Remote Sensing of Environment 288 (2023) 113488

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

# Harnessing data assimilation and spatial autocorrelation for forest inventory

Qing Xu [a], Bo Li [b], Ronald E. McRoberts [c], Zengyuan Li [d,e], Zhengyang Hou [f,*]

[a] Key Laboratory of National Forestry and Grassland Administration/Beijing for Bamboo & Rattan Science and Technology, International Center for Bamboo and Rattan, Beijing 100102, China
[b] Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA
[c] Department of Forest Resources, University of Minnesota, St. Paul, MN, USA
[d] Institute of Forest Resource Information Techniques, Chinese Academy of Forestry, Beijing 100091, China
[e] Key Laboratory of Forestry Remote Sensing and Information System, National Forestry and Grassland Administration, Beijing 100091, China
[f] The Key Laboratory for Silviculture and Conservation of Ministry of Education, Beijing Forestry University, Beijing 100083, China

ARTICLE INFO

Edited by Dr. Marie Weiss

*Keywords:*
Prediction variance
Variance decomposition
Spatially structured
Heterogeneous variance
Spatial autocorrelation
Data assimilation

ABSTRACT

Spatially explicit uncertainties in forest above-ground biomass predictions for population units are underestimated if spatial structure in the form of residual spatial autocorrelation and heteroscedasticity is ignored. Methods that consider the spatial structure of biomass model residuals are needed to comprehensively estimate, as well as to effectively reduce, the uncertainty in biomass predictions, for pursuing higher levels of precision for measurement, reporting and verification of forest carbon stocks. The objectives of the study were threefold: (1) to demonstrate a spatial data assimilation (DA) procedure that harnesses small-footprint airborne LiDAR, the best linear unbiased predictor (BLUP) and the spatial structure of biomass model residuals to reduce prediction variances of individual tree biomass and plot-level biomass density; (2) to derive a variance estimator that decomposes the variance into components associated with corresponding error sources; and (3) to compare prediction variances for three methods used to calibrate a height-based allometric model for tree biomass: ordinary least squares (OLS), generalized least squares (GLS), and spatial DA using the BLUP. Five major conclusions are drawn. First, for individual tree biomass predictions, spatial DA decreased prediction variance by 40% and 20% relative to OLS and GLS. Because the decrease in residual variability accounted for 98% of the decrease in prediction variance in total, the assimilation effect was the largest for reducing residual variability. Second, for biomass density predictions, DA decreased prediction variance by 3% and 49% relative to OLS and GLS, with the largest decrease in residual covariance. Accumulated gain in precision for individual tree predictions in the DA procedure was offset by precision loss caused by residual covariance and the variance associated with omission and commission errors while predicting individual tree biomass from the LiDAR data. Third, OLS, which assumed no spatial structure, underestimated prediction variance for LiDAR-predicted biomass density by 48%. Fourth, from the perspective of prediction accuracy, DA reduced the RMSE for individual tree biomass predictions by 11% and 14% and reduced the RMSE for biomass density predictions by 28% and 33% relative to OLS and GLS. Fifth, the omission/commission difference model was effective for correcting the systematic prediction error in the LiDAR-predicted biomass density. Overall, the proposed spatial DA procedure demonstrated great potential for reducing the uncertainty in forest biomass predictions, thereby facilitating more efficient biomass inventories. The procedure can be generalized to other dependent variables of interest given their correlations with new information from LiDAR.

## 1. Introduction

Precise predictions of forest above-ground biomass (AGB) are critical to understanding changes in forest carbon stocks for meeting the net-zero carbon emissions target (IPCC, 2018; Fankhauser et al., 2022). A variety of terrestrial, airborne and spaceborne sources of remotely sensed data have been used with model-based inference to predict forest biomass population parameters such as the mean per unit area and the total for entire investigated areas (Chen et al., 2016; McRoberts et al., 2013; Ståhl et al., 2011, 2016). Often the model-based predictions for

* Corresponding author at: The Key Laboratory for Silviculture and Conservation of Ministry of Education, Beijing Forestry University, Beijing 100083, China.
*E-mail address:* houzhengyang@bjfu.edu.cn (Z. Hou).

https://doi.org/10.1016/j.rse.2023.113488
Received 12 April 2022; Received in revised form 13 December 2022; Accepted 25 January 2023
Available online 21 February 2023

population parameters were more precise than those obtained using design-based inference (Hou et al., 2018). While the uncertainty in forest biomass predictions has been estimated at the population scale across various forest ecosystems, requests to map spatially explicit uncertainties at finer scales are increasingly raised.

For a variable of interest, model-based inference assumes a super-population distribution for a population unit (typically a sample plot) denoted by $i$, or a joint distribution for all population units. Remotely sensed variables $X_i$ are used as auxiliary data to estimate the mean $\mu_i$ and variance $\sigma_i^2$ of the unit-level distribution, as well as the variance of $\widehat{\mu}_i$. A random realization of the distribution is denoted by $y_i$, a value once observed in a sample plot. To infer the mean for the entire investigated area and its variance, parametric and non-parametric regression models have been proposed (Gregoire et al., 2011; McRoberts et al., 2007, 2018; Saarela et al., 2015). The precision of a prediction is usually characterized by variance or mean squared error (MSE) that indicates a deviance from the predicted value to either the expected value or the "true" value. However, a constant and usually small variance for the predicted areal mean may conceal prediction variances that vary for individual population units as contributions of error sources change. Maps depicting the spatial distribution of prediction uncertainties (Persson et al., 2022; Petersson et al., 2017; Saarela et al., 2020) matter because forest sites with larger uncertainties draw opportunities for improvement. After all, a decrease in prediction variance for individual population units may contribute to an increase in estimation precision for population parameters such as the mean or total for the entire investigated area.

To estimate spatially explicit uncertainties for population units, efforts have been initiated to scale up variance components associated with different error sources across spatial scales. While errors associated with model predictions dominate the uncertainty at tree, plot, and population scales, errors in predicting tree heights and number of stems from LiDAR data were found to contribute less to the uncertainty in biomass predictions (Chen et al., 2015; Saarela et al., 2020; Xu et al., 2018). Nevertheless, prediction variances for population units might still be underestimated if allometric models for individual tree biomass do not accommodate the spatial structure of model residuals (Magnussen et al., 2017; McRoberts, 2010). Spatial autocorrelation and heteroscedasticity are common spatial characteristics of the tree-level biomass residuals. Heteroscedasticity is not a spatial characteristic by definition; however, causes of heteroscedasticity include group effects, which may associate with distinct spatial processes for spatial variables. Neglecting these features can lead to biased estimators for residual variances and covariances. When non-zero correlations between trees are erroneously assumed to be zero, residual covariances are not incorporated in the prediction variance of the larger-scale biomass density. To pursue higher levels of precision for measurement, reporting and verification of forest carbon stocks, methods that consider the spatial structure of biomass prediction residuals, particularly spatial autocorrelation and heteroscedasticity, are needed to comprehensively estimate, as well as to effectively reduce, the uncertainty in biomass predictions.

Data assimilation (DA) can be used in combination with the spatial structure of biomass prediction residuals to minimize spatially explicit uncertainties in remote sensing-predicted forest biomass for population units. DA has its mathematical roots in Bayes' theorem which expresses the conditional distribution of a response variable on new information (observations or predictions) from additional sources. DA consists of a group of data fusion methods that combine primitive model predictions with new information to produce a prediction expected to be more precise than the prediction obtained using either the model or the new information alone. Multiple types of DA procedures have been developed for both temporally sequential and temporally invariant applications. Derived as a best linear predictor (BLP), the Kalman filter and its variants are representatives of temporally sequential DA techniques that focus on updating a series of model predictions along a temporal axis (Czaplewski, 1990; Ehlers et al., 2013; Kalman, 1960; Kangas et al., 2020). The best linear unbiased predictor (BLUP) was used by Hou et al. (2019) as a temporal invariant DA technique that calibrated model predictions of forest response variables at a single point in time with observations of other variables correlated with response variables. A diversity of remotely sensed data such as digital aerial images, TanDEM-X InSAR images, SPOT-5 multispectral images and airborne LiDAR data have been recently used as sources of auxiliary information in DA procedures to improve prediction precision of forest response variables for population units (Ehlers et al., 2018; Lindgren et al., 2017; Nyström et al., 2015).

BLUP establishes a theoretical foundation for spatial DA procedures. Although BLUP is mostly used to predict random effects for mixed-effects models, it can be used to derive the Kalman filter, the method of Kriging, credibility theory, and composite estimators (Breidenbach et al., 2018; Cressie, 1990; Gregoire and Walters, 1988; Robinson, 1991). BLUP adjusts the conditional mean of $\boldsymbol{y}_1$ on the observation of $\boldsymbol{y}_2$, for the multivariate normal vector $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2)^T$ (Henderson, 1975). Forest biomass is often assumed to be lognormally distributed because a lognormal distribution has a range from zero to infinity, eliminating the probability to observe negative biomass (White, 1978). For a random variable of tree biomass denoted by $b$, let $y = ln\,(b)$ so that $y$ is normally distributed. Allometric relationship between tree biomass and biophysical attributes is often expressed using exponential or power functions. It usually requires a logarithmic transformation of tree biomass to linearize the relationship before linear model-fitting techniques are applied (Chave et al., 2005; Mascaro et al., 2013). To satisfy the requirement for the normal vector $\boldsymbol{y}$, BLUP is carried out at the logarithmic scale of tree biomass to calibrate predictions of allometric models with predicted attributes for individual trees from fine-resolution remotely sensed data such as small-footprint airborne LiDAR (Vauhkonen et al., 2010; Xu et al., 2014).

In this study, we focus on reducing the spatially explicit prediction variances of forest biomass for population units, by reducing and upscaling individual tree prediction variances to the plot level. The objectives were threefold: (1) to demonstrate a spatial DA procedure that harnesses small-footprint airborne LiDAR, the BLUP and the spatial structure of biomass prediction residuals to reduce prediction variance for tree biomass and plot-level biomass density; (2) to derive a variance estimator for biomass density that decomposes prediction variance into components associated with five error sources: LiDAR height measurements, tree identification algorithms, uncertainty in model parameter estimates, variance and covariance of model residuals; and (3) to compare prediction variances for three methods used to calibrate the height-based allometric model: ordinary least squares (OLS), generalized least squares (GLS), and spatial DA using the BLUP. We envisaged a development whereby small-footprint LiDAR plays a crucial role in providing implicit measures of spatial autocorrelation and heteroscedasticity of biomass prediction residuals while generating explicit predictions for tree-level biophysical attributes.

## 2. Materials

### 2.1. Study area

Lassen National Forest is a US National Forest (NF) of 4300 km$^2$ in northern California. It is bounded by the Sierra Nevada Mountain range to the south, the Modoc Plateau to the east, and California's Central Valley to the west. Characterized by a variety of topographic features, the forest has a diversified tree species composition including Coast Douglas-fir (*Pseudotsuga menziesii*), Ponderosa Pine (*Pinus ponderosa*), Jeffrey Pine (*Pinus jeffreyi*), Lodgepole pine (*Pinus contorta*), White Fir (*Abies concolor*) and Red Fir (*Abies magnifica*). The forest is a major source of lumber products and is managed for timber, forage, water, minerals and other resources. The Lassen NF is included in the California Forest Legacy Program for Carbon Sequestration which was developed to protect environmentally important forestlands that contribute

significantly to the storage and sequestration of carbon, thereby mitigating the effects of global warming.

