

Multivariate spatial models for small area estimation of species-specific forest inventory parameters

Jeffrey W. Doser^{a,*,}, Malcolm S. Itter^{b,1}, Grant M. Domke^c, Andrew O. Finley^{d,e}

^a Department of Forestry and Environmental Resources, North Carolina State University, Raleigh, NC, USA

^b Department of Environmental Conservation, University of Massachusetts, Amherst, MA, USA

^c Northern Research Station, United States Department of Agriculture Forest Service, St. Paul, MN, USA

^d Department of Forestry, Michigan State University, East Lansing, MI, USA

^e Department of Statistics and Probability, Michigan State University, East Lansing, MI, USA

ARTICLE INFO

Keywords:

Forest inventory and analysis
National forest inventory
Bayesian
Model-based inference

ABSTRACT

National Forest Inventories (NFIs) provide statistically reliable information on forest resources at national and other large spatial scales. As forest management and conservation needs become increasingly complex, NFIs are being called upon to provide forest parameter estimates at spatial scales smaller than current design-based estimation procedures can provide. This is particularly true when estimates are desired by species or species groups, which is often required to inform wildlife habitat management, sustainable forestry certifications, or timber product assessments. Here we propose a multivariate spatial model for small area estimation of species-specific forest inventory parameters. The hierarchical Bayesian modeling framework accounts for key complexities in species-specific forest inventory data, such as zero-inflation, correlations among species, and residual spatial autocorrelation. Importantly, by fitting the model directly to the individual plot-level data, the framework enables estimates of species-level forest parameters, with associated uncertainty, across any user-defined small area of interest. A simulation study revealed minimal bias and higher accuracy of the proposed model-based approach compared to the design-based estimator. We applied the model to estimate species-specific county-level aboveground biomass for the 20 most abundant tree species in the southern United States using Forest Inventory and Analysis (FIA) data. County-level biomass estimates from the proposed model had high correlations with design-based estimates, yet the model-based estimates tended to have a slight positive bias relative to design-based estimates, particularly for abundant and managed species. Importantly, the proposed model provided large gains in precision across all 20 species. On average across species, 91.5% of county-level biomass estimates had higher precision compared to the design-based estimates. Future work should explore incorporation of additional auxiliary data sources that can help explain fine-scale variation in biomass of managed species. The proposed framework is an attractive solution for NFI data users to generate species-level forest parameter estimates with reasonable precision at management-relevant spatial scales.

1. Introduction

Monitoring forest resources across broad spatial and temporal scales is increasingly important for forest management to meet both economic and sustainability objectives. National forest inventories (NFIs) are designed to estimate forest parameters on regional to national scales using a large network of permanent inventory plots selected according to a probability sampling design (McRoberts et al., 2010). For example, the United States (US) Department of Agriculture Forest Service Forest Inventory and Analysis (FIA) program conducts the US NFI, which generates a valuable resource for quantifying forest health, timber resources, and forest landowner characteristics across broad regions of

the US (Bechtold and Patterson, 2005). NFIs, like the FIA program, historically use design-based inference to estimate forest parameters, which generally provides sufficiently precise forest parameter estimates across large areas of forestland (e.g., US states) (Westfall et al., 2022). However, NFI programs across the world have experienced increased demand for estimates within smaller spatial, temporal, and biophysical extents than design-based inference and existing sampling intensities can reasonably deliver (Breidenbach and Astrup, 2012; Pringle et al., 2021). Methodological approaches that provide sufficiently precise estimates within small areas of interest using data from a larger sample are collectively referred to as small area estimation (SAE;

* Corresponding author.

E-mail address: jwdoser@ncsu.edu (J.W. Doser).

<https://doi.org/10.1016/j.foreco.2025.123112>

Received 10 March 2025; Received in revised form 22 July 2025; Accepted 20 August 2025

Available online 11 September 2025

0378-1127/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).

Rao and Molina (2015)). Such techniques have seen increasing applications with NFI data in recent years (e.g., Hou et al. (2021), Cao et al. (2022), Finley et al. (2024), Shannon et al. (2024)).

Both design-based and model-based inferential paradigms have been employed in SAE applications with NFI data (Dettmann et al., 2022). Regardless of the inferential paradigm used, most SAE approaches seek to relate the variable of interest to auxiliary information from satellite and airborne remote sensing data through a statistical model. When a strong relationship exists between the auxiliary information and variable of interest, the precision of small area estimates can be substantially improved (Särndal et al., 2003). In addition to incorporating auxiliary information through regression-based frameworks, model-based SAE approaches can explicitly borrow information from outside the area of interest through the use of random effects. Random effect structures can range from simple, normally distributed effects to account for similarities in forest parameters across ecologically-similar areas (e.g., ecoregions), management units, or political boundaries (e.g., White et al. (2021)), to more complex forms that explicitly acknowledge spatial and spatial-temporal dependencies in forest parameters through a spatial covariance structure (e.g., Finley et al. (2024), Shannon et al. (2024)). Recent advances in model-based SAE approaches that accommodate spatial dependence have been shown to provide improved precision of forest parameters and forest owner characteristics at state (Ver Planck et al., 2017; Harris et al., 2021), regional (Cao et al., 2022), and national levels (May et al., 2023; Shannon et al., 2024).

While considerable progress has been made in developing SAE methods suitable for forest inventory, a vast majority of proposed methods focus on estimating a single forest attribute such as total merchantable volume or aboveground biomass (Dettmann et al., 2022). Estimates of forest attributes by species or species class are often of interest due to individual species representing different forest products and values, providing unique benefits to wildlife, and serving as useful indicators of site quality and/or disturbance history (Kershaw et al., 2016). While independent models can be fit for each individual species, such an approach fails to leverage dependencies between species-specific attributes, which can substantially improve precision of model predictions given the importance of species interactions (e.g., competition) in determining stand structure and dynamics (Weiskittel et al., 2011; Clark et al., 2014; Tang et al., 2025). Total biomass within a fixed area is limited by site quality, and as a result the biomass of one species generally cannot increase without a corresponding decrease in the biomass of another (Clark et al., 2017). Additionally, reliable small area estimates of species-specific forest inventory parameters are complicated by large amounts of zero-inflation and highly right skewed data distributions. For example, species tend to be absent from many plots, present at low abundances in many plots, and present at high abundances in only a few plots. In many cases, individual species are never observed in the small number of plots within an area of interest, which prevents the calculation of a standard error estimate using design-based estimation. Furthermore, even for small areas with some non-zero values of species-specific attributes, the distribution of values is typically highly right-skewed and close to zero, which calls into question the reliability of confidence intervals that rely on the central limit theorem. Thus, there is a need for model-based frameworks that deliver precise estimates of forest parameters by species and similar high-dimensional forest attributes.

Here we develop a multivariate spatial model for SAE of species-specific forest inventory parameters. Our hierarchical Bayesian model simultaneously accounts for zero-inflation in individual species responses, correlations among species, and residual spatial autocorrelation—factors that notoriously complicate analysis of forest inventory data (e.g., Finley et al. (2012), Clark et al. (2014), May et al. (2023)). We implement the model at the unit-level (i.e., FIA plot-level), which provides flexibility in generating species-specific small area estimates across user-defined areal units of interest. We

perform a simulation study to assess bias and accuracy of the model relative to alternative design-based estimators. The proposed model is applied to FIA data in the southern US to estimate county-level aboveground live and standing dead tree biomass for the 20 most abundant tree species in the region. The multivariate model generates county-level estimates with substantially improved precision compared to alternative design-based estimators, showcasing the utility of our approach to generate species-level forest parameter estimates from NFI data at management-relevant spatial scales.

2. Materials and methods

2.1. FIA NFI data

FIA operates a nationally-consistent, quasi-systematic survey designed to monitor forest resources across the US. Around 1999, the FIA program transitioned to a systematic grid of permanent sample plots that are evenly distributed in an annual panel system to provide statistically robust estimates of both current forest resources and change over time. Each individual panel consists of a subset of the permanent sample plots measured in the same year that represent complete spatial coverage across the population of interest. Inventory cycles in the southern US consist of 5 or 7 annual panels. The base intensity of permanent ground plots across the continental US is approximately 1 plot per 2428 hectares (Smith, 2002; Bechtold and Patterson, 2005).

On forested FIA plots, trees with a diameter at breast height (d.b.h.) ≥ 12.7 cm are measured on a cluster of four circular 0.02 ha subplots centered on each plot location (Bechtold and Patterson, 2005). Within each subplot, trees 2.54–12.70 cm d.b.h. are measured on a circular 13.5 m² microplot. Tree attributes (e.g., species, alive/dead/harvested status) and quantitative variables (e.g., d.b.h., height, volume) are measured following nationally-standardized protocols. Aboveground biomass for each measured tree is estimated following FIA's National Scale Volume and Biomass Estimators (NSVB) framework (Westfall et al., 2024; Johnson et al., 2025). These individual tree estimates are summarized and expanded to provide a plot-level estimate of aboveground biomass (Mg/ha).

Here we use plot-level aboveground biomass data from $n = 46,710$ FIA plots across the FIA Southern Region of the United States (Fig. 1). We only use the most recent measurement at a given inventory plot and only use those measurements taken since 2015. We use the publicly available perturbed plot locations in which FIA adds a small amount of random noise to the true plot locations to protect private plot locations and preserve ecological integrity of the plots. Our focus is on estimating species-specific aboveground biomass at the county-level for all live and dead stems (i.e., snags leaning less than 45 degrees) ≥ 2.54 cm d.b.h. of the twenty most abundant species in the region (Table 1). We used the rFIA package (Stanke et al., 2020; Doser et al., 2025) in R (R Core Team, 2024) to download and extract the plot-level FIA data.

