The Annals of Applied Statistics
2017, Vol. 11, No. 1, 340–361
DOI: 10.1214/16-AOAS1007

BIOMASS PREDICTION USING A DENSITY-DEPENDENT DIAMETER DISTRIBUTION MODEL[1]

BY ERIN M. SCHLIEP*, ALAN E. GELFAND†, JAMES S. CLARK† AND BRADLEY J. TOMASEK†

*University of Missouri * and Duke University†*

Prediction of aboveground biomass, particularly at large spatial scales, is necessary for estimating global-scale carbon sequestration. Since biomass can be measured only by sacrificing trees, total biomass on plots is never observed. Rather, allometric equations are used to convert individual tree diameter to individual biomass, perhaps with noise. The values for all trees on a plot are then summed to obtain a *derived* total biomass for the plot. Then, with derived total biomasses for a collection of plots, regression models, using appropriate environmental covariates, are employed to attempt explanation and prediction. Not surprisingly, when out-of-sample validation is examined, such a model will predict total biomass well for holdout data because it is obtained using exactly the same derived approach.

Apart from the somewhat circular nature of the regression approach, it also fails to employ the actual observed plot level response data. At each plot, we observe a *random* number of trees, each with an associated diameter, producing a sample of diameters. A model based on this random number of tree diameters provides understanding of how environmental regressors explain abundance of individuals, which in turn explains individual diameters.

We incorporate density dependence because the distribution of tree diameters over a plot of fixed size depends upon the number of trees on the plot. After fitting this model, we can obtain predictive distributions for individual-level biomass and plot-level total biomass. We show that predictive distributions for plot-level biomass obtained from a density-dependent model for diameters will be much different from predictive distributions using the regression approach. Moreover, they can be more informative for capturing uncertainty than those obtained from modeling derived plot-level biomass directly.

We develop a density-dependent diameter distribution model and illustrate with data from the national Forest Inventory and Analysis (FIA) database. We also describe how to scale predictions to larger spatial regions. Our predictions agree (in magnitude) with available wisdom on mean and variation in biomass at the hectare scale.

Received April 2016; revised September 2016.

[1]Supported in part by U.S. National Science Foundation Grants NSF-CDI-0940671 and NSF-EF-1137364 as well as U.S. Census Bureau Grant NSF-SES-1132031, funded through the NSF-Census Research Network (NCRN) program.

Key words and phrases. Allometry, big O behavior, hierarchical models, Markov chain Monte Carlo, Poisson process.

1. Introduction. Forests influence the flow of carbon between land and atmosphere, which identifies forests as a carbon sink. This process in aboveground live biomass in forests is defined by the net storage of carbon within trees after accounting for losses from tree mortality and respiration [Pan et al. (2011)]. A wide range of estimates is reported for carbon sinks in forests, making assessment of associated uncertainty critical. For instance, the estimated annual carbon sink for eastern North America ranges from 0.21 to 0.25 petagrams of carbon per year (Pg C/yr) [Pan et al. (2011)]. Thurner et al. (2014) estimated total carbon in boreal and temperate forests in North America at 28.7 Pg C, with upper and lower 95% confidence limits at $\pm$ 10.8 Pg C. Biomass per unit area for areal units is never directly measured because it requires sacrificing trees. Hence, assigning statistical model-based uncertainty to these estimates of biomass offers a novel challenge. This is further complicated by the fact that uncertainty in the estimates of biomass does not scale with area.

Predicting total aboveground biomass in response to climate at the plot level or, of greater interest, at the hectare scale, is difficult because biomass is never measured at plot level or at larger spatial scales. Direct observations of biomass per unit area would require sacrificing and measuring every tree on a sample plot. Instead, allometric equations are applied to the diameter of each tree on the plot to obtain derived *noiseless* (or *noisy* if error is added to the allometry) biomass per tree. These per-tree values are summed to obtain derived noiseless (or noisy) total biomass for the plot. Derived biomass per plot or per area, which are not actual data, are treated as response variables in regression models to explain and predict geographic variation. Differences in derived biomass over successive inventories are used to estimate aboveground net primary production (ANPP). These regression models can appear attractive when out-of-sample validation is examined. They will predict total biomass well for holdout data because the holdout data is obtained using exactly the same derived approach. Apart from the somewhat circular nature of the regression approach, it also fails to employ the actual observed plot-level response data. At each plot, we observe a *random* number of trees, each with an associated diameter, producing a sample of diameters.

Remote sensing techniques also provide estimates of total biomass. They employ broad spatial coverage but require ground-based estimates for scale [Baccini et al. (2012), Dong et al. (2003), Saatchi et al. (2011)]. For example, Saatchi et al. (2011) develop a power-law functional relationship between biomass and forest height to predict biomass using satellite imagery (e.g., Lidar) of forest structure. Malhi et al. (2006), Blackard et al. (2008) and Wilson, Woodall and Griffith (2013) propose using various imputation methods (e.g., k-nearest neighbor weighted averages) for obtaining continuous coverage of biomass estimates between sampled inventory plots. However, the inventory plots used in these analyses are often small, and the implications of extrapolation and scaling of biomass estimates and uncertainty are unclear. McRoberts et al. (2015) investigate the uncertainty in large-area estimation of tree volume using a Monte Carlo simulation approach and show that

the uncertainty in individual tree volume predictions can have significant effects on large-area volume estimates of uncertainty. Schliep et al. (2016) develop a model for derived total biomass using plot-level covariate data. They model derived total biomass statically at two time points and predict the change in total biomass by differencing.

Our contribution is a process-driven model specification for the observed set of tree diameters at breast height (henceforth diameter) on a plot of a given abundance (the number of trees on the plot). When plots are all of equal size, abundance is equivalent to plot density. We observe a random number of trees and, for each, an associated diameter. Our process-driven model incorporates density dependence, acknowledging that, in addition to the environmental covariates for the plot, the number of individuals on the plot is needed to explain the distribution of diameters on the plot. From a model for diameters, we can obtain predictive distributions for individual-level biomass and plot-level total biomass post-model fitting.

To illustrate the importance of density dependence in modeling total biomass, we offer an illustration comparing our process-based approach with a regression model for total biomass. Predictors include plot-specific covariates such as stand age, mean annual temperature, soil moisture, deficit and possible interactions. For the process-based model, we obtain a posterior predictive distribution of total derived biomass given the observed data, that is, given the numbers of individuals on each plot and the associated set of diameters. For the regression model, it is unclear exactly what "data" we should use for model fitting, again demonstrating our discomfort with this approach and the difficulty of objective comparison via a data set. Should we use derived noiseless total biomass? If we add noise, then inference depends upon which set of realizations of derived biomass are used. Again, this seems to argue for a model to explain what is observed, the abundances and the associated sets of diameters. In any event, for the regression in the comparison below, we generated 100 noisy total biomass datasets and summarize the inference based upon their compilation.

Figure 1 shows predictive distributions of total biomass for two out-of-sample plots using the regression model and the density-dependent model proposed in this work. These two plots are examples of the scenario where similar total biomass (vertical lines at approximately 120 t/ha) results from different combinations of density and tree diameters. Plot 1 has high plot density with 750 trees relative to plot 2 with 507 trees. The predictive distributions from both models agree on the predictive median of total biomass, but disagree on the uncertainty because the regression model has no information on plot density. The predictive distribution from the model developed here has low uncertainty at high plot density. In addition to improved inference associated with the process model, our approach benefits from information on both tree density and diameters. Explicit comparison of the predictive distributions of biomass using the regression model and the density-dependent model proposed in this work is given in Section 4.3.