### 2.2. Field data

To collect detailed field measurements of trees, a field campaign was conducted in the Lassen NF by the USDA Forest Service (PNW-FIA Field Manual, 2015) during the summers of 2013 and 2014. Tree-level biophysical attributes were collected for 8313 trees on 144 randomly located sample plots (Fig. 1) within multiple ecological strata. Sample plots were circular with radii of 16.93 m and were distributed in pairs with centers 100 m apart. Both attribute variables and spatial coordinates were collected for each individual tree. First, locations of plot centers were recorded using a differential GPS unit. Stem mapping was implemented by measuring the distance and azimuth from plot center to each stem center. Coordinates were determined in the office for each stem using triangular geometry. Second, among other attributes, tree species, height, diameter at breast height (DBH, 1.37 m) and crown widths were measured for all trees with height $\geq$ 2 m. Descriptive statistics based on DBH and height of the tallied trees exhibit a considerable variability in the Lassen NF tree size, including some trees with diameters as large as 2 m (Table 1).

Instead of weighing dry mass via destructive sampling, the field-truth tree biomasses were predicted with the published species-specific allometric models using DBH or DBH and height as predictor variables. However, model predictions lack residual variability of the applied allometric models by neglecting how much the actual biomass is different from the regression line (Chojnacky et al., 2014). To better mimic the actual biomass for trees in the Lassen NF, a four-step procedure was followed to account for the underestimated randomness. First, for trees of a species, allometric models were selected from the GlobAllomeTree database (GlobAllomeTree, 2013). Second, for a linearized allometric model $w = \widehat{w} + \varepsilon, \varepsilon \sim N(0, {\sigma_\varepsilon}^2)$, residual variance was estimated from the reported $R^2$ and $var(\widehat{w})$ using the equation ${\sigma_\varepsilon}^2 = \left(\left(1 - R^2\right)/R^2\right) \times var(\widehat{w})$. $var(\widehat{w})$ was estimated from the California forest inventory and analysis data (Appendix A in Xu et al., 2018). Third, for each tree of the species, a random residual sampled in a parametric bootstrap was added to $\widehat{w}$, and 1000 iterations essentially constructed a distribution of biomass for each tree of the species. Fourth, the mean of the distribution was calculated as the field-truth tree biomass, the result of a synthesis that used the actual tree size, and therefore regarded as a semi-synthesis. The semi-synthetic AGBs for trees measured in the Lassen NF were used as the field truth in section 3.2 for building a height-based allometric model that would allow biomass prediction from the LiDAR-predicted tree height.

**Table 1**
Tree size statistics of the Lassen National Forest field dataset.

| Tree dimension | Min | Mean | Max | SD |
|---|---|---|---|---|
| Height (m) | 2.13 | 12.63 | 61.57 | 9.06 |
| DBH (cm) | 5.08 | 24.65 | 206.80 | 19.74 |

### 2.3. Airborne LiDAR data

Airborne LiDAR data were collected for the Lassen NF in October 2013 using an ALS50 LiDAR system mounted in a Cessna 208-B Grand Caravan flown at 900 m above ground level. With a scan angle of 15 degrees from nadir (30-degree field of view), a laser pulse rate of 105 kHz and a sidelap of 60%, the LiDAR system settings and flight parameters yielded fine resolution data of >8 pulses/m$^2$ over terrestrial surfaces. Multiple echoes were recorded for each pulse.

Preprocessing of LiDAR point clouds consisted of point classification,

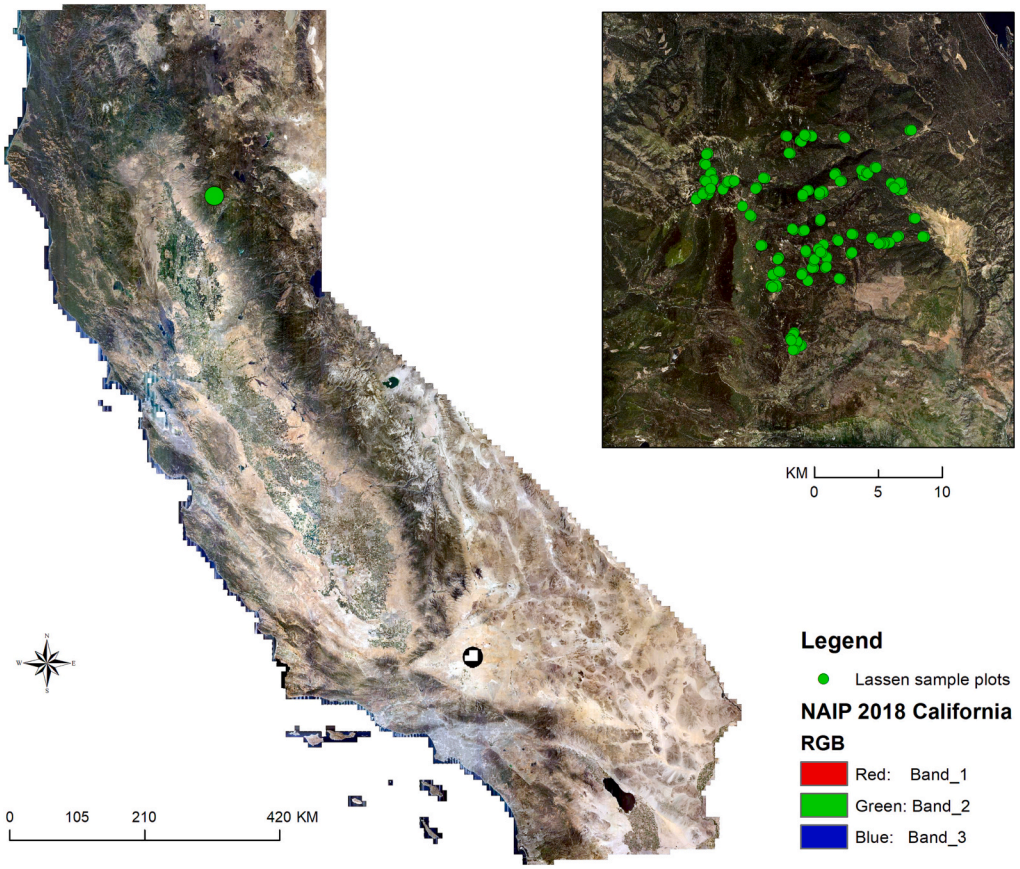


**Fig. 1.** Lassen National Forest in California with 144 sample plots in the inset.

generation of a canopy height model (CHM) and dynamic smoothing of the raw CHM. The progressive triangular irregular network (TIN) densification algorithm by Axelsson (1999) was used to classify raw LiDAR points into ground points and non-ground points. A digital terrain model (DTM) was constructed with the ground points using the TIN approach (Isenburg et al., 2006). Forest heights were predicted by subtracting the DTM from orthometric heights of raw points and were filtered by removing multi-return points and points below 2 m. Buildings in forests were identified and removed using planarity algorithm. The pit-free CHM algorithm proposed by Khosravipour et al. (2014) was applied to construct a raw CHM at the resolution of 0.25 m, which was later smoothed by a fully dynamic Gaussian filter (Xu et al., 2018), with kernel size and sigma determined by the relative height of each focal pixel in a neighborhood.

## 3. Methods

### 3.1. Overview

Based on a spatial model for individual tree biomass fitted in section 3.2 and a set of tree attributes predicted from airborne LiDAR in section 3.3, this study demonstrated a spatial DA procedure in section 3.4 which utilized the method of BLUP to improve biomass prediction precision. To understand the effects of DA, BLUP was compared with the methods of OLS and GLS in terms of variance of the predicted individual tree biomass using the Taylor series expansion in section 3.5, as well as variance of the predicted biomass density in section 3.6.

### 3.2. A spatial height-based allometric model for individual trees using generalized least squares

Since allometric models of the relationship between biomass of individual trees and measurements of tree attributes are often expressed using exponential functions (Jenkins et al., 2003) that can be linearized by taking the natural logarithm of both sides, a similar model (Eq. 1) that related biomass to tree height was formulated using the Lassen NF field measurements,

$$ln(b_i) = \beta_0 + \beta_1 \sqrt{h_i} + e_i \tag{1}$$

where $b_i$ is the semi-synthetic AGB used as the field truth, $h_i$ is the field-measured height of the $i$-th tree with $i = 1, 2, 3, \ldots, k$, $k$ is the total number of the field-measured trees, $\beta_0$ and $\beta_1$ are regression coefficients to be estimated, and $e_i$ is model residual for the $i$-th tree, which is assumed to follow a normal distribution with mean zero and variance $\sigma_i^2$. As the spread of the OLS residuals changed over the range of the field-measured tree height, heterogeneous residual variance $\sigma_i^2$ for the $i$-th tree was expressed as a parametric power function of tree height formulated in Eq. 2. Due to the paired plots in the Lassen NF, parameter $\delta_s$ was used to estimate the effect of pair-specific characteristics for the $s$-th pair of plots.

$$var(e_i) = \sigma_i^2 = \left|\sqrt{h_i}\right|^{2\delta_s} \tag{2}$$

Following the measure of spatial autocorrelation in the OLS residuals by the Moran's I, residual covariance between any pair of trees $i$ and $j$ was expressed in Eq. 3 using an autocovariance function with an exponential correlation function $\rho$ assumed,

$$cov(e_i e_j) = \begin{cases} \sqrt{\sigma_i^2 - \tau_i^2}\sqrt{\sigma_j^2 - \tau_j^2}\,\rho(d_{ij}|\Phi), & d_{ij} > 0 \text{ and } i,j \in s \\ \sigma_i^2, & d_{ij} = 0 \text{ and } i,j \in s \\ 0, & i \in s, j \notin s \end{cases} \tag{3}$$

where $d_{ij}$ is the distance between trees $i$ and $j$, with locations denoted by $(x_{1i}, x_{2i})$ and $(x_{1j}, x_{2j})$, $\tau_i^2$ is nugget, $\Phi$ is spatial decay parameter that controls the effective range. The effective range is the distance where the correlation $\rho$ drops to 0.05. For the exponential correlation function, when $\rho = e^{(-d_{ij}/\Phi)} = 0.05$, the effective range $d_{ij} = -ln(0.05) \times \Phi \approx 3\Phi$. When trees $i$ and $j$ are from a different pair of plots, covariance is assumed to be zero.

Heteroscedasticity and spatial dependence in residuals were characterized in the variance-covariance matrix of residuals denoted by $\boldsymbol{\Sigma}_{k\times k}$. It was a block matrix with variances at the diagonal elements, non-zero covariances between trees on the same pair of plots at the off-diagonal elements, and zero covariances for trees between different pairs of plots. The method of GLS provided a minimum-variance unbiased estimator (Eq. 4) for the regression vector $\boldsymbol{\beta} = (\beta_0, \beta_1)$, where the method of restricted maximum likelihood (REML) was used to simultaneously estimate the parameters specifying the structure of $\Sigma$.

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}y \tag{4}$$

where $\boldsymbol{X}$ is the design matrix consisted of a 1-column and $\sqrt{h}$, $y$ is the response variable $ln(b)$ in Eq. 1, and $\boldsymbol{\Sigma}$ is the variance-covariance matrix of residuals.

