

Hierarchical models for small area estimation using zero-inflated forest inventory variables: comparison and implementation

Grayson W. White ^{a,b,c}, Andrew O. Finley^{a,b}, Josh K. Yamamoto^d, Jennifer L. Green^b, Tracey S. Frescino^e, David W. MacFarlane^a, Hans-Erik Andersen ^f, and Grant M. Domke^g

^aDepartment of Forestry, Michigan State University, East Lansing, MI, USA; ^bDepartment of Statistics & Probability, Michigan State University, East Lansing, MI, USA; ^cDepartment of Mathematics & Statistics, Reed College, Portland, OR, USA; ^dRedcastle Resources, Inc., Salt Lake City, UT, USA; ^eRocky Mountain Research Station, USDA Forest Service, Riverdale, UT, USA; ^fPacific Northwest Research Station, USDA Forest Service, Seattle, WA, USA; ^gNorthern Research Station, United States Department of Agriculture Forest Service, St. Paul, MN, USA

Corresponding author: Grayson W. White (email: gwhite@reed.edu)

Abstract

National Forest Inventory (NFI) data are typically limited to sparse networks of sample locations due to cost constraints. While design-based estimators provide reliable forest parameter estimates for large areas, there is increasing interest in model-based small area estimation (SAE) methods to improve precision for smaller spatial, temporal, or biophysical domains. SAE methods can be broadly categorized into area- and unit-level models, with unit-level models offering greater flexibility, making them the focus of this study. Ensuring valid inference requires satisfying model distributional assumptions, which is particularly challenging for NFI variables that exhibit positive support and zero-inflation, such as forest biomass, carbon, and volume. Here, we evaluate nine candidate estimators, including two-stage unit-level hierarchical Bayesian models, single-stage Bayesian models, and two-stage frequentist models, for estimating forest biomass at the county level in Nevada and Washington, United States. Estimator performance is assessed using repeated sampling from simulated populations and unit-level cross-validation with FIA data. Results show that small area estimators incorporating a two-stage approach to account for zero-inflation, county-specific random intercepts and residual variances, and spatial random effects yield the most accurate and well-calibrated county-level estimates, with spatial effects providing the greatest benefits when spatial autocorrelation is present in the underlying population.

Key words: model-based inference, Gaussian processes, Bayesian, forest biomass, simulation study

1. Introduction

National Forest Inventories (NFIs) play a critical role in collecting data and monitoring forest trends to assess resource availability, health, composition, and other economic and ecological attributes across spatial and temporal scales. Traditionally, NFIs have been designed to provide precise estimates at broad spatial scales, such as state- or nation-level assessments of forest attributes like timber volume and biomass. However, there is growing interest in obtaining more precise biomass estimates at finer spatial scales, such as the county level (Prisley et al. 2021; Wiener et al. 2021; U.S. Senate 2023). This rising demand, coupled with the widespread availability of high-resolution remotely sensed data and other auxiliary information, has prompted users to develop and apply innovative small area estimation (SAE) methods that integrate NFI data with remote sensing products (Cao et al. 2022; May et al. 2023; Finley et al. 2024; Nothdurft et al. 2025).

Despite the wide variety of SAE methods, they can generally be categorized into two main approaches: area-level and unit-level methods (Rao and Molina 2015). Both aim to estimate the same population parameter of interest but differ in their use of data. In area-level estimation, survey unit response variables are aggregated within each area. These aggregates are referred to as direct estimates and are typically generated using a design-based estimator. Direct estimates are then set as the area-level response variable in a regression model that might include area-level summaries of predictor variables and structured random effects. The goal of areal models is to use auxiliary information to smooth noisy direct estimates. In forest inventory, area-level models have been used extensively to estimate forest attributes (Cao et al. 2022; May et al. 2023; Shannon et al. 2024).

In contrast, unit-level approaches retain response variables at the individual unit (in our case, forest inventory plot) level,

making it possible to leverage precise unit locations to estimate fine-scale spatial relationships more effectively. These unit-level attributes are set as the response variable and coupled with spatial and/or temporally aligned predictor variables, possibly along with structured random effects, in a predictive model. This model is then used to predict for all unobserved units, and predictions are subsequently aggregated to any user-defined area of interest. Unit-level models have been applied to forest inventory contexts in a variety of studies (Breidenbach and Astrup 2012; Finley et al. 2024; Kangas et al. 2025; Shannon et al. 2025). Given their advantages and flexibility, we focus entirely on unit-level models in this study.

Model-based estimators rely entirely on the assumed data-generating process and the selection of an appropriate model. As a result, particular care must be taken in specifying SAE models and conducting rigorous model checking. One of the most effective approaches to evaluating SAE models is through the use of simulated populations that closely mimic the true—but only partially observed—population of interest. Simulated populations allow for examination of how inference varies under different estimators and for comparisons of resulting estimates against “true” values. To ensure that this evaluation is meaningful, it is essential to generate simulated populations using processes that are not closely aligned with the models under assessment (Tzavidis et al. 2018; White et al. 2025a).

Beyond assessment using simulated populations, models can be evaluated through cross-validation using observed data (e.g., leave-one-out or k -fold cross-validation). However, in SAE studies, the primary population parameters of interest exist at the area level. Unit-level models can be assessed using cross-validation at the unit level (i.e., iteratively hold-out one or more observations, predict for those holdouts, and compare the predictions to the holdout true values); however, we must be careful to verify that the unit-level assessments align with how well the estimator performs once predictions are aggregated to the desired areas of interest.

This study evaluates a range of unit-level SAE models for estimating average biomass at the county level in Nevada and Washington, United States (US). A key challenge in this context arises when estimating biomass across areas with a mix of forest and nonforest landcover. Specifically, biomass estimates exhibit a mixture of continuous positive values and true zeros, a phenomenon referred to here as zero-inflation. While the term zero-inflation is commonly used in the statistical literature to describe a discrete distribution with an excessive number of zeros, in this case, biomass follows a continuous distribution with an additional zero component.

Various model-based approaches have been developed to address zero-inflation in SAE. Notably, Pfeiffermann et al. (2008) introduced a two-stage mixture model to account for zero-inflation in the response variable, exploring both frequentist and Bayesian inference. Their findings suggest that mean squared error (MSE) estimation is more straightforward in the Bayesian paradigm because Markov chain Monte Carlo (MCMC) naturally propagates uncertainty across model stages. Expanding on this work, Chandra and Sud (2012) applied the same two-stage model in a frequentist setting and introduced a parametric bootstrap-based MSE estimator.

In forest inventory applications, zero-inflation has received relatively limited attention, despite true zeros commonly occurring in forest inventory attributes. Finley et al. (2011) developed a two-stage model for zero-inflation in continuous forest attributes such as biomass, volume, and age, employing a hierarchical Bayesian framework with Gaussian process-based spatial random effects. Their approach enables unit-level predictions of forest attributes along with uncertainty quantification, though they did not directly produce small area estimates. More recently, White et al. (2025b) applied the zero-inflated SAE model from Chandra and Sud (2012) to NFI data in Nevada, generating county-level biomass estimates. Their study compared the zero-inflated estimator to other commonly used small area estimators, including estimators based on the Battese–Harter–Fuller unit-level model and the Fay–Herriot area-level model (Fay and Herriot 1979; Battese et al. 1988). Their simulation results indicate the zero-inflated estimator improves point estimates and produces competitive MSE estimates, though further refinements remain possible (see fig. 2 in White et al. 2025b).

In this study, we compare and extend model-based SAE approaches that account for zero-inflation, applying them to data from the US Department of Agriculture (USDA) Forest Service Forest Inventory and Analysis (FIA) Program (the NFI program of the US) paired with remotely sensed auxiliary data products for Nevada and Washington, as described in Section 2.1. Specifically, we evaluate nine model-based approaches, including the zero-inflated estimator from Chandra and Sud (2012) and eight hierarchical Bayesian estimators of increasing complexity. These Bayesian estimators include both single-stage models that do not explicitly address zero-inflation and two-stage models designed to account for a preponderance of zeros. With counties within each state defining our small areas of interest, we investigate the effects of incorporating county-varying intercepts, county-varying coefficients, county-specific residual variances (CRVs), and space-varying intercepts. Section 2.2 introduces these models and Sections 2.3 and 2.4 provide further details on the frequentist and Bayesian models, respectively.

Rather than defaulting to the most complex specification, we sequentially introduce model components and evaluate their effect on estimator performance. This approach allows us to identify when added complexity improves estimation and when simpler models suffice. Relative to existing literature, Finley et al. (2011) considered spatial effects but did not include county-specific terms, while Pfeiffermann et al. (2008) and Chandra and Sud (2012) did not incorporate spatial dependencies.

In Section 2.5, we introduce nine estimators based on the models introduced previously and discuss details of model implementation. To evaluate the introduced estimators, we conduct a simulation study following the methodology of White et al. (2025a) and a FIA data application, as described in Sections 2.6 and 2.7, respectively. Metrics for comparison of estimators in the simulation study and FIA data application are introduced in Section 2.8. Section 3 presents the simulation results and applies the estimators to FIA data, using cross-validation to assess unit-level model performance.

Table 1. Description of the candidate models considered for estimating county-level forest biomass.

Estimator	Description
F ZI CVI	A frequentist two-stage estimator. The first stage model is a generalized linear mixed model with a county-varying intercept, and the second stage model is a linear mixed model with a county-varying intercept.
B CVI	A Bayesian single-stage estimator based on a linear mixed model with a county-varying intercept.
B CVC	The same as B CVI, but with county-varying coefficients.
B ZI CVI	A Bayesian two-stage estimator. The first stage model is a generalized linear mixed model with a county-varying intercept, and the second stage model is a linear mixed model with a county-varying intercept.
B ZI CVC	The same as B ZI CVI, but with county-varying coefficients.
B ZI CVI CRV	The same as B ZI CVI, but with county-specific residual variances.
B ZI CVC CRV	The same as B ZI CVC, but with county-specific residual variances.
B ZI CVI SVI CRV	The same as B ZI CVI CRV, but with an added spatial random effect, modeled as a Nearest Neighbor Gaussian process (NNGP) on the intercept.
B ZI CVC SVI CRV	The same as B ZI CVC CRV, but with an added spatial random effect, modeled as a NNGP on the intercept.

Note: Abbreviations are frequentist (F); Bayesian (B); zero-inflated (ZI); county-varying intercept (CVI); county-varying coefficient (CVC); county-specific residual variance (CRV); space-varying intercept (SVI).

Finally, [Section 4](#) summarizes our findings and discusses their implications, and [Section 5](#) presents key takeaways and summarizes directions for future research in SAE methods for forest inventory.

2. Methods

2.1. Data

The motivating data are from the FIA Program and consist of inventory plot estimates of live aboveground tree biomass density (Mg/ha). These estimates were obtained from the most recent measurement for each plot in the FIA database, downloaded on 8 February 2023, for the states of Nevada and Washington ([Burrill et al. 2023](#)). Nevada was selected because it has a distinctive biomass distribution: much of the state's arid environment has little to no biomass, punctuated by sky islands with nonzero forest biomass. Washington was selected because it exhibits large differences in biomass across counties, ranging from very high biomass densities on the Olympic Peninsula to near-zero biomass in counties east of the Cascade Range. At the time of our data query, we accessed the most recent measurement of each FIA plot, with most measurements occurring in the 10-year interval from 2010 to 2019. These data were collected from a panel of plots measured annually across a systematic grid of hexagons approximately 2500 ha in size. Biomass estimates were restricted to live trees only and include all trees with diameters of 2.54 cm or greater. This resulted in a sample size of n equal to 11 848 and 7094 FIA plots for Nevada and Washington, respectively.