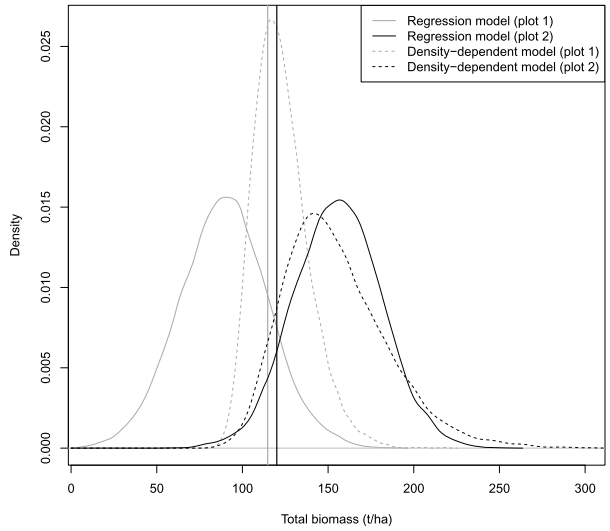

FIG. 1. *Predictive distributions for derived total biomass for two out-of-sample plots using a regression model* (*solid curves*) *and the density-dependent model* (*dashed curves*) *developed here. Plot* 1 (*grey*) *has* 750 *trees*, *whereas plot* 2 (*black*) *has* 507 *trees. The two plots have similar median total biomass*, *denoted by the vertical lines*, *despite having different plot densities. The uncertainty in the predictive distributions from the regression model is the same for the two plots*, *but differs for our model.*

Unlike the traditional approach, the process-driven model for the density-dependent sample of diameters is generative. At the first stage we specify a Poisson distribution for the number of trees on the plot. At the second stage we assume conditionally independent diameters where the distribution for the diameters depends upon the number of trees observed on the plot. More precisely, the observable data will be of the form $\{N_i, \boldsymbol{\mathcal{X}}_i, \mathbf{W}_i, i = 1, \ldots, I\}$, where N_i is the number of stems observed on plot i, $\boldsymbol{\mathcal{X}}_i = \{X_{i,1}, \ldots, X_{i,N_i}\}$ is the observed set of diameters, and $\mathbf{W}_i$ are plot-level covariates. Because each sample plot in our analysis has the same area, we refer to N_i as the density of plot i, which is abundance per plot area. We model N_i given $\mathbf{W}_i$ as Poisson distributed. Then we model the N_i diameters to be conditionally independent (see below in this regard). Using bracket notation for distributions, when we condition on N_i, this takes the form $[X_{i,j}|N_i, \mathbf{W}_i]$. Adjusting for possible left truncation, we model $X_{i,j}|N_i, \mathbf{W}_i$ as Gamma distributed. The modeling is presented in full detail in Section 3.

Typically, estimates of biomass are needed for larger areas than represented by a sample plot. For example, a Forest Inventory and Analysis (FIA) plot in the eastern United States is too small to be of general interest due to the sporadic occurrence of large trees in such small plots. Forest biomass is typically reported at the more relevant scale of metric tons per hectare (t/ha), despite the fact that there are not even derived biomasses at the hectare scale. We indicate how our

modeling approach can be used to scale biomass prediction from smaller plots to the hectare scale. Large variance in biomass is expected from small plots due to the considerable variability in the number of individuals on these plots and the resultant implications of density dependence on diameter distribution. However, such uncertainty would not apply to larger areas.

In Section 2 we summarize the national Forest Inventory and Analysis (FIA) program data used in these analyses along with plot-level covariates. In Section 3 we outline the density-dependent distribution model for diameters. We specify the model, provide model inference on diameters, and formally investigate the behavior of total biomass at the plot level as a function of observed plot density. We define the first-stage Poisson distribution for the number of trees observed in Section 4 and investigate the behavior of plot biomass conditionally and marginally with respect to N. We also formalize the regression approach to biomass. Then, upon marginalizing our model, we can explicitly show the difference between predictive distributions for total biomass. In Section 5 we apply the density-dependent model to predict biomass at the hectare scale. We conclude in Section 6 with a summary and directions for future work.

2. FIA and climate data. We analyze data obtained from FIA plots in the eastern United States, consisting of 23,028 plots and roughly 1 million trees. Due to various management and harvesting efforts across the region, only plots that were coded as "natural stands" in the FIA database were used in the analysis. Natural stands are defined as those which have no clear evidence of artificial regeneration. Between 1997 and 2011, each plot in the region is surveyed twice. In this analysis, we used the second survey. Stand age is included in the FIA data and is used in the model as a plot-level variable.

FIA applies a nationally consistent sampling protocol using a quasi-systematic design covering all ownerships across the United States, resulting in a national sampling intensity of one plot per 2428 hectares [Bechtold and Patterson (2005)]. An FIA plot consists of four circular subplots each having a radius of 7.32 meters (m). The distance between subplot centroids is 35.58 m. The aggregate area of the four subplots is 673.34 m^2 or $\approx$0.067 hectares (ha). Trees greater than 12.7 centimeters (cm) in diameter are included in the survey. Stems less than 12.7 cm are measured in subsets of subplots, but they contribute little to biomass per plot. Therefore, the analyses presented here contain only measurements of stems greater than 12.7 cm. We consider the observed plot density and the sample of diameters to include trees observed on all four subplots.

The covariates included in our model are stand age, mean annual temperature, soil moisture and hydrothermal deficit. Stand age affects the distribution of diameters and productivity [Camarero et al. (2015)]. The Moderate Resolution Imaging

Spectroradiometer (MODIS)[1] and Parameter-elevation Regressions and Independent Slopes Model (PRISM)[2] were used to derive climate variables that can have important impacts on forest growth [Brzostek et al. (2014)]. Specifically, we used 20-year climate normals for the years 1990–2010 for mean annual temperature. Hydrothermal deficit is a regional climate variable that summarizes the intensity and duration of stress caused by moisture. It is calculated as the number of degree hours accumulated for months with a negative water balance. While we recognize that there is uncertainty in these model-obtained covariates, there is also uncertainty in both the climate that a plot experiences and how it responds to climate, and this uncertainty is unknown. Since there is no way of determining whether they are relevant for plots, the climate variables provide a rough index and are assumed error-free. Soil moisture is obtained from the FIA soil quality database[3] and is based on land form, topographic position, and soil.

3. Density-dependent diameter distributions and their implications for biomass prediction. Following the Introduction, we propose a generative model in the form

$$[N_i|\mathbf{W}_i][\boldsymbol{\mathcal{X}}_i|N_i,\mathbf{W}_i], \tag{3.1}$$

where $[N_i|\mathbf{W}_i]$ is the conditional distribution of plot density, N_i, given covariate information, $\mathbf{W}_i$, and $[\boldsymbol{\mathcal{X}}_i|N_i,\mathbf{W}_i]$ is the distribution of diameters conditional on both plot density and covariates.

Suppressing i and the covariates, we make the assumption that diameters are conditionally independent given N. Thus, we only need to specify a *diameter* distribution of the form $[X_j|N,\boldsymbol{\theta}_x]$ after which the joint distribution for the diameters becomes $\Pi_{\{X_j\in\boldsymbol{\mathcal{X}}\}}[X_j|N,\boldsymbol{\theta}_x]$. With regard to the conditional independence assumption, while tree diameters may be spatially dependent, we only have a collection of diameters with no spatial referencing. We have no information upon which to attempt to model dependence between the observed diameters. It is the case that geo-coded locations of individuals are available for some FIA plots, but, since the plots are small and are comprised of four noncontiguous subplots, model-based dependence between individuals seems beyond our ability to assess. The remainder of this section is devoted to exploring the behavior of biomass under a density-dependent diameter distribution.

The general intuition associated with density dependence is that, with increasing N, we expect small trees and less variation in tree diameters, or, with N decreasing, larger trees are expected with more variability in tree diameters. Again,

[1]Wan, Z., S. Hook and G. Hulley. 2015. MOD11A1 MODIS/Terra Land Surface Temperature/Emissivity Daily L3 Global 1km SIN Grid V006. NASA EOSDIS Land Processes DAAC. http://doi.org/10.5067/MODIS/MOD11A1.006.

[2]PRISM Climate Group, Oregon State University, http://prism.oregonstate.edu, created 4 Feb 2004.

[3]http://www.fia.fs.fed.us/program-features/indicators/soils/index.php.

the data consist of a set of tree diameters for each plot over a collection of plots all of the same plot area. To avoid specifying an upper bound on the diameter, that is, for a model defined over $[0, \infty]$, a natural choice is the Gamma distribution. Ignoring covariates for the moment, consider a Gamma(α, β) distribution. Suppose that α is free of N but β depends on N, that is, $\beta(N)$. Then, if $\beta(N)$ is increasing in N, it is immediate under the Gamma distribution that we will achieve the desired behaviors for $[X|N]$. For example, suppose $\beta(N) = \beta N^{\delta}$ with $\delta > 0$ yielding $E_N(X) = \frac{\alpha}{\beta N^{\delta}}$. Then $E_N(X) = O(N^{-\delta})$. Similarly, we have $\mathrm{Var}_N(X) = O(N^{-2\delta})$.

3.1. *Assessing the Gamma model.* Here, we assess the performance of a Gamma distribution model for tree diameter. Because some data sets have a minimum sampling diameter, a left truncation may be appropriate. If so, then one could apply a translation $g : [L, \infty] \to [0, \infty]$ defined as

$$\widetilde{X} = g(X) = X - L, \tag{3.2}$$

where L is the lower threshold for diameter.