Different from GLS, the method of OLS assumes homoscedasticity so that $var(e_i) = \sigma^2$ for all $i$, and no autocorrelation so that $cov(e_i, e_j) = 0$ for $d_{ij} > 0$. The OLS estimator for $\boldsymbol{\beta}$ is therefore $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'y$.

### 3.3. Predicting attributes of correctly identified trees and omission/commission errors from LiDAR

Stem locations, heights and crown diameters of individual trees were predicted from LiDAR based on an analysis for individual tree crowns delineated on a map of a rasterized and smoothed canopy height model (CHM). Individual trees were identified using the watershed segmentation algorithm that relied on the detection of local minima (Pitkänen et al., 2004). CHM was inverted so that local maxima (where treetops were located) became local minima. Around each local minimum, a segment for a tree crown comprised all pixels whose paths of steepest descent terminated at this minimum. Segments of excessively small areas were merged into neighboring segments to avoid splitting intact crowns. Attributes of individual trees were predicted from tree crown measurements. Tree height was the maximum height for all pixels associated with the crown. Stem locations were the XY coordinates of the center of the crown pixel with the maximum height. The maximum crown diameter was predicted by determining the diameter of a convex polygon, the crown diameter perpendicular to the maximum was also measured.

It was essential to categorize the segmentation results into three groups: (1) trees correctly identified by LiDAR, (2) trees omitted by LiDAR (omission errors), (3) objects erroneously identified as trees (commission errors) and split crowns resulting in redundant heights and repeated count of trees (commission errors). For this purpose, field-measured trees were matched with the LiDAR-identified trees based on the similarity in the three-dimensional space, taking both stem location and tree height into consideration (Xu et al., 2014). A correctly identified tree was the field-measured tree which had the shortest distance to a LiDAR-identified tree if their height difference was within 6 m. Unpaired field-measured trees were omission errors and unpaired LiDAR-identified trees were commission errors.

### 3.4. Spatial DA using the BLUP by incorporating LiDAR-predicted attributes

Consider a multivariate normal random vector $\boldsymbol{y}$ that can be partitioned into random vectors $\boldsymbol{y}_{1n\times 1}$ that consists of logarithmic biomass for the LiDAR-identified trees, and $\boldsymbol{y}_{2k\times 1}$ that consists of logarithmic biomass for the field-measured trees. Residuals of both $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ follow normal distributions with their expectations denoted by $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, variances denoted by $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ and covariance denoted by $\boldsymbol{V}_{12}$. In case

we have observed the value of $\boldsymbol{y}_2$ and want to predict the value of $\boldsymbol{y}_1$, the BLUP of $\boldsymbol{y}_1$ is

$$BLUP(\boldsymbol{y}_1) = \widehat{\boldsymbol{y}}_1 = \widehat{\boldsymbol{\mu}}_1 + \widehat{\boldsymbol{V}}_{12}\widehat{\boldsymbol{V}}_2^{-1}(\boldsymbol{y}_2 - \widehat{\boldsymbol{\mu}}_2) = (\boldsymbol{X}_L\widehat{\boldsymbol{\beta}}) + \widehat{\boldsymbol{V}}_{12}\widehat{\boldsymbol{V}}_2^{-1}(\boldsymbol{y}_2 - (\boldsymbol{X}_F\widehat{\boldsymbol{\beta}})). \tag{5}$$

In practice, $\widehat{\boldsymbol{\mu}}_1$ and $\widehat{\boldsymbol{\mu}}_2$ were predicted values obtained using the spatial allometric model fitted in section 3.2. $\boldsymbol{X}_L$ is the design matrix consisting of the LiDAR-predicted tree heights, $\boldsymbol{X}_F$ is the design matrix consisting of the field-measured tree heights, and $\widehat{\boldsymbol{\beta}}$ is the estimate of the regression coefficient vector obtained using Eq. 4. A distance matrix for $\boldsymbol{y}$ was constructed with locations denoted by $x_1 = \begin{pmatrix} x_{1y_1} \\ x_{1y_2} \end{pmatrix}$ and $x_2 = \begin{pmatrix} x_{2y_1} \\ x_{2y_2} \end{pmatrix}$, and elements of the variance-covariance matrix $\boldsymbol{V}_{(n+k)\times(n+k)} = \begin{pmatrix} \boldsymbol{V}_1 & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_2 \end{pmatrix}$ were estimated using the autocovariance function, the power variance function, and the estimated parameters $\Phi$,$\sigma_i^2$, $\tau_i^2$ and $\delta_s$.

The BLUP calibrated spatial model predictions denoted by $\widehat{\boldsymbol{\mu}}_1 = \boldsymbol{X}_L\widehat{\boldsymbol{\beta}}$, with a correction term for new information that played the same role as the "innovation" in the Kalman filter. Both tree heights and stem locations predicted from LiDAR were incorporated. When back-transforming $BLUP(\boldsymbol{y}_1)$ to obtain biomass predictions after DA, a bias correction (Baskerville, 1972) was applied before exponentiation, to compensate for the systematic error associated with the transformation using Eq. 6,

$$\widehat{b}_i = exp\left[BLUP(\boldsymbol{y}_1) + \frac{\sigma_i^2}{2}\right] \tag{6}$$

where $\widehat{b}_i$ is the assimilated biomass predictions for the $i$-th tree.

When constructing the matrix $\boldsymbol{V}$ (Fig. 2), it is crucial to understand that it is the variance-covariance matrix of a multivariate normal vector $\boldsymbol{y}$ that was partitioned into $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$. First, $\boldsymbol{V}$ consisted of four primary elements $\boldsymbol{V}_1$, $\boldsymbol{V}_2$, $\boldsymbol{V}_{12}$ and $\boldsymbol{V}_{21}$ in the form of $\begin{pmatrix} \boldsymbol{V}_1 & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_2 \end{pmatrix}$; second, both $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$ were block matrices consisting of variance-covariance matrices for all pairs of sample plots, denoted by $\boldsymbol{\Sigma}_s$, with $\boldsymbol{\Sigma}_{s.\ y1}$ constructed for the $s$-th pair of plots in $\boldsymbol{y}_1$ and $\boldsymbol{\Sigma}_{s.\ y2}$ for a corresponding pair in $\boldsymbol{y}_2$; and third, $\boldsymbol{V}_{12}$ was a block matrix consisting of the covariance matrix between $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ for each pair of sample plots; fourth, $\boldsymbol{V}_{21}$ is the transpose of $\boldsymbol{V}_{12}$.

### 3.5. Estimating prediction variance in the assimilated biomass of individual trees

The Taylor series expansion, a classic method to assess model uncertainty was used to estimate prediction variance in the LiDAR-predicted individual tree biomass. The allometric model was back-transformed to the original scale before the Taylor series expansion was applied (Gertner et al., 1995; Xu et al., 2018). After exponentiating both sides of Eq. 1, the right side was denoted by $g(h,\beta,e)$, a biomass function of three variables or error sources, the randomness of which contributed to the variance of biomass predictions. Parameter uncertainty associated with the power variance function and the auto-covariance function was not considered because of the common assumption in the applications of spatial prediction that this source of uncertainty is small (Thompson, 2012).

$$\boldsymbol{V} = \begin{bmatrix} \Sigma_{1.y1} & 0 & 0 & \Sigma_{1.y1y2} & 0 & 0 \\ 0 & \Sigma_{2.y1} & 0 & 0 & \Sigma_{2.y1y2} & 0 \\ 0 & 0 & \Sigma_{3.y1} & 0 & 0 & \Sigma_{3.y1y2} \\ \Sigma_{1.y2y1} & 0 & 0 & \Sigma_{1.y2} & 0 & 0 \\ 0 & \Sigma_{2.y2y1} & 0 & 0 & \Sigma_{2.y2} & 0 \\ 0 & 0 & \Sigma_{3.y2y1} & 0 & 0 & \Sigma_{3.y2} \end{bmatrix}$$

**Fig. 2.** Example structure of the variance-covariance matrix $\boldsymbol{V}$.

Using the Taylor Series expansion in several variables, the actual biomass value was approximated using a finite series of first-order polynomials that had a value similar to the truth in a neighborhood around $\widehat{h}_i$,$\widehat{\boldsymbol{\beta}}$ and $\widehat{e}_i$, each of which was a specific value of $h_i$, $\boldsymbol{\beta}$ and $e_i$ (Eq. 7). After re-arranging terms and taking the variance of both sides, the variance estimator was decomposed into three components that included the measurement uncertainty (in the LiDAR-predicted height), the parameter uncertainty and the residual variability, respectively (Eq. 8), based on the assumptions that effects of the error sources were independent. The variance of $g(h,\boldsymbol{\beta},e)$ was zero because it was a constant.

$$b_i \cong g(\widehat{h},\widehat{\beta},\widehat{e}) + \frac{dg}{dh}\times(h_i - \widehat{h}_i) + \frac{dg}{d\boldsymbol{\beta}}\times(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) + \frac{dg}{de}\times(e_i - \widehat{e}_i) \tag{7}$$

$$var(\widehat{b}_i) \cong \left(\frac{dg}{dh}\right)^2 \times var(h_i - \widehat{h}_i) + \left(\frac{dg}{d\boldsymbol{\beta}}\right)^2 \times var(\widehat{\boldsymbol{\beta}}) + \left(\frac{dg}{de}\right)^2 \times var(\widehat{e}_i) \tag{8}$$

where $\widehat{h}$ is the LiDAR-predicted tree height, $\frac{dg}{dh}$, $\frac{dg}{d\boldsymbol{\beta}}$ and $\frac{dg}{de}$ are partial derivatives of the $g$ function with respect to $h$, $\boldsymbol{\beta}$ and $e$, respectively.

Although the general form of the variance estimator was given in Eq. 8, the detailed expression differed among OLS, GLS and BLUP. In case of OLS that assumed no residual correlation and homogenous residual variance, detailed estimators for the three variance components were introduced in Xu et al. (2018) where each component was expressed as a function of LiDAR-predicted tree height. Differences among OLS, GLS and BLUP estimators lay in three aspects. First, residual variance-covariance matrix for GLS was specified in Eq. 3 while it was formulated in Eq. 9 for BLUP, which was supposed to be smaller than that obtained using OLS or GLS, because the correction term $\boldsymbol{V}_{12}\boldsymbol{V}_2^{-1}\boldsymbol{V}_{12}'$ is non-negative. Second, BLUP partial derivatives $\frac{dg}{dh}$, $\frac{dg}{d\boldsymbol{\beta}}$ and $\frac{dg}{de}$ were different in that the $g$ function carried a correction term specified in Eq. 5 while for GLS the $g$ function was Eq. 1 with both sides exponentiated. Third, GLS and BLUP estimated the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ using Eq. 10.

$$var(\widehat{e}) = var(y_1 - \widehat{y}_1) = \boldsymbol{V}_1 - \boldsymbol{V}_{12}\boldsymbol{V}_2^{-1}\boldsymbol{V}_{12}' \tag{9}$$

$$var(\widehat{\boldsymbol{\beta}}) = \sigma_i^2(\boldsymbol{X}_F{}'\boldsymbol{V}_2{}^{-1}\boldsymbol{X}_F)^{-1} \tag{10}$$

where $\boldsymbol{X}_F$ is design matrix consisting of the field-observed tree heights.