2.2. FIA sampling design and design-based estimators

The annual FIA sampling design implemented in the late 1990s is based on a systematic tessellation of ~2400 ha hexagons superimposed over the US based on a random starting location (Bechtold and Patterson, 2005; Westfall et al., 2022). Plot locations within each hexagon were determined following a pseudo-random approach. For cells with an existing inventory plot from previous iterations of the FIA program, a set of predefined rules were used to select a single plot. For cells without any pre-existing plots, plot locations were selected randomly within each cell. Grid cells were then assigned to annual panels to ensure the subset of panels sampled in any given year provided approximately uniform spatial coverage of the population.

As is commonly done in forest inventory, FIA uses post-stratified design-based estimators to estimate forest parameters of interest in an attempt to reduce variance of direct estimates without post-stratification

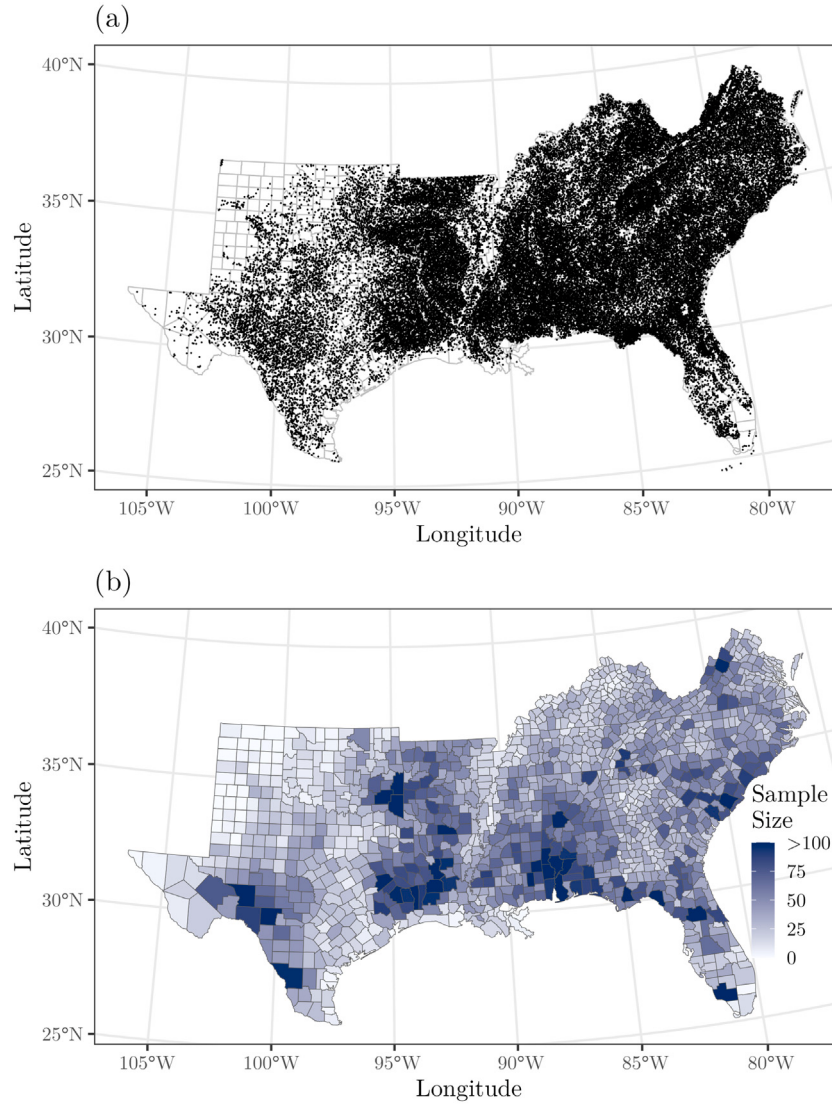


Fig. 1. Plot locations (a) and distribution of $n = 46,710$ Forest Inventory and Analysis (FIA) plots across the 1306 counties (small areas; b) in the southern United States. Note the plot locations are the publicly available perturbed locations in which FIA adds a small amount of random noise to the true plot locations.

(Westfall et al., 2011). Post-stratification proceeds by using wall-to-wall digital imagery to classify locations into a set of strata (e.g., forest/nonforest, canopy cover proportion classes). Stratum weights are then calculated based on the proportion of all pixels in the population assigned to the given stratum, and population estimates are obtained as a weighted average of the stratum-specific estimates (see Westfall et al. (2022) for a full description of FIA post-stratification techniques). While useful for estimating national and regional level forest parameters, such post-stratified estimators are limited in SAE applications, particularly when generating estimates by species due to individual stratum not being present within each SAE, highly zero-inflated and right-skewed distributions of species-specific forest parameters, and very small sample sizes within each stratum/SAE combination. Accordingly, for comparison with our model-based estimates, we use a simple design-based direct estimator (i.e., the sample mean). More specifically, letting $y_{i,j,k}$ denote the aboveground biomass for species j in plot i that falls in county k , our design-based estimate for county-level biomass is

$$\hat{\mu}_{j,k} = \frac{1}{n_k} \sum_{i=1}^{n_k} y_{i,j,k}, \quad (1)$$

where n_k is the number of inventory plots in county k . The associated variance for (1) is

$$\hat{\sigma}_{j,k}^2 = \frac{1}{n_k(n_k - 1)} \sum_{i=1}^{n_k} (y_{i,j,k} - \hat{\mu}_{j,k})^2. \quad (2)$$

2.3. Auxiliary data

We extracted climate, forest cover, and remotely-sensed elevation data for use as auxiliary information in the subsequently described SAE framework. We extracted all climate variables from TerraClimate (Abatzoglou et al., 2018), which provides high-resolution (1/24 degree or approximately 4 km) climate normals from 1981–2010 across the globe. In particular, we extracted monthly climate normals for maximum temperature, minimum temperature, total precipitation, and vapor pressure deficit. From these monthly values, we calculated maximum temperature (TMAX) as the maximum of the monthly maximum temperature normals, minimum temperature (TMIN) as the minimum of the monthly minimum temperature normals, total precipitation (PPT) as the sum of the monthly precipitation normals, and vapor pressure deficit (VPD) as the average of the growing season (April

Table 1

Statistics summarizing the bias and precision of model-based county-level estimates relative to the design-based direct estimates for the twenty species modeled in the FIA analysis. The Pearson correlation coefficient (ρ) is used as a simple measure of correspondence. Bias indicates the model-based estimate minus the design-based estimate, averaged across all counties. % Improv. indicates the percentage of the 1304 counties with a lower model-based CV compared to the direct estimate CV.

Common name	Scientific name	ρ	Avg bias	% Improv.
American beech	<i>Fagus grandifolia</i> Ehrh.	0.78	0.30	77.4
Black oak	<i>Quercus velutina</i> Lam.	0.79	0.19	93.6
Chestnut oak	<i>Quercus montana</i> Wild.	0.93	0.56	82.7
Laurel oak	<i>Quercus laurifolia</i> Michx.	0.88	0.21	86.0
Live oak	<i>Quercus virginiana</i> Mill.	0.89	0.36	76.4
Loblolly pine	<i>Pinus taeda</i> L.	0.95	6.32	73.2
Longleaf pine	<i>Pinus palustris</i> Mill.	0.72	0.57	68.9
Mockernut hickory	<i>Carya tomentosa</i> Sarg.	0.72	0.28	98.9
Northern red oak	<i>Quercus rubra</i> L.	0.82	0.26	94.3
Pignut hickory	<i>Carya glabra</i> Mill.	0.84	0.38	96.4
Post oak	<i>Quercus stellata</i> Wang.	0.92	0.49	91.7
Red maple	<i>Acer rubrum</i> L.	0.88	0.18	97.2
Scarlet oak	<i>Quercus coccinea</i> Muenchh.	0.85	0.22	95.2
Shortleaf pine	<i>Pinus Elliottii</i> Engelm.	0.91	0.17	92.6
Slash pine	<i>Pinus echinata</i> Mill.	0.96	0.64	76.4
Southern red oak	<i>Quercus falcata</i> var. <i>falcata</i> Michx.	0.69	0.39	94.2
Sweetgum	<i>Liquidambar styraciflua</i> L.	0.86	1.13	97.8
Water oak	<i>Quercus nigra</i> L.	0.86	0.52	97.9
White oak	<i>Quercus alba</i> L.	0.86	1.97	96.2
Yellow poplar	<i>Liriodendron tulipifera</i> L.	0.89	1.31	97.7

through October) monthly vapor pressure deficit normals. We calculated tree canopy cover (TCC) at each FIA plot center in 2019 using the National Land Cover Database (NLCD) 2021 Forest Service Science data set (Houseman et al., 2023). Finally, we determined the elevation at the center of each FIA plot using bare-earth terrain height data from Amazon Web Service Terrain Tiles (accessed on 13 June 2025 from <https://registry.opendata.aws/terrain-tiles>).

2.4. Modeling framework

We seek a modeling framework that can provide improved precision of species-specific forest parameter estimates by incorporating plot-level predictor variables correlated with aboveground biomass, exploiting correlations between co-occurring species, and borrowing information from spatially proximate plots. We accomplish this via a multivariate log-normal hurdle model that accommodates the large zero-inflation and heavily right-skewed distribution of species-level biomass. The proposed model first jointly estimates the presence or absence of each species across the set of $n = 46,710$ FIA plots to establish the distribution of each species across the region. We subsequently estimate species-specific biomass only at those plots where the individual species is present. We refer to these sub-models as Stage 1 and Stage 2, respectively. Estimates from the model are then used to predict species-specific biomass across a dense grid of cells over the region of interest, which we ultimately summarize to generate species-level estimates across the 1306 counties in the southern US. In the subsequent model presentation, we focus on the case of estimating species-specific biomass in the southern US, but note that the model could be used to predict any species-specific forest parameter that takes positive-only values (e.g., basal area, volume, carbon). We used the following R packages for data downloading, processing, and manipulating auxiliary data: *sf* (Pebesma and Bivand, 2023), *tidyverse* (Wickham et al., 2019), *climateR* (Johnson, 2024), and *elevatr* (Hollister et al., 2023).