The sample of truncated diameters for plot i, $\widetilde{\boldsymbol{X}}_i$, given plot density and covariate information, is modeled using a Gamma regression model; that is, for diameter $\widetilde{X}_{i,j} \in \widetilde{\boldsymbol{X}}_i$,

$$\widetilde{X}_{i,j} | N_i, \mathbf{W}_i \sim \mathrm{Gamma}(\alpha, \beta_i). \tag{3.3}$$

Here, α is specified globally and β_i is a plot-specific parameter given a regression form. The regression model is specified as

$$\log(\beta_i) = \delta \log(N_i) + \nu_i + \mathbf{W}_i' \boldsymbol{\phi}, \tag{3.4}$$

where δ controls the amount of density dependence exerted by N_i, ν_i is a plot random effect, $\mathbf{W}_i$ is a set of plot-level covariates, and $\boldsymbol{\phi}$ is a vector of coefficients. Again, the diameter distribution is expected to be concentrated at small values when plot density is high and more diffuse across diameters when plot density is low. Alternative specifications are possible. However, the choice above captures expected behavior and fits the observed diameter data well.

Inference for the Gamma regression model is obtained in the Bayesian framework. To complete the Bayesian specification, we assign prior distributions to the remaining model parameters and obtain posterior inference using Markov chain Monte Carlo (MCMC). Conjugate, weak priors are assigned when possible. The shape parameter of the Gamma distribution, α, is assigned a vague Gamma prior with shape = 1 and rate = 1/2. The parameter δ, which controls the behavior of the distribution of individual diameters as a function of plot density, is assigned a noninformative normal prior distribution with mean 0 and variance 100^2 and $\boldsymbol{\phi} \sim \mathrm{MVN}(\mathbf{0}, 100^2 \mathbf{I}_p)$, where p is the number of covariates in $\mathbf{W}_i$ and $\mathbf{I}_p$ is a $p \times p$ identity matrix. The plot random effects are modeled as $\nu_i \overset{i.i.d.}{\sim} \mathrm{N}(\mu_\nu, \sigma_\nu^2)$

TABLE 1
Posterior mean estimates and 95% credible intervals for the density-dependent diameter model

Parameter	Mean	Credible interval
α	1.066	(1.063, 1.070)
δ	0.241	(0.235, 0.248)
μ_ν	−3.081	(−3.103, −3.058)
σ_ν^2	0.096	(0.093, 0.099)
ϕ_1 stand age	−0.318	(−0.323, −0.312)
ϕ_2 temperature	−0.150	(−0.157, −0.143)
ϕ_3 deficit	0.022	(0.017, 0.028)
ϕ_4 moisture	0.006	(0.001, 0.012)
ϕ_5 stand age^2	0.077	(0.074, 0.081)
ϕ_6 stand age × deficit	0.010	(0.004, 0.018)
ϕ_7 stand age × moisture	0.015	(0.009, 0.020)
ϕ_8 deficit × moisture	0.042	(0.035, 0.048)
ϕ_9 temperature × moisture	−0.092	(−0.097, −0.087)

for $i = 1, \ldots, I$. We assign μ_ν and σ_ν^2 conjugate, noninformative prior distributions, where $\mu_\nu \sim \mathrm{N}(0, 100^2)$ and $\sigma_\nu^2 \sim \mathrm{IG}(2, 2)$. This specification is a form of hierarchical centering [Gelfand, Sahu and Carlin (1995)] to enhance convergence of the sampling algorithm. The MCMC algorithm is a Metropolis-within-Gibbs algorithm where the parameters α, δ, ν_i and $\boldsymbol{\phi}$ are each updated using a Metropolis–Hastings step, and μ_ν and σ_ν^2 are sampled directly from their full conditional distribution.

Of the 23,028 plots in the FIA dataset, 90% were used for model fitting, and the remaining 10% were used for out-of-sample prediction. Stand age, mean annual temperature, soil moisture, and hydrothermal deficit were included as plot-level main effects in the model, each of which is centered and scaled. We also included a quadratic term for stand age to allow for nonlinear relationships. Interactions included in the model were stand age × moisture, stand age × deficit, moisture × deficit, and temperature × moisture. Because FIA omits trees with diameters less than 12.7 cm, diameters were truncated at this value.

We fitted the Gamma regression model to the FIA data given the N_is and $\mathbf{W}_i$s. The MCMC was run for 20,000 iterations. The first 10,000 iterations were discarded as burn-in. Due to the conditional independence of diameters on a plot and the linear structure of the model, issues of convergence for the MCMC are not anticipated. Standard diagnostics, such as the Gelman and Rubin R statistic, confirm this. Table 1 gives the posterior mean and 95% credible interval estimates for the Gamma model parameters. The posterior mean estimate of δ is 0.24, indicating that the distribution of diameters is more concentrated at small values when N is large. Negative regression coefficients indicate that diameter is increasing with

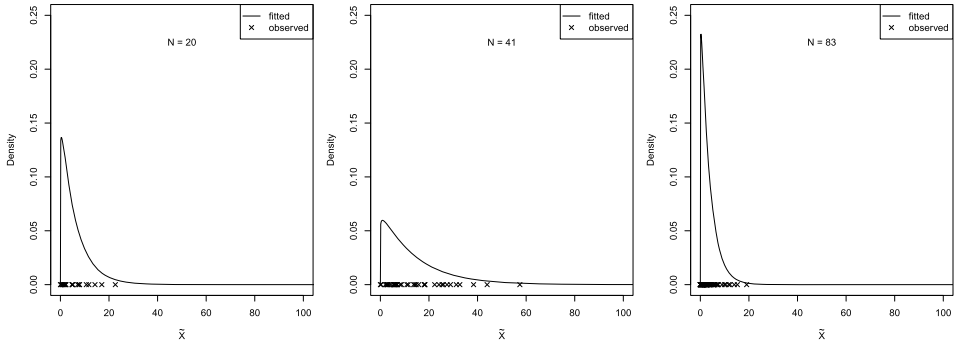

FIG. 2. *Posterior mean fitted Gamma densities for three random plots with observed diameters.*

increases in the covariate, whereas positive regression coefficients indicate that diameter is decreasing with increases in the covariate. Therefore, in general, diameter is increasing with stand age for young stands and decreasing with stand age for old stands. Additionally, the distribution of diameters is more concentrated at small diameters for stands with high temperature, low moisture, and low deficit. Figure 2 shows the posterior mean distribution of diameters with the observed samples for three randomly chosen plots representing low plot density, moderate plot density, and high plot density. The posterior mean distribution of diameters is obtained by averaging the Gamma density functions across 1000 draws from the posterior distribution of α and β_i for each plot. In each, the fitted distribution appears to capture the observed diameters. The diameter distribution for the high density plot with 83 stems is very concentrated at small values, whereas the plots with fewer stems have much broader distributions to match their observed empirical behavior.

Figure 3 shows the posterior mean predictive distribution of diameters with the observed diameter samples for three randomly selected out-of-sample plots, again representing low, moderate, and high plot density. These figures show that the

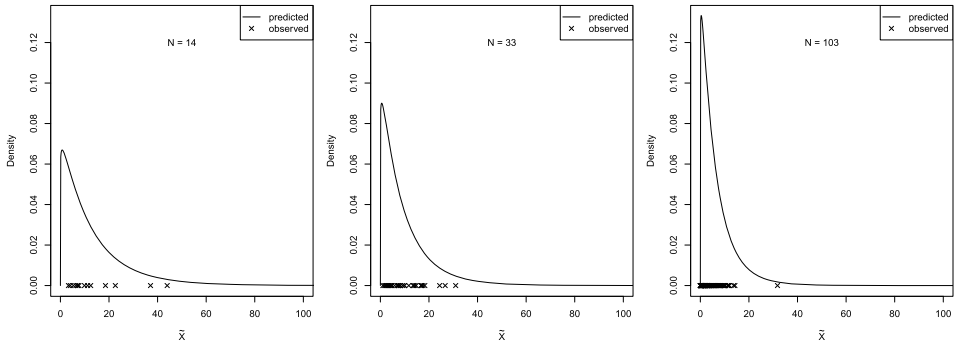

FIG. 3. *Posterior mean predicted Gamma densities for three random out-of-sample plots with observed diameters.*

model predicts out-of-sample reasonably well. In particular, the high density plot with $N = 103$ stems is highly concentrated around smaller values, matching the empirical set of diameters for the plot.