Except for these changes, $var(h_i - \widehat{h}_i)$ in the measurement uncertainty component remained unchanged. It was estimated following a three-step procedure. First, for known locations where both field-measured and LiDAR-predicted heights were matched, differences between the field-measured and LiDAR-predicted heights were binned according to classes of the LiDAR-predicted heights. Second, the empirical variance of the height differences was calculated for each bin. Third, natural cubic splines were fitted to the empirical variances with the mean of the LiDAR-predicted heights as the predictor variable. Modeling results were checked to ensure non-negative values for $var(h_i - \widehat{h}_i)$ for all trees in the Lassen NF, especially for those trees taller than the height of the tallest field-measured tree.

### 3.6. Estimating prediction variance in the LiDAR-predicted biomass density

To predict biomass density, simple aggregations of biomass predictions for the LiDAR-identified trees within plots will lead to a complex blend of under-estimation due to omission errors and over-

estimation due to commission errors. A reasonable method to predict biomass density is to subtract the biomass associated with all commission errors from the LiDAR-identified trees and to add the biomass associated with all omission errors (Eq. 11). While $\sum_{i=1}^{n}\widehat{b}_{i,p}$ was the summarized biomass predictions (Eq. 6) for all the LiDAR-identified trees in a plot including commission errors, $\sum_{o=1}^{l}\widehat{b}_{o,p}$ and $\sum_{c=1}^{m}\widehat{b}_{c,p}$ were not realized by estimating the number of omission errors denoted by $l$ and the number of commission errors denoted by $m$, respectively. Instead $\left(\sum_{o=1}^{l}\widehat{b}_{o,p}-\sum_{c=1}^{m}\widehat{b}_{c,p}\right)\times\frac{10000}{A_p}$, the difference between biomass density associated with omission errors and biomass density associated with commission errors was modeled as a whole, for simplicity, to predict the omission/commission difference from LiDAR explanatory variables.

A three-step procedure was applied at the plot level to construct a multiple linear regression model for the omission/commission difference. First, as the response variable, the empirical differences were the field-truth biomass of the omission errors minus the predicted biomass (Eq. 6) of the commission errors because omission and commission errors were known for sample plots after tree matching described in section 3.3. For candidate explanatory variables, a set of statistics for attributes of the LiDAR-identified tree crowns (height, crown area and distance to the nearest crown) were calculated. Second, explanatory variables were selected using the variance inflation factor (VIF) (Fox, 2016) that evaluated the severity of multicollinearity, and the stepwise selection that chose the model by Bayesian information criterion (BIC). Third, regression coefficients were estimated using OLS and $\left(\sum_{o=1}^{l}\widehat{b}_{o,p}-\sum_{c=1}^{m}\widehat{b}_{c,p}\right)\times\frac{10000}{A_p}$ was predicted as a correction term to compensate for omission errors while removing commission errors from the aggregation of biomass predictions for the LiDAR-identified trees.

$$\widehat{B}_p=\left(\sum_{i=1}^{n}\widehat{b}_{i,p}+\left(\sum_{o=1}^{l}\widehat{b}_{o,p}-\sum_{c=1}^{m}\widehat{b}_{c,p}\right)\right)\times\frac{10000}{A_p} \tag{11}$$

where $\widehat{B}_p$ is the LiDAR-predicted biomass density for the $p$-th plot whose area is $A_p$. Biomass predictions of the LiDAR-identified trees, omission and commission errors are denoted by $\widehat{b}_i,\widehat{b}_o,\widehat{b}_c$, the numbers of which are denoted by $n$, $l$, $m$ respectively.

By taking variances of both sides of Eq. 11, the variance of $\widehat{B}_p$ was a sum of two variance components specified in Eq. 12, based on the assumption that commission errors, omission errors and the errors in biomass of the LiDAR-identified trees are independent. The first component was a sum of estimated variances and covariances of the LiDAR-identified trees in each sample plot (Eq. 8 and the variance-covariance matrix specified in Eq. 3). It varied when the covariance structure of residuals differed. OLS assumed independent residuals, meaning zero covariances for off-diagonal elements of the variance-covariance matrix. GLS and BLUP both assumed spatially correlated residuals. While the GLS variance-covariance matrix was modeled with the distance matrix constructed from the field-measured tree locations, the BLUP procedure calibrated the variance-covariance matrix with the LiDAR-updated tree locations.

The second component was the variance of the omission/commission differences, which was estimated as the root mean squared error (RMSE) of model predictions using a leave-one-out cross validation. While scaling up prediction uncertainties from trees to plots, except for the three error sources of the LiDAR-predicted biomass for individual trees (measurement uncertainty, parameter uncertainty and residual variability), two additional error sources needed to be accommodated: residual covariance as well as omission and commission errors associated with the tree identification algorithm.

$$\begin{aligned}var\left(\widehat{B}_p\right)&=\sum_{i=1}^{n}\sum_{j=1}^{n}cov\left(\widehat{b}_{i,p},\widehat{b}_{j,p}\right)\times\left(\frac{10000}{A_p}\right)^2+var\left(\widehat{B}_{o,p}-\widehat{B}_{c,p}\right)\\&=\left(\sum_{i=1}^{n}var\left(\widehat{b}_{i,p}\right)+\sum_{i\neq j}^{n}\sum_{j\neq i}^{n}cov\left(\widehat{b}_{i,p},\widehat{b}_{j,p}\right)\right)\times\left(\frac{10000}{A_p}\right)^2+var\left(\widehat{B}_{o,p}-\widehat{B}_{c,p}\right)\end{aligned} \tag{12}$$

where $\widehat{B}_{o,p}$ and $\widehat{B}_{c,p}$ are predicted biomass density associated with omission and commission errors.

## 4. Results and discussion

### 4.1. Allometric models

The three allometric models fitted using the OLS, GLS and BLUP methods are summarized in Table 2. GLS produced slightly different estimates for regression coefficients than OLS. In practice, the OLS and GLS estimates for $\boldsymbol{\beta}$ will be different for any dataset. In case of residual spatial autocorrelation and heteroscedasticity, GLS estimates are the best linear unbiased estimators (BLUE) while OLS estimates are not. Different variance structures were assumed for OLS and GLS. OLS had an extremely large variance of 0.60 for the distribution of residuals. For GLS, the estimated variance function parameter $\delta_s$ had a range from −0.61 to 0.47, resulting in a left-skewed distribution of heterogeneous residual variances (Fig. 3a), with mean 0.69 and variance 0.23. The hypothesis of heteroscedasticity for biomass residuals among sample plots in the Lassen NF was verified. DA based on the BLUP had a narrower distribution of non-constant residual variances than GLS, with mean 0.51 and variance 0.13 (Fig. 3b). BLUP decreased the mean residual variance by 26% relative to GLS and by 15% relative to OLS. Although the variance function can be further improved by alternative modeling techniques (Pinheiro and Bates, 2000; Galecki and Burzykowski, 2013), optimization of the variance function was not an objective of this study.

The effective range of the spatial autocorrelation was estimated to be 209 m, beyond which residuals were considered uncorrelated. Both empirical and fitted semi-variogram and correlogram are illustrated in Fig. 4 with binned means for the empirical values. To have a firmer handle on the residual spatial autocorrelation structure, a model that used trees from all plots was additionally fitted. Group effects on residual variances were not accommodated due to the difficulty for the model to converge. A similar range was obtained but this model failed to improve heteroscedasticity and was thus not included in the following analysis. Since the distance between any two pairs of plots was greater than 209 m, trees from another pair of plots had zero covariance regardless of whether they were considered in parameter estimation. The magnitude of spatial dependency is commonly expressed in terms of the nugget-sill ratio $\frac{\tau^2}{\sigma_i^2}$, a relative share of the nugget in the total variance. The small ratio ($< 0.25$) implies a strong spatial dependence of the variable because a large part of variance is introduced spatially, while the large ratio ($> 0.75$) often indicates a weak spatial dependency. The nugget-sill ratio in the Lassen dataset was 0.59, suggesting a moderate spatial dependence.

**Table 2**
Parameter estimates for allometric models fitted using the OLS, GLS and BLUP methods.

| Methods | $\widehat{\beta}_0$ | $\widehat{\beta}_1$ | $E(\sigma_i^2)$ | Φ |
|---|---|---|---|---|
| OLS | 0.03 | 1.41 | 0.60 | |
| GLS | 0.40 | 1.36 | 0.69 | 69.75 |
| BLUP | 0.40 | 1.36 | 0.51 | 69.75 |

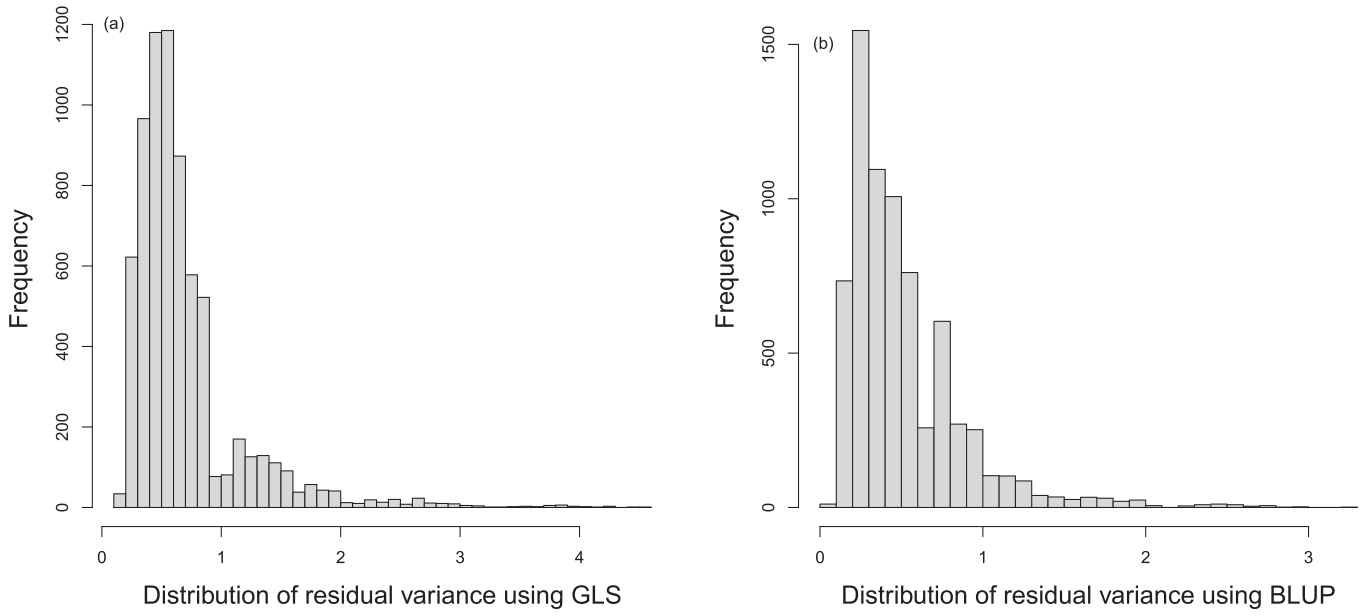


**Fig. 3.** (a) Distribution of the non-constant residual variance verified the assumption of heteroscedasticity. (b) BLUP decreased residual variance by 26% relative to GLS.