Estimators and simulations, described in [Sections 2.2](#) and [2.6](#), respectively, were informed using five auxiliary variables: National Land Cover Dataset Analytical Tree Canopy Cover 2016 (hereafter tcc) ([Yang et al. 2018](#)); LANDFIRE 2010 Digital Elevation Model (hereafter $elev$) ([LANDFIRE 2022](#)); US Geological Survey Terrain Ruggedness Index (hereafter tri) ([LANDFIRE 2022](#)); PRISM mean annual precipitation, 30 year normals (1991–2020) (hereafter ppt) ([Daly et al. 2002](#)); and LANDFIRE 2014 tree/nontree lifeform mask (hereafter tnt) ([Rollins 2009](#); [Picotte et al. 2019](#)). The tcc variable is a mea-

sure of average tree canopy cover in a given pixel, $elev$ gives elevation, tri gives terrain ruggedness, ppt gives 30-year mean precipitation, and tnt is a binary indicator distinguishing between pixels with and without trees. These auxiliary variables were resampled to 90 m resolution and available wall-to-wall in both states. At locations with FIA plots, these variables were matched with the corresponding plot and then used as predictors in the models' regression components and to inform simulated population generation.

2.2. Model-based estimation

We evaluate nine candidate model-based estimators for estimating average biomass density at the county level across Nevada and Washington. The first estimator employs a frequentist approach using a two-stage regression. The remaining eight adopt a Bayesian framework, incorporating both one-stage and two-stage regression structures. A summary of these estimators is provided in [Table 1](#).

To improve adherence to normality assumptions and ensure positive support for back-transformed predictions, models were fit using transformed response variables. While logarithmic transformation of nonzero biomass is common in such contexts, we found it overly aggressive in this application, producing left-skewed transformed distributions. Instead, we applied root-based transformations, selecting state-specific powers to better align with regional biomass distributions: a fourth-root transformation for Washington and a square-root transformation for Nevada. In exploring transformation options, we focused on the distribution of nonzero response values, assuming that the zero-inflated model components would account for excess zeros. Compared to log-based approaches, root-based transformations offer the added benefit that zeros remain unchanged on both the transformed and back-transformed scales.

Although all models are fit at the unit level, our primary inferential target is the average biomass density at the county level, denoted μ_j , where j indexes counties within each state. Let ℓ denote the spatial location of a plot in county j (i.e., $\ell \in j$). Two response variables are modeled: $z(\ell)$, a Bernoulli indica-

tor of biomass presence ($z(\ell) = 1$ for nonzero biomass, $z(\ell) = 0$ otherwise), and $y(\ell)$, the (transformed) continuous biomass value at plots where $z(\ell) = 1$. Predictors are indexed by location as $\mathbf{x}(\ell)$ and $\mathbf{v}(\ell)$, with random effects defined at the county level. This notation also facilitates the introduction of a continuous spatial process random effect, $w(\ell)$, in later models. The relationship between ℓ and j is assumed throughout and is not restated in each model. In the sections that follow, we describe the candidate unit-level models in detail and outline how each is used to estimate these small area population parameters of interest.

2.3. Models for frequentist two-stage estimation

The frequentist estimator (F ZI CVI) adopts a two-stage framework originally developed by Chandra and Sud (2012) and later applied to forest inventory data by White et al. (2025b). The first stage models biomass presence using a Bernoulli mixed model with county-level random intercepts, while the second stage models transformed nonzero biomass values with a linear mixed model, also including county-level random intercepts. Together, these stages provide an estimator and a corresponding MSE estimator suitable for zero-inflated continuous responses.

The continuous response model is

$$(1) \quad y(\ell) = \beta_0 + \tilde{\beta}_{0,j} + \mathbf{x}(\ell)^\top \boldsymbol{\beta} + \varepsilon(\ell),$$

where β_0 is the intercept, $\tilde{\beta}_{0,j} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_{\tilde{\beta}_0}^2)$, $\mathbf{x}(\ell)$ is a $p \times 1$ vector of predictors with regression coefficients $\boldsymbol{\beta}$, and $\varepsilon(\ell) \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \tau^2)$.

The Bernoulli stage for biomass presence is

$$(2) \quad \log\left(\frac{p(\ell)}{1 - p(\ell)}\right) = \alpha_0 + \tilde{\alpha}_{0,j} + \mathbf{v}(\ell)^\top \boldsymbol{\alpha},$$

where $p(\ell)$ denotes the probability of nonzero biomass, i.e., $p(\ell) = \Pr(z(\ell) = 1)$, where $z(\ell)$ is the Bernoulli indicator defined above, α_0 is the intercept, $\tilde{\alpha}_{0,j} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_{\tilde{\alpha}_0}^2)$, and $\boldsymbol{\alpha}$ is a $q \times 1$ coefficient vector for predictors $\mathbf{v}(\ell)$.

F ZI CVI provides a useful benchmark for handling zero-inflated continuous responses with county-level effects. However, uncertainty quantification and extension to more complex model structures can be cumbersome in the frequentist setting. To address these challenges, we next consider a sequence of Bayesian estimators that build from a common unit-level regression framework. These models begin with a simple county-varying intercept specification and progressively incorporate county-varying slopes, explicit two-stage formulations for zero-inflation, CRVs, and spatial random effects.

2.4. Models for Bayesian estimation

The Bayesian estimators summarized in Table 1 extend the frequentist framework by incorporating hierarchical modeling and MCMC-based inference. All models share a common unit-level regression structure but vary in how they ad-

dress zero-inflation, county-level heterogeneity, and spatial dependence. To clarify the role of each modeling feature, we present the estimators in order of increasing complexity, beginning with the baseline B CVI model and sequentially adding components.

All parameters are estimated unless otherwise noted. Prior distributions and hyperparameter values are given in Table A1. Priors were chosen to be weakly informative and commonly used in hierarchical modeling, with the goal of avoiding undue influence on posterior inference.

2.4.1. B CVI

The simplest model is the county-varying intercept specification, the Bayesian analogue to eq. (1),

$$(3) \quad y(\ell) = \beta_0 + \tilde{\beta}_{0,j} + \mathbf{x}(\ell)^\top \boldsymbol{\beta} + \varepsilon(\ell),$$

with $\beta_0 \sim \mathcal{N}(0, \sigma_{\beta_0}^2)$, $\tilde{\beta}_{0,j} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_{\tilde{\beta}_0}^2)$, $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \sigma_{\boldsymbol{\beta}}^2 \mathbf{I}_p)$, and $\varepsilon(\ell) \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \tau^2)$. This model allows county-level differences in mean biomass while assuming predictor effects are constant across counties.

2.4.2. B CVC

The B CVC model extends B CVI to allow regression coefficients to vary across counties, so that

$$(4) \quad y(\ell) = \beta_0 + \tilde{\beta}_{0,j} + \mathbf{x}(\ell)^\top (\boldsymbol{\beta} + \tilde{\boldsymbol{\beta}}_j) + \varepsilon(\ell),$$

where $\tilde{\boldsymbol{\beta}}_j = (\tilde{\beta}_{1,j}, \dots, \tilde{\beta}_{p,j})^\top$ with $\tilde{\beta}_{k,j} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_{\tilde{\beta}_k}^2)$ for $k = 1, \dots, p$. This flexibility captures differences in how predictors relate to biomass across counties.

Moving to zero-inflated models, the Bayesian analogue of the frequentist Bernoulli stage in eq. (2), used for all two-stage Bayesian models, is

$$(5) \quad \log\left(\frac{p(\ell)}{1 - p(\ell)}\right) = \alpha_0 + \tilde{\alpha}_{0,j} + \mathbf{v}(\ell)^\top \boldsymbol{\alpha},$$

with $\alpha_0 \sim \mathcal{N}(0, \sigma_{\alpha_0}^2)$, $\tilde{\alpha}_{0,j} \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \sigma_{\tilde{\alpha}_0}^2)$, and $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \sigma_{\boldsymbol{\alpha}}^2 \mathbf{I}_q)$. All zero-inflated models use eq. (5) to define $z(\ell)$.

2.4.3. B ZI CVI

The first two-stage Bayesian model mirrors the frequentist zero-inflated specification in Section 2.3. The continuous response is modeled conditional on $z(\ell)$ as

$$(6) \quad y(\ell) = z(\ell) \left(\beta_0 + \tilde{\beta}_{0,j} + \mathbf{x}(\ell)^\top \boldsymbol{\beta} + \varepsilon_1(\ell) \right) + (1 - z(\ell)) \varepsilon_2,$$

where $\varepsilon_1(\ell) \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \tau_1^2)$ and $\varepsilon_2 \stackrel{\text{iid}}{\sim} \mathcal{N}(0, \tau_2^2)$, with τ_2^2 fixed at 10^{-6} following Finley et al. (2011). This model accommodates

excess zeros, while retaining a county-varying intercept structure.

2.4.4. B ZI CVC

B ZI CVI is extended to allow predictor effects to vary across counties, giving

$$(7) \quad y(\ell) = z(\ell) \left(\beta_0 + \tilde{\beta}_{0,j} + \mathbf{x}(\ell)^\top \left(\boldsymbol{\beta} + \tilde{\boldsymbol{\beta}}_j \right) + \varepsilon_{1,j}(\ell) \right) + (1 - z(\ell)) \varepsilon_2,$$

which is useful when relationships between biomass and predictors differ across counties.

2.4.5. B ZI CVI CRV

Residual variance often increases with mean biomass, producing heteroscedasticity across counties. To address this, B ZI CVI is extended to allow CRVs so that

$$(8) \quad y(\ell) = z(\ell) \left(\beta_0 + \tilde{\beta}_{0,j} + \mathbf{x}(\ell)^\top \boldsymbol{\beta} + \varepsilon_{1,j}(\ell) \right) + (1 - z(\ell)) \varepsilon_2,$$

with $\varepsilon_{1,j}(\ell) \sim \mathcal{N}(0, \tau_{1,j}^2)$.

2.4.6. B ZI CVC CRV

Combining the previous two extensions yields a model with both county-varying coefficients and CRVs, expressed as

$$(9) \quad y(\ell) = z(\ell) \left(\beta_0 + \tilde{\beta}_{0,j} + \mathbf{x}(\ell)^\top \left(\boldsymbol{\beta} + \tilde{\boldsymbol{\beta}}_j \right) + \varepsilon_{1,j}(\ell) \right) + (1 - z(\ell)) \varepsilon_2,$$

which accounts for county-varying relationships between predictors and biomass and heteroscedasticity in the residuals.

