



Hierarchical Bayesian Small Area Estimation Using Weakly Informative Priors in Ecologically Homogeneous Areas of the Interior Western Forests

Grayson W. White¹, Kelly S. McConville^{2*}, Gretchen G. Moisen³ and Tracey S. Frescino³

¹ RedCastle Resources, Inc., Salt Lake City, UT, United States, ² Mathematics Department, Reed College, Portland, OR, United States, ³ Rocky Mountain Research Station, U.S. Department of Agriculture (USDA) Forest Service, Ogden, UT, United States

OPEN ACCESS

Edited by:

Annika Kangas,
Natural Resources Institute Finland
(Luke), Finland

Reviewed by:

Göran Ståhl,
Swedish University of Agricultural
Sciences, Sweden
Andrew Finley,
Michigan State University,
United States

*Correspondence:

Kelly S. McConville
mcconville@reed.edu

Specialty section:

This article was submitted to
Forest Management,
a section of the journal
Frontiers in Forests and Global
Change

Received: 03 August 2021

Accepted: 25 October 2021

Published: 15 December 2021

Citation:

White GW, McConville KS, Moisen GG
and Frescino TS (2021) Hierarchical
Bayesian Small Area Estimation Using
Weakly Informative Priors in
Ecologically Homogeneous Areas of
the Interior Western Forests.
Front. For. Glob. Change 4:752911.
doi: 10.3389/ffgc.2021.752911

The U.S. Forest Inventory and Analysis Program (FIA) collects inventory data on and computes estimates for many forest attributes to monitor the status and trends of the nation's forests. Increasingly, FIA needs to produce estimates in small geographic and temporal regions. In this application, we implement area level hierarchical Bayesian (HB) small area estimators of several forest attributes for ecosubsections in the Interior West of the US. We use a remotely-sensed auxiliary variable, percent tree canopy cover, to predict response variables derived from ground-collected data such as basal area, biomass, tree count, and volume. We implement four area level HB estimators that borrow strength across ecological provinces and sections and consider prior information on the between-area variation of the response variables. We compare the performance of these HB estimators to the area level empirical best linear unbiased prediction (EBLUP) estimator and to the industry-standard post-stratified (PS) direct estimator. Results suggest that when borrowing strength to areas which are believed to be homogeneous (such as the ecosection level) and a weakly informative prior distribution is placed on the between-area variation parameter, we can reduce variance substantially compared the analogous EBLUP estimator and the PS estimator. Explorations of bias introduced with the HB estimators through comparison with the PS estimator indicates little to no addition of bias. These results illustrate the applicability and benefit of performing small area estimation of forest attributes in a HB framework, as they allow for more precise inference at the ecosubsection level.

Keywords: forest inventory, empirical best linear unbiased prediction, remote sensing, post-stratification, indirect estimation, probabilistic graphical model, weakly informative priors, ecoregion

1. INTRODUCTION

The USDA Forest Service Forest Inventory and Analysis Program (FIA) collects a sample of inventory data nationwide to monitor status and trends in forested ecosystems at scales relevant for strategic-level planning. Increasingly, this network of valuable inventory plots is being called upon to answer questions relevant to forest land management which is below the spatial and temporal scales for which the sample was originally designed. Information is needed on resources lost and recovery rates within disturbance boundaries, on significant change in carbon sources and sinks, as well as on the state of the forests within individual counties, districts, or other small management

units. There is strong interest in exploring methods to integrate extant inventory data with remotely sensed data through models that expand the capacity to estimate forest attributes over smaller domains in space and time.

A standard estimator that combines inventory data with remotely sensed data is the generalized regression estimator (Cassel et al., 1976) in which the inventory data are modeled and predicted over the domain of interest and then the observed data and predictions are aggregated to construct an estimator. Post-stratification (PS), a common estimation technique for national forest inventories (NFI) such as FIA, is a special case of the generalized regression estimator that incorporates a single, categorical auxiliary variable into the estimator (Särndal et al., 1992). Since the generalized regression estimator only makes use of data within the domain of interest, it is called a *direct estimator*. And, although leveraging auxiliary data typically improves the precision of a direct estimator, it tends to still not achieve adequate levels of precision when the sample size of the inventory data in the domain is small (Rao and Molina, 2015). Therefore, we consider here *indirect estimators* with their defining characteristic of borrowing strength from data outside the domain of interest. These domains are often classified as *small areas* and we will use the terms *domain* and *small area* interchangeably. When the indirect estimator explicitly relies on a model to link the data in the desired small area with data in other related small areas it is called a *small area estimator*. These linking models can be built either at the area level or unit (i.e., plot) level, depending on data availability and the strength of the relationships between the inventory data and remote sensing data at these two resolutions. We study area level models here because the inventory and remotely-sensed data we consider have strong linear trends at the area level and violate normality assumptions at the unit level. These estimators are constructed under either a frequentist framework where the quantities of interest are fixed, unknown values or a Bayesian framework where they are considered random variables. Key advantages of the Bayesian approach are that it allows the modeler to directly consider uncertainty between the small areas and to obtain distributions, not just point estimates and standard error estimates, for the parameters of interest.

A frequently utilized, and frequentist-based, indirect estimator is the empirical best linear unbiased prediction (EBLUP) estimator, which uses a linking model with random area-specific effects to borrow strength from related areas (Rao and Molina, 2015). The suitability of area and unit level EBLUP estimators to the small area applications found in NFIs have been studied extensively (Goerndt et al., 2011; Breidenbach and Astrup, 2012; Magnussen et al., 2017; Mauro et al., 2017; Coulston et al., 2021). This paper considers the Bayesian analog to the EBLUP, a hierarchical Bayesian (HB) estimator. These HB estimators are not commonly used in forest inventory research; however, they have been applied in a variety of other application areas ranging from poverty mapping to agriculture to transportation to employment (You et al., 2003; Vaish et al., 2010; Wang et al., 2012; Molina et al., 2014) to name a few. Within the NFI literature, Ver Planck et al. (2018) explored an area level HB estimator for estimating forest attributes and did

find improvements in precision over the Horvitz-Thompson (HT) direct estimator.

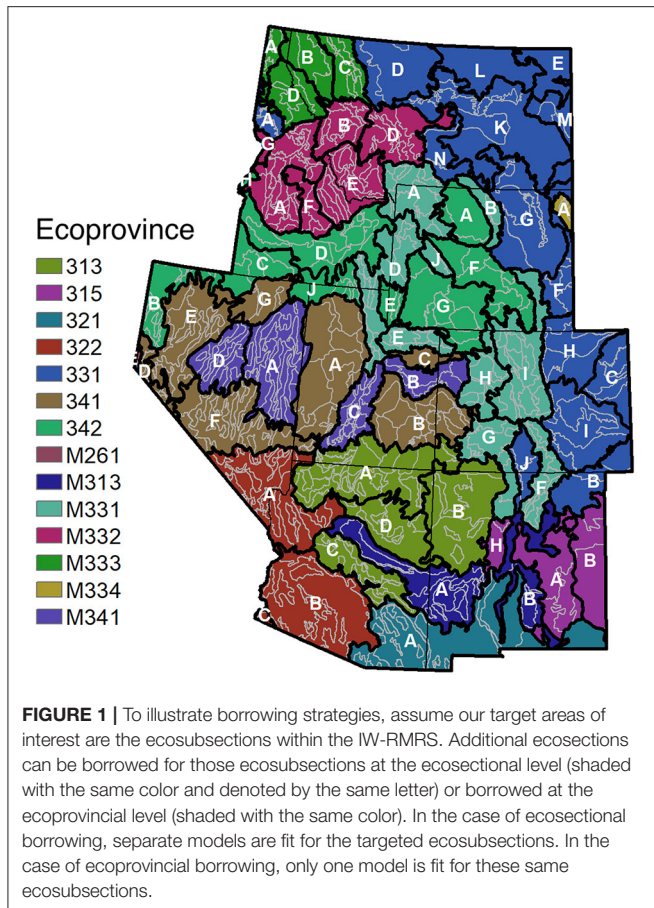
In this paper, we explore the performance of the PS, the area level EBLUP, and the area level HB estimators at estimating the mean value of four response variables: basal area (m^2 per hectare), count of trees per hectare, above-ground biomass (kg per hectare), and net volume of trees (m^3 per hectare), excluding rotten or form defects, across the Interior West (IW) of the US. We generate estimates within the subregion (ecosubsection) of a hierarchical system of ecological divisions. For both the EBLUP and HB approaches, we consider the impact of borrowing strength from two resolutions from upper hierarchical levels: ecosubsection and ecoprovince, at scales of thousands of acres and millions of acres, respectively. Leveraging the flexibility of the HB, we study the impacts of varying how prior information on the homogeneity of the modeled small areas is incorporated into the estimator. We find that when borrowing strength to the ecosubsection level and including weakly informative prior information about small area homogeneity with the area level HB estimator we can reduce variance substantially compared to other common estimators. Explorations of potential bias through comparison with the post-stratified estimator display almost no introduction of bias with this estimator. However, since we do not know the true mean of the response variable of interest, caution is warranted when making strong conclusions about bias.

2. METHODS

2.1. Region of Study and FIA Data

This manuscript focuses on estimating the mean of several key forest attributes for the ecosubsections in the IW region of the United States (**Figure 1**), which encompasses the states of Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, and Wyoming. The inventory data were collected by FIA using a geographically-based systematic sampling design, where each plot represents about 2,500 ha of land (Bechtold and Patterson, 2005) and cover a 10 year measurement cycle from 2007 to 2017. This sample of 86,065 inventory plots were downloaded on February 6, 2019 from the FIA database, version FIADB_1.8.9.99 (last updated Dec 3, 2018). Our analyses include the use of four variables from the FIA database as response variables: basal area (m^2 per hectare), count of trees per hectare, above-ground biomass (kg per hectare), and net volume (m^3 per hectare). For remotely-sensed auxiliary variables, we consider a forest/non-forest classification used for post-stratifying in the IW (Blackard et al., 2008) and the 2016 National Land Cover Database percent tree canopy cover map (Yang et al., 2018), which has a spatial resolution of 30 m. Although each FIA plot consists of 4 subplots, response variables represent the aggregation of information at the plot level, and just the FIA plot center was intersected with the two auxiliary data layers. As input to the area-level estimators described below, both the percent canopy cover and proportions of forest and non-forest classes are averaged to the small area level for the IW. FIA data retrievals and processing of auxiliary data were done through the R package FLESTA (Frescino et al., 2015).