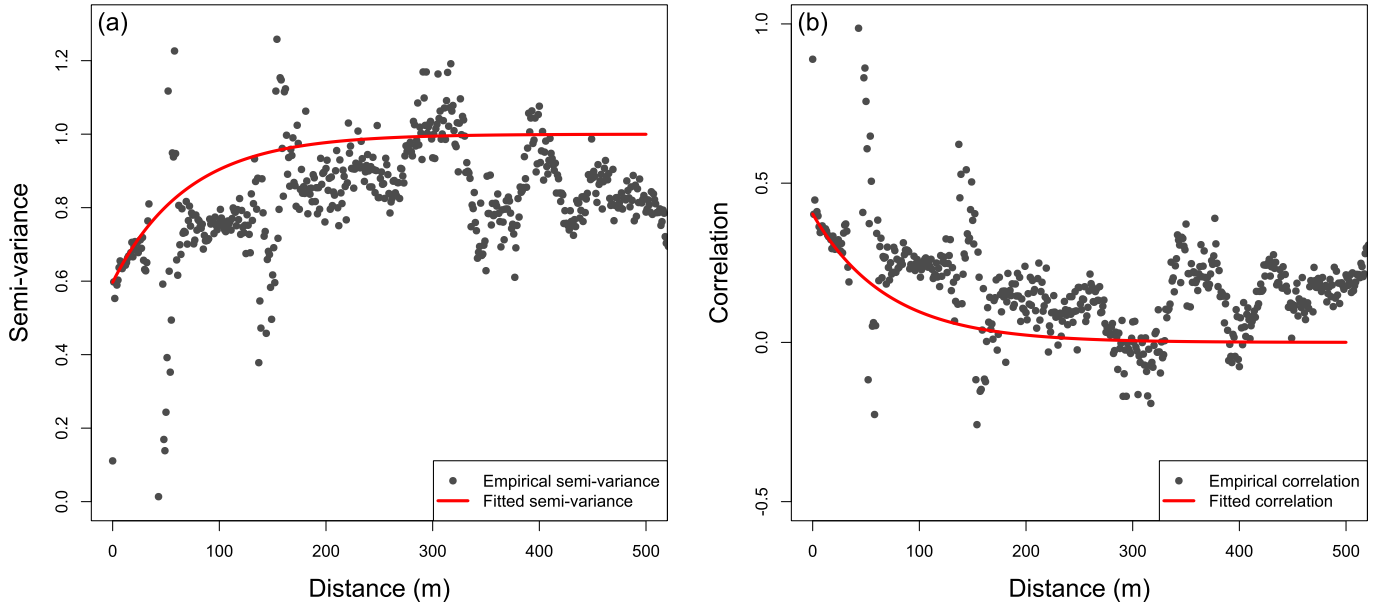


**Fig. 4.** (a) A standardized semi-variogram and (b) a correlogram show the effective range as 209 m.

**Table 3**
Prediction accuracies for the LiDAR-predicted biomass of individual trees (categorized by quantiles) using OLS, GLS and BLUP.

| Methods | RMSE% | | | | *SPE%*[*] | | | |
|---|---|---|---|---|---|---|---|---|
| | overall | <25th | 25-75th | >75th | overall | <25th | 25-75th | >75th |
| OLS | 185.14 | 354.57 | 113.81 | 107.13 | 2.76 | −163.85 | −0.68 | 4.16 |
| GLS | 191.34 | 446.04 | 129.78 | 110.54 | −7.81 | −234.55 | −19.80 | −4.74 |
| BLUP | 164.89 | 216.17 | 87.51 | 95.60 | 2.14 | −95.49 | 0.89 | 2.84 |

[*] SPE: Systematic prediction error.

### 4.2. Effects of assimilating the LiDAR-predicted attributes on prediction accuracy

Table 3 summarizes prediction accuracies for individual tree biomasses using OLS, GLS and the BLUP-based DA. Compared with GLS, OLS overestimated prediction accuracy in that covariances between residuals of individual tree biomass predictions were not accommodated in the estimation of regression coefficients. GLS, therefore, more comprehensively estimated residual variance which, in turn, led to an increase in RMSE. As expected, the BLUP-based DA, produced residual variance estimates less than GLS estimates, and even less than OLS estimates which was not expected. BLUP, therefore, increased biomass prediction accuracies for individual trees by 11% and 14%, relative to OLS and GLS, respectively.

BLUP effectively increased prediction accuracy for all three subsets of the data as well. Young trees were represented by trees with biomass less than the 1st quartile, mature trees by those with biomass no less than the 1st quartile but less than the 3rd quartile, and old-growth trees by those of biomass no less than the 3rd quartile. The DA effect was greatest for young trees, decreasing the RMSE by 39% and 52% compared with OLS and GLS, respectively. The DA effects were 23% and 33% reductions in RMSE for mature trees, and 11% and 14% reductions in RMSE for old-growth trees, relative to OLS and GLS. These DA effects are illustrated in Fig. 5 which depicts studentized individual tree residuals on the logarithmic scale for the three methods. While the OLS residuals exhibited heteroscedasticity, the GLS residuals were more homogeneously distributed, thereby justifying the logarithmic transformation. The BLUP residuals further reduced heteroscedasticity, particularly for young trees whose systematic prediction error (SPE) was reduced by 42%.

Spatially explicit biomass density predictions are more useful than individual tree biomass predictions for getting a big-picture view of the geographical distribution of forest biomass and its future change over time. Table 4 summarizes prediction accuracies for the LiDAR-predicted biomass density and its two components: biomass predictions of the LiDAR-identified trees and omission/commission difference (Eq. 11). In line with findings for individual trees, the BLUP-based DA procedure increased prediction accuracy by 28% and 33% relative to OLS and GLS, respectively.

Biomass associated with the LiDAR-identified trees was remarkably overpredicted using all three methods. A rational explanation could be associated with commission errors that contributed largely to the systematic prediction error at the plot level. Although commission errors should include non-tree objects erroneously identified as trees by LiDAR, the majority of them in the Lassen NF were essentially split crowns for trees in dominant canopy layers. While an intact crown was divided into multiple parts by the watershed segmentation algorithm, only one part could be paired with the field-measured tree according to the shortest distance in the three-dimensional space, leaving the remaining crown parts categorized as commission errors with redundant tree heights. Although the number of commission errors accounted for 25% of the field-measured trees in the Lassen NF, the biomasses associated with commission errors were much larger than the biomasses of omission errors (38% of the field-measured trees) that usually consisted of small trees in subcanopy layers. As a means of solution, the omission/commission difference model simulated the integrated effects of compensating for omission errors and removing commission errors. The model was found effective for correcting the overprediction caused by commission errors because after the correction, RMSE for the LiDAR-predicted biomass density was reduced by 36%, 38%, and 41%, the systematic prediction error was reduced by 77%, 65% and 83% respectively, relative to the LiDAR-identified biomass (Table 4).

**Table 4**
Prediction accuracies for the LiDAR-predicted biomass density, omission/commission difference and biomass of the LiDAR-identified trees using the OLS, GLS and BLUP methods (SPE stands for systematic prediction error).

| Response variables | Methods | RMSE | RMSE% | SPE | SPE% |
|---|---|---|---|---|---|
| LiDAR-predicted-AGB | OLS | 203.76 | 49.73 | 7.95 | 1.94 |
| LiDAR-predicted-AGB | GLS | 217.69 | 53.13 | −28.33 | −6.92 |
| LiDAR-predicted-AGB | BLUP | 146.82 | 35.83 | 6.20 | 1.51 |
| Omcom-diff* | OLS | 45.30 | −104.07 | <0.01 | <0.01 |
| Omcom-diff | GLS | 56.06 | −102.31 | <0.01 | <0.01 |
| Omcom-diff | BLUP | 64.71 | −146.26 | <0.01 | <0.01 |
| LiDAR-identified-AGB | OLS | 319.68 | 78.02 | −33.9 | −8.28 |
| LiDAR-identified-AGB | GLS | 350.45 | 85.53 | −81.97 | −20.00 |
| LiDAR-identified-AGB | BLUP | 248.89 | 60.74 | −36.84 | −8.99 |

* Omission/commission difference.

### 4.3. Effects of DA on prediction variance of the LiDAR-predicted biomass for individual trees

For biomass predictions for individual trees, proportions of the three variance components that included measurement uncertainty, parameter uncertainty and residual variability, are illustrated in Fig. 6 which reveals differences in the contributions of the three error sources among the three methods. Taking OLS as the reference, GLS increased the contribution of residual variability for trees with height <25 m but decreased the contribution of residual variability for taller trees. DA effect in reducing the proportion of residual variability was visible across the entire height rangeand was the most pronounced at the two ends. The trade-off mainly occurred between residual variability and measurement uncertainty, because parameter uncertainty was negligible for both the magnitude of proportions and the magnitude of changes. Although the proportion of residual variability was reduced by the BLUP, it was still the component that contributed the most to the variability of the individual tree biomass predictions.