3.2. *The behavior of biomass.* To obtain a derived biomass for a tree, we use an allometric model, which transforms a diameter to biomass. The literature here is substantial [e.g., Henry et al. (2013), Jenkins et al. (2003), Lambert, Ung and Raulier (2005), Picard, Saint-André and Henry (2012), Whittaker and Woodwell (1968)]. We work with the widely used allometric form aX^b. More precisely, for each random diameter, we create the random variable $Y = aX^b$, where a and b are known allometric coefficients. For now, assume these coefficients are the same for all species. We can interpret Y as the noiseless derived random tree-level biomass associated with the random diameter, X. In practice, this will not be the actual biomass because there is error in the allometry. If we write the allometric model on the log scale, it is customary to introduce error additively (and independently), yielding $\log(Y) = \log(a) + b\log(X) + \epsilon$. Variances for ϵ have been extracted [Jenkins et al. (2003), Lambert, Ung and Raulier (2005), Molto, Rossi and Blanc (2013), Ung, Bernier and Guo (2008)] from the limited data collected on trees from allometric studies. The coefficient of variation, defined as $\frac{\sqrt{\text{var}}}{\text{mean}}$, is one such metric for reporting uncertainty and discussed with regard to total derived biomass in Chave et al. (2004) and Chave et al. (2014). Uncertainty at the tree level has been investigated in terms of measurement error, allometric model specification, the sampling protocol of stems in a plot, and the representativeness of small plots for a forest landscape [Chave et al. (2004, 2014)]. Therefore, on the original scale, we might specify $Y = aX^b e$ with lognormal error $e = \exp^{\epsilon}$. We will refer to such Ys as noisy tree-level biomasses.

Under the Gamma model specified above for diameter, we are interested in the implications for mean and variance of biomass as N grows large, that is, the order of $\mathrm{E}_N(Y)$ and $\mathrm{Var}_N(Y)$. First, consider the case with no ϵ such that $e = 1$ w.p. 1. Then $\mathrm{E}_N(Y) = a\mathrm{E}_N(X^b)$. Standard moment calculations for the Gamma distribution reveal that $\mathrm{E}_N(Y) = O(N^{-b\delta})$ and $\mathrm{Var}_N(Y) = a^2\,\mathrm{Var}(X^b) = O(N^{-2b\delta})$.

Next, consider total derived biomass, B, where $B = \sum_{j=1}^{N} Y_j$ and the Y_j are conditionally independent. From the above calculations, $\mathrm{E}_N(B) = O(N^{1-b\delta})$ and $\mathrm{Var}_N(B) = O(N^{1-2b\delta})$. We can also obtain the large N behavior of the coefficient of variation (CV); that is, $\mathrm{CV}_N(B) = \frac{\sqrt{N\,\mathrm{Var}_N(Y)}}{N\mathrm{E}_N(Y)} = N^{-1/2}\frac{O(N^{(1-2b\delta)/2})}{O(N^{1-b\delta})} = O(N^{-1})$.

As a function of plot density, we therefore anticipate various "large" N behaviors for mean, variance, and CV of total derived biomass.[4] In particular, if $b\delta < 1$, then we expect to see increasing total derived biomass in N, while if $b\delta > 1$, we

[4]For these asymptotics to apply in practice, we either need $L = 0$ in (3.2) or else to consider $\tilde{X} = X - L$. Otherwise, $Y \geq aL^b$, and so $B \geq NaL^b$.

expect decreasing total derived biomass. The latter is not a contradiction. With increasing plot density, we expect smaller trees on average, and thus total derived biomass need not increase with N. Similar remarks apply to the variance in total derived biomass. Now, if $2b\delta < 1$, variance increases in N, but if $2b\delta > 1$, variance decreases in N. As N increases, we expect plots containing smaller trees with less variability between the plots. Interestingly, for the CV, we expect $O(N^{-1})$ behavior regardless of δ.

The analysis of the previous section suggests estimates for δ with regard to the diameter distribution. The allometry suggests estimates for b. Together, we can infer about the mean and variance behavior of total derived biomass as a function of plot density, N.

Returning to the case where error is introduced in the allometry, if e is independent of X (and the distribution of e does not depend upon N), then the order results are unaffected by including e. However, it is often assumed that the errors depend upon X such that bigger diameters produce bigger errors in the allometry. This suggests $\mathrm{Var}(e)$ to be of the form $\tau^2(X)$ where, say, $\tau^2(X) = X^\psi \tau^2$ with $\psi > 0$. Now, $\mathrm{E}_N(Y) = a\mathrm{E}(\mathrm{E}_{e|X}(X^b e))$ will still be $O(N^{-b\delta})$ with, say, $\mathrm{E}(e) = 1$. However, $\mathrm{Var}_N(Y) = a^2(\mathrm{E}_N(\mathrm{Var}(X^b e|X)) + \mathrm{Var}_N(\mathrm{E}(X^b e|X))) = a^2(\mathrm{E}_N(X^{2b+\psi}) + \mathrm{Var}(X^b))$. Here, the first term will be $O(N^{-(2b+\psi)\delta})$, while the second will be $O(N^{-2b\delta})$. Since $\psi > 0$, the order of the variance will still be $O(N^{-2b\delta})$. Evidently, the order behavior of CV will also be unchanged.

In the literature, other allometric forms exist, for example, $\log(Y) = b_0 + b_1 \log(X) + b_2(\log(X))^2$. Alternatively, sometimes tree height, h, is introduced (though it is acknowledged that height is less accurately measured and less affected by competition than diameter) to yield a form $\log(Y) = a(X^2 h)^b$; that is, biomass is derived from the *volume* of the trunk. A long list of allometric forms relating the variables is supplied in the GlobAllomTree manual (http://www.globallometree.org/media/cms_page_media/6/tarifs_en_web_May23_1.pdf) in Table 5.1, page 105. Similar large N calculations to the above can be attempted for some of these allometric forms.

3.3. *Predicting total biomass.* We employ the allometric model that depends only on diameter such that

$$\log(Y) = \log(a) + b\log(X) + \epsilon, \tag{3.5}$$

where ϵ has mean zero normal error with constant variance that is independent of X and N. Here, X is diameter given in centimeters (cm) and Y is noisy biomass given in kilograms, which we convert to metric tons (t). In the FIA data for natural stands in the eastern United States, hardwoods are the most common and, among them, hard maple/oak/hickory/beech are the most prevalent species group. For this species group, Jenkins et al. (2003) provide coefficient estimates of $\log(\hat{a}) = -2.0127$, $\hat{b} = 2.4342$ and a variance estimate of 0.0559. These allometric model estimates are used in the remainder of this work for all biomass calculations.

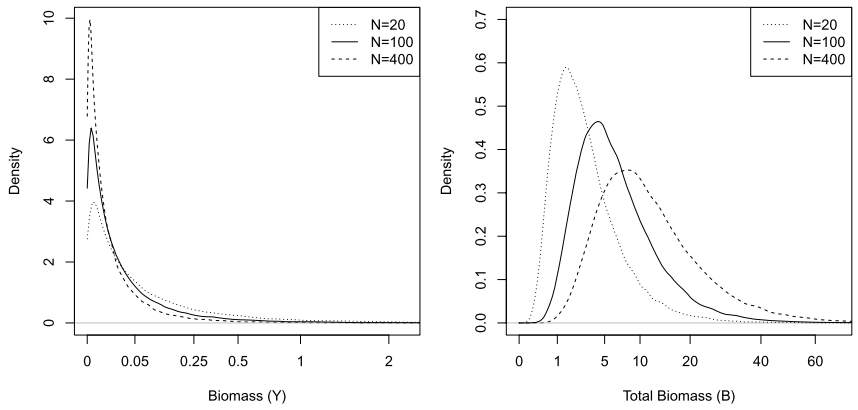

FIG. 4. *Posterior predictive distribution of individual biomass, Y, (left) and total biomass, B, (right) reported in metric tons (t) for FIA size plots with* $N = 20, 100,$ *and* 400. *In each distribution, the covariates,* $\mathbf{W}$, *are set to their centered values.*

Figure 4 (left) shows the posterior predictive distribution of $[Y|N, \mathbf{W}]$ for various values of N and $\mathbf{W}$ fixed. The distribution of Y becomes concentrated at small values of Y for large N; that is, individual biomass becomes much smaller and more localized as the density of the plot increases. Figure 4 (right) shows the posterior predictive distribution of $[B|N, \mathbf{W}]$ again for various N and fixed $\mathbf{W}$. Estimates of total biomass at the plot level tend to increase as N increases, as well as uncertainty, although the rate of the increase decreases with N.

