen, P., Gobakken, T., Næsset, E., Ståhl, G., Korhonen, L., 2023. Estimation of boreal forest biomass from ICESat-2 data using hierarchical hybrid inference. https://doi.org/10.48550/ARXIV.2307.04497

[27] R. Davidson, J.G. MacKinnon, Estimation and Inference in Econometrics, Oxford University Press, 1993.

[28] S. Saarela, S. Holm, S.P. Healey, P.L. Patterson, Z. Yang, H.-E. Andersen, R.O. Dubayah, W. Qi, L.I. Duncanson, J.D. Armston, T. Gobakken, E. Næsset, M. Ekström, G. Ståhl, Comparing frameworks for biomass prediction for the global ecosystem dynamics investigation, Remote Sens. Environ. 278 (2022) 113074, doi:10.1016/j.rse.2022.113074.

[29] K.T. Wallenius, A conditional covariance formula with applications, Am. Stat. 25 (1971) 32–33.

[30] C.-E. Särndal, B. Swensson, J. Wretman, Model Assisted Survey Sampling, Springer Science & Business Media, 2003.

[31] G. Ståhl, J. Heikkinen, H. Petersson, J. Repola, S. Holm, Sample-based estimation of greenhouse gas emissions from forests—A new approach to account for both sampling and model errors, For. Sci. 60 (2014) 3–13, doi:10.5849/forsci.13-005.

[32] G. Ståhl, S. Holm, T.G. Gregoire, T. Gobakken, E. Næsset, R. Nelson, Model-based inference for biomass estimation in a LiDAR sample survey in Hedmark County, NorwayThis article is one of a selection of papers from extending forest inventory and monitoring over space and time, Can. J. For. Res. 41 (2011) 96–107, doi:10.1139/X10-161.

[33] G. Ståhl, S. Saarela, S. Schnell, S. Holm, J. Breidenbach, S.P. Healey, P.L. Patterson, S. Magnussen, E. Næsset, R.E. McRoberts, T.G. Gregoire, Use of models in large-area forest surveys: comparing model-assisted, model-based and hybrid estimation, For. Ecosyst. 3 (2016) 5, doi:10.1186/s40663-016-0064-9.

[34] Cassel, C.-M., Sarndal, C.-E., Wretman, J.H., 1977. Foundations of inference in survey sampling.

[35] W. Feller, An introduction to probability theory and its applications, Wiley Ser. Probab. Math. Stat. 1 (1971) 343–366.

[36] N. Lindgren, K. Nyström, S. Saarela, H. Olsson, G. Ståhl, Importance of calibration for improving the efficiency of data assimilation for predicting forest characteristics, Remote Sens. 14 (2022) 4627, doi:10.3390/rs14184627.

[37] R. Chambers, R. Clark, An Introduction to Model-Based Survey Sampling With Applications, Oxford University Press, 2012.

[38] P.L. Patterson, S.P. Healey, G. Ståhl, S. Saarela, S. Holm, H.-E. Andersen, R.O. Dubayah, L. Duncanson, S. Hancock, J. Armston, J.R. Kellner, W.B. Cohen, Z. Yang, Statistical properties of hybrid estimators proposed for GEDI—NASA's global ecosystem dynamics investigation, Environ. Res. Lett. 14 (2019) 065007, doi:10.1088/1748-9326/ab18df.

[39] S. Saarela, S. Holm, A. Grafström, S. Schnell, E. Næsset, T.G. Gregoire, R.F. Nelson, G. Ståhl, Hierarchical model-based inference for forest inventory utilizing three sources of information, Ann. For. Sci. 73 (4) (2016) 895–910.