Stage 1: Binary sub-model

We first jointly estimate the presence or absence of all species to accommodate the large propensity of locations where a species does not occur (and thus has zero biomass). Let s_i denote the spatial coordinates of the i th forest plot for each of the $i = 1, \dots, 46,710$ plots

(e.g., the easting and northing coordinate for the center subplot of a FIA inventory plot i). Define $z_j(s_i)$ as the presence (1) or absence (0) of species j at plot i for each of the $j = 1, \dots, J$ species, where here $J = 20$. We model $z_j(s_i)$ following

$$z_j(s_i) \sim \text{Bernoulli}(\psi_j(s_i)), \quad (3)$$

where $\psi_j(s_i)$ is the probability species j occurs at plot i . We model $\psi_j(s_i)$ as a function of multiple climate variables and a spatially varying intercept. More specifically, we have

$$\text{logit}(\psi_j(s_i)) = \beta_{0,j} + w_j^*(s_i) + \beta_{1,j} \cdot \text{TMIN}_i + \beta_{2,j} \cdot \text{TMIN}_i^2 + \beta_{3,j} \cdot \text{TMAX}_i + \beta_{4,j} \cdot \text{TMAX}_i^2 + \beta_{5,j} \cdot \text{PPT}_i + \beta_{6,j} \cdot \text{PPT}_i^2, \quad (4)$$

where $\beta_{0,j}$ is the overall intercept for species j , $\beta_{1,j}$ and $\beta_{2,j}$ are linear and quadratic effects of minimum temperature (TMIN), $\beta_{3,j}$ and $\beta_{4,j}$ are linear and quadratic effects of maximum temperature (TMAX), $\beta_{5,j}$ and $\beta_{6,j}$ are linear and quadratic effects of total precipitation, and $w_j^*(s_i)$ is a spatially varying intercept for species j at plot i to accommodate residual spatial structure in the species distribution. We included minimum/maximum temperature and total precipitation to help establish individual species distributions given their importance in determining species ranges across the southern US. We include linear and quadratic effects to allow the relationships to peak at intermediate levels of one or more of the climate variables in relation to the climatic niche of the individual species.

We model all regression coefficients and the non-spatial intercept for each species hierarchically (i.e., as random effects; Dorazio and Royle (2005), Gelfand et al. (2005)). More specifically, for each $t = 0, \dots, p$, we model $\beta_{j,t}$ according to

$$\beta_{j,t} \sim \text{Normal}(\mu_{\beta_t}, \tau_{\beta_t}^2), \quad (5)$$

where μ_{β_t} is the mean effect across all species, and $\tau_{\beta_t}^2$ is the variability in the effect across all species. By treating species-specific effects as random effects, we borrow strength across species and obtain more precise estimates for both rare and common species (Zipkin et al., 2009).

We seek to jointly model the species-specific spatially varying intercepts to account for residual correlations between species and their occurrence patterns. We use a spatial factor modeling approach (Hogan and Tchernis, 2004), a dimension reduction technique that can efficiently model correlations among a large number of species. This

approach has seen recent applications in forest inventory for integrating highly-multivariate LiDAR data with classical forest inventory data (Taylor-Rodriguez et al., 2019), ultimately providing substantial gains in computational run time in comparison to alternative multivariate modeling approaches (e.g., the linear model of coregionalization). We decompose $w_j^*(s_i)$ into a linear combination of q latent variables (i.e., factors) and their associated species-specific coefficients (i.e., factor loadings). More specifically, we have

$$w_j^*(s_i) = \lambda_j^T \mathbf{w}(s_i), \quad (6)$$

where λ_j^T is a coefficient vector for species j for each of the q independent spatial factors at plot i ($\mathbf{w}(s_i)$). When q is small relative to the total number of species, this approach provides substantial computational improvements compared to estimating a separate spatial process for each species (Tikhonov et al., 2020; Doser et al., 2023). Such an approach accounts for species correlations not accounted for by the covariates using individual responses (i.e., loadings) to the q spatial factors such that co-occurring species will have similar species-specific factor loadings (Doser et al., 2023).

The standard spatial factor approach predicts the spatial factors using a spatial Gaussian process (GP), which requires calculating the inverse and determinant of a dense $n \times n$ matrix. This computation involves $O(n^3)$ computations (floating point operations or FLOPs) for each of the q spatial factors, which quickly renders the approach impractical for even modest sized forest inventory data sets (e.g., hundreds of locations). Here we estimate the latent spatial factors using a GP approximation called the Nearest Neighbor Gaussian Process (NNGP; Datta et al. (2016), Finley et al. (2019)) that provides substantial improvements in run time, with negligible differences in inference and prediction, compared to a model that uses a full GP. For each $r = 1, \dots, q$ spatial factor, we have

$$w_r(s_i) \sim N(\mathbf{0}, \tilde{C}_r(\phi_r)), \quad (7)$$

where $\tilde{C}_r(\phi_r)$ is the NNGP-derived correlation matrix for the r th spatial process. We use an exponential spatial correlation model, where ϕ_r is a spatial decay parameter controlling the range of the spatial dependence for the r th spatial factor. The number of factors can be chosen using an exploratory approach where models with differing numbers of factors are used, compared using some model assessment criteria (e.g., Widely Applicable Information Criterion), and the number of factors can be set to the value where the model improvement becomes negligible (analogous to a scree plot in principal components analysis). For this analysis, we found $q = 5$ provided a sufficient balance between model fit and model run time.

To complete the Bayesian specification of the model, we use normal priors with mean equal to 0 and variance set to 2.72 for all mean parameters, which results in a vague prior on the probability scale (Northrup and Gerber, 2018). We use vague inverse gamma priors with both the shape and scale parameter set to 0.1 for all variance parameters. We set moderately informative uniform priors to the spatial decay parameters such that the effective spatial range (i.e., the distance at which the correlation between two sites falls to 0.05) is bounded between 50 km and 2000 km. This allows for sufficient flexibility to estimate fine-scale and/or broad-scale spatial autocorrelation. Finally, we assign standard normal priors to the lower triangle of the factor loadings matrix. Following Taylor-Rodriguez et al. (2019), we fix the diagonal elements of the factor loadings matrix to 1 and the upper triangle to 0 to ensure identifiability of the factor loadings and the spatial factors.

Stage 2: Log-normal sub-model

For the second component of the hurdle model, we jointly estimate the species-specific biomass only at the locations where the species is observed to be present. Let $y_j(s_i)$ denote the aboveground biomass for species j at plot i . Analogous to Finley et al. (2012), we predict $y_j(s_i)$ conditional on the presence ($z_j(s_i) = 1$) or absence ($z_j(s_i) = 0$) of

species j at plot i . More specifically, when the species is present at plot i , we have

$$\log(y_j(s_i) | (z_j(s_i) = 1)) \sim \text{Normal}(\mu_j(s_i), \tau_j^2), \quad (8)$$

where $\mu_j(s_i)$ is the average biomass (on the log scale) for species j at plot i , and τ_j^2 is the species-specific non-spatial residual error (e.g., measurement error, microscale variation). We log-transform the variable of interest given the large positive skew in the species-level biomass data (Supplemental Information S1: Figure S1), such that using a log-normal distribution allows us to better adhere to basic linear model assumptions and to ensure positive support upon back-transformation. We model $\mu_j(s_i)$ according to

$$\mu_j(s_i) = \alpha_{0,j} + \eta_j^*(s_i) + \alpha_{1,j} \cdot \text{TCC}_i + \alpha_{2,j} \cdot \text{VPD}_i + \alpha_{3,j} \cdot \text{PPT}_i + \alpha_{4,j} \cdot \text{ELEV}_i + \alpha_{5,j} \cdot \text{ELEV}_i^2, \quad (9)$$

where $\alpha_{0,j}$ is the overall intercept for species j , $\alpha_{1,j}$ is the linear effect of percent tree canopy cover (TCC), $\alpha_{2,j}$ is the linear effect of average growing season vapor pressure deficit (VPD), $\alpha_{3,j}$ is the linear effect of total precipitation (PPT), and $\alpha_{4,j}$ and $\alpha_{5,j}$ are linear and quadratic effects of elevation (ELEV). We included vapor pressure deficit and total precipitation as covariates on expected biomass given their physiological importance for tree growth. Percent tree canopy cover was hypothesized to have a positive relationship with biomass, particularly with common species. Elevation was included given its correlation with distinct forest types across the southern US. The species-specific spatially varying intercept ($\eta_j^*(s_i)$) is modeled using the spatial factor NNGP approach described in (6) and (7). As in Stage 1, we fit the model with five spatial factors.

At sites where each species does not occur, (i.e., $z_j(s_i) = 0$), we model the variable of interest following

$$y_j(s_i) | (z_j(s_i) = 0) \sim \text{Normal}(0, 0.0001). \quad (10)$$

By definition, aboveground biomass of species j is zero at sites where the species does not occur and thus the measurement is not taken. However, we cannot simply fix $y_j(s_i)$ to zero, as this would result in a degenerate likelihood (Finley et al., 2012) that prevents prediction across a grid to form subsequent small area estimates. This approach avoids such degeneracy, ultimately providing stable estimation properties while ensuring estimates of inventory parameters at sites where a species does not occur are essentially estimated at zero.