Figure 5 shows the behavior of $\mathrm{E}_N(B)$, $\mathrm{Var}_N(B)$, and $\mathrm{CV}_N(B)$ as N increases. The asymptotic behavior of both the mean and variance of biomass are computed using the posterior mean estimate of δ. To capture sensitivity to b, we show the behavior using the minimum and maximum species group-specific coefficient estimates, $\hat{b}$, of the allometric models in Jenkins et al. (2003) corresponding to the species groups in the FIA data. Specifically, the coefficient estimates used are $\hat{b} = 2.2592$ and 2.4835, which are for cedar/larch and mixed hardwood, respectively. In terms of asymptotic behavior, there is little difference due to the allometry coefficients. Because $\delta\hat{b} < 1$, the order of $\mathrm{E}_N(B)$ increases with N, whereas since $2\delta\hat{b} > 1$, the order of $\mathrm{Var}_N(B)$ is decreasing in N. Thus, for large N, the increase in expected total biomass at the plot level that results from adding a tree to a plot more than offsets the decrease in expected total biomass at the plot level as a result of the negative shift in the density-dependent diameter distribution. This difference between the increase and decrease, however, decreases with N.

4. Completing the diameter distribution model. We complete the density-dependent diameter distribution model by specifying a first-stage Poisson distribution for the random number of trees observed on a plot. The generative model has

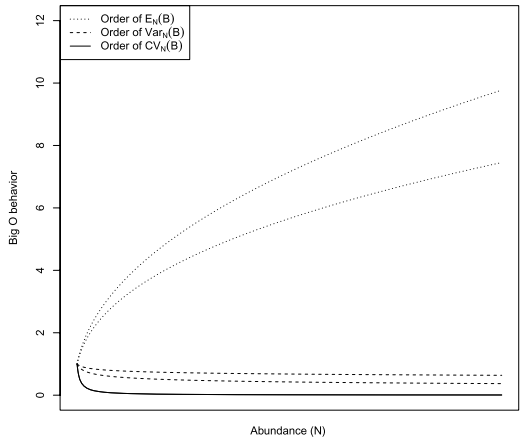

FIG. 5. *Order of* $\mathrm{E}_N(B) = O(N^{1-\delta\hat{b}})$, $\mathrm{Var}_N(B) = O(N^{1-2\delta\hat{b}})$, *and* $\mathrm{CV}_N(B) = O(N^{-1})$ *versus* N *using the minimum and maximum species group-specific coefficient estimates* $\hat{b}$ *of the allometric model* (3.5) *and the posterior mean value of* δ *from Table* 1.

the form

$$
\begin{aligned}
&[N_i|\mathbf{W}_i, \eta_i, \boldsymbol{\theta}][\widetilde{\boldsymbol{\mathcal{X}}}_i|N_i, \mathbf{W}_i, \alpha, \delta, \nu_i, \boldsymbol{\gamma}] \\
&\quad = [N_i|\mathbf{W}_i, \eta_i, \boldsymbol{\theta}]\Pi_{\{\widetilde{X}_{i,j}\in\widetilde{\boldsymbol{\mathcal{X}}}_i\}}[\widetilde{X}_{i,j}|N_i, \mathbf{W}_i, \alpha, \delta, \nu_i, \boldsymbol{\gamma}],
\end{aligned} \tag{4.1}
$$

where $[N_i|\mathbf{W}_i, \eta_i, \boldsymbol{\theta}]$ is the conditional distribution of plot density, N_i, given covariate information, $\mathbf{W}_i$. We specify a Poisson regression model with plot specific intensity,

$$
\begin{aligned}
N_i|\mathbf{W}_i, \eta_i, \boldsymbol{\theta} &\sim \mathrm{Poisson}(\lambda_i) \\
\log(\lambda_i) &= \eta_i + \mathbf{W}_i'\boldsymbol{\theta}.
\end{aligned} \tag{4.2}
$$

The regression model is defined with random effect, η_i, and coefficient vector, $\boldsymbol{\theta}$. Since the N_i are given, the model for plot density and the model for diameter in (4.1) can be fitted separately.

4.1. *Marginalizing over* N. In practice, we want to predict total biomass for a plot of a particular size experiencing a particular set of environmental covariates. To address this, we would not specify an explicit number of trees on the plot. Rather, given a distribution for the number of trees, N, we would marginalize over N to obtain predictive distributions for (marginalized) plot-level biomass. Formally, in the previous section we considered $[B|N, \mathbf{W}, \alpha, \delta, \nu, \boldsymbol{\gamma}]$. Now we investigate $[B|\mathbf{W}, \alpha, \delta, \nu, \boldsymbol{\gamma}, \eta, \boldsymbol{\theta}] = \sum_N [B|N, \mathbf{W}, \alpha, \delta, \nu, \boldsymbol{\gamma}][N|\mathbf{W}, \eta, \boldsymbol{\theta}]$. We are interested in the behavior of total biomass given the model parameters; we are not marginalizing over the model parameters as we would do in predictive inference.

Figure 6 shows the distribution of total biomass conditionally (given N) and marginally for three constant values for λ in (4.2). The covariates, $\mathbf{W}$, and random

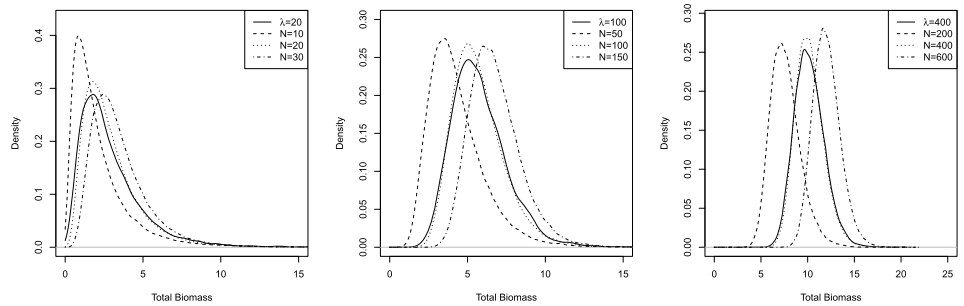

FIG. 6. *Distributions of total biomass* (t) *both conditionally and marginally with respect to plot density*, N, *for three different values of* λ; *that is*, *we compare* $[B|\lambda]$ *and* $[B|N]$.

effect, ν, in the Gamma regression model have been set to their centered values, and α and δ are set to their posterior mean estimates (Table 1). For each value of λ, the distribution of total biomass is shown conditionally for three plausible values of N. The distribution of total biomass depends on N both in terms of the number of trees on the plot and the distribution of diameters.

For each value of λ, the marginal and conditional distributions are similar when $N = \lambda$. However, the variance of the marginal distribution is larger than the variance of the conditional distribution and the difference decreases with λ. When $N < \lambda$, the mean of the conditional distribution of total biomass tends to be less than that of the marginal distribution, while when $N > \lambda$, the mean of the conditional distribution of total biomass tends to be larger. Additionally, the mean of total biomass increases with N and, for large N, the variance in total biomass decreases with N, which agrees with the asymptotic theory presented in Section 3.2.

4.2. *Poisson model inference*. We assign prior distributions to the model parameters in (4.2) and obtain posterior inference. The plot random effects are modeled as $\eta_i \overset{i.i.d.}{\sim} \mathrm{N}(\mu_\eta, \sigma_\eta^2)$ for $i = 1, \ldots, I$, where $\mu_\eta \sim \mathrm{N}(0, 100^2)$ and $\sigma_\eta^2 \sim \mathrm{IG}(2, 2)$. The coefficient vector $\boldsymbol{\theta}$ is assigned $\mathrm{MVN}(\mathbf{0}, 100^2\mathbf{I}_p)$.

The model is fitted using the same 90% of FIA plots as in Section 3. We obtain posterior samples using MCMC. The chains are run for 20,000 iterations, and the first half are disregarded as burn-in. Standard diagnostics resulted in no issues with convergence. Table 2 gives the posterior mean and 95% credible intervals for the parameters of the plot density model. The estimates of $\boldsymbol{\theta}$ indicate that, in general, the number of stems on a plot is increasing with temperature, moisture, and stand age for young stands, while decreasing with deficit and stand age for old stands.