Borrowing strength in small area applications often occurs using political boundaries, such as counties within a state. But in



order to borrow from “similar” domains, there is an opportunity to use any one of many classification systems that exist in the US to help guide borrowing in a more ecologically sensible fashion. As one prominent example, Cleland et al. (2007) delineated ecological units across the conterminous US using biological and physical information such as potential natural vegetation, geology, soils, climate, and hydrology. These ecological units were developed in a nested hierarchical structure. Ecoprovinces identify major vegetation cover types and land forms. Ecosubsections delineate more homogeneous areas within the ecoprovinces based on more detailed physical and biological components of the environment. Ecosubsections provide another step toward homogeneity at an even finer scale, with the number of FIA plots in each ecosubsection ranging from 1 to 2,200 in the Interior West. **Figure 1** illustrates how the nested hierarchical structure of this ecological classification system facilitates borrowing at different ecological scales. We investigate how borrowing strategies affect the performance of the indirect estimators by comparing estimates and standard errors of the area level estimators applied to ecosubsections when borrowing occurs at the ecoprovincial vs. ecosubsectional levels.

2.2. Estimators

We consider the PS estimator, which is a direct estimator, and two indirect estimation approaches, the EBLUP and HB, based on an area level, linear mixed model. For the HB method, we

explore four different estimators, which vary based on how prior knowledge is incorporated to see how that impacts the estimator’s precision and bias. All data analysis is conducted using the statistical software package R (R Core Team, 2020). In particular, the PS estimator is fit with the mase package (McConville et al., 2018), the EBLUP estimators are fit with sae (Molina and Marhuenda, 2015), and the HB estimators are fit with mcmcSae (Boonstra, 2021).

In order to explore these estimators in depth, we now introduce relevant notation. First, suppose we have m small areas we wish to estimate. Next, the indices are as follows: i indexes over units sampled; j indexes over small areas (in our case, ecosubsections); and k indexes over post-strata. Now, recall the goal of producing estimates of the mean of some response variable y , such as trees per hectare, in a small area. So, let μ_{y_j} be the population mean of the study variable in ecosubsection j in the IW. To denote the estimator produced for μ_{y_j} we use $\hat{\mu}_{y_j}$ with a superscript denoting which estimator is being used. We also use $\hat{V}(\hat{\mu}_{y_j})$ to denote the estimator of the variance of $\hat{\mu}_{y_j}$. The set s_j of size n_j includes all units sampled within ecosubsection j . We use the shorthand “iid” when referring to independent and identically distributed random variables and “ind” for independent random variables.

2.2.1. Direct Estimation via Post-stratification

We implement the PS estimator, which is commonly used by FIA and other NFIs, and is considered a direct estimator of μ_{y_j} since it only uses the inventory and auxiliary data within ecosubsection j . With the set of weights, $\{w_{jk}\}_{k=1}^K$, representing the proportion of pixels in each post-stratum for ecosubsection j , the PS estimator of μ_{y_j} is represented as follows:

$$\hat{\mu}_{y_j}^{PS} = \sum_{k=1}^K w_{jk} \hat{\mu}_{y_{jk}}^{HT} \tag{1}$$

and is a weighted average of the post-strata HT estimators, given by $\hat{\mu}_{y_{jk}}^{HT} = n_{jk}^{-1} \sum_{i \in s_{jk}} y_i$ where s_{jk} is the subset of the sample in ecosubsection j that falls in post-stratum k and n_{jk} is the corresponding sample size (Särndal et al., 1992). Since we have equal probability sampling, the post-strata HT estimators equal the post-strata sample means. For the IW, ignoring adjustments for non-response, the post-strata classes are forest and non-forest so $K = 2$. Post-stratification can certainly be conducted using more than 2 classes (e.g., Rintoul et al., 2020) but here we applied post-strata consistent with that used in the IW production inventory processes.

The variance estimator for $\hat{\mu}_{y_j}^{PS}$ is given by:

$$\hat{V}(\hat{\mu}_{y_j}^{PS}) = \frac{1}{n_j} \left(\sum_{k=1}^K w_{jk} n_{jk} \hat{V}(\hat{\mu}_{y_{jk}}^{HT}) + \sum_{k=1}^K (1 - w_{jk}) \frac{n_{jk}}{n_j} \hat{V}(\hat{\mu}_{y_{jk}}^{HT}) \right) \tag{2}$$

(Equation 7.6.6 in Särndal et al., 1992 without the finite population correction) where the HT variance estimator of $\hat{\mu}_{y_{jk}}^{HT}$ is given by

$$\hat{V}(\hat{\mu}_{y_{jk}}^{HT}) = \frac{1}{n_{jk}(n_{jk} - 1)} \sum_{i \in s_{jk}} (y_i - \hat{\mu}_{y_{jk}}^{HT})^2. \quad (3)$$

Since the number of pixels for the post-strata map is significantly larger than the sample size, the finite population correction would be negligible and is therefore omitted from the variance estimator calculation. For the IW, the PS estimator is typically more efficient than the HT estimator since many of the desired response variables are more homogeneous within the forest/non-forest post-strata. [For the forest/non-forest used here, Blackard et al. (2008) report an accuracy of 91% correctly classified based on an independent test set, with errors of omission and commission for forest at 17 and 18%, respectively, and for non-forest at 7 and 6%, respectively]. Therefore, we use the PS estimator in the subsequent indirect estimators when a direct estimator is needed.

2.2.2. Indirect Estimation via Small Area Models

When the sample sizes in the domains of interest are small, direct estimation techniques often do not provide sufficiently small variances, even with the use of auxiliary data, to make informative inferences. Indirect estimators increase the effective sample size by borrowing strength from data outside, with greater gains made when the larger area has the same characteristics, in terms of the response variables and their relationships with the auxiliary data, as the small areas of interest.

One common technique for borrowing strength is to explicitly use a linking model with a random-area specific effect, in addition to the sampling model which describes the data generation. Combining the linking model and the sampling model results in a mixed model approach to estimating the parameters of interest in the small areas. We consider a linear mixed model, which can be estimated using the EBLUP or using HB when additional assumptions are made on model parameters.

2.2.2.1. The Area Level EBLUP Estimator

For our parameters of interest, μ_{y_j} , we assume the following linking model:

$$\mu_{y_j} = \beta_o + \beta_1 \bar{X}_j + v_j \quad (4)$$

where \bar{X}_j is the average percent tree canopy cover for ecosubsection j and the area-specific random effects satisfy the following conditions:

$$v_j \stackrel{\text{iid}}{\sim} N(0, \sigma_v^2).$$

And, we assume the PS estimators were generated from the following data generation model:

$$\hat{\mu}_{y_j}^{PS} = \mu_{y_j} + \epsilon_j \quad (5)$$

where $\epsilon_j \stackrel{\text{iid}}{\sim} N(0, \sigma_j^2)$. Inserting Equation (4) into Equation (5) gives the following area level mixed model, also known as the Fay-Herriot model (Fay and Herriot, 1979):

$$\hat{\mu}_{y_j}^{PS} = \beta_o + \beta_1 \bar{X}_j + v_j + e_j \quad (6)$$

where

$$v_j \stackrel{\text{iid}}{\sim} N(0, \sigma_v^2), \quad e_j \stackrel{\text{iid}}{\sim} N(0, \sigma_j^2), \quad \text{and} \quad v_j \perp\!\!\!\perp e_j.$$

To obtain an estimator of μ_{y_j} from this model, we use an EBLUP approach. This requires estimating the within-area and between area variances and the model coefficients. For $j = 1, 2, \dots, m$, the within-area variations, σ_j^2 are set to $\hat{V}(\hat{\mu}_{y_j}^{PS})$, the estimated variances of the PS estimates. The between-area variation, σ_v^2 , is estimated using a method of moments estimator (Ch 6.1.2 in Rao and Molina, 2015) and the estimated model coefficients $\hat{\beta}_o$ and $\hat{\beta}_1$ are the EBLUPs of β_o and β_1 , respectively. The equation for the variance estimator of the EBLUP estimator of μ_{y_j} is given in the **Appendix**.

The EBLUP estimator of μ_{y_j} can be expressed as a weighted average of the direct estimator and an area level regression-synthetic estimator:

$$\hat{\mu}_{y_j}^{EBLUP} = \hat{\gamma}_j \hat{\mu}_{y_j}^{PS} + (1 - \hat{\gamma}_j)(\hat{\beta}_o + \hat{\beta}_1 \bar{X}_j) \quad (7)$$

where

$$\hat{\gamma}_j = \frac{\hat{\sigma}_v^2}{\hat{V}(\hat{\mu}_{y_j}^{PS}) + \hat{\sigma}_v^2}. \quad (8)$$

Notice that the EBLUP estimator is a composite of an indirect and a direct estimator where the weighting term accounts for local variation. In particular, $\hat{\gamma}$ is the ratio of between-area variation and total variation. When the small areas are fairly heterogeneous, the EBLUP will rely more heavily on the direct, PS estimator, which only relies on data within the small area of interest. The estimator leans more on outside information when the variance estimator of the PS estimator is large compared to the variability between the small areas. In this case, it relies on the fixed effect component of the estimated regression line, which is called a regression-synthetic estimator.