As opposed to the hierarchical approach used in Stage 1, here we assign independent normal priors with a mean of 0 and variance of 10 to the species-specific regression coefficients, as preliminary analyses revealed the shared prior was not adequate to describe variation in the parameters across species. We assigned inverse gamma priors with the scale parameter set to 2 and shape parameter set to 1 to species-specific residual variance parameters. We use the same priors and identifiability constraints for the spatial decay parameters and factor loadings matrix as described for Stage 1.

2.5. Model implementation

We estimated all model parameters in a Bayesian framework using efficient Markov chain Monte Carlo (MCMC) algorithms. We implemented Stage 1 using the `sfJSDM` function in the `spOccupancy` R package (Doser et al., 2022). We ran three chains of the Stage 1 model for 200,000 MCMC iterations with a burn-in period of 140,000 iterations and a thinning rate of 30, yielding 6000 posterior samples. We implemented Stage 2 using the `sfMsAbund` function in the `spAbundance` R package (Doser et al., 2024), running the model for three chains each of 100,000 MCMC iterations, a burn-in period of 40,000, and a thinning rate of 30, yielding 6000 posterior samples. Convergence was assessed using visual assessment of traceplots and the potential scale reduction factor (i.e., $\hat{R}$, Brooks and Gelman (1998)).

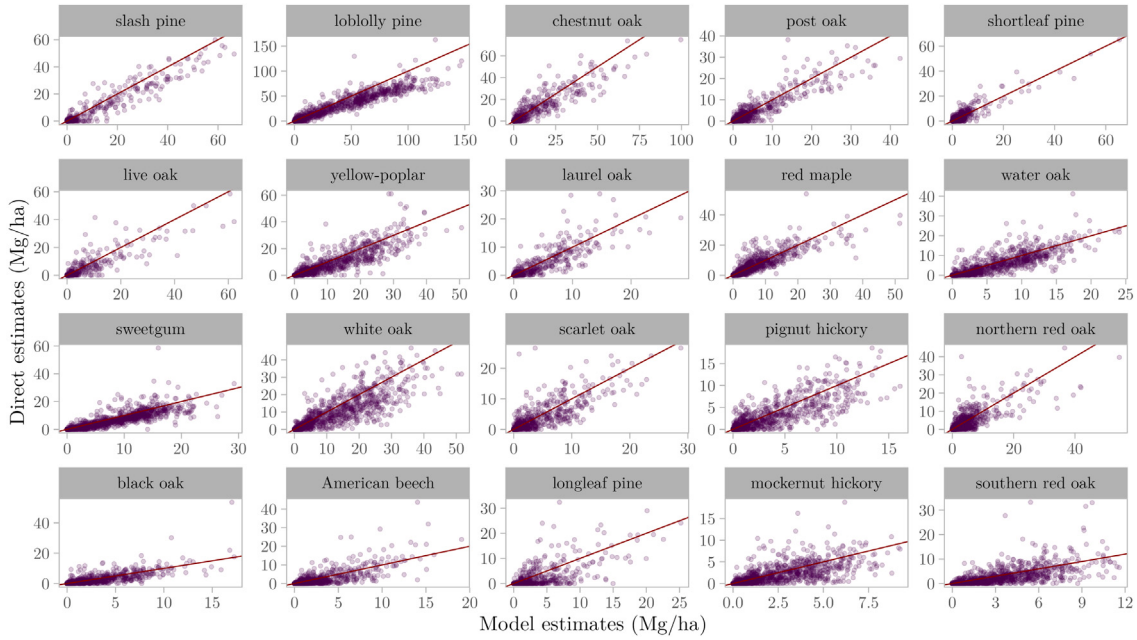


Fig. 2. Comparison of small area estimates of species-specific biomass for the 20 species from the proposed multivariate spatial model and design-based direct estimates. Species are ordered in descending order of the correlation between direct and model-based estimates, with slash pine (*Pinus elliottii*, top left) having the highest correlation and southern red oak (*Quercus falcata* var. *falcata*, bottom right) having the lowest correlation.

2.6. Prediction and small area estimation

Our inferential objective is to estimate species-specific biomass density (Mg/ha) across the 1306 counties in the southern US. We accomplish this task by predicting biomass for each species across a dense grid of n_0 prediction locations overlaid across the study region. For our application, we overlaid a 1×1 km grid of points across the southern US, which resulted in a total of $n_0 = 2,198,845$ prediction locations with a median of 1515 prediction locations falling into a single county. Prediction at each location proceeds in two stages, analogous to the model fitting process, during which we first predict the presence or absence of each species at each prediction site, and subsequently predict the amount of biomass for each species that is predicted to be present at a site. Importantly, because there is uncertainty in the set of locations where a given species is predicted to be present, we must fully account for this uncertainty when predicting biomass to avoid inflating the precision of our resulting estimates (Finley et al., 2012). The Bayesian framework readily enables such uncertainty propagation via posterior predictive inference.

Let s_0 denote a generic location across the grid of n_0 prediction locations. For each MCMC sample l , we predict the presence (1) or absence (0) of species j at site s_0 following

$$z_j^{(l)}(s_0) \sim \text{Bernoulli}(\psi_j^{(l)}(s_0)), \quad (11)$$

where $\psi_j^{(l)}(s_0)$ is the predicted occurrence probability at location s_0 for species j that is calculated according to (4) using the parameter estimates at MCMC iteration l obtained from fitting the model to the observed data. The prediction of biomass for species j at prediction location s_0 and MCMC iteration l is then estimated conditional on $z_j^{(l)}(s_0)$ such that

$$y_j^{(l)}(s_0) \sim \begin{cases} \text{Normal}(0, 0.0001), & z_j^{(l)}(s_0) = 0 \\ \exp(\text{Normal}(\mu_j^{(l)}(s_0), \tau_j^{2(l)})), & z_j^{(l)}(s_0) = 1, \end{cases} \quad (12)$$

where $\mu_j^{(l)}(s_0)$ is the predicted mean biomass (on the log scale) calculated according to (9). Once biomass is predicted at each of the $m = 1, \dots, n_{0,k}$ prediction locations in county k , county-level biomass density

(Mg/ha) is obtained by simply taking the average of all point-level predictions, i.e.,

$$\bar{y}_{j,k}^{(l)} = \frac{1}{n_{0,k}} \sum_{m=1}^{n_{0,k}} y_j^{(l)}(s_m). \quad (13)$$

This process results in a full posterior distribution of biomass density predictions for each species in each county. These samples can then be summarized using any measure of central tendency (e.g., median, mean) to yield the final model-based estimate, along with an associated measure of uncertainty (e.g., standard deviation, 95% credible interval, coefficient of variation).

2.7. Model assessment and comparison to design-based estimates

We compared county-level biomass estimates from the proposed multivariate spatial model to the design-based direct estimates (i.e., sample mean) previously defined in (1). Ideally, the proposed model will provide small area estimates that show high correlation with the design-based direct estimates, but are more precise. To assess improvements in precision, we compared the design-based coefficient of variation (i.e., the standard error divided by the sample mean) to the model-based coefficient of variation (i.e., posterior standard deviation divided by the posterior median) for biomass density estimates for each species in each county. The design-based estimator and the Bayesian model-based approach rely on fundamentally different statistical paradigms, so this comparison reflects how each quantifies uncertainty for its estimate of the population parameter—in this case, species-specific biomass in each county. See Gregoire (1998) for a more complete discussion regarding the comparison of model-based and design-based estimates. As an assessment of correspondence between the design-based and model-based estimates, we calculated simple correlation coefficients for each species. We additionally calculated the average bias between model-based county-level estimates and the design-based county-level estimates. Note that these bias estimates do not necessarily represent bias from the true, unknown species-specific county-level biomass values and rather reflect differences of the model-based estimate from the design-based estimates, which themselves could be biased.

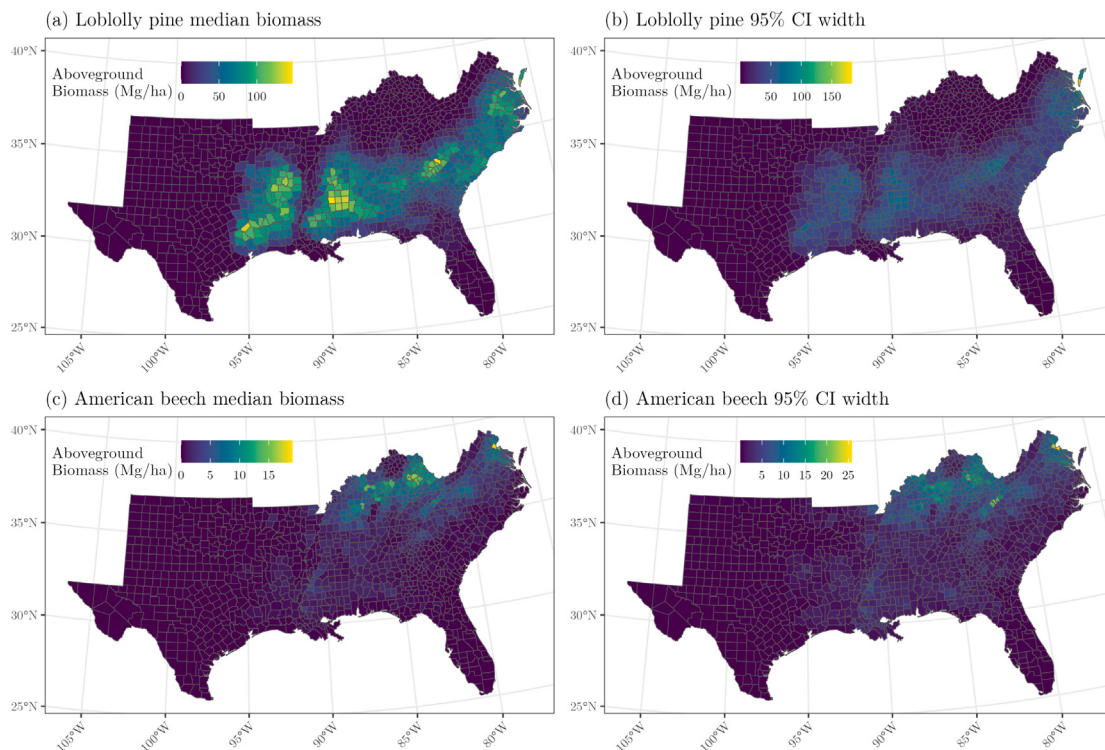


Fig. 3. Model-based estimates (posterior medians) of aboveground biomass (Mg/ha; a, c) and associated 95% credible intervals (b, d) for the most abundant species (loblolly pine (*Pinus taeda*)), and least abundant species (American beech (*Fagus grandifolia*)) included in our analysis.