4.3. *Predictive distribution of total biomass*. For a new plot with covariates $\mathbf{W}_{new}$, we consider the predictive distribution of total biomass, B_{new}. Let $\boldsymbol{\Theta}_P$ and $\boldsymbol{\Theta}_G$ denote the parameters of the Poisson regression model and Gamma regression

TABLE 2
Posterior mean estimates and 95% credible intervals for the Poisson regression model for plot density

Parameter	Mean	Credible interval
μ_η	3.389	(3.381, 3.397)
σ_η^2	0.230	(0.224, 0.235)
θ_1 stand age	0.101	(0.096, 0.109)
θ_2 temperature	0.027	(0.019, 0.033)
θ_3 deficit	−0.033	(−0.043, −0.026)
θ_4 moisture	0.008	(0.001, 0.016)
θ_5 stand age^2	−0.066	(−0.071, −0.061)
θ_6 stand age × deficit	−0.090	(−0.097, −0.084)
θ_7 stand age × moisture	0.035	(0.026, 0.042)
θ_8 deficit × moisture	0.029	(0.020, 0.038)
θ_9 temperature × moisture	0.022	(0.013, 0.031)

model, respectively, defined in the previous sections. The predictive distribution of stand density, N_{new}, is of the form

$$\big[N_{new}|\mathbf{W}_{new}, \{N_i\}, \{\mathbf{W}_i\}\big] = \int [N_{new}|\mathbf{W}_{new}, \mathbf{\Theta}_P]\big[\mathbf{\Theta}_P|\{N_i\}, \{\mathbf{W}_i\}\big]\, d\mathbf{\Theta}_P.$$

Then, given N_{new}, the density-dependent predictive distribution of truncated diameters $\widetilde{X}_{new,j}$ for $j = 1, \ldots, N_{new}$ can be written

$$\begin{aligned}&\big[\widetilde{X}_{new,j}|N_{new}, \mathbf{W}_{new}, \{\widetilde{\mathcal{X}}_i\}, \{N_i\}, \{\mathbf{W}_i\}\big]\\ &\quad = \int [\widetilde{X}_{new,j}|\mathbf{W}_{new}, N_{new}, \mathbf{\Theta}_G]\big[\mathbf{\Theta}_G|\{\widetilde{\mathcal{X}}_i\}, \{N_i\}, \{\mathbf{W}_i\}\big]\, d\mathbf{\Theta}_G.\end{aligned}$$

From this, we can derive the predictive distribution of individual biomass, $Y_{new,j}$, using the allometric model (3.5), and total biomass for the stand, B_{new}, by summing $Y_{new,j}$ for $j = 1, \ldots, N_{new}$; that is, we can obtain $[B_{new}|N_{new}, \mathbf{W}_{new}, \{\widetilde{\mathcal{X}}_i\}, \{N_i\}, \{\mathbf{W}_i\}]$.

Marginalizing over N_{new}, the predictive distribution of total biomass becomes

$$\begin{aligned}&\big[B_{new}|\mathbf{W}_{new}, \{\mathcal{X}_i\}, \{N_i\}, \{\mathbf{W}_i\}\big]\\ &\quad = \int \big[B_{new}|N_{new}, \mathbf{W}_{new}, \{\mathcal{X}_i\}, \{N_i\}, \{\mathbf{W}_i\}\big]\big[N_{new}|\mathbf{W}_{new}, \{N_i\}, \{\mathbf{W}_i\}\big]\, dN_{new}.\end{aligned}$$

Inserting the above predictive distributions, we obtain the marginal predictive distribution of total biomass as

$$
\begin{aligned}
&\big[B_{new}|\mathbf{W}_{new},\{\mathcal{X}_i\},\{N_i\},\{\mathbf{W}_i\}\big]\\
&\quad=\int\int\Big(\int[B_{new}|N_{new},\mathbf{W}_{new},\boldsymbol{\Theta}_G][N_{new}|\mathbf{W}_{new},\boldsymbol{\Theta}_P]\,dN_{new}\Big)\\
&\qquad\times\big[\boldsymbol{\Theta}_G|\{\mathcal{X}_i\},\{N_i\},\{\mathbf{W}_i\}\big]\big[\boldsymbol{\Theta}_P|\{N_i\},\{\mathbf{W}_i\}\big]\,d\boldsymbol{\Theta}_G\,d\boldsymbol{\Theta}_P\\
&\quad=\int\int[B_{new}|\mathbf{W}_{new},\boldsymbol{\Theta}_G,\boldsymbol{\Theta}_P]\\
&\qquad\times\big[\boldsymbol{\Theta}_G|\{\mathcal{X}_i\},\{N_i\},\{\mathbf{W}_i\}\big]\big[\boldsymbol{\Theta}_P|\{N_i\},\{\mathbf{W}_i\}\big]\,d\boldsymbol{\Theta}_G\,d\boldsymbol{\Theta}_P,
\end{aligned}
\tag{4.3}
$$

where

$$
[B_{new}|\mathbf{W}_{new},\boldsymbol{\Theta}_G,\boldsymbol{\Theta}_P]=\int[B_{new}|N_{new},\mathbf{W}_{new},\boldsymbol{\Theta}_G][N_{new}|\mathbf{W}_{new},\boldsymbol{\Theta}_P]\,dN_{new}.
$$

The regression model discussed in the Introduction which uses derived total biomass as the response variable can also be used to obtain predictive distributions of total biomass. Letting $\boldsymbol{\Theta}_R$ denote the parameters of the regression model, the model for total biomass can be written in the form $\prod_i[B_i|\mathbf{W}_i,\boldsymbol{\Theta}_R]$. With a prior on $\boldsymbol{\Theta}_R$, we obtain $[\boldsymbol{\Theta}_R|\{B_i\},\{\mathbf{W}_i\}]$ and the predictive distribution of biomass for a new plot as

$$
\big[B_{new}|\mathbf{W}_{new},\{B_i\},\{\mathbf{W}_i\}\big]=\int[B_{new}|\mathbf{W}_{new},\boldsymbol{\Theta}_R]\big[\boldsymbol{\Theta}_R|\{B_i\},\{\mathbf{W}_i\}\big]\,d\boldsymbol{\Theta}_R.
\tag{4.4}
$$

The predictive distributions for B_{new} shown in (4.3) and (4.4) are functionally very different and will be different in practice as well (Figure 1).

5. Biomass prediction at the hectare scale in the eastern US. Prediction of biomass at the hectare scale, for example, metric tons per hectare (t/ha), is of interest because it represents an area large enough to depict a range of diameter classes. Shugart, Saatchi and Hall (2010) highlight that aggregating from 0.1 ha plots to 1 ha plots provides a significant amount of variability reduction in biomass inference. FIA plots are small, yielding inference too noisy for developing conservation and management decisions. Note that the scaling is additive; that is, we infer about biomass for a plot of area $k|A|$ from plots of area $|A|$ by summing the biomass of k independent plots of area $|A|$. In particular, since FIA plots are approximately 0.067 hectares, summing over 16 plots will give us predictive biomass at roughly one hectare scale.

5.1. *Estimation at the hectare scale.* We create aggregate one hectare area observations of tree densities and diameter distributions using k-means clustering. The FIA plots are clustered based on stand age and the climate variables introduced in Section 2. The result of k-means clustering is a partition of n observations into k

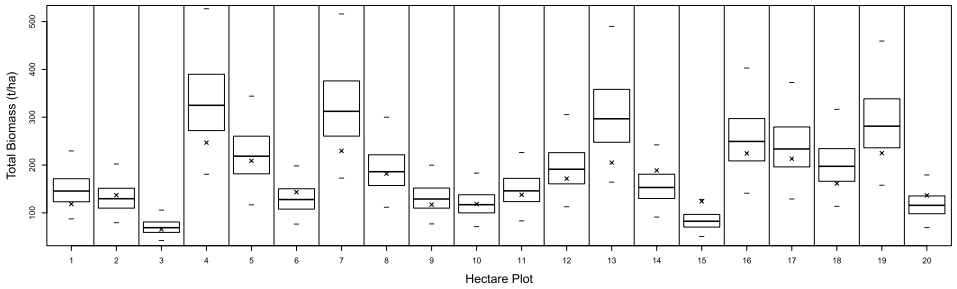

FIG. 7. *Estimates of total biomass* (*t/ha*) *for* 20 *randomly chosen out-of-sample hectare plots. Dashes indicate upper and lower limits of the* 95% *credible intervals, and the circles indicate derived total biomass computed by applying the error-free allometric model to the observed diameters for the plot and summing over all trees.*

clusters which minimize the distance in covariate space between each observation and the cluster mean. Using this algorithm, we created 500 clusters, of which 483 had 16 or more FIA plots. We then approximate hectare area observations by combining 16 FIA plots (yielding 1.072 ha) from a cluster. When clusters contain many plots, we can create multiple one hectare areal units per cluster. Upon truncation, the observed set of diameters become an aggregated sample of diameters on $\mathbb{R}^+$; that is, they consist of the total number of trees, again referred to as plot density, and the empirical diameter distribution for the roughly one hectare areal unit. Since the clustering is done in covariate space, the covariates within each set of 16 FIA plots will be very similar. The covariates at the hectare scale are computed as the mean of the 16 FIA plot-level covariate values.