2.4.7. B ZI CVI SVI CRV

Residual spatial variation in biomass can arise from disturbance history, climate gradients, or other unobserved spatial drivers. To capture these, a spatial random effect is added to the intercept, leading to

$$(10) \quad y(\ell) = z(\ell) \left(\beta_0 + \tilde{\beta}_{0,j} + \mathbf{x}(\ell)^\top \boldsymbol{\beta} + w(\ell) + \varepsilon_{1,j}(\ell) \right) + (1 - z(\ell)) \varepsilon_2,$$

where $w(\ell)$ is a spatial random effect that adjusts the intercept based on residual spatial dependence. We estimate $w(\ell)$ using the Nearest Neighbor Gaussian Process (NNGP; [Datta et al. 2016](#); [Finley et al. 2019](#)), which provides substantial improvements in run time with negligible differences in inference and prediction compared to a full Gaussian process. In brief, the vector of random effects $\mathbf{w} = (w(\ell_1), \dots, w(\ell_n))^\top$ is

distributed multivariate normal with mean zero and covariance $\sigma_w^2 \mathbf{R}(\phi)$, where $\mathbf{R}(\phi)$ is the NNGP-derived correlation matrix based on an exponential correlation function with decay parameter ϕ .

2.4.8. B ZI CVC SVI CRV

The most general Bayesian estimator considered combines all features—county-varying coefficients, CRVs, and a spatial random effect on the intercept—so that

$$(11) \quad y(\ell) = z(\ell) \left(\beta_0 + \tilde{\beta}_{0,j} + \mathbf{x}(\ell)^\top \left(\boldsymbol{\beta} + \tilde{\boldsymbol{\beta}}_j \right) + w(\ell) + \varepsilon_{1,j}(\ell) \right) + (1 - z(\ell)) \varepsilon_2,$$

where $w(\ell)$ is defined as in [eq. \(10\)](#). This specification offers the greatest flexibility, simultaneously addressing zero-inflation, county-level heterogeneity, heteroscedasticity, and residual spatial dependence.

2.5. Model implementation and comparison

2.5.1. Frequentist two-stage estimator

The R ([R Core Team 2024](#)) package `saeczi` ([Yamamoto et al. 2025](#)) implements methods presented by [Chandra and Sud \(2012\)](#) to fit the F ZI CVI model. However, in order to implement back-transformation of the response variable and bias correction due to back-transformation, we implement the model without direct use of `saeczi`. Instead, model parameters for F ZI CVI described in [Section 2.3](#) were estimated using restricted maximum likelihood via `lme4` ([Bates et al. 2015](#)).

As described in [Section 1](#), to generate estimates for a small area of interest, we predict biomass and probability of nonzero biomass using [eq. \(1\)](#) and [eq. \(2\)](#), respectively, over a fine grid of prediction locations. For generic prediction location ℓ^* in county j these predictions are

$$(12) \quad \hat{y}^*(\ell^*) = \hat{\beta}_0 + \hat{\beta}_{0,j} + \mathbf{x}(\ell^*)^\top \hat{\boldsymbol{\beta}},$$

$$(13) \quad \hat{p}(\ell^*) = \frac{\exp \left(\hat{\alpha}_0 + \hat{\alpha}_{0,j} + \mathbf{v}(\ell^*)^\top \hat{\boldsymbol{\alpha}} \right)}{1 + \exp \left(\hat{\alpha}_0 + \hat{\alpha}_{0,j} + \mathbf{v}(\ell^*)^\top \hat{\boldsymbol{\alpha}} \right)},$$

where the hats indicate point estimates. The estimate for μ_j is then the average of the product of these predictions over the grid of prediction locations

$$(14) \quad \hat{\mu}_j = \frac{1}{n_j^*} \sum_{\ell^* \in U_j} g^{-1}(\hat{y}^*(\ell^*)) \hat{p}(\ell^*),$$

where $g^{-1}(\cdot)$ is the bias-corrected inverse of the function used to transform the response variable when fitting the model, and U_j is the set of n_j^* prediction locations within county j . To correct for bias introduced from applying the inverses of the root-based transformations we follow [Gregoire et al. \(2008\)](#) and apply bias corrections based on the n th noncentral moment of the Gaussian distribution for the n th root

transformation. In our case, the square- and fourth-root bias-corrected back-transformations are

$$(15) \quad g^{-1}(\hat{y}^*(\ell^*)) = \hat{y}^*(\ell^*)^2 + \hat{\tau}^2,$$

and

$$(16) \quad g^{-1}(\hat{y}^*(\ell^*)) = \hat{y}^*(\ell^*)^4 + 6\hat{y}^*(\ell^*)^2\hat{\tau}^2 + 3\hat{\tau}^4,$$

respectively. The MSE of $\hat{\mu}_j$ is estimated via a parametric bootstrap introduced by [Chandra and Sud \(2012\)](#) and explored further in [White et al. \(2025b\)](#).

2.5.2. Bayesian estimators

Parameter inference for the Bayesian models in [Section 2.4](#) was based on MCMC. Gibbs and Metropolis–Hastings steps were implemented in C++ to efficiently sample from posterior distributions. Code, additional information about the algorithms, and example analyses using simulated data are given in [Finley \(2025\)](#). A list of prior distributions and hyperparameter values is provided in [Appendix A](#). Posterior inference is based on $M = 3000$ thinned post-convergence samples (1000 from each of three chains), following convergence diagnostics and thinning rules from [Gelman et al. \(2013\)](#), with a focus on adequate convergence of regression coefficients and variance parameters.

Inference about biomass at prediction locations and subsequent county-level means is based on samples from the posterior predictive distribution. For example, under the most general B ZI CVC SVI CRV model ([eq. \(11\)](#)), for a generic prediction location ℓ^* in county j we generate one posterior predictive sample per retained MCMC iteration $s = 1, \dots, M$. Specifically,

$$(17) \quad z^{(s)}(\ell^*) \sim \text{Bernoulli} \left(\frac{\exp \left(\alpha_0^{(s)} + \tilde{\alpha}_{0,j}^{(s)} + \mathbf{v}(\ell^*)^\top \boldsymbol{\alpha}^{(s)} \right)}{1 + \exp \left(\alpha_0^{(s)} + \tilde{\alpha}_{0,j}^{(s)} + \mathbf{v}(\ell^*)^\top \boldsymbol{\alpha}^{(s)} \right)} \right),$$

and, given $z^{(s)}(\ell^*)$,

$$(18) \quad y^{(s)}(\ell^*) \sim \mathcal{N} \left(z^{(s)}(\ell^*) \left(\beta_0^{(s)} + \tilde{\beta}_{0,j}^{(s)} + \mathbf{x}(\ell^*)^\top \left(\boldsymbol{\beta}^{(s)} + \tilde{\boldsymbol{\beta}}_j^{(s)} \right) \right) + w^{(s)}(\ell^*), z^{(s)}(\ell^*) \tau_{1,j}^{2(s)} + (1 - z^{(s)}(\ell^*)) \tau_2^{2(s)} \right).$$

Given M posterior predictive samples from [eq. \(18\)](#), county-level averages are computed as

$$(19) \quad \mu_j^{(s)} = \frac{1}{n_j^*} \sum_{\ell^* \in U_j} h^{-1} \left(y^{(s)}(\ell^*) \right),$$

for $s = 1, \dots, M$ and where $h^{-1}(\cdot)$ is the inverse of the function used to transform the response variable when fitting the model. Unlike the frequentist approach where we back-transform the estimated mean of the response variable, $y(\ell^*)$, in the Bayesian approach we back-transform the posterior predictive samples, $y^{(s)}(\ell^*)$, making any bias correction unnecessary ([Stow et al. 2006](#)). Posterior means and credible intervals for μ_j are then obtained directly from these samples, with the Bayesian analogue of [eq. \(14\)](#) given by

$$(20) \quad \hat{\mu}_j = \frac{1}{M} \sum_{s=1}^M \mu_j^{(s)}.$$

Table 2. Auxiliary variables used in model fitting and simulated population generation.

Predictor	Gaussian	Bernoulli	Simulated population
tcc	WA, NV	WA, NV	WA, NV
elev	WA, NV	WA, NV	WA, NV
tri	NV	NV	NV
ppt	WA	WA	WA
tnt	none	WA	WA, NV

The MSE of $\hat{\mu}_j$ is estimated from the variance of the posterior distribution of μ_j .

2.5.3. Variable selection

The five auxiliary variables described in [Section 2.1](#) were used throughout the analysis. In particular, they informed the generation of simulated populations, were included as predictors when fitting models to those populations, and were subsequently applied in the analysis of FIA data.

For simulated population generation, variable selection was informed by domain knowledge about regional biomass drivers. In both Nevada and Washington, tcc and elev were selected as core predictors due to their expected influence on biomass distribution. In Nevada, we additionally included tri to reflect the ecological relevance of sky islands in structuring biomass presence, while in Washington we included ppt to capture the east–west precipitation gradient across the Cascade Range, which strongly influences biomass patterns. To ensure distinct imputation strategies for different vegetation conditions, we stratified the simulated populations in both states by tnt.

For the candidate estimators applied to both simulated and FIA data, we used the same predictor variables within each state to maintain information consistency with the corresponding data-generating processes. Initially, tnt was excluded from the models. While this exclusion had little effect in Nevada, in Washington it led to poor predictive performance in the Bernoulli models. Consequently, we included tnt as a predictor in all Bernoulli models for Washington.

[Table 2](#) summarizes the use of auxiliary variables across model types and data sources.

2.6. Simulation study setup

We conducted a simulation study to evaluate the performance of the estimators introduced in [Section 2.2](#). Simulation studies are especially valuable in SAE contexts because they enable comparisons against known population values, providing direct insight into estimator accuracy and uncertainty quantification. To ensure a fair and realistic evaluation, we followed the methodology proposed by [White et al. \(2025a\)](#).

Simulated populations were constructed using a k -nearest neighbors (k NN) imputation algorithm applied to auxiliary variables, with neighbors weighted by bootstrap inclusion probabilities. This approach imputes forest inventory attributes for every pixel in the population, yielding a synthetic

yet realistic population over which true area-level parameters (e.g., county means) are known. The procedure is described in detail in Algorithm 1 of [White et al. \(2025a\)](#).

We generated separate simulated populations for Nevada and Washington. As shown in [Table 2](#), the kNN algorithm used `tcc`, `elev`, and `tri` in Nevada, and `tcc`, `elev`, and `ppt` in Washington. In both states, `tnt` was used to stratify between treed and nontreed areas. All predictors were centered and scaled before matching to ensure balanced influence. Population generation was implemented using the `kbaabb` R package ([White et al. 2024](#)).

To evaluate estimator performance, we drew $d = 100$ simple random samples from each simulated population. Each sample was selected at the state level, with the number of observations per county matched to the number of FIA plots observed in that county. This preserved the original FIA sampling intensity across space. Estimators were applied to each sample, and performance metrics were computed across the d samples.

2.7. FIA data analysis setup

In addition to the simulation study, we evaluated estimator performance using observed FIA data in Nevada and Washington through a predictive validation approach based on 10-fold cross-validation. This complements the simulation study by evaluating model behavior under realistic conditions, without assuming a known population generating process.