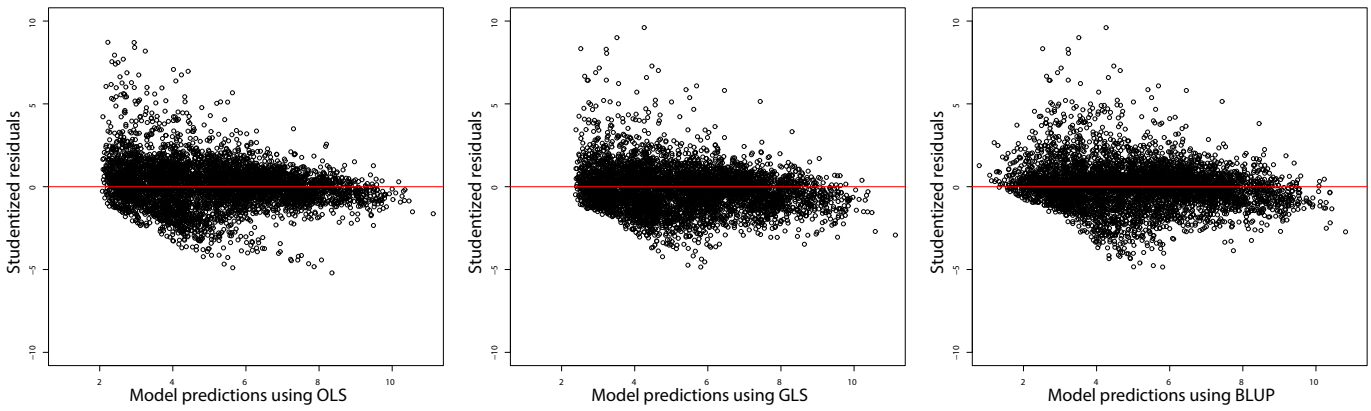


**Fig. 5.** Studentized residuals at the logarithmic scale for the OLS, GLS and BLUP methods.

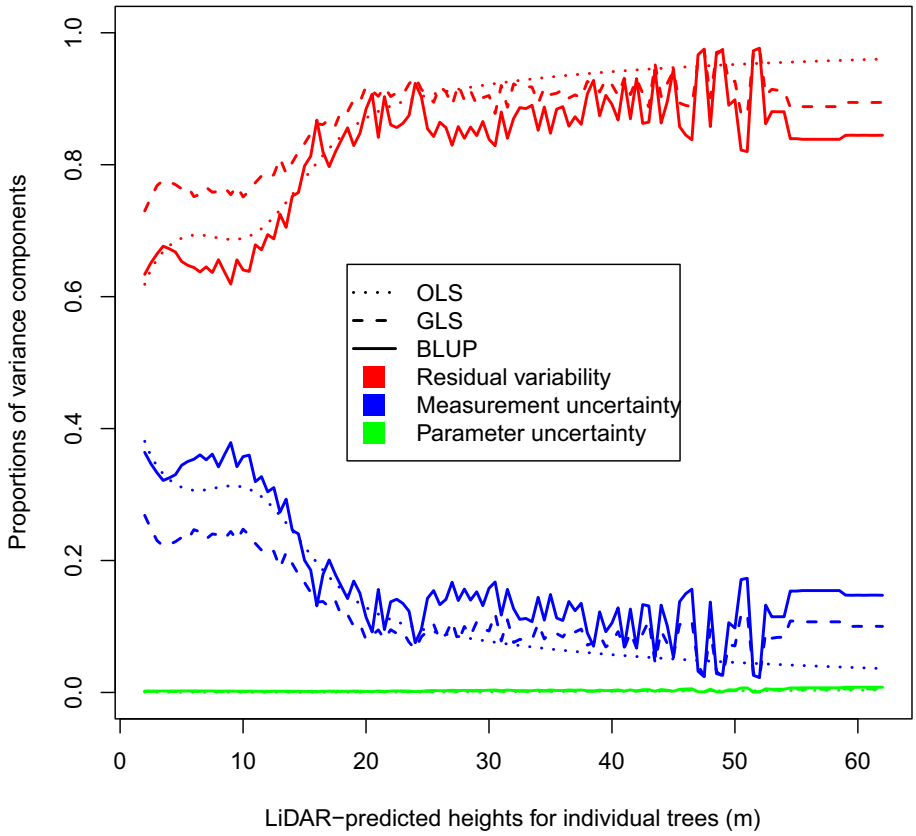


**Fig. 6.** Proportions of the three variance components (measurement uncertainty, parameter uncertainty and residual variability) among the OLS, GLS and BLUP methods.

For individual tree biomass, DA was effective in reducing all three variance components, resulting in the decrease in the overall prediction variance. Following the order of parameter uncertainty, measurement uncertainty and residual variability, three variance components were accumulated one-by-one in Fig. 7, with mean values across the entire range of tree heights illustrated in the inset for a more comprehensive understanding. First, BLUP decreased prediction variance by 40% and 20% with respect to OLS and GLS. Compared with OLS, it decreased the three variance components by 28%, 12% and 42%, respectively (Table 5), which was more evident in the mean values in the inset. The magnitude of the decrease appeared to be positively correlated with tree size. Second, the decrease in residual variability was the largest among the three components, and it played a deterministic role in the ultimate decrease in prediction variance, because it accounted for 98% of the decreased variance in total. The decrease in measurement uncertainty and parameter uncertainty, respectively accounted for 1.6% and 0.4% of the overall reduction in variance. Therefore, the effect of DA was the largest for reducing residual variability for individual trees.

Although both OLS and GLS offer unbiased estimators for $\boldsymbol{\beta}$, GLS was more efficient than the OLS, when there was residual spatial autocorrelation. The efficiency was revealed by the variance of the GLS estimate, $\widehat{\boldsymbol{\beta}}$, which was 16% smaller than the variance of the OLS estimate (Table 5). This indicates that for OLS, both the confidence interval and the *p*-value reported were smaller than the desired level. OLS optimistically estimated the variance of $\widehat{\boldsymbol{\beta}}$.

Cumulative coefficients of variation (CV) for the biomass of individual trees are illustrated in Fig. 8, with mean values illustrated in the inset and presented in Table 5. As a ratio of standard deviation to mean, CV is often used to measure relative variability of biomass predictions, because it facilitates estimation of the magnitude of uncertainty relative to the biomass prediction. First, BLUP decreased CV by 16% and 9% relative to OLS and GLS, respectively. Second, larger CVs for parameter uncertainty and measurement uncertainty were found for younger trees after DA. Because BLUP improved biomass predictions for younger trees by correcting for a large over-estimation (Fig. 5 and Table 3), the decrease in the CV denominator was larger than the decrease in the numerator, resulting in the increase in the CV. Third, DA consistently decreased CV except for trees at the two ends of the height axis. The reasons might be related to negative autocorrelations, where sapling or subcanopy trees grew at the boundary of large tree crowns, resulting in a mis-specified correlation. Compared with Xu et al. (2018) who reported a mean CV of 135% for the LiDAR-predicted biomass of individual trees using OLS, this study achieved 83%, 77% and 70% using the OLS, GLS and BLUP methods, respectively. Smaller prediction variance was associated with the use of local allometric models specifically for the Lassen NF where tree locations were accessible from in-situ measurements rather than a California-wide regional model.

### *4.4. Effects of DA on prediction variance of the LiDAR-predicted biomass density*

At the plot level, the effects of DA are more like an adjustment for the structure of the variance components, although it indeed reduced prediction variance, on average, by 3% and 49% relative to OLS and GLS. Figs. 9 and Fig. 10 illustrate the accumulated prediction variance and coefficient of variation by components respectively, following the order of parameter uncertainty, measurement uncertainty, residual variance, residual covariance and variance associated with omission and commission errors from bottom to top. Fig. 11 presents the proportions of variance components that have the same order from bottom to top and the same color scheme as the previous two figs.

A number of findings are worthy of note. First, prediction variance was underestimated by 48% when residual correlation was erroneously

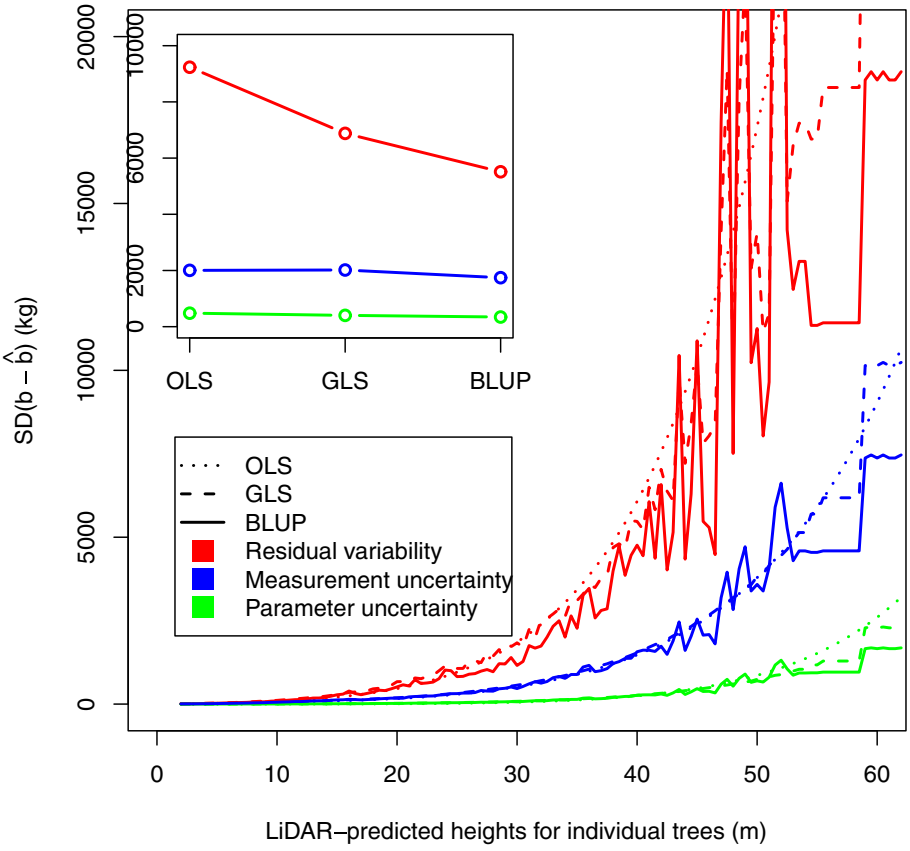


**Fig. 7.** Cumulative variance components in the order of parameter uncertainty (green), measurement uncertainty (blue) and residual variability (red) illustrate a general decrease in the overall prediction variance after DA. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

**Table 5**
Mean values of prediction variance and coefficient of variation for the LiDAR-predicted biomass of individual trees*.

| Method | Mean SD | | | | Mean CV (%) | | | |
|---|---|---|---|---|---|---|---|---|
| | $sd_p$ | $sd_h$ | $sd_r$ | $sd_o$ | $cv_p$ | $cv_h$ | $cv_r$ | $cv_o$ |
| OLS | 481 | 1942 | 9011 | 9234 | 2.7 | 28 | 77 | 83 |
| GLS | 403 | 1976 | 6565 | 6881 | 3.4 | 27 | 72 | 77 |
| BLUP | 344 | 1708 | 5198 | 5513 | 3.6 | 29 | 63 | 70 |
| Change (%) | ↓ | ↓ | ↓ | ↓ | ↑ | ↑ | ↓ | ↓ |
| | 28 | 12 | 42 | 40 | 0.9 | 0.1 | 14 | 13 |

* SD stands for standard deviation; CV stands for coefficient of variation. Subscripts *p*, *h*, *r*, *o* stand for parameter, height, residual and overall, respectively. For change in SD, it is ratio of the difference to the OLS; for change in CV, it is difference itself.

assumed to be zero. The underestimated residual covariance illustrated by the blue section in Fig. 11 (b) was comprehensively estimated using the GLS method, but it was found to be the largest component among the five variance components. Second, compared with GLS, the decrease in BLUP total variance was attributed to both residual variance and covariance, with the decrease in covariance contributing the greatest. Therefore, the effect of DA was the largest in reducing residual covariance for the LiDAR-predicted biomass density. Third, the accumulated gain in per-tree precision was largely offset by the error sources at the plot level. Nevertheless, unexpectedly, the prediction variance after DA was slightly smaller than that of the optimistically estimated OLS, with the mean SD dropped from 126.80 to 122.60. This may suggest that the BLUP method properly utilized the captured spatial structure of residuals, thus improving the precision of the LiDAR-predicted biomass density.