5.2. *Results*. The clustering described above resulted in 1212 hectare area plots, where 1090 were used for fitting the model and the remaining 122 were withheld for out-of-sample prediction. Parameter inference was obtained using MCMC. The model was run for 50,000 iterations, and the first 10,000 were discarded as burn-in.

Predictive distributions of total biomass are obtained from the density-dependent diameter distribution model as outlined in Section 4.3. Figure 7 shows boxplots of the posterior predictive distribution of total biomass for 20 randomly chosen out-of-sample hectare plots. The bounds of the 95% credible intervals of the predictive distributions are also shown. The $\times$ indicates the noiseless derived total biomass computed by applying the error-free allometric model to the observed diameters for the hectare plot and summing over all trees.

We present estimates of total biomass over climate space in Figures 8 and 9. Posterior mean estimates and pointwise 95% credible intervals are shown in Figure 8 versus each of the covariates, assuming all other covariates are at their average value. The range of values along the x-axis spans the range of the variable in the observed data shown by the histogram. These figures indicate that total biomass is

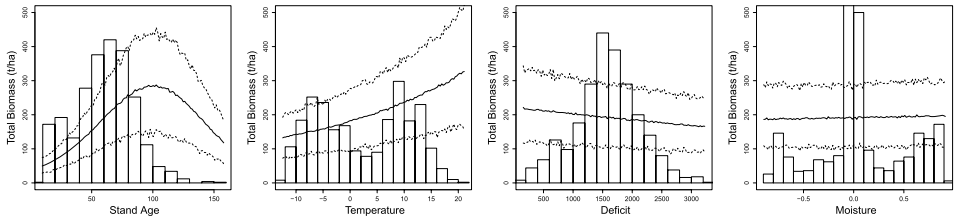

FIG. 8. *Estimates of total biomass* (*t/ha*) *versus the covariates stand age*, *temperature*, *deficit and moisture. The solid line indicates the posterior mean*, *and the dashed indicates the upper and lower* 95% *pointwise credible interval.*

increasing with stand age for young stands and decreasing with stand age for old stands. Total biomass is also increasing with temperature while decreasing with deficit. Finally, we anticipate a very small increase in total biomass with moisture.

In Figure 9, mean estimates of total biomass are shown for varying levels of covariates bivariately, again where all other covariates are at their average value. The $\times$ on each plot shows the observed covariate pairs. The left panel shows stand age versus deficit, indicating that, when all other covariates are average, total biomass is largest for moderately old stands with small deficit. The effect of stand age appears stronger than deficit since the increase in total biomass is greater in the horizontal direction than the vertical direction. For moderately old stands, total biomass doesn't appear to vary significantly across moisture levels. The third panel shows deficit versus standardized moisture. This reveals that, for average stand age and temperature, total biomass is greatest when deficit is low and moisture is high. Similarly, for average stand age and deficit, total biomass is greatest when temperature is low and moisture is high. However, the variability in total biomass for the interactions between deficit and moisture and temperature and moisture are small relative to the interactions with stand age.

Table 3 gives the posterior means and 95% credible intervals for the model parameters. All $\boldsymbol{\theta}$ coefficients for the Poisson regression model and $\boldsymbol{\phi}$ coefficients for the Gamma regression model are significantly different from 0, as indicated by the credible intervals. Since the covariates are centered and scaled, when all

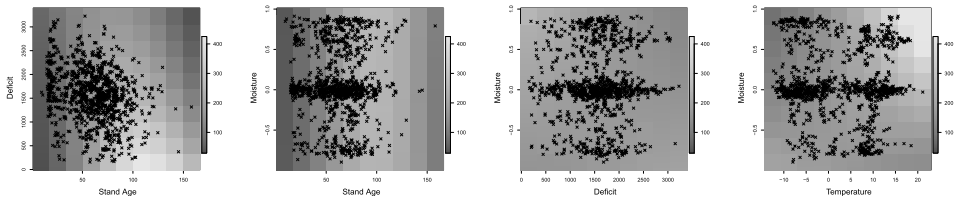

FIG. 9. *Estimates of total biomass* (*t/ha*) *versus the covariate interactions for* (*left*) *deficit and stand age*, (*middle-left*) *moisture and stand age*, (*middle-right*) *moisture and deficit*, *and* (*right*) *moisture and temperature.*

TABLE 3
Posterior mean estimates and 95% credible intervals for the density-dependent diameter distribution model parameters using data at the hectare scale

Parameter	**Mean**	**Credible interval**
α	0.992	(0.989, 0.995)
δ	0.290	(0.278, 0.304)
μ_ν	−4.141	(−4.228, −4.067)
σ_ν^2	0.018	(0.017, 0.020)
ϕ_1 stand age	−0.302	(−0.311, −0.292)
ϕ_2 temperature	−0.150	(−0.160, −0.139)
ϕ_3 deficit	0.028	(0.018, 0.038)
ϕ_4 moisture	−0.001	(−0.010, 0.008)
ϕ_5 stand age^2	0.085	(0.078, 0.092)
ϕ_6 stand age × deficit	0.015	(0.005, 0.024)
ϕ_7 stand age × moisture	0.014	(0.005, 0.023)
ϕ_8 deficit × moisture	0.041	(0.029, 0.052)
ϕ_9 temperature × moisture	−0.093	(−0.104, −0.082)
μ_η	6.228	(6.211, 6.246)
σ_η^2	0.062	(0.056, 0.067)
θ_1 stand age	0.095	(0.079, 0.109)
θ_2 temperature	0.023	(0.006, 0.044)
θ_3 deficit	−0.022	(−0.042, −0.003)
θ_4 moisture	0.018	(0.005, 0.037)
θ_5 stand age^2	−0.068	(−0.077, −0.057)
θ_6 stand age × deficit	−0.100	(−0.116, −0.086)
θ_7 stand age × moisture	0.041	(0.027, 0.059)
θ_8 deficit × moisture	0.023	(0.007, 0.042)
θ_9 temperature × moisture	0.024	(0.005, 0.037)

covariates are at their average value (set equal to 0), the expected number of trees per hectare is $\exp^{6.228} = 507$. In terms of the main effects, when all other covariates are at their average value, the number of trees per hectare increases with stand age for young stands and decreases with stand age for older stands. The number of trees per hectare also increases with temperature and moisture but decreases with deficit. The interactions indicate that older plots will have more trees when both deficit is low and moisture is high. Additionally, the decrease in expected number of trees resulting from high deficit will be less for plots with high moisture.

Under the Gamma regression model, diameter increases with stand age for young stands and decreases with stand age for older stands. Diameter also increases with temperature and moisture and decreases with deficit. The negative effect of deficit on tree diameter is less for younger stands than older stands. Additionally, the positive effect of moisture is greater for younger stands, stands with less deficit, and stands with greater temperature.

6. Summary and future work. We have developed a plot-level process-driven model for the observed number of individuals and the associated set of diameters. Prediction of total biomass is achieved a posteriori using allometric models. Ecological theory suggests that the distribution of diameters should be density dependent, and thus, is incorporated into our modeling. Through simulation and real data, we have demonstrated the inference performance of our total biomass prediction. Usual regression models for total biomass may be unsatisfying since they are fitted to derived, not observed, total biomasses. Moreover, these models are unable to use the available observed plot-level information in their specifications.

As future work, for diameters having associated species labels, we will build species-specific density-dependent models. We will obtain predictions of total biomass by summing over individual biomass predictions from species-specific allometric models [see, e.g., Jenkins et al. (2003), Lambert, Ung and Raulier (2005)]. In application, many species groups will not be present on many plots. However, if some species groups tend to present larger trees while others tend to present smaller trees and if there is consequential variation in the allometric models, then common allometry may not predict total biomass well.

We briefly outline a strategy for addressing this problem with just two species groups, retaining the modeling framework developed in Section 3 above. Now, for every observed diameter $X_{i,j}$ we have a binary variable, $S_{i,j}$, indicating which of the two species groups is associated with this diameter. As a result, we add a logistic regression modeling stage for $S_{i,j}$, where $P(S_{i,j} = 1)$ depends upon the diameter, $X_{i,j}$, in addition to N_i and $\mathbf{W}_i$. Again, biomass is predicted post-model fitting. Now, given N and $\mathbf{W}$, we first sample an X. Then we sample an S given X and obtain a biomass, Y, according to the allometry determined by S. By summing over the N Ys, we obtain a predicted total biomass.