2.2.2.2. The Area Level Hierarchical Bayesian Estimator

So far, we have explored common frequentist approaches to small area estimation. However, the primary focus of this paper is to study the performance of the HB for small area estimation. Under the Bayesian paradigm, the parameter of interest, μ_{y_j} , and other model parameters, are treated as random variables instead of fixed, unknown values. Leveraging Bayes' Theorem, this technique synthesizes information gained from the data via a likelihood function with prior knowledge about the parameter of interest and model parameters to obtain a posterior distribution for the parameters:

$$P(\mu_{y_j}, \beta_o, \beta_1, \sigma_v^2 \mid \text{data}) \propto P(\text{data} \mid \mu_{y_j}, \beta_o, \beta_1, \sigma_v^2) \cdot P(\mu_{y_j}, \beta_o, \beta_1, \sigma_v^2) \quad (9)$$

A marginal posterior distribution for μ_{y_j} is found by integrating out the model parameters or by Markov chain Monte Carlo (MCMC) methods. Typically the posterior mean of

the distribution, $E[\mu_{y_j} | \text{data}]$, serves as the estimator of the parameter, with precision provided by the posterior variance, $\text{Var}[\mu_{y_j} | \text{data}]$.

For the area level HB estimator, we start with Equation (6), as was done for the area level frequentist EBLUP, and apply a HB approach. This transformation involves rewriting the data generation model, also referred to as the likelihood function, as a conditional normal distribution where we condition on the parameter of interest and model parameters:

$$\hat{\mu}_{y_j}^{PS} | \mu_{y_j}, \beta_0, \beta_1, \sigma_v^2 \sim N(\mu_{y_j}, \hat{V}(\hat{\mu}_{y_j}^{PS}))$$

and the distribution of μ_{y_j} as a conditional normal distribution where we condition on the model parameters:

$$\mu_{y_j} | \beta_0, \beta_1, \sigma_v^2 \sim N(\beta_0 + \bar{X}_j \beta_1, \sigma_v^2).$$

The HB approach also requires specifying prior distributions for β_0 , β_1 , and σ_v^2 . For the model coefficients, we assume a flat prior:

$$f(\beta_0, \beta_1) \propto 1.$$

For the between-area variation parameter, we consider two prior distributions, an uninformative improper uniform distribution:

$$f(\sigma_v^2) \propto 1$$

and a unit-scale half-Cauchy distribution:

$$\sigma_v \sim \text{half-Cauchy}(\text{scale} = 1).$$

Note that the half-Cauchy distribution is applied to the between-area standard deviation, not the between-area variance. Lastly, we assume the model parameters are independent, namely, $f(\beta_0, \beta_1, \sigma_v^2) = f(\beta_0)f(\beta_1)f(\sigma_v^2)$.

Now that the HB model has been specified, we can attain the small area estimator and variance estimator. For the estimator in ecosubsection j , the Bayes estimator for μ_{y_j} is:

$$\hat{\mu}_{y_j}^{HB} = E[\mu_{y_j} | \hat{\mu}_{y_j}^{PS}]. \tag{10}$$

For the variance in ecosubsection j , the variance of the posterior distribution is used:

$$\hat{V}(\hat{\mu}_{y_j}^{HB}) = V(\mu_{y_j} | \hat{\mu}_{y_j}^{PS}). \tag{11}$$

The estimator and variance estimator are obtained through MCMC methods with the `mcmcscsae` R package (Boonstra, 2021). Using MCMC methods allow for posterior distributions to be well-approximated by sampling from a probability distribution. We use 1,000 sampling iterations (the length of each Markov Chain), 3 Markov Chains, and a burn-in period length of 250 to obtain the results of each HB model we fit.

Figure 2 represents the area level HB estimator as a probabilistic graphical model (PGM). This diagrammatic view can be helpful in understanding the relationships between the parameter of interest, the data, model parameters, and other key

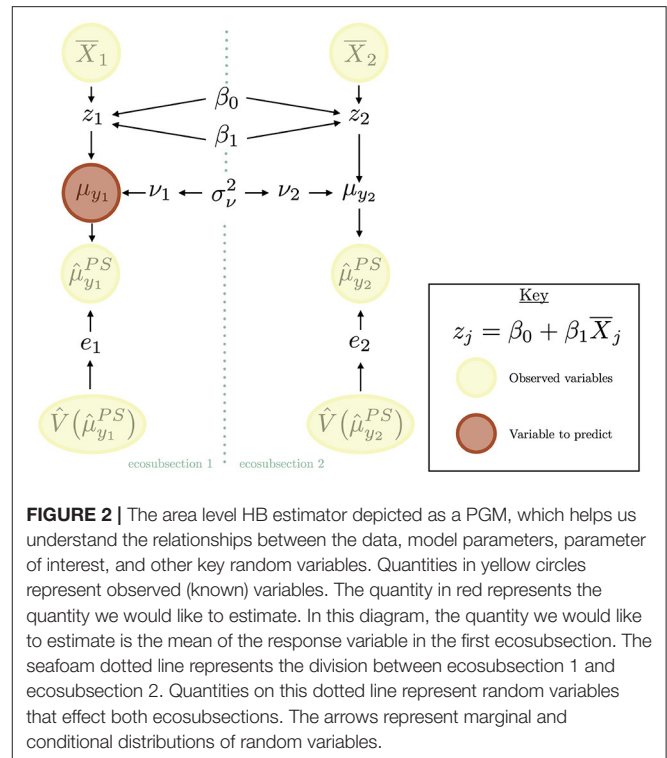


FIGURE 2 | The area level HB estimator depicted as a PGM, which helps us understand the relationships between the data, model parameters, parameter of interest, and other key random variables. Quantities in yellow circles represent observed (known) variables. The quantity in red represents the quantity we would like to estimate. In this diagram, the quantity we would like to estimate is the mean of the response variable in the first ecosubsection. The seafoam dotted line represents the division between ecosubsection 1 and ecosubsection 2. Quantities on this dotted line represent random variables that effect both ecosubsections. The arrows represent marginal and conditional distributions of random variables.

random variables included in the model. For example, the arrows in **Figure 2** can give us the distribution for $\hat{\mu}_{y_j}^{PS}$ and show us that it depends on the parameter of interest (μ_{y_j}) and model parameters (β_0 , β_1 , and σ_v^2). Not only can we quickly see how distributions are conditioned through the use of a PGM, we can also less formally view how variables are related to each other and gain a deeper understanding of how strength is borrowed for this area level HB estimator. If we remove the formality of some parameters representing random variables, we can even use **Figure 2** to visualize how strength is borrowed with the area level EBLUP. Recall that the area level EBLUP is specified with the same linking model and thus strength is borrowed from the same places. Thus, **Figure 2** not only depicts the components of the area level HB model, but also the area level EBLUP, albeit in a less formal way.

2.3. Methods Summary

We use seven estimators—the PS estimator, two area level EBLUPs, and four area level HB estimators—to produce estimates for the average of basal area (m^2 per hectare), tree count per hectare, above-ground biomass (kg per hectare), and net volume (m^3 per hectare). The EBLUPs and HB estimators use one explanatory variable, the average percent tree canopy cover of the ecosubsection, to produce estimates. Estimation occurs at the ecosubsection level, and thus we have produced 11,928 estimates (seven estimators, four response variables, and 426 ecosubsections). The model-based estimators are fit either within an ecoprovince or an ecosection, and hence each ecosubsection only borrows strength out to either the ecoprovince or ecosection

level, not the entire IW region. In order to assess the quality of these estimators, we summarize the findings over the entire study region and for particular regions.

The data span the entire IW; however, we are forced to exclude a small portion of ecosubsections from our analyses. These ecosubsections contain either no or very close to no sampled areas with non-zero values for the variables of interest: that is, areas which are in extremely non-forested areas. These areas have to be excluded due to their within-area variance being zero or so close to zero that the software does not recognize that the number was positive.