Cross-validation was performed at the unit level. FIA plots were partitioned into 10 folds, and each model was fit to 9 folds and used to predict on the remaining (held-out) fold. This process was repeated until each fold had served as a test set. For each held-out observation, predictions were generated on the original response scale using the model-specific inverse transformations. For frequentist models, we applied the bias-corrected inverse transformation $g^{-1}(\cdot)$ to the fitted values. For Bayesian models, posterior predictive samples were transformed using the inverse transformation $h^{-1}(\cdot)$, and posterior predictive means were used as point predictions. This procedure produced a complete set of unit-level predictions for each estimator, which were then used to evaluate performance across folds.

2.8. Metrics for comparison

Estimator performance was evaluated using two complementary approaches: repeated samples from simulated populations ([Section 2.6](#)) and 10-fold cross-validation on FIA data ([Section 2.7](#)). In both settings, we assessed performance using root mean square error (RMSE), bias, and coverage, applied to either county-level estimates or unit-level predictions depending on the context.

2.8.1. Simulation study

In the simulation study, the focus is on county-level parameters. Let μ_j denote the true county-level mean biomass in county j , and let $\hat{\mu}_{ji}$ be the estimate from the i th sample, for i

$= 1, \dots, d$, where d is the number of simple random samples drawn from the simulated population.

The root mean square error (RMSE) for county j is

$$(21) \quad \text{RMSE}(\hat{\mu}_j) = \sqrt{\frac{1}{d} \sum_{i=1}^d (\hat{\mu}_{ji} - \mu_j)^2}.$$

Bias is defined as the difference between the mean of the estimates and the true value,

$$(22) \quad \text{Bias}(\hat{\mu}_j) = \frac{1}{d} \sum_{i=1}^d \hat{\mu}_{ji} - \mu_j.$$

We also evaluated the bias of the RMSE estimator $\widehat{\text{RMSE}}(\hat{\mu}_j)$ relative to the empirical RMSE defined above. For county j , this is

$$(23) \quad \text{Bias}(\widehat{\text{RMSE}}(\hat{\mu}_j)) = \frac{1}{d} \sum_{i=1}^d \widehat{\text{RMSE}}(\hat{\mu}_{ji}) - \text{RMSE}(\hat{\mu}_j),$$

where $\widehat{\text{RMSE}}(\hat{\mu}_{ji})$ is the estimated RMSE for county j from the i th sample.

Coverage is the proportion of estimated uncertainty intervals that contain the true value. Let I_{ji} be the uncertainty interval for $\hat{\mu}_{ji}$. Coverage for county j is

$$(24) \quad \text{Coverage}(\hat{\mu}_j) = \frac{1}{d} \sum_{i=1}^d \mathbb{1}_{I_{ji}}(\mu_j),$$

where the indicator function is 1 if $\mu_j \in I_{ji}$ and 0 otherwise. For frequentist estimators, I_{ji} is defined as $\hat{\mu}_{ji} \pm 1.96 \cdot \widehat{\text{RMSE}}(\hat{\mu}_{ji})$. For Bayesian estimators, I_{ji} is based on the 0.025 and 0.975 posterior quantiles of μ_j .

2.8.2. FIA data analysis

For the FIA data, performance was evaluated at the unit level using 10-fold cross-validation. Let $y(\ell^*)$ denote the observed biomass value at held-out location ℓ^* and $\hat{y}(\ell^*)$ the corresponding model prediction.

The root mean square prediction error (RMSPE) for all held-out locations ℓ^* is

$$(25) \quad \text{RMSPE}(\hat{y}(\ell^*)) = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}(\ell_i^*) - y(\ell_i^*))^2},$$

where n is the number of held-out locations (all locations are held out once over the 10 folds), $\hat{y}(\ell^*)$ is given by $g^{-1}(\hat{y}^*(\ell^*)) \hat{p}(\ell^*)$ for the frequentist estimator and by the posterior predictive mean $\frac{1}{M} \sum_{s=1}^M h^{-1}(y^{(s)}(\ell^*))$ for the Bayesian estimators.

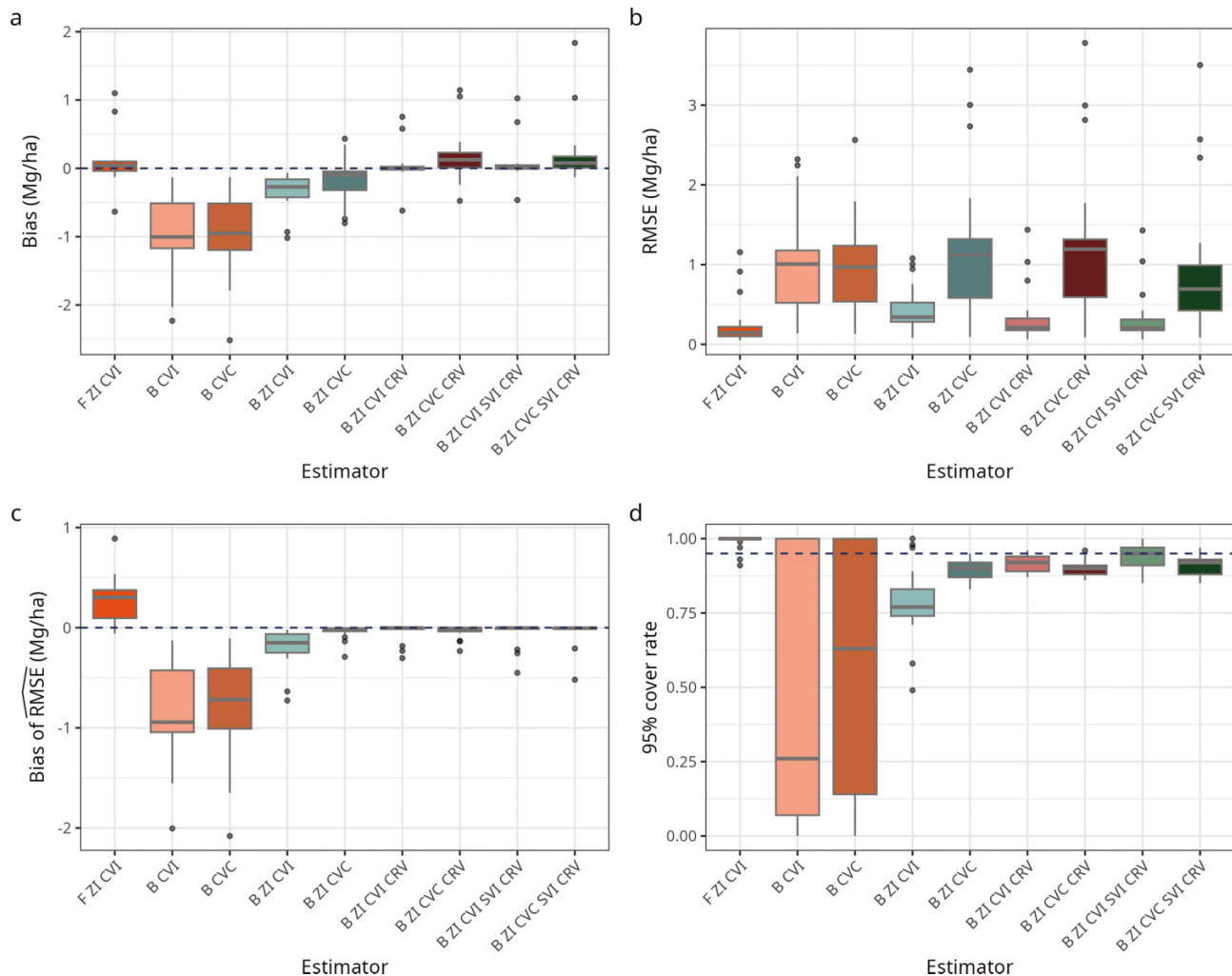
Bias is computed across all held-out locations ℓ^* as

$$(26) \quad \text{Bias}(\hat{y}(\ell^*)) = \frac{1}{n} \sum_{i=1}^n (\hat{y}(\ell_i^*) - y(\ell_i^*)).$$

Coverage is based on the proportion of credible intervals that contain the true held-out value. For Bayesian estimators, the interval is

$$(27) \quad I_i = [Q_{0.025}(h^{-1}(y(\ell_i^*))), Q_{0.975}(h^{-1}(y(\ell_i^*)))] .$$

Fig. 1. Estimator performance metrics in Nevada. The x-axis and fill correspond to estimator, and the y-axis corresponds to the value of the performance metric. Each point represents the performance metric in a given county for a particular estimator. (a) Empirical bias, (b) empirical root mean square error (RMSE), (c) bias of the RMSE estimator, and (d) empirical 95% uncertainty interval coverage. Abbreviations: root mean square error (RMSE); frequentist (F); zero-inflated (ZI); county-varying intercept (CVI); Bayesian (b); county-varying coefficient (CVC); county-specific residual variance (CRV); space-varying intercept (SVI).



Frequentist unit-level prediction intervals were not computed, as the bootstrap procedure from [Chandra and Sud \(2012\)](#) is not designed for this purpose.

3. Results

3.1. Simulation study

The nine estimators introduced in [Section 2](#) were evaluated using repeated simple random samples drawn from the simulated populations described in [Section 2.6](#). Model fitting was conducted separately for Nevada and Washington to reflect state-specific differences in data-generating processes. Estimator performance varied between states, likely due to ecological contrasts: Nevada's forests are confined to high-elevation sky islands, while Washington's forests are concentrated west of the Cascade Range, where precipitation and productivity are higher.

For Nevada, [Fig. 1a](#) shows that single-stage models have the largest empirical bias. Estimators that include CRVs perform