We used four-fold cross-validation to provide an assessment of the proposed model's ability to predict species-level small area estimates in small areas that were not used to fit the model. We randomly split the 1306 counties into four approximately equal sets. We subsequently fit the proposed model four times, where each model was fit using the inventory plots found within three of the four sets of counties. We subsequently predicted species-specific biomass density within the counties that were not used when fitting the model and estimated bias between the resulting model-based estimates and the design-based estimates. We averaged the estimated bias value for each species across the four hold-out sets for a single measure of bias for each individual species. Our approach to generating hold out sets based on counties aligned with our desire to produce county-level estimates, and thus provided a more realistic assessment of model predictive performance compared to randomly holding out individual inventory plots.

2.8. Simulation study

In addition to model assessment in the FIA case study, we performed a simulation study to quantify the bias and accuracy of the proposed model-based approach to species-specific SAE. We simulated a single population of 41,501 population units distributed across a 1.75×1.75 km grid over the state of North Carolina. Each population unit represents a potential sampling unit. We simulated biomass of ten species across all population units following the multivariate spatial model described in Section 2.4, where we first simulated the presence/absence of each species at each population unit conditional on the climate variables and four spatial factors, then subsequently simulated the biomass for each location where the species occurs conditional on climate, elevation, tree canopy cover, and four spatial factors. The spatial factors were simulated to represent a mix of fine-scale and broad-scale spatial autocorrelation, a frequent phenomenon in forest inventory data (e.g., May and Finley (2025)). This simulation approach resulted in data with similar characteristics to the observed FIA data (e.g., highly zero-inflated and right-skewed), with species occurring on

average at approximately ~29% (minimum 2% and maximum 87%) of the population units in the region.

From this population, we sampled 100 replicate data sets each comprising 1000 population units selected at random from the population. For each of the replicate data sets, we calculated species-specific average biomass (Mg/ha) in 100 small areas (i.e., counties) across the state using the proposed model and design-based estimator. For both estimation approaches, we calculated bias as the difference between the estimate and true population value for each species and county, as well as the root mean square error (RMSE) as a measure of average accuracy. Complete details on the simulation study are provided in Supplemental Information S2.

3. Results

Estimates of species-specific aboveground biomass from the multivariate spatial model across the 1306 counties in the southern US showed close correspondence to design-based direct estimates across all twenty species (Fig. 2). The average correlation coefficient between the model-based estimates and design-based estimates across the 20 species was 0.85, ranging from 0.96 for slash pine to 0.69 for southern red oak (Table 1). County-level estimates for the most abundant (loblolly pine) and least abundant (American beech) species are shown in Fig. 3, along with associated 95% credible intervals to characterize uncertainty in the estimates. Maps of estimates and their uncertainty for all 20 species are provided in Supplemental Information S1.

Despite the correspondence between model-based and design-based estimates, model-based estimates had a slight positive bias in county-level biomass estimates relative to the design-based estimates (Table 1). Patterns in bias were often species-specific, with multiple species showing increases in bias as sample size increased (e.g., water oak, laurel oak, slash pine; Fig. 4). Residual assessments revealed moderate deviations from normality in the form of a heavy left tail (Supplemental Information S3) for many species, with the deviations being the most substantial for species with the largest amount of positive bias in

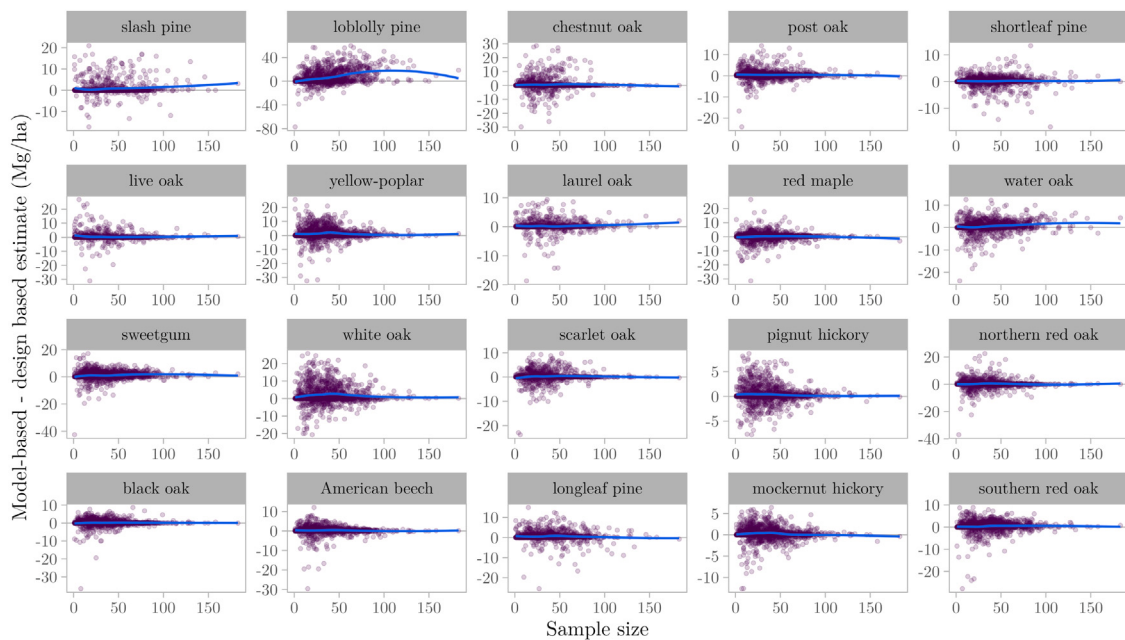


Fig. 4. Relationship between sample size and bias in the model-based county-level estimates relative to the design-based estimates. Species are ordered in descending order of the correlation between direct and model-based estimates, with slash pine (top left) having the highest correlation and southern red oak (bottom right) having the lowest correlation. Note the sample sizes indicate all FIA plots in a county, regardless of whether the species was detected at the plot.

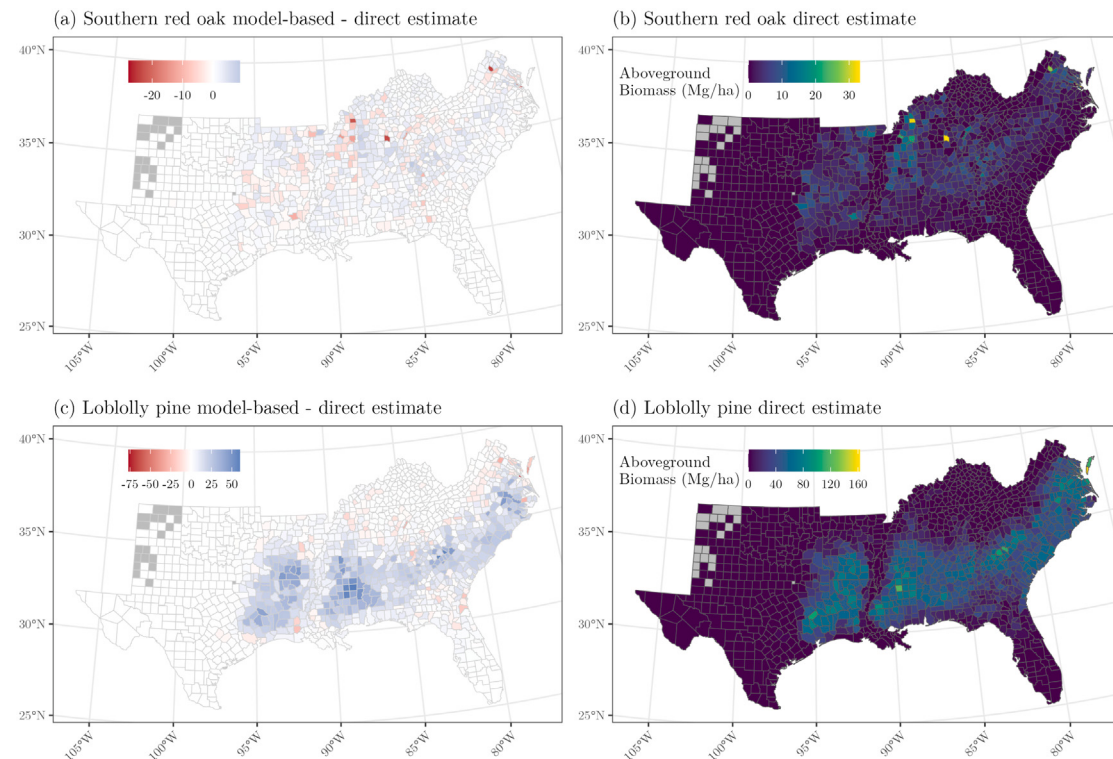


Fig. 5. Differences between the design-based and model-based estimates for southern red oak (a) and loblolly pine (c) along with maps of the design-based direct estimate (b, d). Gray corresponds to counties with no FIA plots.