3. RESULTS

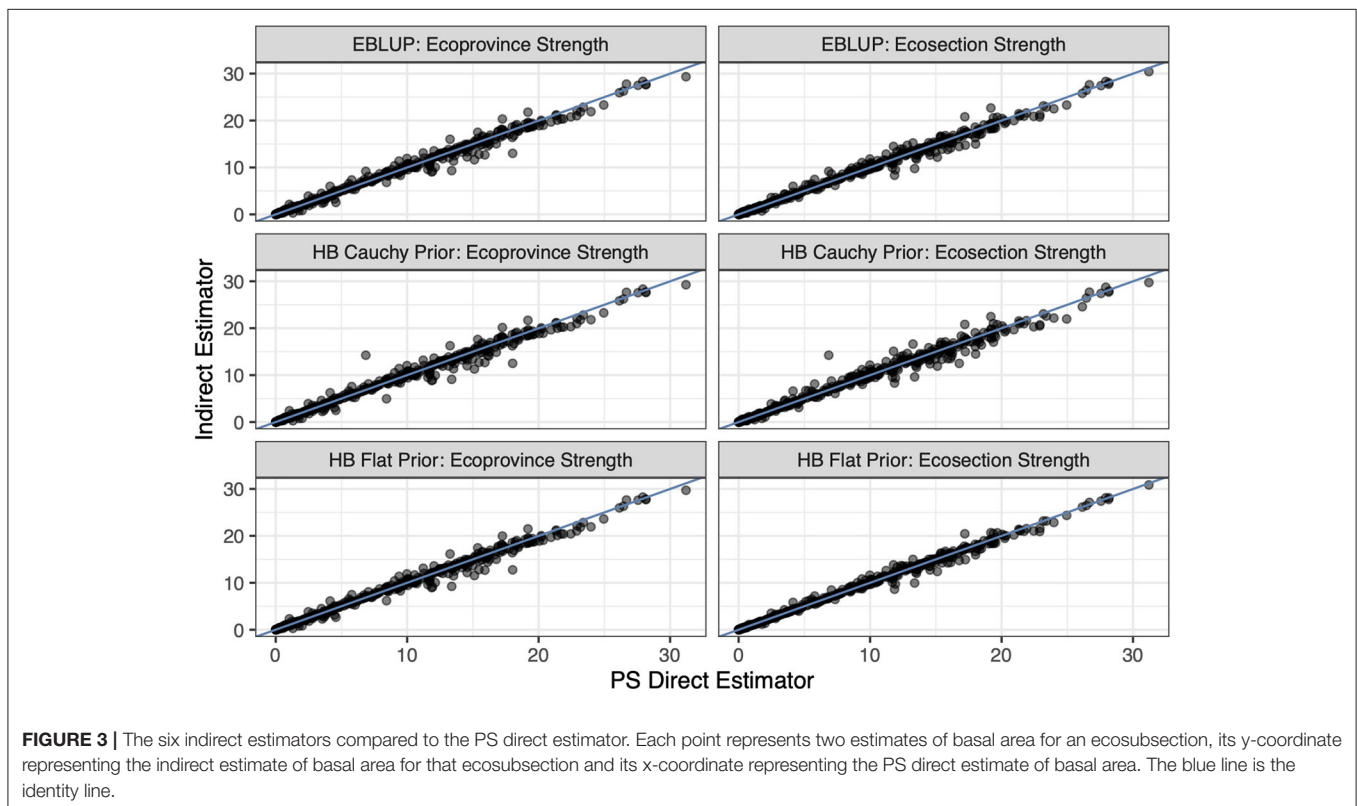
3.1. Estimator Performance

The indirect estimators that we implement perform similarly, on average, to the PS estimator. **Figure 3** displays each indirect estimator's estimate on the y-axis and the PS estimate on the x-axis for the basal area response variable and **Figure 4** does so for the count per hectare response variable. Notably, **Figure 3** shows a strong linear relationship between the indirect and direct estimates (which is also observed for volume and biomass) whereas in **Figure 4** this linear relationship begins to deteriorate for larger values of average canopy cover. This is due to the relationship between the explanatory variable (average canopy cover) and the PS estimator of the average tree count per hectare exhibiting more variability for those larger values, violating the model assumption of homoskedasticity. **Figure 5** displays this larger variability for the tree count per hectare variable

and showcases that the PS estimates consistently fall below the regression line for the largest average canopy cover values. This violation of the homoskedasticity assumption seems to have introduced bias into our indirect estimates. This represents a good cautionary tale that while indirect estimators can provide significant reductions in variance, they can be biased when the model is incorrectly specified. For the remainder of the paper, we focus on basal area, where the linear model specification seems most appropriate.

Figures 3, 4 also display that the flat prior HB estimator produces very similar estimates to the EBLUP for both the estimators that borrow strength out to the ecosubsection level and to the ecoprovince level. **Figure 6** displays this relationship in further detail. Notably, the flat prior HB estimates and standard errors are very similar to the EBLUP. This is expected as we add no prior information to the flat prior HB estimators. By adding no information and specifying the same model we should and do see extremely similar results.

While it is reassuring for the flat prior HB estimator to reinforce the results of the EBLUP, the full benefits of the HB estimators are not gained without careful thought into how prior information is incorporated. In our case, we specify a half-Cauchy prior with scale of one, which is considered a weakly informative prior, on the between-area variation parameter. This distribution places more probability mass over smaller values for our between-area variation, signifying that we expect the between-area variation to be low. This prior is commonly used for the between-area standard deviation parameter in hierarchical models, especially when the number of small areas



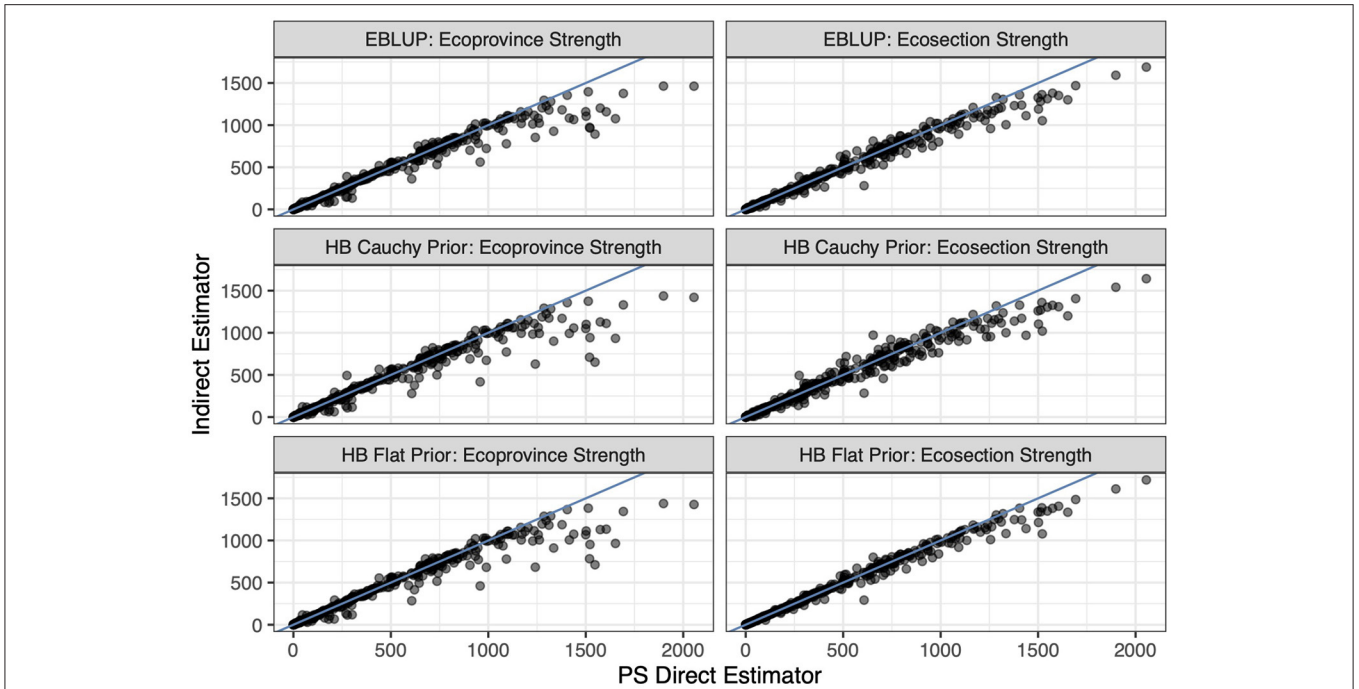


FIGURE 4 | The six indirect estimators compared to the PS direct estimator. Each point represents two estimates of tree count per hectare for an ecosubsection, its y-coordinate representing the indirect estimate of tree count per hectare for that ecosubsection and its x-coordinate representing the PS direct estimate of tree count per hectare. The blue line is the identity line.

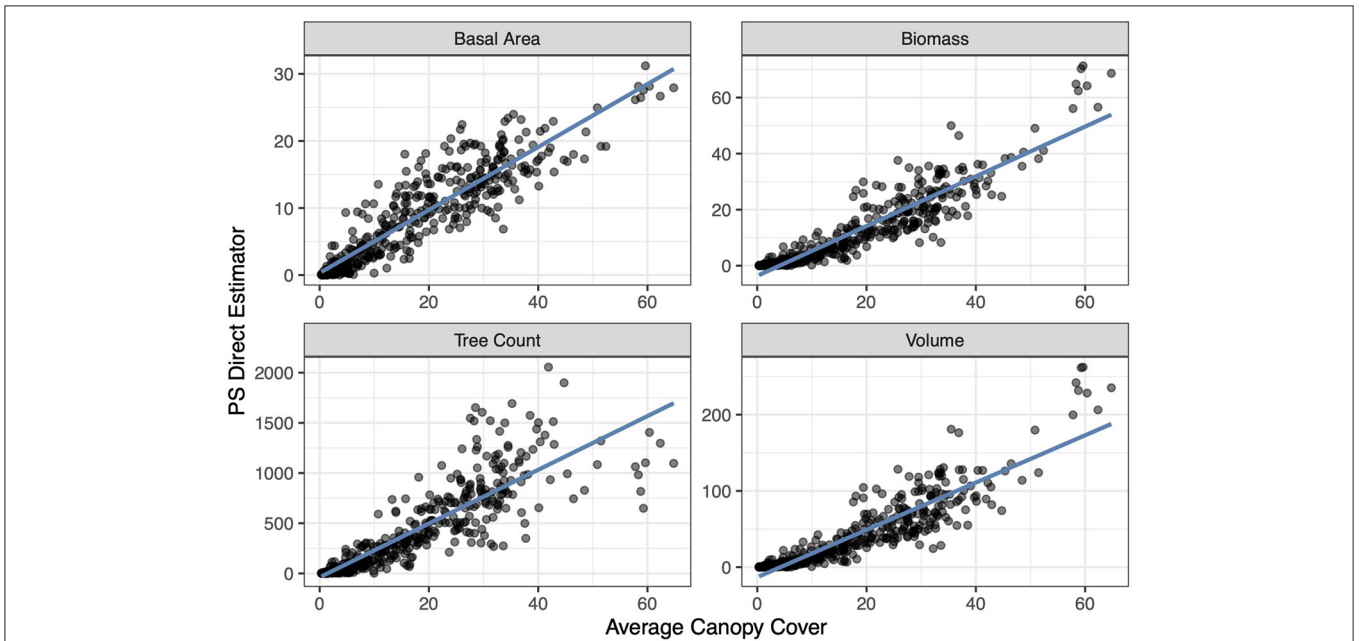


FIGURE 5 | The relationship between the response variables and the explanatory variable (average canopy cover) for each response variable at the ecosubsection level across the IW. The x-coordinate represents the population value of average canopy cover based on remotely sensed data in a given ecosubsection, and the y-coordinate represents the post-stratified estimate of a response variable in a given ecosubsection. The blue line is the ordinary least squares regression line.

is small and so the data provide little information about the group-level variance (Gelman, 2006). Further, we know that ecosections should be more homogeneous than ecoprovinces

and this prior information should reinforce the homogeneity we see in the data. We still chose to place a half-Cauchy prior on the between-area variation when borrowing out to the