The proposed spatial DA procedure showcased an application of incorporating the LiDAR-predicted stem locations into the BLUP for reducing the uncertainty in biomass predictions for population units. This procedure made fuller use of the LiDAR positioning information, so that the LiDAR-predicted tree location other than tree size, also mattered in a more efficient biomass inventory. As the cost for collecting small footprint airborne LiDAR keeps decreasing, it is likely that more and more forests have tree locations and heights predicted from LiDAR. Combined with field observations, like sample plot networks for national forest inventory (NFI) and sample plots for inventories for forest planning and operations, the effective range, within which trees are correlated can be understood for different types of forests and species compositions. The spatial DA procedure is therefore expected to improve biomass precision according to the intensity of spatial autocorrelation and heteroscedasticity. Different from conventional remote sensing-based biomass studies, the field-truth tree biomasses used in this study were more realistic because they accounted for the deviance from the predicted mean biomass to the true values of tree biomasses given the same species, DBH and height. As an attempt to compensate for the underestimated uncertainty, the simulation of tree level residuals for the field truth surely enlarged the estimated uncertainty in biomass at either

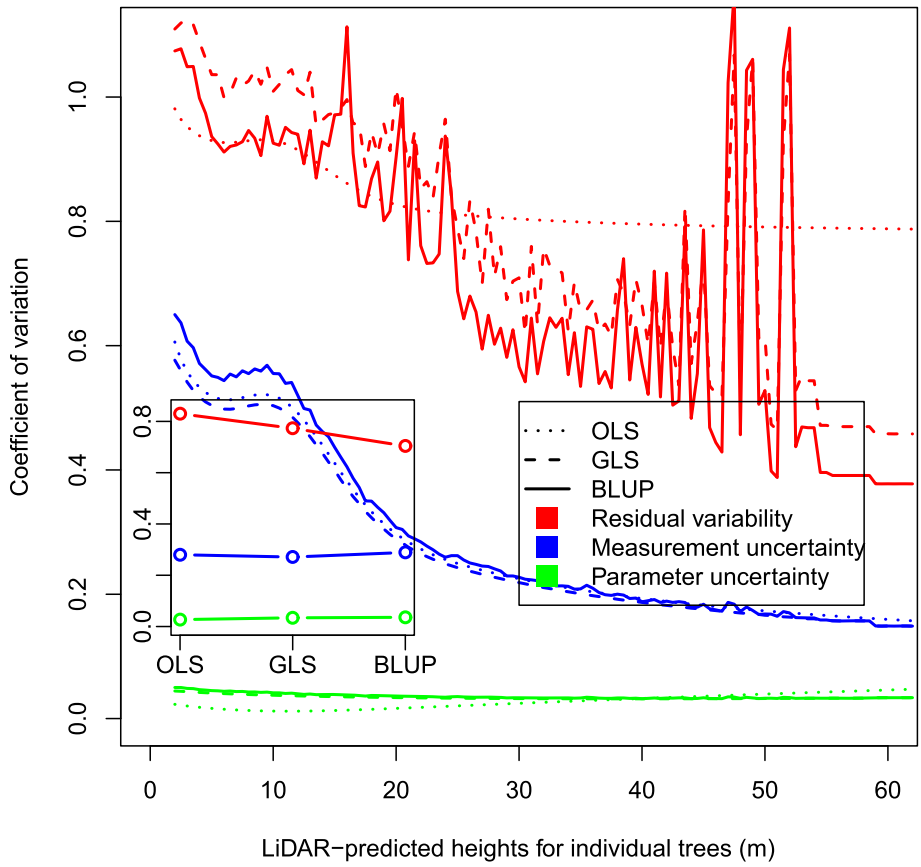


**Fig. 8.** The DA effect was the largest for residuals, which substantially decreased the post-DA coefficient of variation.

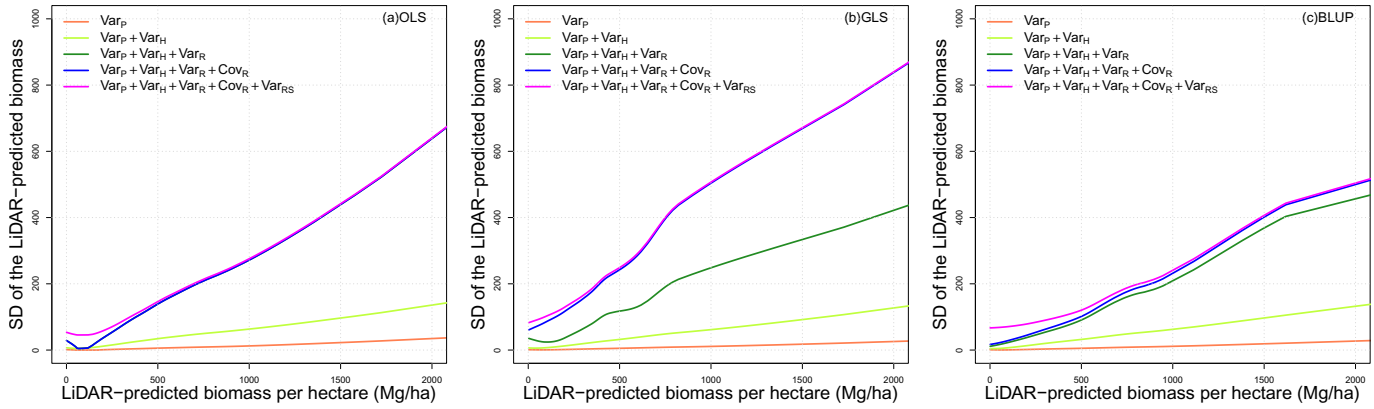


**Fig. 9.** DA decreased prediction variance in the LiDAR-predicted biomass density, relative to both OLS and GLS.

the tree or the plot level. Further experiments are needed to investigate the effects of such an attempt on the results following the methods of OLS, GLS and the DA using the BLUP.

## 5. Conclusions

This study proposed a spatial DA procedure that harnesses small footprint airborne LiDAR, the spatial structure of forest biomass prediction residuals and the BLUP to enhance forest inventory predictions for biomass and carbon. Small-footprint airborne LiDAR played a crucial role in providing predictions of 3-dimensional attributes of individual trees for updating the variance-covariance matrix required in the procedure.

Five major conclusions were drawn from the comparisons of results for the proposed method and results for the OLS and GLS methods. First, for biomass of individual trees, the spatial DA decreased prediction variance by 40% and 20% and decreased CV by 16% and 9%, relative to OLS and GLS. Because the decrease in residual variability accounted for 98% of the decreased prediction variance in total, the DA effect was the largest for reducing residual variability. Second, for biomass density, DA decreased prediction variance by 3% and 49% with respect to OLS and GLS, the DA effect was the largest for reducing residual covariance. Though the gain in precision for individual tree predictions was accumulated at a larger scale, it was offset by the larger-scale precision loss caused by residual covariance and the variance associated with omission and commission errors. Third, the OLS method that assumes no spatial structure underestimated prediction variance for the LiDAR-predicted biomass density by 48%. Fourth, from the perspective of prediction

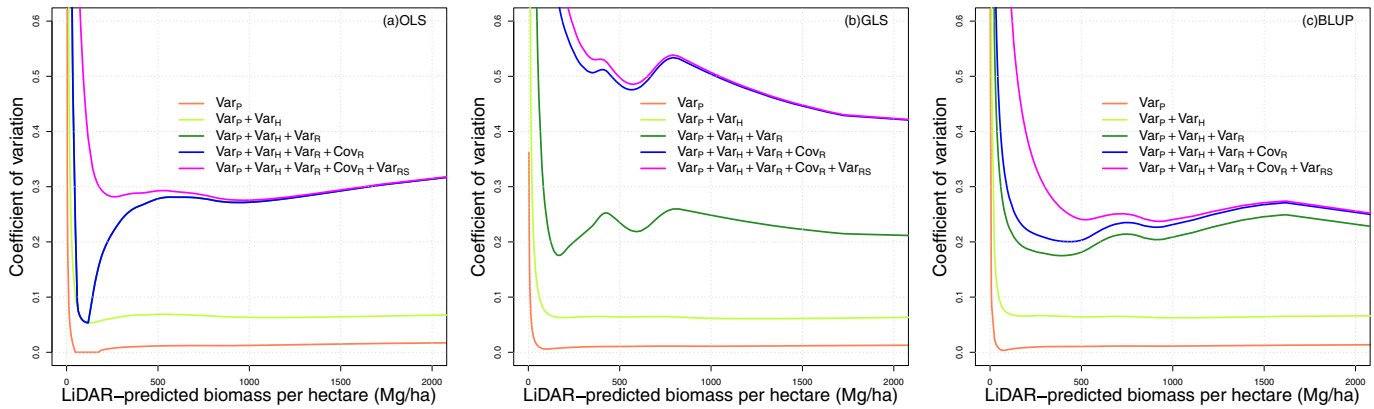


**Fig. 10.** The decrease in both residual variance and residual covariance helps reducing the coefficient of variation for the LiDAR-predicted biomass density.

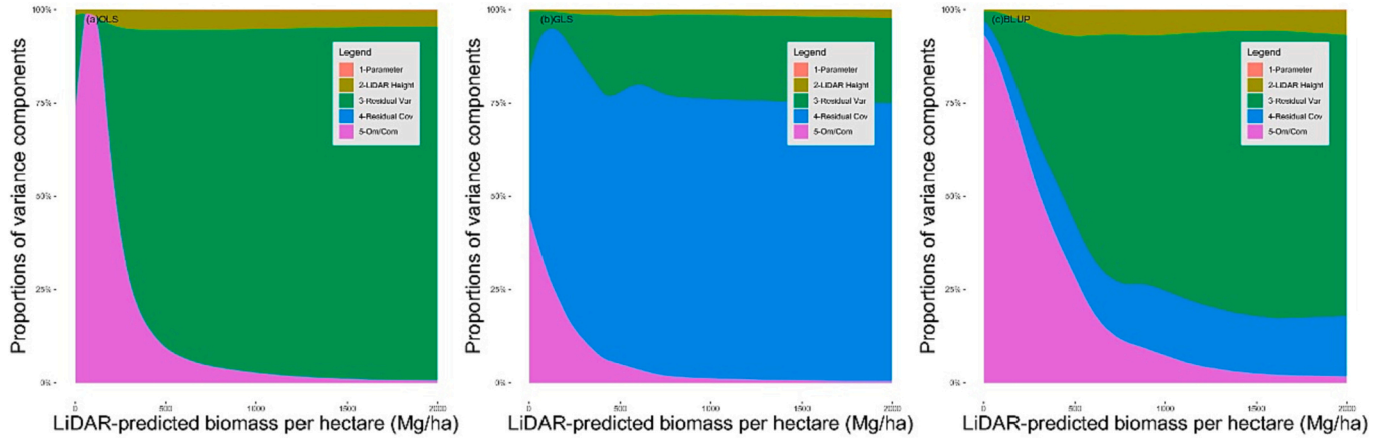


**Fig. 11.** Proportions of variance components are adjusted after DA, with residual covariance estimated using GLS largely reduced.

accuracy, DA decreased the RMSE for biomass of individual trees by 11% and 14% and decreased the RMSE for biomass density by 28% and 33%, relative to OLS and GLS. Fifth, it was effective to use the omission/commission difference model for correcting the systematic prediction error in the LiDAR-predicted biomass density.

Overall, the proposed spatial DA procedure demonstrated great potential for reducing the uncertainty in forest biomass predictions, thereby facilitating more efficient biomass inventories. It can be generalized to other dependent variables of interest given their correlations with new information from LiDAR.

## CRediT authorship contribution statement

**Qing Xu:** Conceptualization, Methodology, Software, Writing – original draft. **Bo Li:** Validation, Writing – review & editing. **Ronald E. McRoberts:** Methodology, Writing – review & editing. **Zengyuan Li:** Writing – review & editing. **Zhengyang Hou:** Conceptualization, Methodology, Writing – review & editing.

## Declaration of Competing Interest

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

## Data availability

The authors do not have permission to share data.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 32001252), the International Center for Bamboo and Rattan (Grant No. 1632020029 and Grant No. 1632021024), the National Social Science Fund of China (Grant No. 22BTJ005) and the NASA Carbon Monitoring System (proposal number: 14-CMS14-0048).