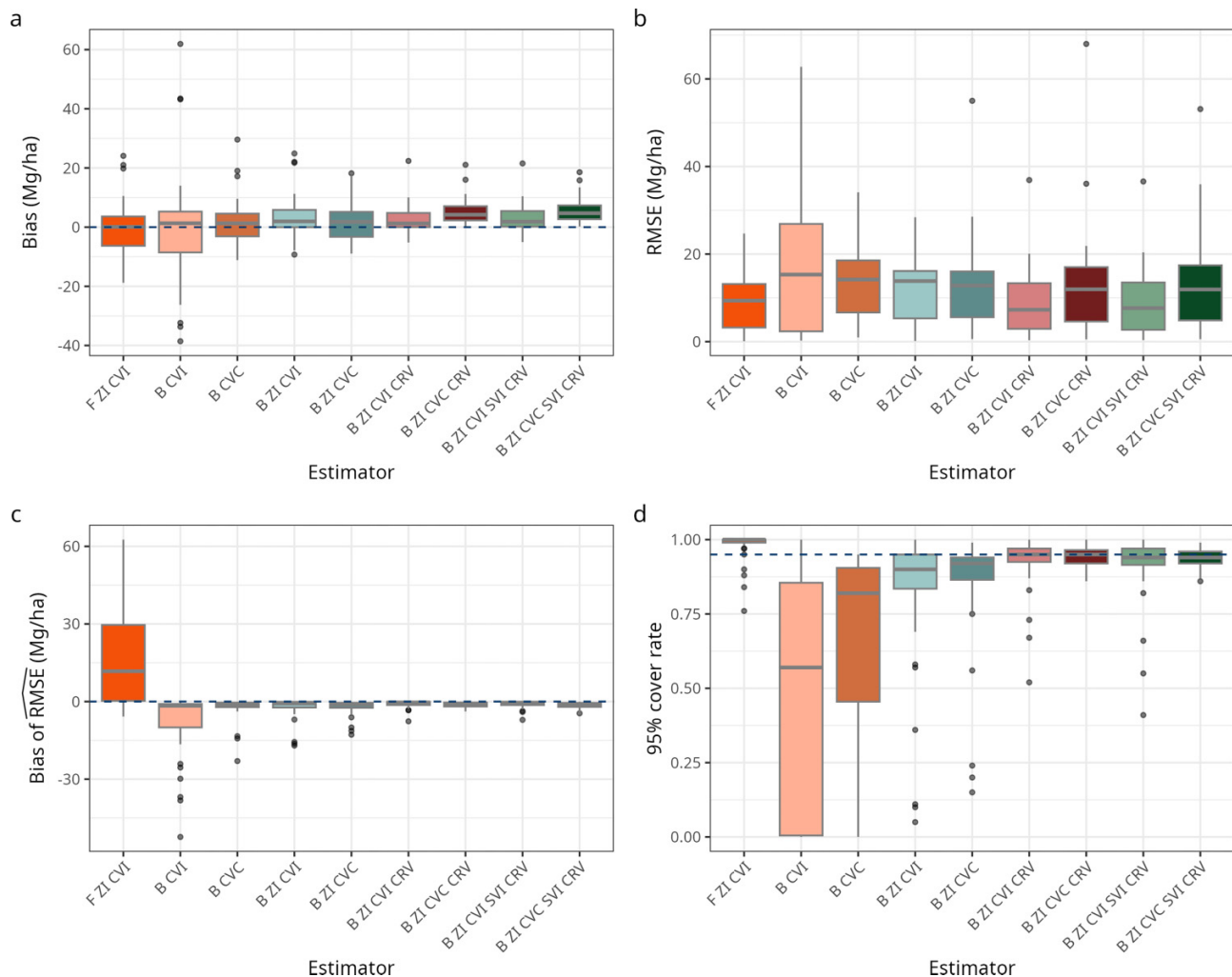
better than those that omit them, and incorporating county-varying coefficients typically increases bias. The least biased estimators are B ZI CVI CRV and B ZI CVI SVI CRV, which yield nearly identical results.

In [Fig. 1b](#), single-stage estimators B CVI and B CVC, as well as two-stage estimators that include CVC terms, consistently show the largest RMSE. The two-stage estimators with county-varying intercepts and county-specific residual variance terms, along with the F ZI CVI, produce the smallest RMSE, followed closely by B ZI CVI.

[Figure 1c](#) shows that B CVI and B CVC underestimate RMSE, consistent with poor model specification. B ZI CVI also underestimates RMSE despite accounting for zero inflation, while F ZI CVI consistently overestimates it. The B ZI models with CRV and/or SVI terms are generally accurate but occasionally produce negative outliers.

[Figure 1d](#) illustrates uncertainty interval coverage. B CVI and B CVC show widespread undercoverage (a result of their large negative RMSE bias), while F ZI CVI consistently displays

Fig. 2. Estimator performance metrics in Washington. The x-axis and fill correspond to estimator, and the y-axis corresponds to the value of the performance metric. Each point represents the performance metric in a given county for a particular estimator. (a) Empirical bias, (b) empirical root mean square error (RMSE), (c) bias of the RMSE estimator, and (d) empirical 95% uncertainty interval coverage. Abbreviations: root mean square error (RMSE); frequentist (F); zero-inflated (ZI); county-varying intercept (CVI); Bayesian (b); county-varying coefficient (CVC); county-specific residual variance (CRV); space-varying intercept (SVI).



overcoverage (a function of its large positive RMSE bias). Estimators incorporating zero inflation, CRV, and SVI perform best overall, with B ZI CVI SVI CRV achieving median coverage closest to 0.95.

For Washington, Fig. 2a shows that the estimators incorporating the two-stage structure all perform similarly. Here too, even the single-stage estimators perform well; however, B CVI shows a marginally larger range of values compared with other models.

In Fig. 2b, the F ZI CVI, B ZI CVI CRV, and B ZI CVI SVI CRV estimators have the lowest median RMSE values. Other two-stage Bayesian estimators perform similarly, with single-stage estimators performing slightly worse, particularly B CVI.

Figure 2c shows that F ZI CVI and B CVI perform poorly when estimating their own variability, whereas most other estimators—particularly those with CRV terms—have minimal bias and accurately estimate RMSE.

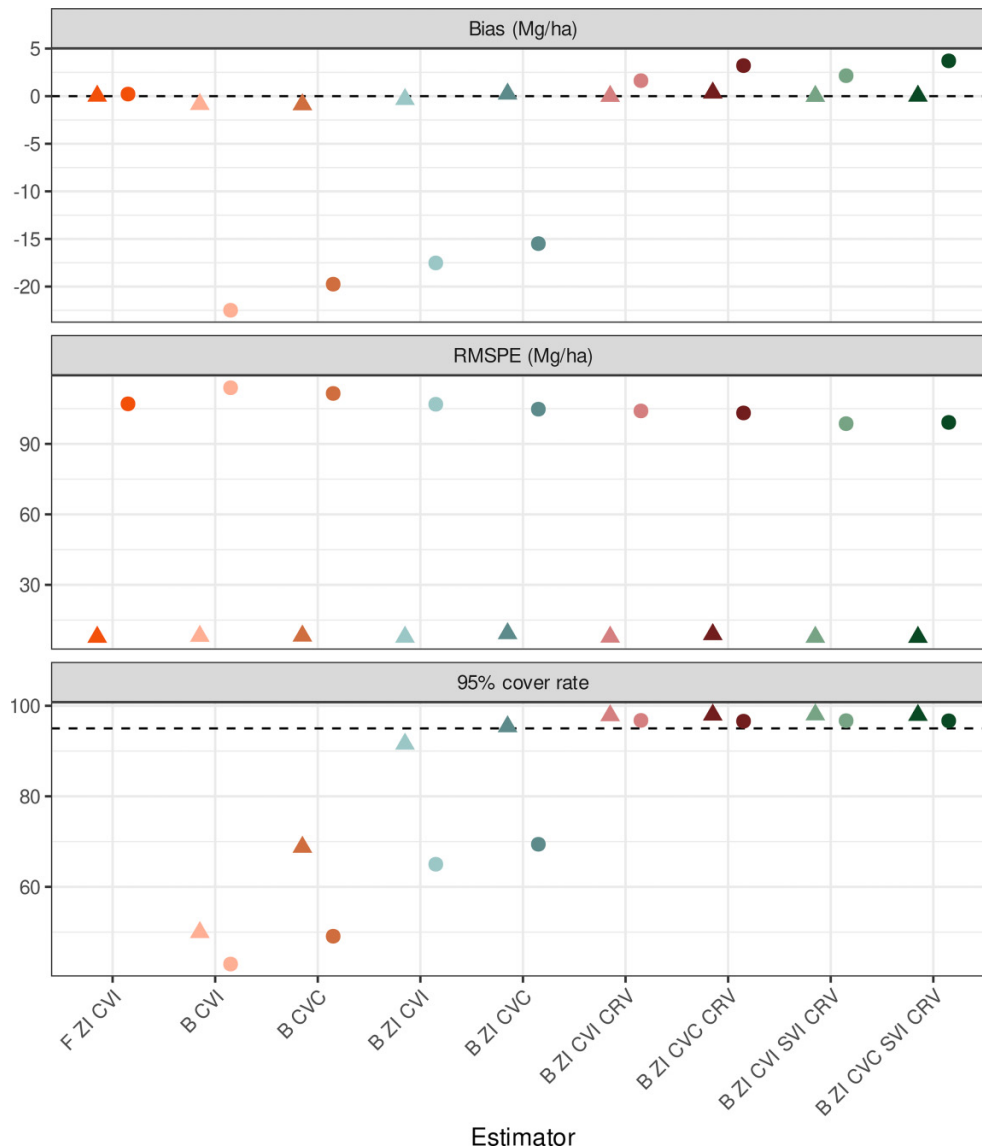
Figure 2d shows that B CVI and B CVC have the lowest coverage rates, while all two-stage models provide better coverage. Those incorporating both CVC effects and CRV terms, such as B ZI CVC SVI CRV, exhibit the best performance. F ZI CVI shows substantial overcoverage, again a function of its large positive RMSE bias.

3.2. FIA data application

Estimator performance for FIA data was evaluated using 10-fold cross-validation. Based on these results and the simulation study, the B ZI CVI SVI CRV model—which performed well in both analyses—was then used to generate county-level biomass estimates.

Cross-validation was conducted at the unit level, with results summarized in Fig. 3. Patterns are broadly consistent with the simulation study. In Nevada, bias was relatively small across all estimators, with two-stage models perform-

Fig. 3. Unit-level cross-validation results by state. Panels display empirical bias (Mg/ha), root mean square prediction error (RMSPE; Mg/ha), and 95% uncertainty interval coverage for each estimator. Nevada (NV) is shown with filled triangles (▲) and Washington (WA) with filled circles (●). Abbreviations: frequentist (F); zero-inflated (ZI); county-varying intercept (CVI); Bayesian (B); county-varying coefficient (CVC); county-specific residual variance (CRV); space-varying intercept (SVI). Dashed horizontal line indicates nominal 95% coverage.



ing marginally better than single-stage models. In Washington, F ZI CVI showed the smallest bias, followed by the two-stage Bayesian models with county-varying intercepts and residual variances (B ZI CVI CRV and B ZI CVI SVI CRV).