county-level estimates. In general, the model-based estimates tended to provide more spatially smooth estimates relative to fine-scale heterogeneity represented in the direct estimates, as would be expected given the sharing of information between plots across space via the spatial autocorrelation structure and auxiliary variables in the model-based estimates (Supplemental Information S4). Large differences between model-based and design-based estimates occurred as a result of: (1)

outlier counties within the species primary range that had a very high direct estimate in which the model provided substantially lower estimates more in line with surrounding counties (e.g., Fig. 5a,b); (2) locations with high biomass where the model-based estimates tended to extend estimates of high biomass farther in space than the direct estimates (e.g., Fig. 5c,d); (3) spatially disparate counties with positive biomass according to the direct estimate but often substantially lower

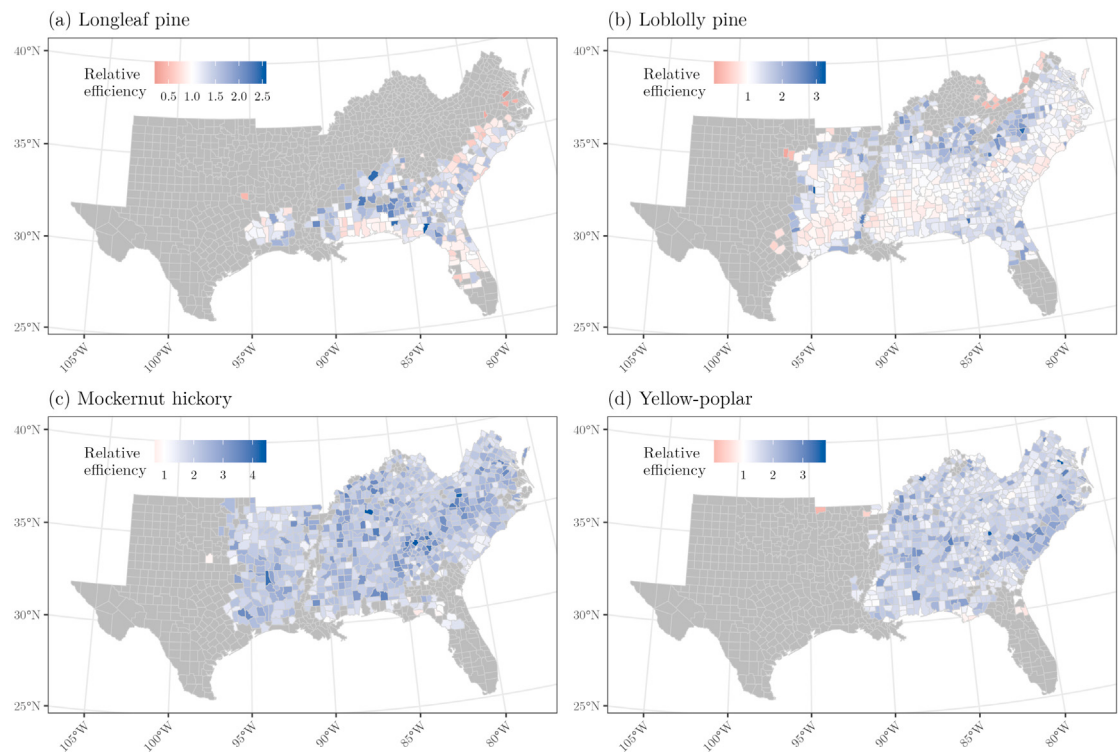


Fig. 6. Relative efficiency of the model-based estimate coefficient of variation relative to the direct design-based estimate coefficient of variation for the two species with the smallest precision gains from the model-based approach (longleaf pine [*Pinus palustris*] and loblolly pine [*Pinus taeda*]) and the two species with the largest precision gains from the model-based approach (mockernut hickory [*Carya tomentosa*] and water oak [*Quercus nigra*]). Relative efficiency values greater than 1 (blue) indicate counties where the model-based estimate had higher precision, while values less than one (red) indicate counties where the model-based estimate had lower precision. Gray corresponds to counties where the design-based estimate is 0 and there is no design-based standard error.

Table 2
Comparison of model-based and design-based county-level estimates for the four-fold cross-validation assessment in the FIA analysis. Bias indicates the model-based estimate minus the design-based estimate, averaged across all counties. Values are averaged across the four-folds.

Common name	Correlation coefficient (ρ)	Average bias
American beech	0.67	0.40
Black oak	0.71	0.25
Chestnut oak	0.84	0.53
Laurel oak	0.80	0.31
Live oak	0.77	0.36
Loblolly pine	0.92	5.98
Longleaf pine	0.54	0.68
Mockernut hickory	0.65	0.35
Northern red oak	0.79	0.36
Pignut hickory	0.78	0.41
Post oak	0.81	0.55
Red maple	0.82	0.20
Scarlet oak	0.80	0.24
Shortleaf pine	0.84	0.33
Slash pine	0.91	0.51
Southern red oak	0.57	0.61
Sweetgum	0.77	1.12
Water oak	0.76	0.68
White oak	0.79	2.03
Yellow poplar	0.82	1.32

biomass predicted from the model-based estimate. These patterns of correspondence between model-based and design-based estimates with slight positive bias in the model-based estimates were consistent when considering the entire data set as well as the hold-out assessments from the four-fold cross-validation (Table 2).

The model-based county-level estimates were consistently more precise across the twenty species compared to the design-based estimates.

Across all species, 91.5% of county-level biomass estimates had higher precision (in terms of the coefficient of variation) compared to the design-based estimates, with mockernut hickory (98.9%) having the largest percentage improvements and longleaf pine (68.9%) having the lowest percentage improvements (Table 1). Note that precision comparisons are only possible in counties where a species was observed on at least one FIA plot, as the design-based estimates do not have standard errors for counties where a species was not observed (i.e., all observations are zero).

Visualizing the relative efficiency of model-based county-level biomass estimates compared to the design-based direct estimates highlights the spatial extent of improvement the proposed model provides (Figs. 6, 7). The multivariate spatial model led to widespread improvements across species in essentially the entirety of the southern US study region (Fig. 7). Noticeably, the multivariate spatial model had inferior performance to the design-based estimates in the western portion of the region as well as in southern Florida. Similar patterns emanate in maps of relative efficiency for individual species (Supplemental Information S1). For example, live oak, which has a complex spatial distribution spanning much of the southern and eastern coast of the study region, has large precision gains across Florida and much of Georgia, while precision is slightly decreased in much of the western portion of its range, particularly in several isolated counties (Supplemental Information S1: Figure S26).

3.1. Simulation study

The proposed model-based estimator and the direct estimator did not have any clear directional bias across the ten species, with the sign of the bias varying across the ten species (Supplemental Information S2: Table 1). Variation in bias across counties tended to be larger for the model-based estimator than the direct estimator (Supplemental

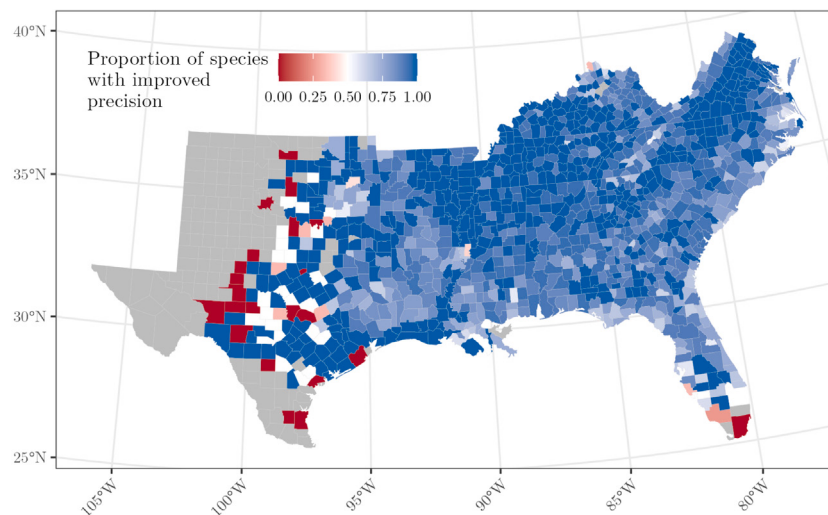


Fig. 7. Proportion of species observed at least once in a given county for which the multivariate spatial model has higher precision relative to the design-based estimate. Gray corresponds to counties where none of the twenty species were observed.

Information S2: Figure S1). Importantly, the model-based estimator had consistently higher accuracy (lower RMSE) compared to the direct estimator (Supplemental Information S2: Figure S2).

4. Discussion

While NFIs have long provided statistically reliable information on forest resources at national and other large spatial scales, there is increased demand for estimates within smaller spatial scales (Dettmann et al., 2022; Prisley et al., 2021). Standard design-based approaches are limited in their ability to provide reasonable precision of small area estimates, particularly when estimating forest parameters by species or species groups. Here we developed a multivariate spatial modeling framework to enable small area estimation of species-specific forest inventory parameters. In a simulation study, we found less bias and higher accuracy from our model compared to a design-based estimator. We applied our model to estimate county-level aboveground biomass for the 20 most abundant tree species in the southern US, which resulted in substantial precision gains relative to the design-based direct estimates. More broadly, our framework can be applied to estimate species-level forest parameters (e.g., biomass, carbon, volume) within management-relevant spatial scales to support diverse forest management objectives.