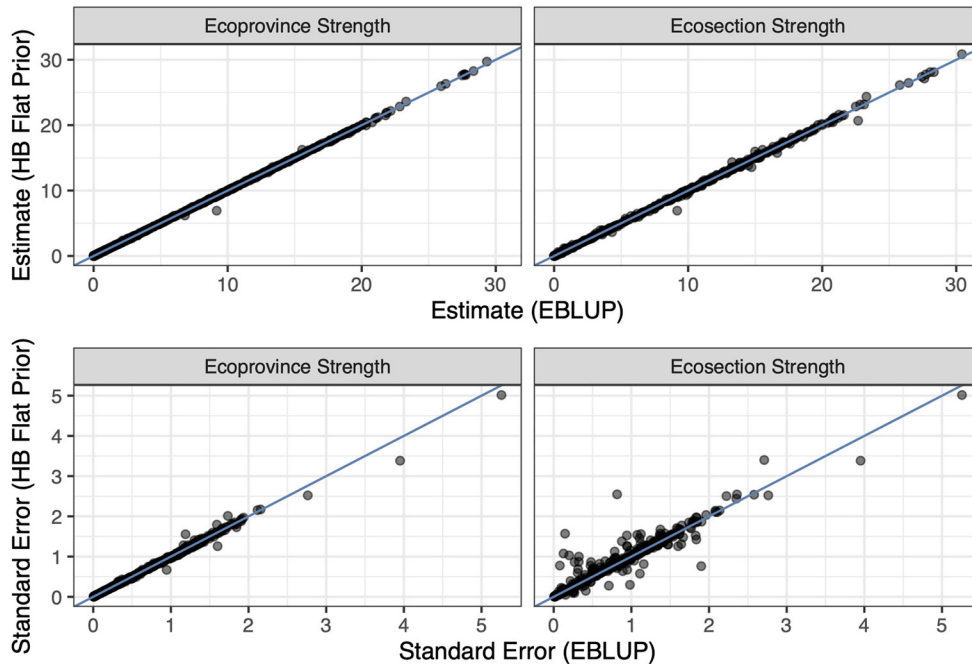


FIGURE 6 | The flat prior HB estimators compared with the EBLUP estimators at each level of strength. The top left plot displays the estimates for each estimator at the ecoprovince level, the top right plot displays the estimates for each estimator at the ecosection level, the bottom left plot displays the standard errors for each estimator at the ecoprovince level, and the bottom right plot displays the standard errors for each estimator at the ecosection level. Each plot contains either estimates or standard errors for the basal area response variable. The blue line in each plot is the identity line.

ecoprovince level, as these regions are defined by ecologists as more homogeneous than the rest of the study region (McNab et al., 2007).

Figure 7 displays the reduction in variance when we change the prior on the between-area variation from flat to half-Cauchy in both the ecosection and ecoprovince approaches. The variance is reduced much more significantly when we use a half-Cauchy prior for estimators that borrow strength to the ecosection level because the number of small areas is smaller. In particular, one can observe that most areas where variance is reduced a large amount have less ecosubsections that they borrow strength from (light purple dots).

Outside of this graphical representation, we can look numerically at the mean and median percent reduction in variance when moving from a flat prior HB estimator to one with the half-Cauchy prior. Table 1 displays both the mean and median percent reduction in variance of basal area for the ecosection and ecoprovince level HB estimators. Borrowing to the more homogeneous ecosection level with the half-Cauchy prior on the between-area variation leads to the greater reductions in variance. While this reduction in variance is compelling, it is possible that the weakly informative prior introduced bias to the estimator.

To understand where bias may be introduced in our estimates, Figure 8 displays the estimates made by the HB estimators with a half-Cauchy prior compared to those made by the PS estimator,

an estimator that is unbiased under resampling regardless of model accuracy. Here, we see a high level of agreement between the two estimators, which suggests the HB estimators are not systematically biased. However, it is important to note that it is the PS estimator, under resampling, that is unbiased, not a given PS estimate. We also saw strong agreement between the estimators from the half-Cauchy prior and those from the flat prior, signifying a robustness to the choice of prior distribution for the between-area variation.

We can also investigate indications of bias numerically, with the percent relative difference (PRD) metric. The PRD between two estimators is defined as followed:

$$PRD(\hat{\mu}_1, \hat{\mu}_2) = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\mu}_2} \cdot 100\%.$$

When we examine the PRD between the PS estimator and the half-Cauchy prior HB estimator for the basal area response variable we see that the average PRDs are -0.007% and 0.756% at the ecoprovince and ecosection level, respectively. The median PRDs between for these estimators are -0.12% and -0.225% at the ecoprovince and ecosection level, respectively. The low PRD values provide additional evidence that we are not introducing much systematic bias with the use of the auxiliary data and prior on the between-area variation parameter.

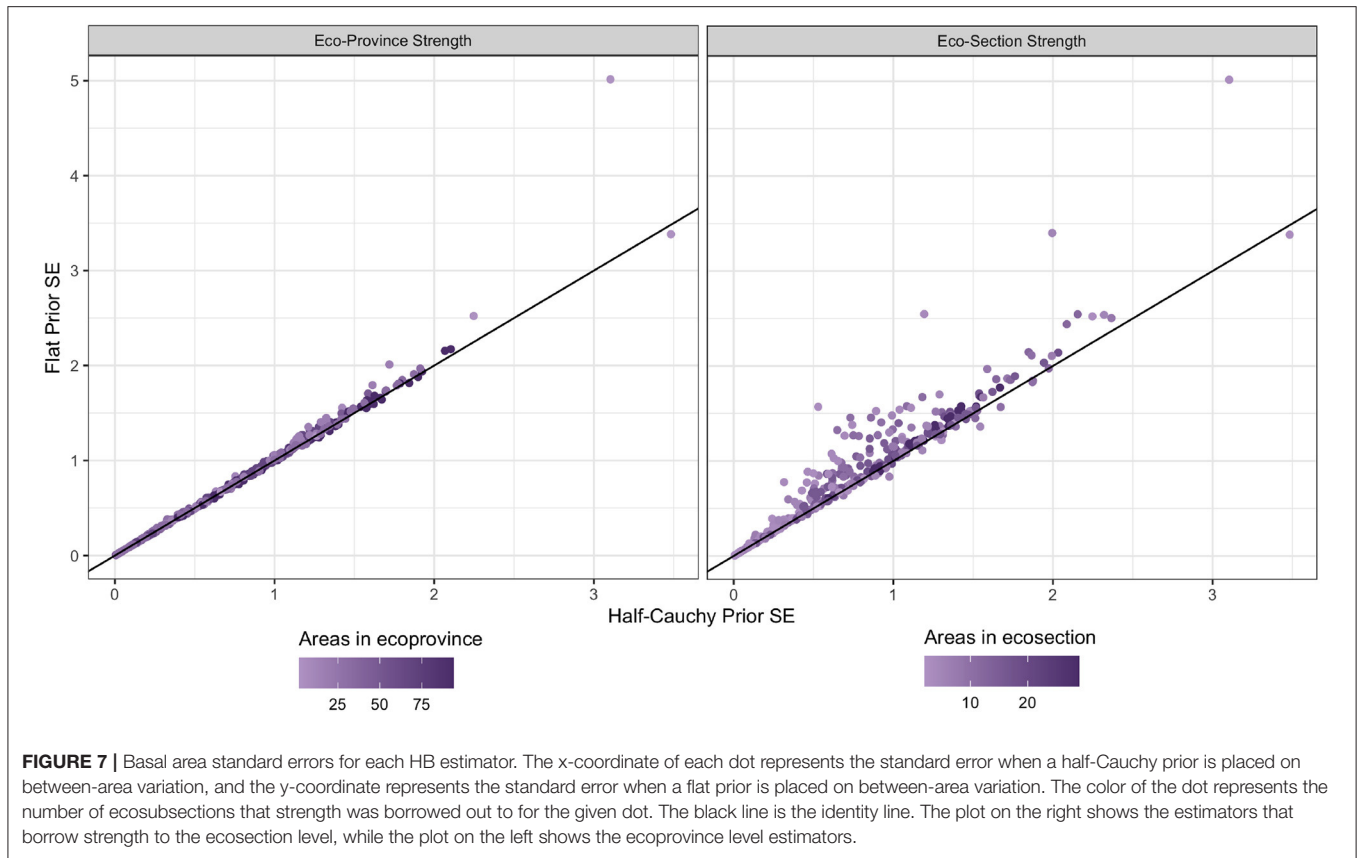


FIGURE 7 | Basal area standard errors for each HB estimator. The x-coordinate of each dot represents the standard error when a half-Cauchy prior is placed on between-area variation, and the y-coordinate represents the standard error when a flat prior is placed on between-area variation. The color of the dot represents the number of ecosubsections that strength was borrowed out to for the given dot. The black line is the identity line. The plot on the right shows the estimators that borrow strength to the ecosection level, while the plot on the left shows the ecoprovince level estimators.

TABLE 1 | Percent reduction in variance of basal area estimates from flat prior to half-cauchy.

Strength	Metric	Percent reduction
Ecosection	Mean	14.580
	Median	7.313
Ecoprovince	Mean	3.384
	Median	2.792

3.2. Case Study: The South Central Highlands (M331G) and the Utah High Plateau (M341C)

We now explore the effects of adding prior information to the HB estimators at a micro level: by examining two ecosections. The South Central Highlands and the Utah High Plateau ecosections both exist in mountainous ecoprovinces in the IW. **Figure 9** displays both ecosections. These two ecosections are located relatively close to each other in the IW, yet the addition of the half-Cauchy prior when we borrow strength to the ecosection level has a very different effect within each ecosection. To understand how the estimators perform differently across these two ecosections, we explore the mean estimates for basal area, and corresponding standard errors, within these two ecosections.

Figure 10 displays the HB estimates and illustrates that, for both ecosections, the basal area estimate is about the same for

both priors on the between-area variation parameter. This again showcases a robustness to how the prior information is specified for between-area variation.

While the estimate values show high agreement, it should be noted that the standard error estimates changed more drastically when we changed the prior on between-area variation, as seen in **Figure 11**. Interestingly, the standard errors in ecosection M331G are reduced significantly when the half-Cauchy prior is used compared to the flat prior, while the standard errors in ecosection M341C hardly change. This is likely due to a couple of factors. First of all, the estimated variance of the PS estimates in ecosection M331G is lower than the estimated variance of the PS estimates in M341C (32.044 and 44.384, respectively). That is, based on the data, there is less between-area variation in ecosection M331G. By placing a prior which has high probability density for small values of σ_v^2 we have reinforced the pattern seen in the data. Additionally, M331G is borrowing from less small areas and therefore will lean more on the weakly informative prior which preferences smaller values for the between-area variation.