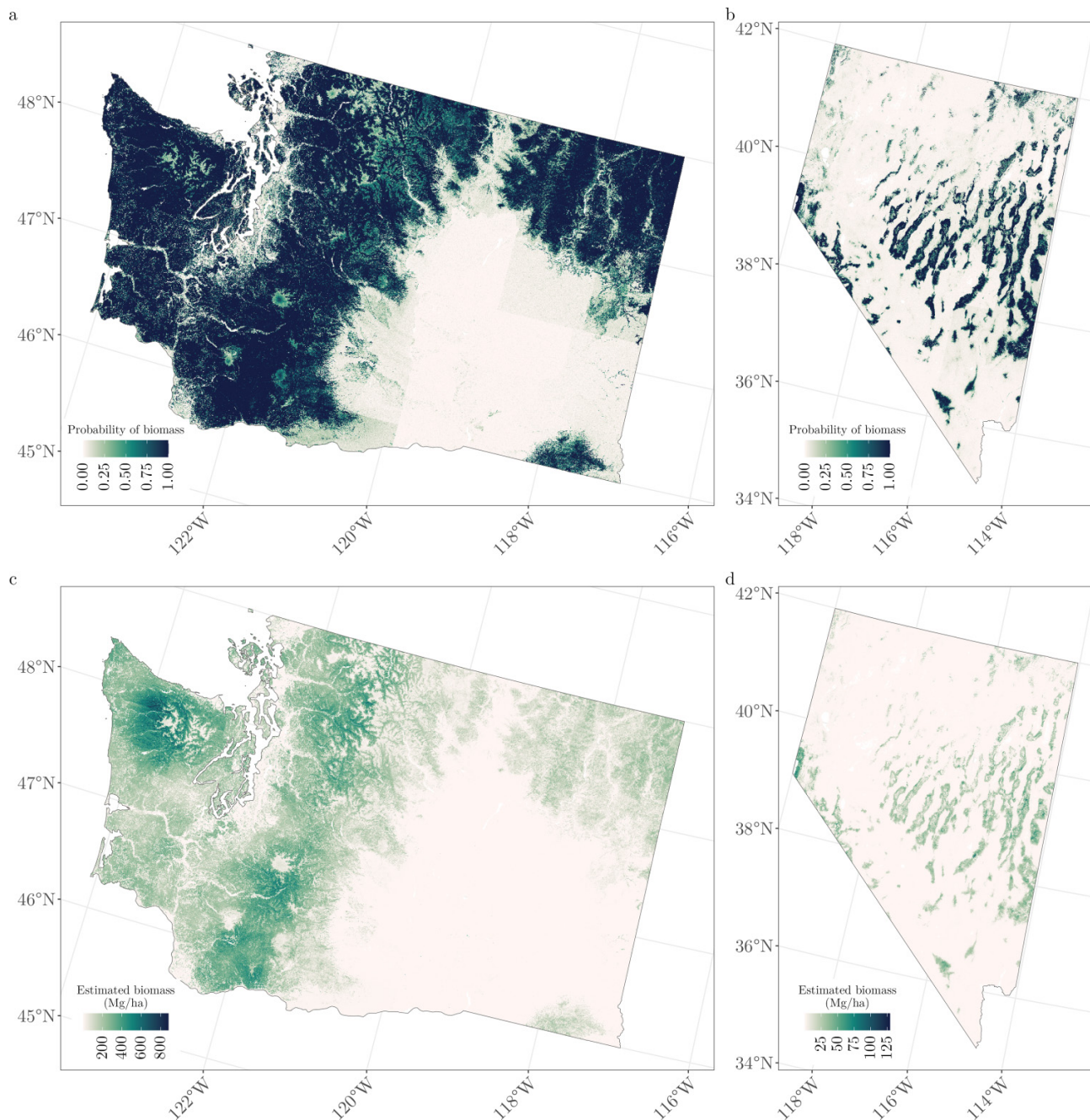
For RMSPE (Fig. 3), prediction accuracy was similar across models within each state. In Washington, the Bayesian models with a space-varying intercept term (SVI models) showed marginal improvements in predictive performance.

Coverage rates generally improved with model complexity, with the best performance observed for models that included county-varying residual terms. Coverage is not reported for the frequentist estimator because unit-level prediction intervals are not available in closed form or via bootstrap for the two-stage model.

Simulation results in Section 3.1 and FIA cross-validation results (Fig. 3) show that B ZI CVI SVI CRV performed well in both states and across candidate models, so this model was used to generate pixel- and county-level biomass estimates.

Model parameter estimates for this model are provided in Tables A2 and A3 and Figs. A1 and A2 for Nevada and Washington, respectively. Because our focus is on predictive performance, and given the challenges of interpreting regression coefficients in the presence of random effects—particularly spatial random effects (e.g., Mäkinen et al. 2022; Zimmerman and Hoef 2022; Gilbert et al. 2024)—we do not emphasize interpretation of individual coefficients. The county- and space-varying effects shown in Figs. A1 and A2 reveal variability and

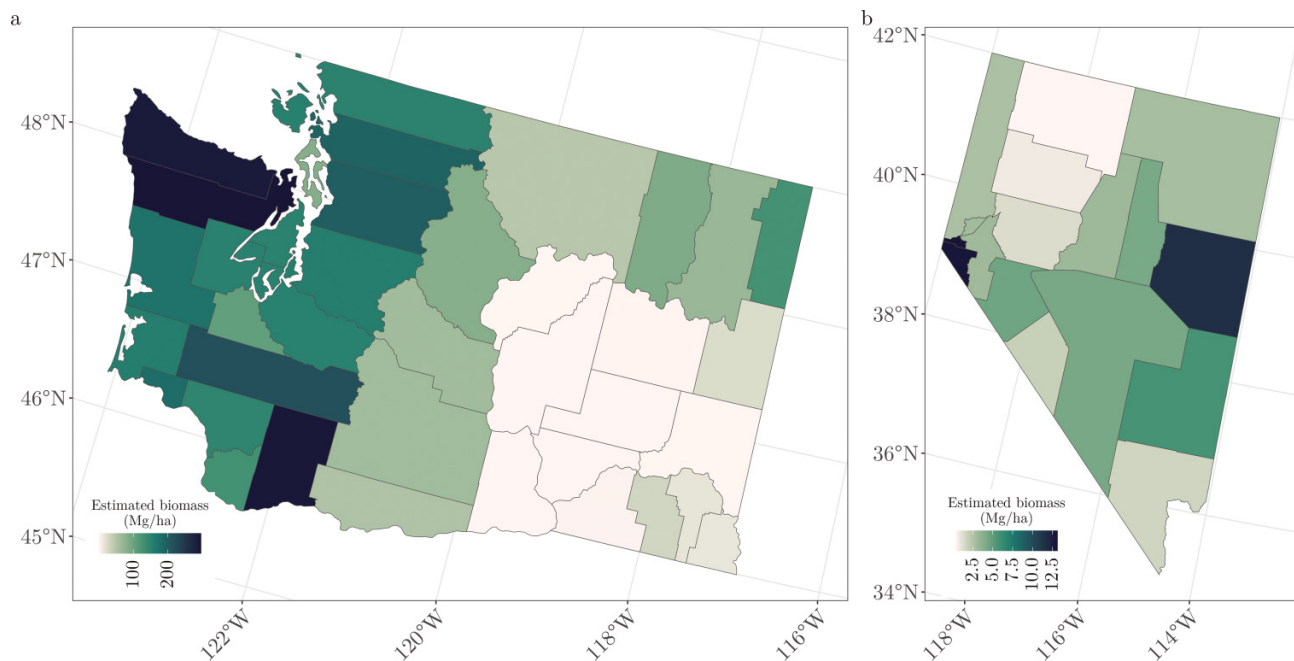
Fig. 4. Pixel-level estimates of biomass probability (*a* and *b*) and estimated biomass (*c* and *d*) in Washington (*a* and *c*) and Nevada (*b* and *d*). Biomass probabilities are derived from the Bayesian Bernoulli model with county-varying intercepts; biomass estimates are derived from the Bayesian zero-inflated model with county-varying intercept, space-varying intercept, and county-specific residual variances. Figure was created using R version 4.5.1 with package ggplot2. The shapefile is courtesy of the FIESTAR package (Frescino et al. 2023). Base map from the ggplot2R package (Wickham 2016).



patterns consistent with the spatial heterogeneity of forest and biomass distributions. In particular, the spatial random effects in Washington, and to a lesser degree in Nevada, identify expected patterns of biomass distribution and compensate for missing predictors and the limitations of county-level random effects. While these figures highlight the potential importance of SVI terms, the formal cross-validation metrics show only marginal improvements in RMSPE for the spatial models.

Figure 4 shows pixel-level biomass probabilities (a–b) and biomass estimates (c–d) for both states. These maps illustrate how the modeling framework supports estimation and uncertainty quantification at arbitrary spatial scales, and how each model stage contributes to the results. In Washington, the pixel-level estimates reflect forest distribution and biomass density, with a clear gradient across the Cascade Range. In Nevada, the pixel-level estimates highlight the sky islands and extensive nonforest areas.

Fig. 5. County-level estimates of average biomass (Mg/ha) in Washington (a) and Nevada (b). Estimates are produced from the Bayesian zero-inflated model with county-varying intercept, space-varying intercept, and county-specific residual variances. Figure was created using R version 4.5.1 with package ggplot2. The shapefile and county boundaries are courtesy of the FI-ESTAR package (Frescino et al. 2023). Base map from the ggplot2R package (Wickham 2016).



As described in Section 2.5.2, the pixel-level posterior distributions (whose means are shown in Figs. 4c and 4d) were summarized to produce the county-level μ_j estimates shown in Fig. 5.

4. Discussion

This study evaluated nine model-based estimators for county-level biomass in Nevada and Washington, focusing on how zero-inflation, county-level effects, residual variance structures, and spatial components influence estimator performance. We used two complementary approaches: repeated sampling from simulated populations, which allowed direct comparison to known county-level values, and unit-level cross-validation using FIA data, which assessed predictive performance under real-world conditions. The latter approach, by necessity, evaluated estimator performance at the unit level rather than for the area-level parameters that are the target of inference.

Across both analyses, two-stage models that explicitly accounted for zero-inflation outperformed single-stage models. Incorporating CRVs consistently reduced bias and RMSE, particularly in Nevada, where biomass is concentrated on isolated sky islands. Adding spatial random effects (SVIs) further improved performance in the FIA analysis, especially in Washington, where biomass patterns are strongly structured by the Cascade Range. In contrast, the benefits of including SVIs were less pronounced in the simulation study, suggesting that the simulated populations may not fully capture the degree of spatial autocorrelation present in the FIA data.

This limitation has been noted previously. White et al. (2024) highlighted that kbaabb-generated populations lack

the degree of spatial autocorrelation commonly observed in FIA data. In our FIA analyses, the estimated effective spatial range for the SVI components was approximately 114 km in Washington and 7 km in Nevada (Tables A3 and A2, respectively), confirming the presence of meaningful residual spatial structure in the observed data. Future simulation work that incorporates realistic spatial dependence—for example, through nonparametric smoothing or spatial processes—would enable more realistic assessments of spatial estimators.

Single-stage estimators showed the highest bias and poorest coverage overall. CVC effects tended to increase bias in Nevada without clear benefits, and in Washington they provided only modest improvements in coverage. The frequentist two-stage estimator performed reasonably well but exhibited a pronounced positive bias in RMSE estimation, leading to overly wide confidence intervals and inflated coverage rates. Refining its MSE estimator would improve its utility as a baseline approach. Future work may include refining or developing new MSE estimators for the frequentist two-stage estimator.

The FIA cross-validation results reinforced the advantages of including CRV and SVI terms. Models combining these components achieved the lowest RMSPE and bias, with coverage rates approaching nominal levels. The frequentist and more complex Bayesian zero-inflated estimators were particularly effective in Nevada, where extensive nonforest areas result in many true zeros that single-stage models cannot accommodate.

Taken together, the simulation and FIA analyses highlight how model structure influences estimator performance and underscore the importance of accounting for key popula-

tion features when applying SAE methods to forest inventory data.

5. Conclusion

This study demonstrates that explicitly accommodating zero-inflation, variance heterogeneity at the county level, and residual spatial structure can be important for producing accurate and well-calibrated biomass estimates across contrasting ecological settings. Bayesian two-stage models provide a flexible framework for integrating these components, offering improved calibration, reduced bias, and robust uncertainty propagation compared with simpler alternatives.