County-level estimates of species-specific aboveground biomass in the southern US showed high correspondence with design-based estimates (Fig. 2) and substantial improvements in estimate precision across the 20 species and 1306 counties (Table 1, Figs. 6, 7). Unlike the design-based direct estimator, the model-based approach can readily provide estimates for small areas without any inventory plots by leveraging information from auxiliary covariates and spatially proximate locations. Further, it is common for individual species to never be observed within the set of inventory plots in a given small area, particularly when the number of plot locations is small. While design-based estimates in such situations are always estimated at 0 with no corresponding standard error, the proposed model does not have such restrictions. The model-based estimator can readily provide uncertainty predictions in small areas where a species is never observed, and is also not forced to estimate the parameter at 0. This is particularly relevant for predicting attributes of rare species that might be missed in relatively small fixed area plots such as those used in FIA but occur in other portions of the small area of interest. In such a situation, the multivariate spatial model can use information from the auxiliary variables in the model, the spatial random effects, and the presence of other species to help estimate the forest parameter of interest.

Additionally, design-based confidence intervals require the estimator's sampling distribution to be approximately normally distributed. This is often argued to be achieved through the central limit theorem when sample sizes are sufficiently large. While some results show a sample size of $n = 25\text{--}30$ is sufficient for an approximately normal population (Hogg, 2020), the distributions of species-specific forest inventory parameters are often highly skewed, zero-inflated, and non-symmetric (Tang et al., 2025; Supplemental Information S1: Figure S1). Such data require even larger sample sizes for the normality assumption to be met, which is often not achievable for small area estimation, and thus confidence intervals for design-based small area estimates of species-specific parameters are likely not reliable.

The application of the proposed model to species-specific biomass estimation in the southern US provides an important step forward for generating reliable species-specific small area estimates from FIA data. Individual species represent distinct values to wildlife, forest products, and functions for overall ecosystem health. Thus, reliable estimation of species-specific forest parameters from publicly available FIA data provides an important asset for forest managers and landowners for strategic planning, assessments of timber supply, and support for sustainable forestry certification. While we focused on estimation at the county-level, a key benefit of our proposed model is that it is implemented at the unit-level (i.e., inventory plot level), ultimately allowing users to generate small area estimates in any areal unit of interest larger than the individual plot size. Furthermore, the proposed model is not restricted to estimating average parameter values across a small area. Estimates of total biomass, carbon, or other variable of interest, with associated uncertainty, can readily be calculated by simply multiplying the average small area estimate calculated in (13) by the total area of the small area.

Despite the large improvements in estimate precision, the model-based estimates did consistently predict higher biomass relative to the design-based estimates, with the degree of positive bias varying substantially across species (Table 1, Fig. 4). Bias was largest for loblolly pine, the most common and abundant species in the region (Fig. 5c,d). Overestimation was fairly consistent within the core of loblolly pine's range, with the largest areas of overestimation occurring in regions of high loblolly pine biomass. One potential reason for this is the limited ability of the covariates in Stage 2 of the model to predict areas of high biomass for loblolly pine, which results in the spatial component of the model attempting to explain such variation and providing more smooth estimates of high biomass in the surrounding region. This phenomenon is likely the main driver of the general

tendency across all species to have, on average, higher model-based estimates compared to the direct estimates. Biomass of loblolly pine and other managed species in the study area (e.g., slash pine, longleaf pine) is heavily driven by management practices that are localized in space, and thus the ability to predict fine-scale biomass variation within a model-based framework may be limited for intensively managed species, unless such information can be incorporated into the model via additional auxiliary variables. More generally across all species, inclusion of covariates that characterize forest productivity (e.g., soil characteristics, disturbance history) as well as publicly available LiDAR data may provide more accurate estimates, particularly if covariates can explain fine-scale variation not accounted for by broad climatic and environmental variables. Additionally, use of the true FIA plot coordinates instead of the fuzzed plot coordinates used in this analysis would likely improve associations between species-specific biomass and the auxiliary variables, resulting in more accurate small area estimates. Such tasks represent important next steps to improve species-specific biomass estimates from FIA data such as those presented here.

While the higher estimates from the proposed model relative to the design-based estimates indicate areas of improvement for the model in the FIA case study, direct estimates of species-specific county-level biomass themselves may not accurately represent true characteristics of the county. When sample sizes within counties are small, the direct estimate solely relies on the small set of plots to generate an estimate (i.e., the sample mean). Sampling theory dictates that any individual direct estimate based on a small sample size may not itself be particularly close to the true value (i.e., given the large standard error of such an estimate). In the southern US case study, 25% of the counties contained less than 17 FIA plots, and thus such estimates could be substantially far from the true biomass density in the county. Thus, large discrepancies between model-based and design-based estimates in counties with low sample sizes does not inherently suggest inferior performance of the model-based approach in these areas, but rather represents a discrepancy between two estimates of some unknown, true population value. This concept complicates comparisons of accuracy between design-based and model-based estimates. Simulation studies such as the one implemented here provide an alternative approach for comparing model-based and direct estimators where the true underlying population is known.

Despite the advances our proposed model provides in estimate precision, there are certain limitations that require future consideration. First, the log transformation of biomass occasionally resulted in extremely large uncertainty for the small area estimates (e.g., western Texas for live oak and water oak, Supplemental Information S1: Figures S6 and S19). Given this only occurred on the boundaries of our study region, this could be related to limitations of the isotropic and stationary exponential covariance function used to estimate the spatial processes. The use of more complex covariance functions (e.g., anisotropic functions) could help mitigate this pattern, which have been shown to provide improvements over isotropic functions in ecological applications (e.g., [Finley \(2011\)](#)). Additionally, counties along the edge of a species distribution within the study area often had more minimal precision gains relative to counties within the core of its range (e.g., laurel oak, live oak, slash pine). This could similarly reflect limitations of the isotropic exponential covariance function, or could result from localized drivers of biomass at these range edges that could not be explained by the covariates included in the analysis. Finally, care must be used when determining the size of the grid needed for generating the small area estimates ([Finley et al., 2024](#)). If the grid is not fine enough to represent spatial variation adequately throughout the small area of interest, the resulting estimates can be unstable. To ensure small area estimates are stable, we recommend predicting across multiple grid resolutions for at least a subset of the small areas of interest and comparing the resulting estimates. Because increasing the resolution of the grid will also increase the computational burden of the prediction task, one should choose the lowest resolution grid at which the estimates stabilize upon a single value.

NFIs are increasingly being called upon to provide insights on forest resources at small spatial scales. Forest parameter estimates by species are often desired at these scales to inform wildlife habitat protection, assessments of timber supply and products, and strategic management planning. Given the sample sizes that are feasibly collected by NFIs, classic design-based estimates are not able to provide sufficient precision to meet these needs. The proposed multivariate spatial model developed here provides more precise estimates of species-specific forest parameters from NFIs, ultimately generating more opportunities for managers, landowners, and researchers to address small-scale management and research questions with NFI data.

CRediT authorship contribution statement

Jeffrey W. Doser: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Malcolm S. Itter:** Writing – review & editing, Methodology, Data curation. **Grant M. Domke:** Writing – review & editing, Validation. **Andrew O. Finley:** Writing – review & editing, Validation, Supervision, Resources, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare no conflict of interest.

Acknowledgments

We thank two anonymous reviewers for their insightful comments that improved the manuscript. We acknowledge the computing resources provided by North Carolina State University High Performance Computing Services Core Facility (RRID:SCR_022168). This work was supported by USDA Forest Service (FS), USDA FS/NCASI Partnership on Small Area Estimation, National Science Foundation (NSF), United States DEB-2213565 and DEB-1946007, and National Aeronautics and Space Administration (NASA) CMS Hayes-2023. The findings and conclusions in this publication are those of the authors and should not be construed to represent any official USDA or US Government determination or policy.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at <https://doi.org/10.1016/j.foreco.2025.123112>.

Data availability

All code and data used in this manuscript are available at <https://doi.org/10.5281/zenodo.16995883> (Doser, 2025).

References

- Abatzoglou, J.T., Dobrowski, S.Z., Parks, S.A., Hegewisch, K.C., 2018. Terraclimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. *Sci. Data* 5 (1), 1–12.
- Bechtold, W.A., Patterson, P.L., 2005. The Enhanced Forest Inventory and Analysis Program–National Sampling Design and Estimation Procedures, vol. 80, USDA Forest Service, Southern Research Station.
- Breidenbach, J., Astrup, R., 2012. Small area estimation of forest attributes in the Norwegian National Forest Inventory. *Eur. J. For. Res.* 131, 1255–1267.
- Brooks, S.P., Gelman, A., 1998. General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Statist.* 7 (4), 434–455.
- Cao, Q., Dettmann, G.T., Radtke, P.J., Coulston, J.W., Derwin, J., Thomas, V.A., Burkhart, H.E., Wynne, R.H., 2022. Increased precision in county-level volume estimates in the United States National Forest Inventory with area-level small area estimation. *Front. For. Glob. Chang.* 5, 769917.
- Clark, J.S., Gelfand, A.E., Woodall, C.W., Zhu, K., 2014. More than the sum of the parts: forest climate response from joint species distribution models. *Ecol. Appl.* 24 (5), 990–999.