4. DISCUSSION

We consider six indirect, area level small area estimators and one direct estimator across the IW region of the United States. The two HB estimators with flat priors on between-area variation

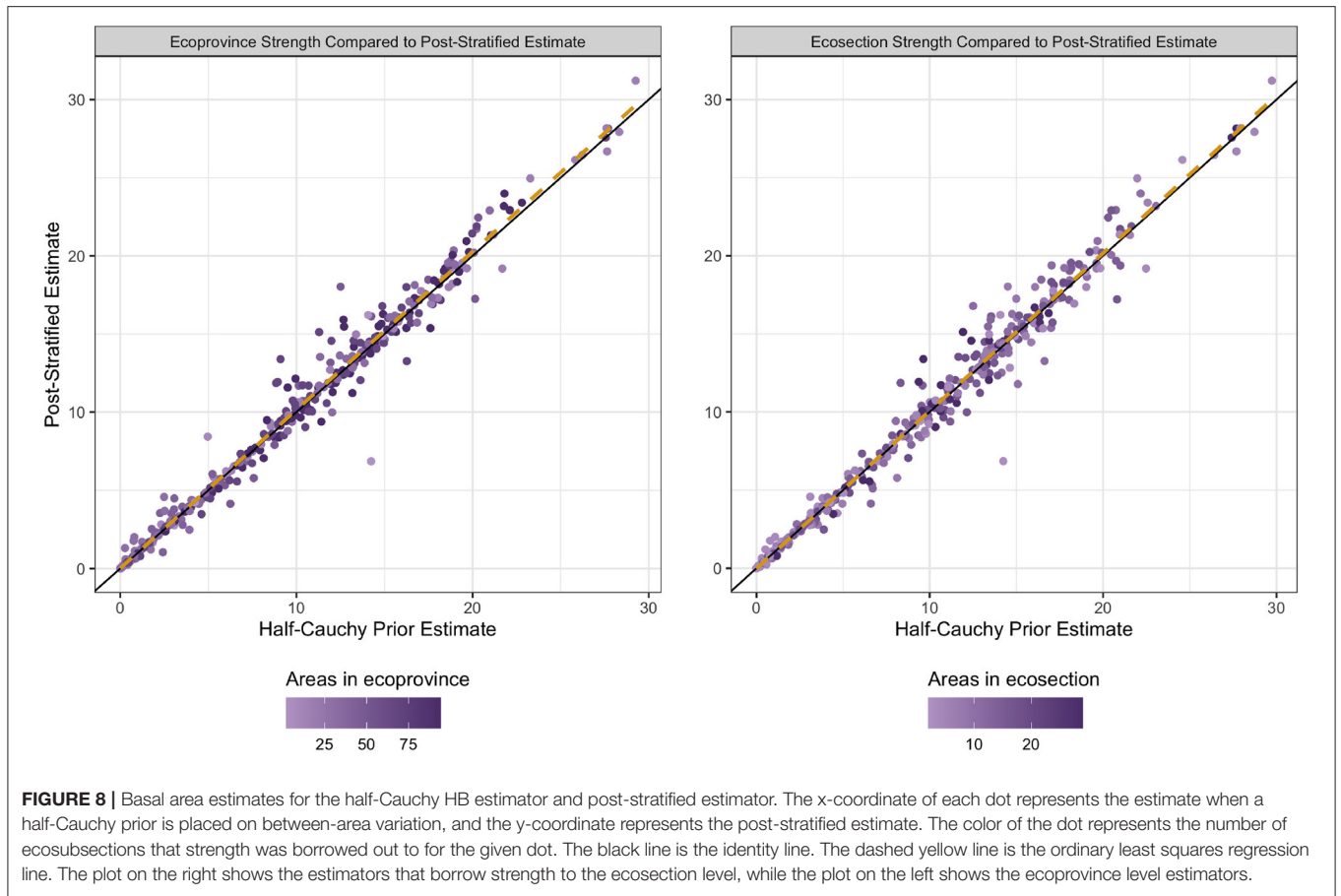


FIGURE 8 | Basal area estimates for the half-Cauchy HB estimator and post-stratified estimator. The x-coordinate of each dot represents the estimate when a half-Cauchy prior is placed on between-area variation, and the y-coordinate represents the post-stratified estimate. The color of the dot represents the number of ecosubsections that strength was borrowed out to for the given dot. The black line is the identity line. The dashed yellow line is the ordinary least squares regression line. The plot on the right shows the estimators that borrow strength to the ecosession level, while the plot on the left shows the ecoprovince level estimators.

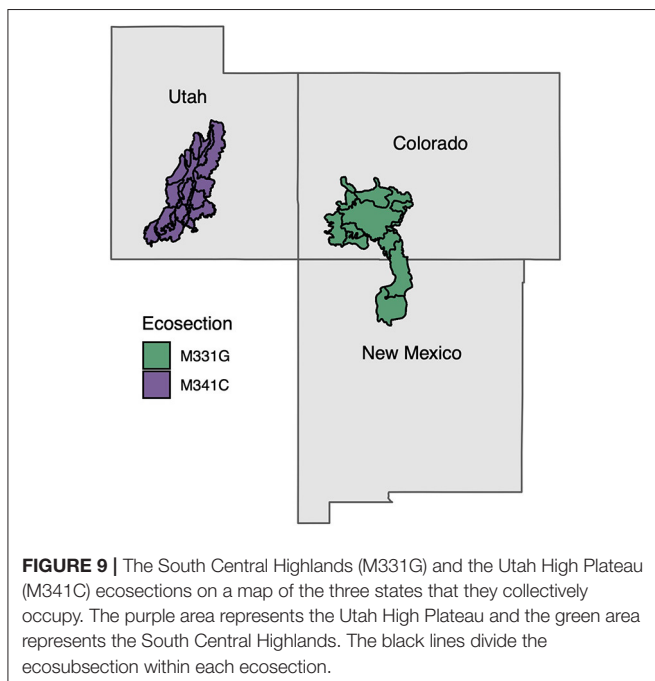
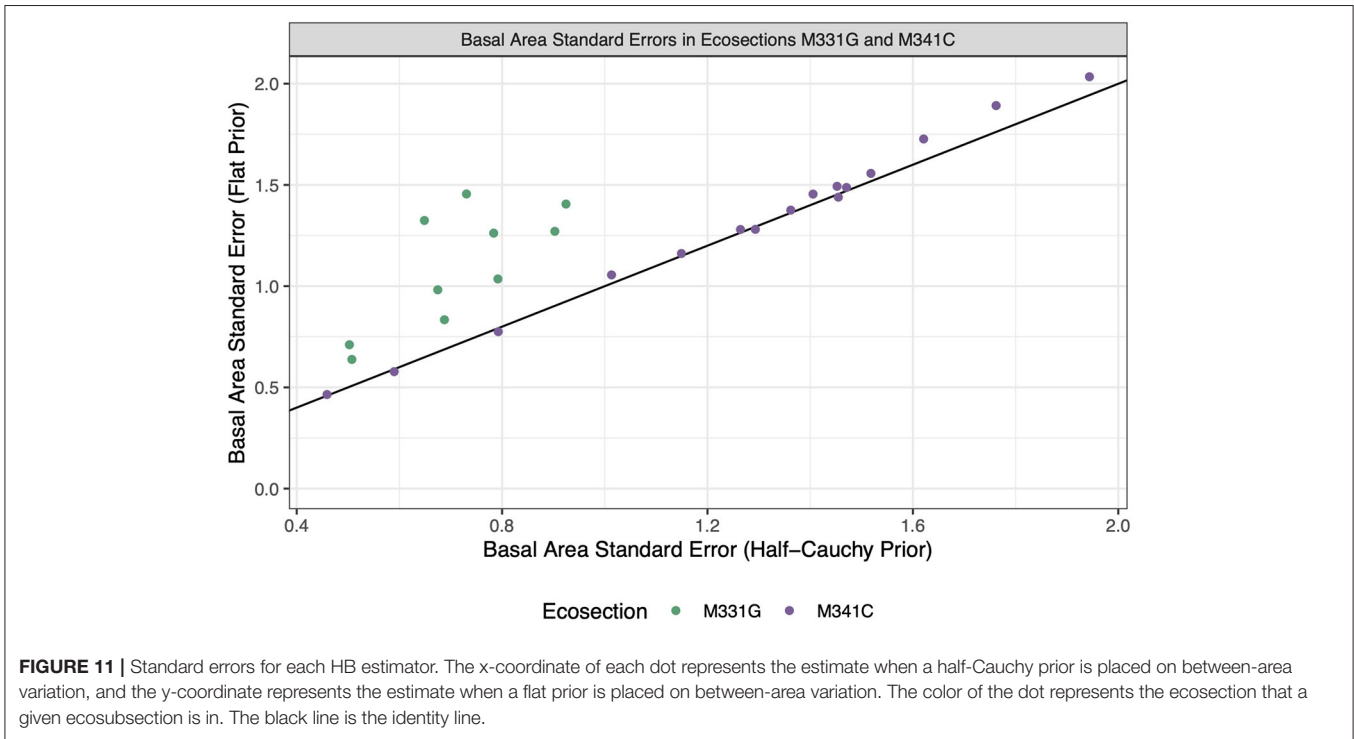
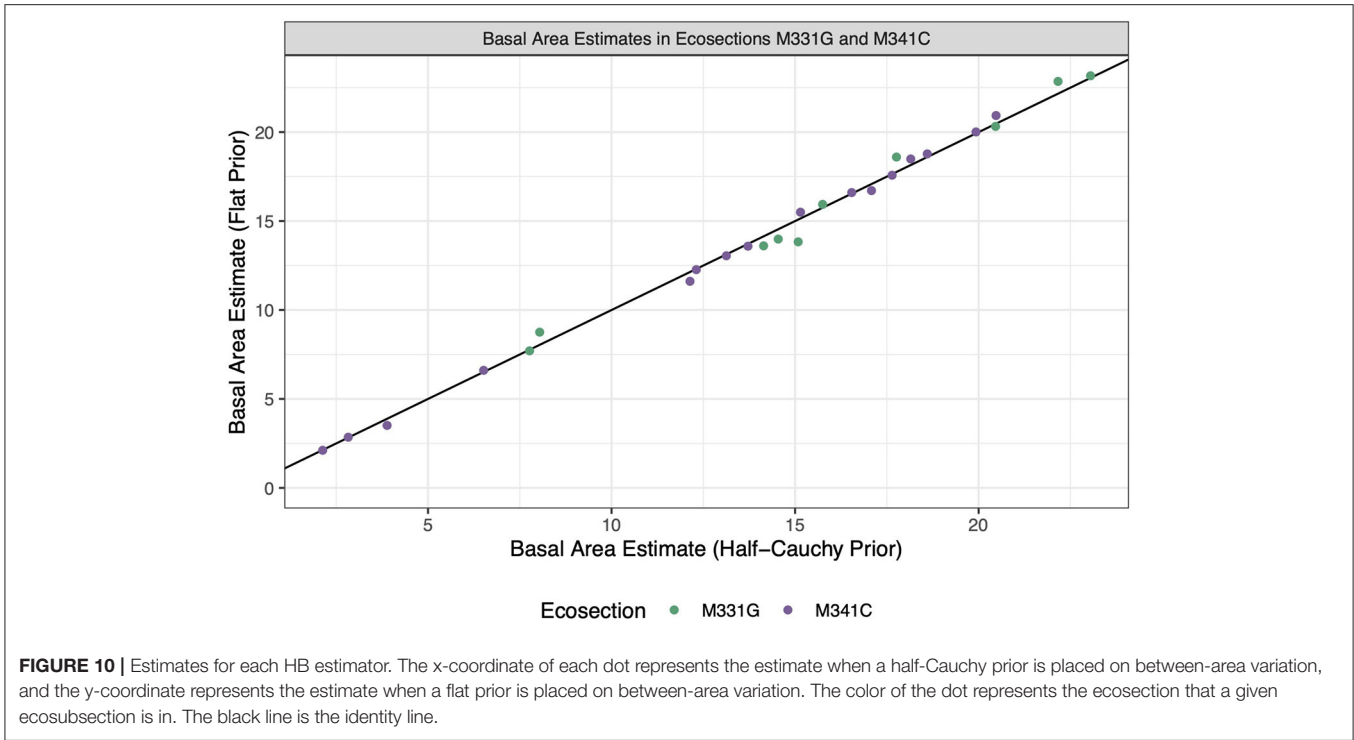