The simulation and FIA cross-validation analyses provided complementary perspectives. While similar trends in estimator performance were seen between the simulated county-level and FIA unit-level assessments, the county-level assessments are better suited to model selection because the target parameters are defined at the county level. Together, these approaches offer a more complete view of estimator behavior under both controlled and real-world conditions.

The results offer practical guidance for producing reliable county-level biomass estimates, which are increasingly needed for resource monitoring, wildfire risk assessment, and local-scale planning. By improving population generation methods to better reflect spatial structure and extending these models to additional spatial and temporal domains, future work can further strengthen the role of model-based SAE methods in forest inventory and environmental monitoring.

Acknowledgements

The authors thank the Forest Inventory & Analysis Program for the data.

Article information

History dates

Received: 6 June 2025

Accepted: 2 February 2026

Accepted manuscript online: 18 February 2026

Version of record online: 9 April 2026

Copyright

© 2026 Authors White, Finley, Yamamoto, Green, and MacFarlane. This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/) (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

Data availability

The data include confidential plot data, which cannot be shared publicly. FIA data can be accessed through the FIA DataMart (<https://apps.fs.usda.gov/fia/datamart/datamart.html>). Requests for data used here or other requests including confidential data should be directed to FIA's Spatial Data Services (<https://www.fs.usda.gov/research/programs/fia/sds>).

Author information

Author ORCIDs

Grayson W. White <https://orcid.org/0000-0003-4993-2792>

Hans-Erik Andersen <https://orcid.org/0000-0002-5546-9522>

Author contributions

Conceptualization: GWW, AOF

Data curation: GWW, TSF, GMD

Formal analysis: GWW, AOF

Funding acquisition: AOF, GMD

Investigation: GWW, AOF

Methodology: GWW, AOF, JLG

Project administration: AOF, JLG, DWM, HEA

Resources: AOF

Software: GWW, AOF, JKY, TSF

Supervision: AOF, JLG, DWM

Validation: GWW, AOF

Visualization: GWW, AOF

Writing – original draft: GWW, AOF, JLG, TSF

Writing – review & editing: GWW, AOF, JKY, JLG, TSF, DWM, HEA, GMD

Competing interests

JKY was previously full-time employed by RedCastle Resources, Inc. GWW is part-time employed by RedCastle Resources, Inc. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

- Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., et al. 2015. Package 'lme4'. *Convergence*, **12**(1): 2.
- Battese, G.E., Harter, R.M., and Fuller, W.A. 1988. An error-components model for prediction of county crop areas using survey and satellite data. *J. Am. Stat. Assoc.* **83**(401): 28–36. doi:[10.1080/01621459.1988.10478561](https://doi.org/10.1080/01621459.1988.10478561).
- Breidenbach, J., and Astrup, R. 2012. Small area estimation of forest attributes in the Norwegian National Forest Inventory. *Eur. J. Forest Res.* **131**(4): 1255–1267. doi:[10.1007/s10342-012-0596-7](https://doi.org/10.1007/s10342-012-0596-7).
- Burrill, E., Christensen, G., Conkling, B., DiTommaso, A., Lepine, L., Perry, C., et al. 2023. Forest inventory and analysis database: database description and user guide for phase 2 (version: 9.1). U.S. Department of Agriculture, Forest Service. 1066p. <https://www.fia.fs.usda.gov/library/database-documentation/index>.
- Cao, Q., Dettmann, G.T., Radtke, P.J., Coulston, J.W., Derwin, J., Thomas, V.A., et al. 2022. Increased precision in county-level volume estimates in the United States National Forest Inventory with area-level small area estimation. *Front. Forest Glob. Change*, **5**: 769917. doi:[10.3389/ffgc.2022.769917](https://doi.org/10.3389/ffgc.2022.769917).
- Chandra, H., and Sud, U.C. 2012. Small area estimation for zero-inflated data. *Commun. Stat. Simulat. Comput.* **41**: 632–643. doi:[10.1080/03610918.2011.598991](https://doi.org/10.1080/03610918.2011.598991).
- Daly, C., Gibson, W.P., Taylor, G.H., Johnson, G.L., and Pasteris, P. 2002. A knowledge-based approach to the statistical mapping of climate. *Clim. Res.* **22**(2): 99–113. doi:[10.3354/cr022099](https://doi.org/10.3354/cr022099).
- Datta, A., Banerjee, S., Finley, A.O., and Gelfand, A.E. 2016. Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *J. Am. Stat. Assoc.* **111**(514): 800–812. doi:[10.1080/01621459.2015.1044091](https://doi.org/10.1080/01621459.2015.1044091).
- Fay, R.E.I., and Herriot, R.A. 1979. Estimates of income for small places: an application of james-stein procedures to census data. *J. Am. Stat. Assoc.* **74**(366a): 269–277. doi:[10.1080/01621459.1979.10482505](https://doi.org/10.1080/01621459.1979.10482505).

- Finley, A.O. 2025. Models for zero-inflated data. Available from https://github.com/finleya/zi_models [accessed 29 January 2025].
- Finley, A.O., Andersen, H.-E., Babcock, C., Cook, B.D., Morton, D.C., and Banerjee, S. 2024. Models to support forest inventory and small area estimation using sparsely sampled lidar: a case study involving g-lidat lidar in tanana, alaska. *J. Agr. Biol. Environ. Stat.* **29**(4): 695–722. doi:10.1007/s13253-024-00611-3.
- Finley, A.O., Banerjee, S., and MacFarlane, D.W. 2011. A hierarchical model for quantifying forest variables over large heterogeneous landscapes with uncertain forest areas. *J. Am. Stat. Assoc.* **106**(493): 31–48. doi:10.1198/jasa.2011.ap09653.
- Finley, A.O., Datta, A., Cook, B.D., Morton, D.C., Andersen, H.E., and Banerjee, S. 2019. Efficient algorithms for Bayesian nearest neighbor Gaussian processes. *J. Comput. Graph. Stat.* **28**(2): 401–414. doi:10.1080/10618600.2018.1537924.
- Frescino, T.S., Moisen, G.G., Patterson, P.L., Toney, C., and White, G.W. 2023. FIESTA: a forest inventory estimation and analysis R package. *Ecogr. Softw. Note*.
- Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. 2013. Bayesian data analysis. 3rd ed. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.
- Gilbert, B., Ogburn, E.L., and Datta, A. 2024. Consistency of common spatial estimators under spatial confounding. *Biometrika*, **112**(2): asae070. doi:10.1093/biomet/asae070.
- Gregoire, T.G., Lin, Q.F., Boudreau, J., and Nelson, R. 2008. Regression estimation following the square-root transformation of the response. *Forest Sci.* **54**(6): 597–606. doi:10.1093/forestscience/54.6.597.
- Kangas, A., Myllymäki, M., and Packalen, P. 2025. Small area estimators in a simulation test. *Can. J. Forest Res.* **55**: 1–17. doi:10.1139/cjfr-2024-0070.
- LANDFIRE. 2022. Elevation, LANDFIRE 2.0.0, U.S. Department of the Interior, Geological Survey, and U.S. Department of Agriculture. <https://landfire.gov/topographic/elevation>.
- Mäkinen, J., Numminen, E., Niittynen, P., Luoto, M., and Vanhatalo, J. 2022. Spatial confounding in bayesian species distribution modeling. *Ecography*, **2022**(11): e06183. doi:10.1111/ecog.06183.
- May, P., McConville, K.S., Moisen, G.G., Bruening, J., and Dubayah, R. 2023. A spatially varying model for small area estimates of biomass density across the contiguous United States. *Remote Sens. Environ.* **286**: 113420. doi:10.1016/j.rse.2022.113420.
- Nothdurft, A., Sarkkela, V., Ofner-Graff, T., Tockner, A., Gollob, C., Ritter, T., et al. 2025. Small area estimation of growing stock timber volume, Basal Area, mean stem diameter, and stem density for mountain forests in Austria. *Can. J. Forest Res.* **55**: 1–20. doi:10.1139/cjfr-2024-0302.
- Pfeffermann, D., Terry, B., and Moura, F. 2008. Small area estimation under a two-part random effects model with application to estimation of literacy in developing countries. *Surv. Methodol.* **34**.
- Picotte, J.J., Dockter, D., Long, J., Tolk, B., Davidson, A., and Peterson, B. 2019. LANDFIRE remap prototype mapping effort: developing a new framework for mapping vegetation classification, change, and structure. *Fire*, **2**(2): 35. doi:10.3390/fire2020035.
- Prisley, S., Bradley, J., Clutter, M., Friedman, S., Kempka, D., Rakestraw, J., and Sonne Hall, E. 2021. Needs for small area estimation: perspectives from the us private forest sector. *Front. Forests Glob. Change*, **4**: 746439. doi:10.3389/ffgc.2021.746439.
- R Core Team. 2024. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
- Rao, J., and Molina, I. 2015. Small area estimation. 2nd ed. Wiley, ISBN: 978-1-118-73578-7.
- Rollins, M.G. 2009. LANDFIRE: a nationally consistent vegetation, wildland fire, and fuel assessment. *Int. J. Wildland Fire*, **18**(3): 235–249. doi:10.1071/WF08088.
- Shannon, E.S., Finley, A.O., Domke, G.M., May, P.B., Andersen, H.-E., Gaines, G.C., III, and Banerjee, S. 2024. Toward spatio-temporal models to support national-scale forest carbon monitoring and reporting. *Environ. Res. Lett.* **20**(1): 014052. doi:10.1088/1748-9326/ad9e07.
- Shannon, E.S., Finley, A.O., May, P.B., Domke, G.M., Andersen, H.-E., Gaines, G.C., III, et al. 2025. Leveraging National Forest Inventory data to estimate forest carbon density status and trends for small areas. *Forest Ecol. Manage.* **596**: 122999. doi:10.1016/j.foreco.2025.122999.
- Stow, C.A., Reckhow, K.H., and Qian, S.S. 2006. A Bayesian approach to retransformation bias in transformed regression. *Ecology*, **87**(6): 1472–1477. doi:10.1890/0012-9658(2006)87[1472:ABATRB]2.0.CO;2.
- Tzavidis, N., Zhang, L.-C., Luna, A., Schmid, T., and Rojas-Perilla, N. 2018. From start to finish: a framework for the production of small area official statistics. *J. Roy. Stat. Soc. Ser. A (Stat. Soc.)* **181**(4): 927–979. doi:10.1111/rssa.12364.
- U.S. Senate 2023. S. Rep. 118-83 - Department of the Interior, Environment, and Related Agencies Appropriations Bill.
- White, G.W., Wieczorek, J.A., Cody, Z.W., Tan, E.X., Chistolini, J.O., McConville, K.S., et al. 2025a. A method for empirically assessing small area estimators via bootstrap-weighted k-nearest-neighbor artificial populations, with applications to forest inventory. *Forestry*, cpaf071. doi:10.1093/forestry/cpaf071.
- White, G.W., Wieczorek, J.A., Frescino, T.S., and McConville, K.S. 2024. kbaabb: generates an artificial population based on the KBAABB methodology. R package version 0.0.0.9000. <https://github.com/grays-onwhite/kbaabb>.
- White, G.W., Yamamoto, J.K., Elsyad, D.H., Schmitt, J.F., Korsgaard, N.H., Hu, J.K., et al. 2025b. Small area estimation of forest biomass via a two-stage model for continuous zero-inflated data. *Can. J. Forest Res.* **55**: 1–19.
- Wickham, H. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York.
- Wiener, S.S., Bush, R., Nathanson, A., Pelz, K., Palmer, M., Alexander, M.L., et al. 2021. United states forest service use of forest inventory data: examples and needs for small area estimation. *Front. Forests Glob. Change*, **4**: 763487. doi:10.3389/ffgc.2021.763487.
- Yamamoto, J., Elsyad, D., White, G., Schmitt, J., Korsgaard, N., McConville, K., and Hu, K. 2025. saeczi: small area estimation for continuous zero inflated data. R package version 0.2.0.9000, commit f580319049b152c890033c770913b7a296ce63cb.
- Yang, L., Jin, S., Danielson, P., Homer, C., Gass, L., Bender, S.M., et al. 2018. A new generation of the United States National Land Cover Database: Requirements, research priorities, design, and implementation strategies. *ISPRS J. Photo. Remote Sens.* **146**: 108–123. doi:10.1016/j.isprsjprs.2018.09.006.
- Zimmerman, D.L., and Hoef, J.M.V. 2022. On deconfounding spatial confounding in linear models. *Am. Stat.* **76**(2): 159–167. doi:10.1080/00031305.2021.1946149.