- Clark, J.S., Nemerut, D., Seyednasrollah, B., Turner, P.J., Zhang, S., 2017. Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data. *Ecol. Monograph* 87 (1), 34–56.
- Datta, A., Banerjee, S., Finley, A.O., Gelfand, A.E., 2016. Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *J. Amer. Statist. Assoc.* 111 (514), 800–812.
- Dettmann, G.T., Radtke, P.J., Coulston, J.W., Green, P.C., Wilson, B.T., Moisen, G.G., 2022. Review and synthesis of estimation strategies to meet small area needs in forest inventory. *Front. For. Glob. Chang.* 5, 813569.
- Dorazio, R.M., Royle, J.A., 2005. Estimating size and composition of biological communities by modeling the occurrence of species. *J. Amer. Statist. Assoc.* 100 (470), 389–398.
- Doser, J.W., 2025. Code and data for Multivariate spatial models for small area estimation of species-specific forest inventory parameters. Zenodo. <https://doi.org/10.5281/zenodo.1695883>.
- Doser, J.W., Finley, A.O., Banerjee, S., 2023. Joint species distribution models with imperfect detection for high-dimensional spatial data. *Ecology* 104 (9), e4137.
- Doser, J.W., Finley, A.O., Kéry, M., Zipkin, E.F., 2022. spOccupancy: An R package for single-species, multi-species, and integrated spatial occupancy models. *Methods Ecol. Evol.* 13 (8), 1670–1678.
- Doser, J.W., Finley, A.O., Kéry, M., Zipkin, E.F., 2024. spAbundance: An R package for single-species and multi-species spatially explicit abundance models. *Methods Ecol. Evol.* 15 (6), 1024–1033.
- Doser, J.W., Stanke, H., Finley, A.O., 2025. rFIA: Estimation of forest variables using the FIA database. <https://cran.r-project.org/package=rFIA>.
- Finley, A.O., 2011. Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. *Methods Ecol. Evol.* 2 (2), 143–154.
- Finley, A.O., Andersen, H.-E., Babcock, C., Cook, B.D., Morton, D.C., Banerjee, S., 2024. Models to support forest inventory and small area estimation using sparsely sampled lidar: A case study involving G-LiHT LiDAR in Tanana, Alaska. *J. Agric. Biol. Environ. Stat.* 1–28.
- Finley, A.O., Banerjee, S., MacFarlane, D.W., 2012. A hierarchical model for quantifying forest variables over large heterogeneous landscapes with uncertain forest areas. *J. Amer. Statist. Assoc.* 106 (493), 31–48.
- Finley, A.O., Datta, A., Cook, B.D., Morton, D.C., Andersen, H.E., Banerjee, S., 2019. Efficient algorithms for Bayesian nearest neighbor Gaussian processes. *J. Comput. Graph. Statist.* 28 (2), 401–414.
- Gelfand, A.E., Schmidt, A.M., Wu, S., Silander, Jr., J.A., Latimer, A., Rebelo, A.G., 2005. Modelling species diversity through species level hierarchical modelling. *J. R. Stat. Soc. Ser. C. Appl. Stat.* 54 (1), 1–20.
- Gregoire, T.G., 1998. Design-based and model-based inference in survey sampling: appreciating the difference. *Can. J. Forest Res.* 28 (10), 1429–1447.
- Harris, V., Caputo, J., Finley, A., Butler, B.J., Bowlick, F., Catanzaro, P., 2021. Small-area estimation for the USDA Forest Service, National Woodland Owner Survey: Creating a fine-scale land cover and ownership layer to support county-level population estimates. *Front. For. Glob. Chang.* 4.
- Hogan, J.W., Tchernis, R., 2004. Bayesian factor analysis for spatially correlated data, with application to summarizing area-level material deprivation from census data. *J. Amer. Statist. Assoc.* 99 (466), 314–324.
- Hogg, R.V., 2020. Probability and Statistical Inference, vol. 993, Pearson.
- Hollister, J., Shah, T., Nowosad, J., Robitaille, A.L., Beck, M.W., Johnson, M., 2023. Elevatr: Access elevation data from various APIs. R package version 0.99.0.
- Hou, Z., Domke, G.M., Russell, M.B., Coulston, J.W., Nelson, M.D., Xu, Q., McRoberts, R.E., 2021. Updating annual state-and county-level forest inventory estimates with data assimilation and FIA data. *Forest Ecol. Manag.* 483, 118777.
- Houseman, I., Schleeweis, K., Heyer, J., Ruefenacht, B., Bender, S., Megown, K., Goetz, W., Bogle, S., 2023. National Land Cover Database Tree Canopy Cover Methods. US Department of Agriculture, Forest Service, Geospatial Technology and Applications Center, Salt Lake City, Utah, https://data.fs.usda.gov/geodata/rastergateway/treecanopycover/docs/TCC_v2021-4.Methods.pdf.
- Johnson, M., 2024. ClimateR: climateR. R package version 0.3.5.
- Johnson, L.K., Mahoney, M.J., Domke, G.M., Beier, C.M., 2025. New allometric models for the USA create a shift in forest carbon estimation, modeling, and mapping. *Forest Ecol. Manag.* 589, 122751.
- Kershaw, Jr., J.A., Ducey, M.J., Beers, T.W., Husch, B., 2016. Forest Mensuration. John Wiley & Sons.
- May, P.B., Finley, A.O., 2025. Spatial-temporal models for forest inventory data. arXiv e-prints, pages arXiv-2503.
- May, P., McConville, K.S., Moisen, G.G., Bruening, J., Dubayah, R., 2023. A spatially varying model for small area estimates of biomass density across the contiguous United States. *Remote Sens. Environ.* 286, 113420.
- McRoberts, R.E., Tomppo, E.O., Næsset, E., 2010. Advances and emerging issues in national forest inventories. *Scand. J. For. Res.* 25 (4), 368–381.
- Northrup, J.M., Gerber, B.D., 2018. A comment on priors for Bayesian occupancy models. *PLoS One* 13 (2), e0192819.
- Pebesma, E., Bivand, R., 2023. Spatial Data Science: With Applications in R. Chapman and Hall/CRC.
- Prisley, S., Bradley, J., Clutter, M., Friedman, S., Kempka, D., Rakestraw, J., Sonne Hall, E., 2021. Needs for small area estimation: Perspectives from the US private forest sector. *Front. For. Glob. Chang.* 4, 746439.
- R Core Team, 2024. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
- Rao, J.N., Molina, I., 2015. Small Area Estimation. John Wiley & Sons.
- Särndal, C.-E., Swensson, B., Wretman, J., 2003. Model Assisted Survey Sampling. Springer Science & Business Media.
- Shannon, E., Finley, A.O., Domke, G.M., May, P.B., Andersen, H.-E., Gaines, III, G.C., Banerjee, S., 2024. Toward spatio-temporal models to support national-scale forest carbon monitoring and reporting. *Environ. Res. Lett.* 20 (1), 014052.
- Smith, W.B., 2002. Forest inventory and analysis: a national inventory and monitoring program. *Environ. Pollut.* 116, S233–S242.
- Stanke, H., Finley, A.O., Weed, A.S., Walters, B.F., Domke, G.M., 2020. rFIA: An R package for estimation of forest attributes with the US Forest Inventory and Analysis database. *Environ. Model. Softw.* 127, 104664.
- Tang, B., Frye, H.A., Silander, Jr., J.A., Gelfand, A.E., 2025. Zero-inflated multivariate tobit regression modeling. *J. Statist. Plann. Inference* 236, 106229.
- Taylor-Rodriguez, D., Finley, A.O., Datta, A., Babcock, C., Andersen, H.-E., Cook, B.D., Morton, D.C., Banerjee, S., 2019. Spatial factor models for high-dimensional and large spatial data: An application in forest variable mapping. *Statist. Sinica* 29, 1155.
- Tikhonov, G., Duan, L., Abrego, N., Newell, G., White, M., Dunson, D., Ovaskainen, O., 2020. Computationally efficient joint species distribution modeling of big spatial data. *Ecology* 101 (2), e02929.
- Ver Planck, N.R., Finley, A.O., Huff, E.S., 2017. Hierarchical Bayesian models for small area estimation of county-level private forest landowner population. *Can. J. Forest Res.* 47 (12), 1577–1589.
- Weiskittel, A.R., Hann, D.W., Kershaw, Jr., J.A., Vanclay, J.K., 2011. Forest Growth and Yield Modeling. John Wiley & Sons.
- Westfall, J.A., Coulston, J.W., Gray, A.N., Shaw, J.D., Radtke, P.J., Walker, D.M., Weiskittel, A.R., MacFarlane, D.W., Affleck, D.L., Zhao, D., et al., 2024. A National-Scale Tree Volume, Biomass, and Carbon Modeling System for the United States. Gen. Tech. Rep. WO-104, US Department of Agriculture, Forest Service, Washington, DC, p. 104.
- Westfall, J.A., Coulston, J.W., Moisen, G.G., Andersen, H.-E., 2022. Sampling and Estimation Documentation for the Enhanced Forest Inventory and Analysis Program: 2022. Gen. Tech. Rep. NRS-GTR-207, –US Dept of Agriculture, Forest Service, Northern Research Station, p. 129p.
- Westfall, J.A., Patterson, P.L., Coulston, J.W., 2011. Post-stratified estimation: within-strata and total sample size recommendations. *Can. J. Forest Res.* 41 (5), 1130–1139.
- White, G.W., McConville, K.S., Moisen, G.G., Frescino, T.S., 2021. Hierarchical Bayesian small area estimation using weakly informative priors in ecologically homogeneous areas of the interior Western forests. *Front. For. Glob. Chang.* 4, 752911.
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al., 2019. Welcome to the tidyverse. *J. Open Source Softw.* 4 (43), 1686.
- Zipkin, E.F., DeWan, A., Andrew Royle, J., 2009. Impacts of forest fragmentation on species richness: a hierarchical approach to community modelling. *J. Appl. Ecol.* 46 (4), 815–822.