FIGURE 9 | The South Central Highlands (M331G) and the Utah High Plateau (M341C) ecosessions on a map of the three states that they collectively occupy. The purple area represents the Utah High Plateau and the green area represents the South Central Highlands. The black lines divide the ecosession within each ecosession.

mimic the two analogous area level EBLUP estimators in both estimates and variances. When supplying the HB estimators with a half-Cauchy prior for the between-area variation parameter, we see a reduction in variance when borrowing strength out to both the ecoprovince and ecosession level, with more reduction observed at the latter level of strength.

Table 2 displays the relative efficiency of each estimator implemented in this article compared the standard HT direct estimator for the basal area response variable. We define relative efficiency of a given estimator as the variance estimator of that estimator divided by the variance estimator of a direct estimator. The first column of **Table 2** makes it clear that incorporating informative auxiliary data into a direct estimator, the PS estimator in this case, does improve its efficiency. These improvements mimic FIA's production process with just 2 post strata assigned at the plot level, not at the subplot level. However, greater gains can be had by moving to an indirect estimator. In particular, the HB estimator with a half-Cauchy prior on between-area variation borrowing strength to the ecosession level has the highest mean and median relative efficiency. Notably, the ecosession-level, half-Cauchy prior, HB estimators relative efficiency is greater than the ecoprovince-level, half-Cauchy prior, HB estimators relative efficiency. This gain in relative



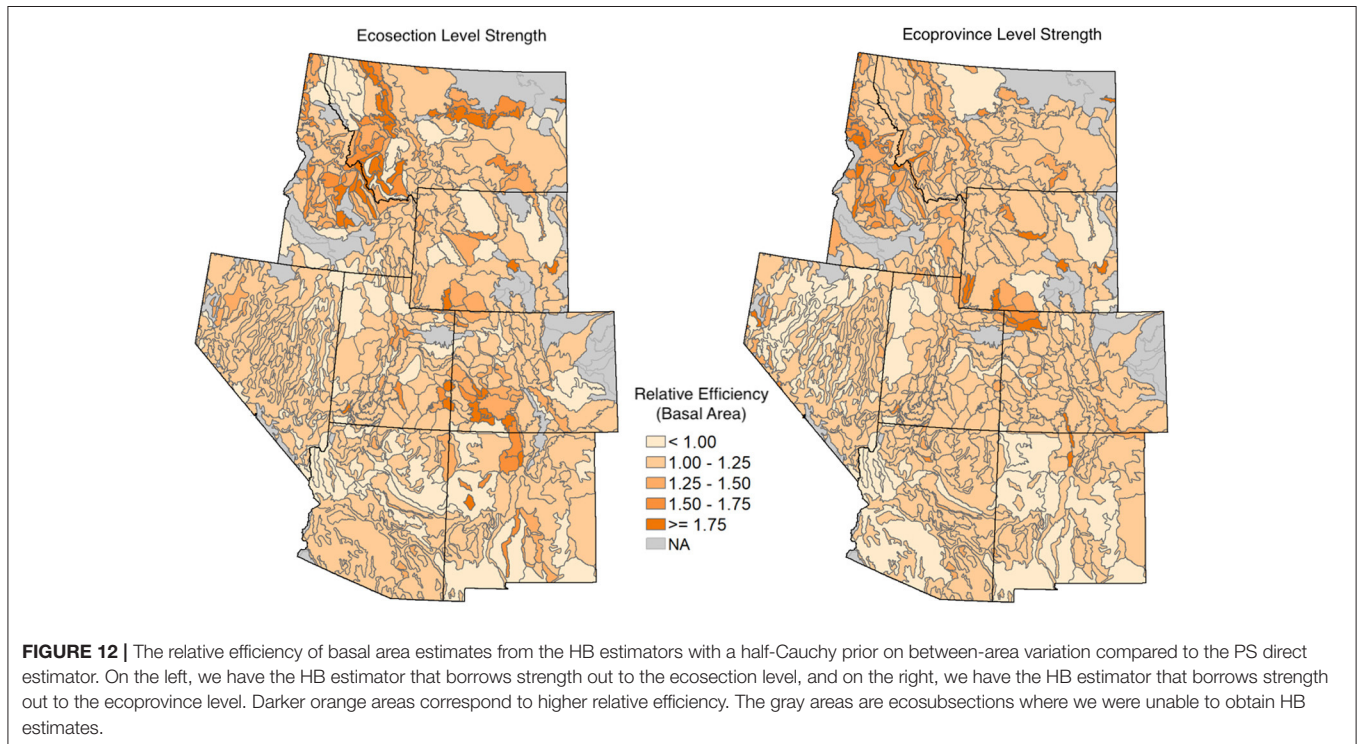
efficiency is likely due to the half-Cauchy prior being a reasonable depiction of the between-area variation of ecosubsections within a given ecosubsection in the IW.

Figure 12 shows the relative efficiency of basal area estimates for both the ecosubsection- and ecoprovince-level half-Cauchy

prior HB estimators compared to the PS direct estimator (as opposed to the Horvitz Thompson in **Table 2**). We can see more drastic improvements in relative efficiency in the more forested Northern parts of the IW, and the relative efficiency is sometimes below the PS estimator in extremely unforested areas.

TABLE 2 | Relative efficiency of each estimator compared to the Horvitz-Thompson (basal area response).

Metric	Post-strat	HB cauchy: ecosection	HB cauchy: ecoprovince	HB flat: ecosection	HB flat: ecoprovince	EBLUP: ecosection	EBLUP: ecoprovince
Mean	1.40	1.87	1.86	1.63	1.80	1.38	1.78
10% quantile	2.00	5.45	3.64	2.65	3.24	3.41	3.03
Median	1.31	1.80	1.74	1.55	1.70	1.59	1.67
90% quantile	1.01	1.10	1.17	1.07	1.15	0.99	1.13



This might be due to artificially low variance estimates for the PS estimator, which can occur when almost all sampled units in an ecosubsection have values of 0 for the response variables. In the case of an estimator that borrows strength, such as the HB estimators, we will likely borrow strength to some areas that have larger direct estimates of response variables, giving us a larger variance.

The efficiency gains of the HB estimators with informative priors on between-area variation over the more common EBLUP and PS (see **Table 2** and **Figure 12**) imply that these estimators can attain the same level of precision but with less sampled plots. However, the benefits of a HB approach do not stop there. Conveniently, the Bayesian paradigm allows for more intuitive inferential statements than provided by frequentist methods. Since the Bayesian methods provide a distribution for our parameter of interest, we can make probabilistic statements about the location of the parameter, whereas the frequentist approach only allows us to talk about the behavior of our method under repeated sampling.

Considering the performance and all the characteristics of these estimators, the results of this work provide some guidance on when to consider which of these estimators. If one only has

a small number of areas that they borrow strength out to, and those areas are believed to have a good amount of homogeneity between them, a HB estimator with a half-Cauchy prior on the between-area variation might be preferred. On the other hand, if one has the ability to borrow strength to a large number of groups that may not be too homogeneous, keeping a flat prior on between-area variation should be considered. This suggests that the HB estimator that borrows to the ecosection level and uses the half-Cauchy prior on between-area variation may be a viable estimator for FIA applications. Further testing with alternative responses and auxiliary data in other parts of the country is warranted.