Appendix A

Table A1 includes the prior distribution and hyperparameter values for all parameters used for the Bayesian models.

Table A1. Prior distributions and hyperparameter values used for the Bayesian models.

Parameter	Prior distribution	Hyperparameter values
α_0	$\mathcal{N}(\mu, \sigma^2)$	$\mu = 0, \sigma^2 = 1000$
$\tilde{\alpha}_{0,j}$	$\mathcal{N}(\mu, \sigma^2)$	$\mu = 0, \sigma^2 = \sigma_{\alpha_0}^2$
α_k	$\mathcal{N}(\mu, \sigma^2)$	$\mu = 0, \sigma^2 = 1000$
$\sigma_{\alpha_0}^2$	$\mathcal{IG}(a, b)$	$a = 2, b = 1$
β_0	$\mathcal{N}(\mu, \sigma^2)$	$\mu = 0, \sigma^2 = 1000$
$\tilde{\beta}_{0,j}$	$\mathcal{N}(\mu, \sigma^2)$	$\mu = 0, \sigma^2 = \sigma_{\beta_0}^2$
β_k	$\mathcal{N}(\mu, \sigma^2)$	$\mu = 0, \sigma^2 = 1000$
$\tilde{\beta}_{k,j}$	$\mathcal{N}(\mu, \sigma^2)$	$\mu = 0, \sigma^2 = \sigma_{\beta_{k,j}}^2$
$\sigma_{\beta_0}^2$	$\mathcal{IG}(a, b)$	$a = 2, b = 1$
$\sigma_{\beta_{k,j}}^2$	$\mathcal{IG}(a, b)$	$a = 2, b = 1$
σ_w^2	$\mathcal{IG}(a, b)$	$a = 2, b = 1$
ϕ	$\mathcal{U}(a, b)$	$a = 0.003, b = 3$
τ^2	$\mathcal{IG}(a, b)$	$a = 2, b = 1$
τ_1^2	$\mathcal{IG}(a, b)$	$a = 2, b = 1$
$\tau_{1,j}^2$	$\mathcal{IG}(a, b)$	$a = 2, b = 1$

Table A2. Posterior means and 95% credible intervals (CrI) for parameters of the Bayesian zero-inflated county-varying intercept space-varying intercept county-specific residual variance model eq. (10) fit to Forest Inventory and Analysis data in Nevada.

Bernoulli model for $z(\ell)$	
Parameter	Posterior mean (95% CrI)
α_0	-6.410 (-7.023, -5.794)
α_{tcc}	0.148 (0.114, 0.180)
α_{elev}	-0.010 (-0.019, 0.000)
α_{tri}	0.268 (0.253, 0.284)
$\sigma_{\alpha_0}^2$	0.389 (0.191, 0.756)
Continuous model for $y(\ell)$	
Parameter	Posterior mean (95% CrI)
β_0	1.450 (0.755, 2.188)
β_{tcc}	0.058 (0.024, 0.094)
β_{elev}	-0.004 (-0.012, 0.003)
β_{tri}	0.080 (0.074, 0.086)
$\sigma_{\beta_0}^2$	0.259 (0.119, 0.519)
σ_w^2	1.502 (1.218, 1.742)
Eff. rng. (km)	6.679 (5.189, 8.434)

Note: The Bernoulli model corresponds to biomass presence $z(\ell)$, and the continuous model corresponds to transformed biomass $y(\ell)$. Estimates for the remaining model parameters are shown in Fig. A1 because they vary over counties or across a continuous spatial surface and are more effectively visualized as maps. The effective range (Eff. rng.) is computed using estimates for ϕ as $-\log(0.05)/\phi$ and represents the distance (km) at which spatial correlation drops to 0.05.

Table A3. Posterior means and 95% credible intervals (CrI) for parameters of the Bayesian zero-inflated county-varying intercept space-varying intercept county-specific residual variance model fit to Forest Inventory and Analysis data in Washington.

Bernoulli model for $z(\ell)$	
Parameter	Posterior mean (95% CrI)
α_0	-4.785 (-5.258, -4.387)
α_{tcc}	0.124 (0.097, 0.151)
α_{elev}	0.014 (-0.006, 0.034)
α_{ppt}	0.102 (0.095, 0.109)
α_{tnt}	1.222 (0.989, 1.460)
$\sigma_{\alpha_0}^2$	0.863 (0.487, 1.424)
Continuous model for $y(\ell)$	
Parameter	Posterior mean (95% CrI)
β_0	1.752 (1.606, 1.893)
β_{tcc}	-0.008 (-0.016, 0.001)
β_{elev}	0.014 (0.008, 0.019)
β_{ppt}	0.024 (0.022, 0.025)
$\sigma_{\beta_0}^2$	0.084 (0.052, 0.134)
σ_w^2	0.157 (0.112, 0.195)
Eff. rng. (km)	114.009 (93.033, 137.224)

Note: The Bernoulli model corresponds to biomass presence $z(\ell)$, and the continuous model corresponds to transformed biomass $y(\ell)$. Estimates for the remaining model parameters are shown in Fig. A2 because they vary over counties or across a continuous spatial surface and are more effectively visualized as maps. The effective range (Eff. rng.) is computed using estimates for ϕ as $-\log(0.05)/\phi$ and represents the distance (km) at which spatial correlation drops to 0.05.

Fig. A1. Posterior means for county-level and space-varying parameters of the Bayesian zero-inflated county-varying intercept space-varying intercept county-specific residual variance model fit to Forest Inventory and Analysis data in Nevada. Each sub-figure corresponds to a single parameter, with the posterior mean displayed above the scale bar. Parameter definitions and fixed-effect estimates are provided in Table A2. Figure was created using R version 4.5.1 with package ggplot2. The shapefile and county boundaries are courtesy of the FIESTAR package (Frescino et al. 2023). Base map from the ggplot2R package (Wickham 2016).

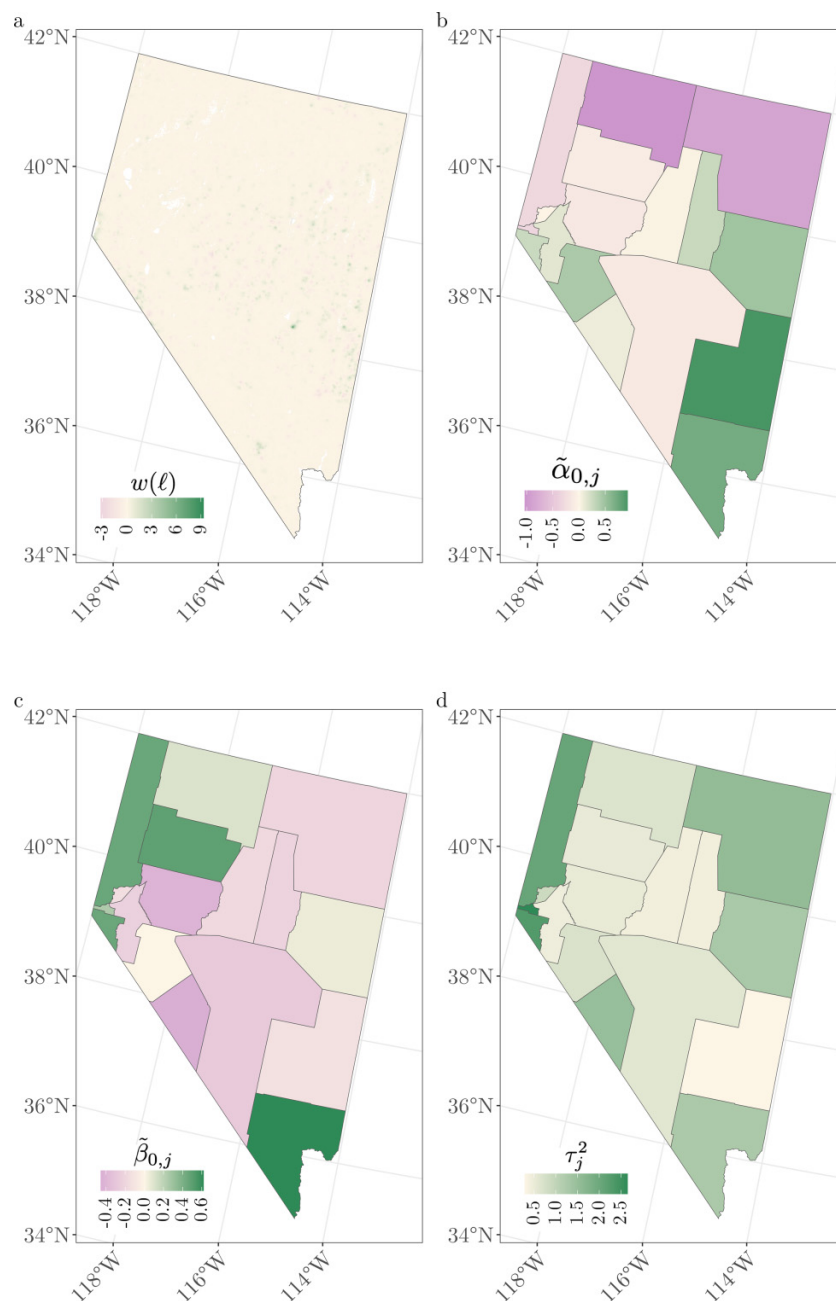


Fig. A2. Posterior means for county-level and space-varying parameters of the Bayesian zero-inflated county-varying intercept space-varying intercept county-specific residual variance model fit to Forest Inventory and Analysis data in Washington. Each subfigure corresponds to a single parameter, with the posterior mean displayed above the scale bar. Parameter definitions and fixed-effect estimates are provided in [Table A3](#). Figure was created using R version 4.5.1 with package ggplot2. The shapefile and county boundaries are courtesy of the FIESTAR package ([Frescino et al. 2023](#)). Base map from the ggplot2R package ([Wickham 2016](#)).