## Appendix A

Evaluation of the Taylor series expansion for the biomass function at specific values of $h_i$, $\boldsymbol{\beta}$ and $e$.

$$ln(b_i) = \beta_0 + \beta_1 \sqrt{h_i} + e_i$$

$$ln(b_i) = f(h_i, \boldsymbol{\beta}) + e_i$$

$$b_i = exp(f(h_i, \boldsymbol{\beta})) \times exp(e_i)$$

$$b_i = g(h_i, \boldsymbol{\beta}, e_i)$$

$$b_i \cong g(\widehat{h}_i, \widehat{\boldsymbol{\beta}}, \widehat{e}_i) + \frac{dg}{dh} \times (h_i - \widehat{h}_i) + \frac{dg}{d\boldsymbol{\beta}} \times (\beta - \widehat{\beta}) + \frac{dg}{de} \times (e_i - \widehat{e}_i)$$

## References


Axelsson, P., 1999. Processing of laser scanner data-algorithms and applications. ISPRS J. Photogramm. Remote Sens. 54, 138–147.

Baskerville, G.L., 1972. Use of logarithmic regression in the estimation of plant biomass. Can. J. For. Res. 2, 49–53.

Breidenbach, J., Magnussen, S., Johannes, R., Rasmus, A., 2018. Unit-level and area-level small area estimation under heteroscedasticity using digital aerial photogrammetry data. Remote Sens. Environ. 212, 199–211.

Chave, J., Andalo, C., Brown, S., Cairns, M.A., Chamber, J.Q., Eamus, D., Fölster, H., Fromard, F., Higuchi, N., Kira, T., Lescure, J.-P., Nelson, B.W., Ogawa, H., Puig, H., Riéra, B., Yamakura, T., 2005. Tree allometry and improved estimation of carbon stocks and balance in tropical forests. Oecologia 145, 87–99.

Chen, Q., Laurin, G.V., Valentini, R., 2015. Uncertainty of remotely sensed aboveground biomass over an African tropical forest: propagating errors from trees to plots to pixels. Remote Sens. Environ. 160, 134–143.

Chen, Q., McRoberts, R.E., Wang, C., Radtke, P.J., 2016. Forest aboveground biomass mapping and estimation across multiple spatial scales using model-based inference. Remote Sens. Environ. 184, 350–360.

Chojnacky, D.C., Heath, L.S., Jenkins, J.C., 2014. Updated generalized biomass equations for North American tree species. Forestry 87, 129–151.

Cressie, N., 1990. The origins of Kriging. Math. Geol. 22, 239–252.

Czaplewski, R.L., 1990. Kalman filter to update forest cover estimates. In: Labau, V.J., Cunia, T., Serv, USDA For (Eds.), State-of-the-art Methodology of Forest Inventory. Pacific Northwest Research Station, vol. 263. GTR PNW, pp. 457–465.

Ehlers, S., Grafström, A., Nyström, K., Olsson, H., Ståhl, G., 2013. Data assimilation in stand-level forest inventories. Can. J. For. Res. 43, 1104–1113.

Ehlers, S., Saarela, S., Lindgren, N., Lindberg, E., Nyström, M., Persson, H.J., Olsson, H., Ståhl, G., 2018. Assessing error correlations in remote sensing-based estimates of forest attributes for improved composite estimation. Remote Sens. 10, 667.

Fankhauser, S., Smith, S.M., Allen, M., et al., 2022. The meaning of net zero and how to get it right. Nat. Clim. Chang. 12, 15–21. https://doi.org/10.1038/s41558-021-01245-w.

Fox, J., 2016. Applied Regression Analysis and Generalized Linear Models, 3rd ed. Sage Publications, Los Angeles.

Galecki, A., Burzykowski, T., 2013. Linear Mixed-Effects Models Using R: A Step-By-Step Approach. Springer.

Gertner, G., Cao, X., Zhu, H., 1995. A quality assessment of a Weibull based growth projection system. For. Ecol. Manag. 71 (3), 235–250.

GlobAllomeTree, 2013. (accessed March 30, 2022). Retrieved from. http://globallometree.org/.

Gregoire, T.G., Walters, D.K., 1988. Composite vector estimator by weighting inversely proportional to variance. Can. J. For. Res. 18, 282–284.

Gregoire, T.G., Ståhl, G., Næsset, E., Gobakken, T., Nelson, R., Holm, S., 2011. Model-assisted estimation of biomass in a LiDAR sample survey in Hedmark County, Norway. Can. J. For. Res. 41, 83–95.

Henderson, C.R., 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31, 423–447.

Hou, Z., McRoberts, R.E., Ståhl, G., Packalen, P., Greenberg, J., Xu, Q., 2018. How much can natural resource inventory benefit from a finer resolution auxiliary data? Remote Sens.Environ. 209, 31–40.

Hou, Z., Mehtätalo, L., McRoberts, R.E., Ståhl, G., Tokola, T., Rana, P., Siipilehto, J., Xu, Q., 2019. Remote sensing-assisted data assimilation and simultaneous inference for forest inventory. Remote Sens. Environ. 234, 1114431.

IPCC, 2018. Special Report on Global Warming of 1.5 °C (eds Masson-Delmotte, V. et al.). WMO.

Isenburg, M., Liu, Y., Shewchuk, J., Snoeyink, J., Thirion, T., 2006. Generating raster DEM from mass points via TIN streaming. In: Geographic Information Science: 4th International Conference, GIScience 2006. Münster, Germany, September, pp. 186–198.

Jenkins, J.C., Chojnacky, D.C., Heath, L.S., Birdsey, R.A., 2003. National-scale biomass estimators for United States tree species. For. Sci. 49 (1), 12–35.

Kalman, R.E., 1960. A new approach to linear filtering and prediction problems. J. Basic Eng. T ASME 82, 35–45 (Series D).

Kangas, A., Gobakken, T., Næsset, E., 2020. Benefits of past inventory data as prior information for the current inventory. For. Ecosyst. 7, 20.

Khosravipour, A., Skidmore, A.K., Isenburg, M., Wang, T.J., Hussin, Y.A., 2014. Generating pit-free canopy height models from airborne LiDAR. Photogramm. Eng. Remote. Sens. 80 (9), 863–872.

Lindgren, N., Persson, H.J., Nyström, M., Nyström, K., Grafström, A., Muszta, A., Willén, E., Fransson, J.E.S., Ståhl, G., Olsson, H., 2017. Improved prediction of forest variables using data assimilation of interferometric synthetic aperture radar data. Can. J. Remote. Sens. 43, 374–383.

Mascaro, G., Deidda, R., Hellies, M., 2013. On the nature of rainfall intermittency as revealed by different metrics and sampling approaches. Hydrol. Earth Syst. Sci. 17, 355–369.

McRoberts, R.E., Tomppo, E.O., Finley, A.O., Heikkinen, J., 2007. Estimating areal means and variances of forest attributes using the k-nearest neighbors technique and satellite imagery. Remote Sens. Environ. 111, 466–480.

McRoberts, R.E., 2010. Probability-and model-based approaches to inference for proportion forest using satellite imagery as ancillary data. Remote Sens. Environ. 114 (5), 1017–1025.

McRoberts, R.E., Næsset, E., Gobakken, T., 2013. Inference for LiDAR-assisted estimation of forest growing stock volume. Remote Sens. Environ. 128, 268–275.

McRoberts, R.E., Næsset, E., Gobakken, T., Chirici, G., Condés, S., Hou, Z., Saarela, S., Chen, Q., Ståhl, G., Walters, B.F., 2018. Assessing components of the model-based mean square error estimator for remote sensing assisted forest applications. Can. J. For. Res. 48, 642–649.

Magnussen, S., Breidenbach, J., Mauro, F., 2017. The challenge of estimating a residual spatial autocorrelation from forest inventory data. Can. J. For. Res. 47 (11), 1557–1566.

Nyström, M., Lindgren, N., Wallerman, J., Grafström, A., Muszta, A., Nyström, K., Bohlin, J., Willén, E., Fransson, J.E.S., Ehlers, S., Olsson, H., Ståhl, G., 2015. Data assimilation in forest inventory: first empirical results. Forests. 6, 4540–4557.

Persson, H.J., Ekström, M., Ståhl, G., 2022. Quantify and account for field reference errors in forest remote sensing studies. Remote Sens. Environ. 283, 113302.

Petersson, H., Breidenbach, J., Ellison, D., Holm, S., Muszta, A., Lundblad, M., Ståhl, G. R., 2017. Assessing uncertainty: sample size trade-offs in the development and application of carbon stock models. For. Sci. 63 (4), 402–412.

Pinheiro, J.C., Bates, D.M., 2000. Mixed-Effects Models in S and S-PLUS. Springer.

Pitkänen, J., Maltamo, M., Hyyppä, J., Yu, X., 2004. Adaptive methods for individual tree detection on airborne laser based canopy height model. Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. 36 (8/W2), 187–191.

PNW-FIA Field Manual, 2015. https://www.fs.fed.us/pnw/rma/fia-topics/documentation/field-manuals/documents/Annual/2015_PFSL_FIA_Field_Manual.pdf (accessed 2017.08.10).

Robinson, G.K., 1991. That BLUP is a good thing: the estimation of random effects. Stat. Sci. 6, 15–32.

Ståhl, G., Holm, S., Gregoire, T.G., Gobakken, T., Næsset, E., Nelson, R., 2011. Model-based inference for biomass estimation in a LiDAR sample survey in Hedmark County,Norway. Can. J. For. Res. 41, 96–107.

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

Saarela, S., Schnell, S., Grafström, A., Tuominen, S., Nordkvist, K., Hyyppä, J., Kangas, A., Ståhl, G., 2015. Effects of sample size and model form on the accuracy of model-based estimators of growing stock volume. Can. J. For. Res. 45 (11), 1524–1534.

Saarela, S., Wästlund, A., Holmström, E., Mensah, A.A., Holm, S., Nilsson, M., Fridman, J., 2020. Mapping aboveground biomass and its prediction uncertainty using LiDAR and field data, accounting for tree-level allometric and LiDAR model errors. For. Ecosyst. 7, 43.

Thompson, S.K., 2012. Sampling, third ed. Wiley, New Jersey, United States.

Vauhkonen, J., Korpela, I., Maltamo, M., Tokola, T., 2010. Imputation of single-tree attributes using airborne laser scanning-based height, intensity, and alpha shape metrics. Remote Sens. Environ. 114 (6), 1263–1276.

White, Q., 1978. Estimation of plant biomass from quadrat data using the lognormal distribution. J. Range Manag. 31 (2), 118–120.

Xu, Q., Hou, Z., Maltamo, M., Tokola, T., 2014. Calibration of area-based diameter distribution with individual tree based diameter estimates using airborne laser scanning. ISPRS J. Photogramm. Remote Sens. 93, 65–75.

Xu, Q., Man, A., Fredrickson, M., Hou, Z., Pitkänen, J., Wing, B., Ramirez, C., Li, B., Greenberg, J.A., 2018. Quantification of uncertainty in aboveground biomass estimates derived from small-footprint airborne LiDAR. Remote Sens. Environ. 216, 514–528.