Further work will include investigations of the unit level HB estimator, particularly with an eye to handling non-Gaussian data. Researchers have explored unit level modeling of non-Gaussian data types, such as zero-inflated data (Krieg et al., 2016) and other non-Gaussian data (Parker et al., 2020a). In particular, Parker et al. (2020b) discusses the benefits of unit level models, both in terms of potential efficiency gains and incorporating various levels of spatial aggregations. We hope to investigate the utility of these unit level models in a forest inventory setting. At both the area and unit level we will also explore extensions to

the HB estimators through spatially structured variance models. Ver Planck et al. (2018) explores area level HB estimators with conditional autoregressive random effects and conditional autoregressive random effects with smoothed sampling variance and found that these spatially structured variance models can help reduce the variance of the estimator. We hope to explore these spatially structured variance models further and investigate how they perform with different prior information supplied.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: the data include confidential plot data, which can not 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.fia.fs.fed.us/tools-data/spatial/index.php>).

REFERENCES

- Bechtold, W. A., and Patterson, P. L. (2005). *The Enhanced Forest Inventory and Analysis Program—National Sampling Design and Estimation Procedures*. USDA Forest Service, Southern Research Station.
- Blackard, J., Finco, M., Helmer, E., Holden, G., Hoppus, M., Jacobs, D., et al. (2008). Mapping US forest biomass using nationwide forest inventory data and moderate resolution information. *Remote Sens. Environ.* 112, 1658–1677. doi: 10.1016/j.rse.2007.08.021
- Boonstra, H. J. (2021). *mcmcsc: Markov Chain Monte Carlo Small Area Estimation*. Available online at: <https://CRAN.R-project.org/package=mcmcsc>
- Breidenbach, J., and Astrup, R. (2012). Small area estimation of forest attributes in the Norwegian National Forest Inventory. *Eur. J. Forest Res.* 131, 1255–1267. doi: 10.1007/s10342-012-0596-7
- Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika* 63, 615–620. doi: 10.1093/biomet/63.3.615
- Cleland, D. T., Freeouf, J. A., Keys, J. E., Nowacki, G. J., Carpenter, C. A., and McNab, W. H. (2007). *Ecological Subregions: Sections and Subsections for the Conterminous United States*. Available online at: <https://www.fs.fed.us/research/publications/misc/73326-wo-gtr-76d-cleland2007.pdf>
- Coulston, J. W., Green, P. C., Radtke, P. J., Prisley, S. P., Brooks, E. B., Thomas, V. A., et al. (2021). Enhancing the precision of broad-scale forestland removals estimates with small area estimation techniques. *Forest. Int. J. Forest Res.* 94, 427–441. doi: 10.1093/forestry/cpaa045
- Fay, R. E. III, 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, 269–277. doi: 10.1080/01621459.1979.10482505
- Frescino, T. S., Patterson, P. L., Moisen, G. G., and Freeman, E. A. (2015). FIESTA—An R estimation tool for FIA analysts,” in *Forest Inventory and Analysis (FIA) Symposium 2015* (Portland, OR: US Department of Agriculture, Forest Service, Pacific Northwest Research Station), 72.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.* 1, 515–534. doi: 10.1214/06-BA117A
- Goerndt, M. E., Monleon, V. J., and Temesgen, H. (2011). A comparison of small-area estimation techniques to estimate selected stand attributes using LiDAR-derived auxiliary variables. *Can. J. Forest Res.* 41, 1189–1201. doi: 10.1139/x11-033
- Hidioglou, M., and You, Y. (2016). Comparison of unit level and area level small area estimators. *Survey Methodol.* 42, 41–61.

AUTHOR CONTRIBUTIONS

GW, KM, GM, and TF: conceptualization. GW and KM: methodology and writing. GW: analysis. TF: data curation. GW and TF: data visualization. GW, KM, and GM: review and editing. All authors have read and agreed to the published version of the manuscript.

FUNDING

This work was supported by the USDA Forest Service, Forest Inventory and Analysis Program (via agreement 19-JV-11221638-112) and by Reed College.

ACKNOWLEDGMENTS

The authors would like to thank the USDA Forest Service, Forest Inventory and Analysis Program for the data.

- Krieg, S., Boonstra, H., and Smeets, M. (2016). Small-area estimation with zero-inflated data – a simulation study. *J. Off. Stat.* 32:963–986. doi: 10.1515/jos-2016-0051
- Magnussen, S., Mauro, F., Breidenbach, J., Lanz, A., and Kändler, G. (2017). Area-level analysis of forest inventory variables. *Eur. J. Forest Res.* 136, 839–855. doi: 10.1007/s10342-017-1074-z
- Mauro, F., Monleon, V. J., Temesgen, H., and Ford, K. R. (2017). Analysis of area level and unit level models for small area estimation in forest inventories assisted with LiDAR auxiliary information. *PLoS ONE* 12:e0189401. doi: 10.1371/journal.pone.0189401
- McConville, K., Tang, B., Zhu, G., Cheung, S., and Li, S. (2018). *MASE: Model-Assisted Survey Estimation*. Available online at: <https://cran.r-project.org/package=mase>
- McNab, W. H., Cleland, D. T., Freeouf, J. A., Keys, J. E., Nowacki, G. J., and Carpenter, C. A. (2007). *Description of “Ecological Subregions: Sections of the Conterminous United States*. United States Department of Agriculture. doi: 10.2737/wo-gtr-76b
- Molina, I., and Marhuenda, Y. (2015). sae: an R package for small area estimation. *R J.* 7, 81–98. doi: 10.32614/RJ-2015-007
- Molina, I., Nandram, B., and Rao, J. (2014). Small area estimation of general parameters with application to poverty indicators: a hierarchical bayes approach. *Ann. Appl. Stat.* 8, 852–885. doi: 10.1214/13-AOAS702
- Parker, P. A., Holan, S. H., and Janicki, R. (2020a). Computationally efficient Bayesian unit-level models for non-Gaussian data under informative sampling. *arXiv [Preprint]*. arXiv:2009.05642
- Parker, P. A., Janicki, R., and Holan, S. H. (2020b). Unit level modeling of survey data for small area estimation under informative sampling: a comprehensive overview with extensions. *arXiv [Preprint]*. arXiv:1908.10488.
- R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online at: <https://www.R-project.org/>
- Rao, J. N., and Molina, I. (2015). *Small Area Estimation*. Hoboken, New Jersey: John Wiley & Sons.
- Rintoul, M. A., Maebius, S., Alvarado, E., Lloyd-Damjanovic, A., Toyohara, M., McConville, K. S., et al. (2020). “An alternative post-stratification scheme to decrease variance of forest attribute estimates in the interior west,” in *Celebrating Progress, Possibilities, And Partnerships: Proceedings of the 2019 Forest Inventory and Analysis (FIA) Science Stakeholder Meeting* (Asheville, NC: U.S. Department of Agriculture Forest Service, Southern Research Station), 268–276. Available online at: <https://www.fs.usda.gov/treesearch/pubs/63184>
- Särndal, C. E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York, NY: Springer-Verlag.

- Vaish, A. K., Chen, S., Sathe, N. S., Folsom, R. E., Chandhok, P., and Guo, K. (2010). Small area estimates of daily person-miles of travel: 2001 National Household Transportation Survey. *Transportation* 37, 825–848. doi: 10.1007/s11116-010-9279-8
- Ver Planck, N. R., Finley, A. O., Kershaw, J. A. Jr., Weiskittel, A. R., and Kress, M. C. (2018). Hierarchical bayesian models for small area estimation of forest variables using LiDAR. *Remote Sens. Environ.* 204, 287–295. doi: 10.1016/j.rse.2017.10.024
- Wang, J. C., Holan, S. H., Nandram, B., Barboza, W., Toto, C., and Anderson, E. (2012). A Bayesian approach to estimating agricultural yield based on multiple repeated surveys. *J. Agric. Biol. Environ. Stat.* 17, 84–106. doi: 10.1007/s13253-011-0067-5
- 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. Photogramm. Remote Sens.* 146, 108–123. doi: 10.1016/j.isprsjprs.2018.09.006
- You, Y., Rao, J. N., and Gambino, J. (2003). Model-based unemployment rate estimation for the Canadian labour force survey: a hierarchical Bayes approach. *Survey Methodol.* 29, 25–32.

Conflict of Interest: GW is 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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 White, McConville, Moisen and Frescino. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

A. APPENDIX

A.1. Area Level EBLUP Variance Estimator

The variance of the Area level EBLUP is expressed by the following equation (Molina and Marhuenda, 2015; Rao and Molina, 2015):

$$\hat{V}(\hat{\mu}_j^{EBLUP}) = g_{1j} + g_{2j} + 2g_{3j} - b \quad (\text{A1})$$

where

$$g_{1j} = \hat{\gamma}_j \hat{V}(\hat{\mu}_{y_j}^{PS}),$$

$$g_{2j} = \hat{\sigma}_v^2 (1 - \hat{\gamma}_j)^2 \mathbf{z}_j' \left(\sum_j \hat{\gamma}_j \mathbf{z}_j \mathbf{z}_j' \right)^{-1} \mathbf{z}_j,$$

$$g_{3j} = 2m \left(\hat{V}(\hat{\mu}_{y_j}^{PS}) \right)^2 \left(\hat{\sigma}_v^2 + \hat{V}(\hat{\mu}_{y_j}^{PS}) \right)^{-3} \left(\sum_j \left(\hat{\sigma}_v^2 + \hat{V}(\hat{\mu}_{y_j}^{PS}) \right)^{-1} \right)^{-2},$$

$$b = 2m \hat{\sigma}_v^2 \left(\sum_j (\hat{\gamma}_j)^2 - \left(\sum_j \hat{\gamma}_j \right)^2 \right) \left(\sum_j \hat{\gamma}_j \right)^{-3} (1 - \hat{\gamma}_j)$$

where

$$\mathbf{z}_j = \begin{bmatrix} 1 \\ \bar{X}_j \end{bmatrix}.$$

One can intuitively think about each $g_{\#j}$ as follows: g_{1j} accounts for within-area variation, g_{2j} accounts for variation in estimating the regression parameter β , and g_{3j} accounts for model-variance estimation (Hidiroglou and You, 2016).