Another interesting future question is whether it is sensible to introduce expected density dependence rather than observed density dependence into the diameter distribution model, that is, replacing $[X_{i,j}|N_i, \mathbf{W}_i]$ with $[X_{i,j}|E(N_i), \mathbf{W}_i]$. There is considerable variability in the N_is but much less in the $E(N_i)$s. Using the latter, less variability in predicting biomass will accrue, but the predictions may not be well centered.

Finally, an important ecological issue is change in biomass, so-called *productivity*, particularly in response to change in climate. We will attempt to modify our modeling to provide prediction for change in total biomass at various temporal scales.

Acknowledgments. We would also like to thank the Editor, Associate Editor and two anonymous referees for their diligent reviews that greatly strengthened this manuscript.

REFERENCES

BACCINI, A., GOETZ, S., WALKER, W., LAPORTE, N., SUN, M., SULLA-MENASHE, D., HACKLER, J., BECK, P., DUBAYAH, R., FRIEDL, M., SAMANTA, S. and HOUGHTON, R. (2012). Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps. *Nature Climate Change* **2** 182–185.

BECHTOLD, W. A. and PATTERSON, P. L. (2005). The enhanced forest inventory and analysis program: National sampling design and estimation procedures. Technical report, US Department of Agriculture Forest Service, Southern Research Station Asheville, NC.

BLACKARD, J., FINCO, M., HELMER, E., HOLDEN, G., HOPPUS, M., JACOBS, D., LISTER, A., MOISEN, G., NELSON, M., RIEMANN, R., RUEFENACHT, B., SALAJANU, D., WEYERMANN, D., WINTERBERGER, K., BRANDEIS, T., CZAPLEWSKI, R., MCROBERTS, R., PATTERSON, P. and TYMCIO, R. (2008). Mapping US forest biomass using nationwide forest inventory data and moderate resolution information. *Remote Sensing of Environment* **112** 1658–1677.

BRZOSTEK, E. R., DRAGONI, D., SCHMID, H. P., RAHMAN, A. F., SIMS, D., WAYSON, C. A., JOHNSON, D. J. and PHILLIPS, R. P. (2014). Chronic water stress reduces tree growth and the carbon sink of deciduous hardwood forests. *Glob. Change Biol.* **20** 2531–2539.

CAMARERO, J. J., GAZOL, A., GALVÁN, J. D., SANGÜESA-BARREDA, G. and GUTIÉRREZ, E. (2015). Disparate effects of global-change drivers on mountain conifer forests: Warming-induced growth enhancement in young trees vs. CO2 fertilization in old trees from wet sites. *Glob. Change Biol.* **21** 738–749.

CHAVE, J., CONDIT, R., AGUILAR, S., HERNANDEZ, A., LAO, S. and PEREZ, R. (2004). Error propagation and scaling for tropical forest biomass estimates. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **359** 409–420.

CHAVE, J., RÉJOU-MÉCHAIN, M., BÚRQUEZ, A., CHIDUMAYO, E., COLGAN, M. S., DELITTI, W. B., DUQUE, A., EID, T., FEARNSIDE, P. M., GOODMAN, R. C., HENRY, M., MARTINEZ-YRIZAR, A., MULLER-LANDAU, H. C., MENCUCCINI, M., NELSON, B. W., NGOMANDA, A., NOGUEIRA, E. M., ORTIZ-MALAVASSI, E., PÉLISSIER, R., PLOTON, P., RYAN, C. M., SALDARRIAGA, J. G. and VIEILLEDENT, G. (2014). Improved allometric models to estimate the aboveground biomass of tropical trees. *Glob. Change Biol.* **20** 3177–3190.

DONG, J., KAUFMANN, R. K., MYNENI, R. B., TUCKER, C. J., KAUPPI, P. E., LISKI, J., BUERMANN, W., ALEXEYEV, V. and HUGHES, M. K. (2003). Remote sensing estimates of boreal and temperate forest woody biomass: Carbon pools, sources, and sinks. *Remote Sensing of Environment* **84** 393–410.

GELFAND, A. E., SAHU, S. K. and CARLIN, B. P. (1995). Efficient parameterisations for normal linear mixed models. *Biometrika* **82** 479–488. MR1366275

HENRY, M., BOMBELLI, A., TROTTA, C., ALESSANDRINI, A., BIRIGAZZI, L., SOLA, G., VIEILLEDENT, G., SANTENOISE, P., LONGUETAUD, F., VALENTINI, R., PICARD, N. and SAINT-ANDRÉ, L. (2013). Globallometree: International platform for tree allometric equations to support volume, biomass and carbon assessment. *IForest-Biogeosciences and Forestry* **6** 326.

JENKINS, J. C., CHOJNACKY, D. C., HEATH, L. S. and BIRDSEY, R. A. (2003). National-scale biomass estimators for United States tree species. *Forest Science* **49** 12–35.

LAMBERT, M., UNG, C. and RAULIER, F. (2005). Canadian national tree aboveground biomass equations. *Canadian Journal of Forest Research* **35** 1996–2018.

MALHI, Y., WOOD, D., BAKER, T. R., WRIGHT, J., PHILLIPS, O. L., COCHRANE, T., MEIR, P., CHAVE, J., ALMEIDA, S., ARROYO, L., HIGUCHI, N., KILLEEN, T. J., LAURANCE, S. G., LAURANCE, W. F., LEWIS, S. L., MONTEAGUDO, A., NEILL, D. A., VARGAS, P. N., PITMAN, N. C. A., QUESADA, C. A., SALOMÃO, R., SILVA, J. N. M., LEZAMA, A. T., TERBORGH, J., MARTÍNES, R. V. and VINCETI, B. (2006). The regional variation of aboveground live biomass in old-growth Amazonian forests. *Glob. Change Biol.* **12** 1107–1138.

McRoberts, R. E., Moser, P., Zimermann Oliveira, L. and Vibrans, A. C. (2015). A general method for assessing the effects of uncertainty in individual-tree volume model predictions on large-area volume estimates with a subtropical forest illustration. *Canadian Journal of Forest Research* **45** 44–51.

Molto, Q., Rossi, V. and Blanc, L. (2013). Error propagation in biomass estimation in tropical forests. *Methods Ecol. Evol.* **4** 175–183.

Pan, Y., Birdsey, R. A., Fang, J., Houghton, R., Kauppi, P. E., Kurz, W. A., Phillips, O. L., Shvidenko, A., Lewis, S. L., Canadell, J. G., Ciais, P., Jackson, R. B., Pacala, S. W., McGuire, A. D., Piao, S., Rautiainen, A., Sitch, S. and Hayes, D. (2011). A large and persistent carbon sink in the worlds forests. *Science* **333** 988–993.

Picard, N., Saint-André, L. and Henry, M. (2012). Manual for building tree volume and biomass allometric equations: From field measurement to prediction. Technical report, FAO/CIRAD.

Saatchi, S. S., Harris, N. L., Brown, S., Lefsky, M., Mitchard, E. T., Salas, W., Zutta, B. R., Buermann, W., Lewis, S. L., Hagen, S., Petrova, S., White, L., Silman, M. and Morel, A. (2011). Benchmark map of forest carbon stocks in tropical regions across three continents. *Proc. Natl. Acad. Sci. USA* **108** 9899–9904.

Schliep, E. M., Gelfand, A. E., Clark, J. S. and Zhu, K. (2016). Modeling change in forest biomass across the eastern US. *Environ. Ecol. Stat.* **23** 23–41. MR3458607

Shugart, H., Saatchi, S. and Hall, F. (2010). Importance of structure and its measurement in quantifying function of forest ecosystems. *Journal of Geophysical Research: Biogeosciences* **115**.

Thurner, M., Beer, C., Santoro, M., Carvalhais, N., Wutzler, T., Schepaschenko, D., Shvidenko, A., Kompter, E., Ahrens, B., Levick, S. R. and Schmullius, C. (2014). Carbon stock and density of northern boreal and temperate forests. *Global Ecology and Biogeography* **23** 297–310.

Ung, C.-H., Bernier, P. and Guo, X.-J. (2008). Canadian national biomass equations: New parameter estimates that include British Columbia data. *Canadian Journal of Forest Research* **38** 1123–1132.

Whittaker, R. H. and Woodwell, G. M. (1968). Dimension and production relations of trees and shrubs in the Brookhaven Forest, New York. *The Journal of Ecology* 1–25.

Wilson, B. T., Woodall, C. W. and Griffith, D. M. (2013). Imputing forest carbon stock estimates from inventory plots to a nationally continuous coverage. *Carbon Balance and Management* **8** 1–15.

Department of Statistics
University of Missouri
Columbia, Missouri 65211
USA
E-mail: schliepe@missouri.edu

Department of Statistical Science
Nicholas School of the Environment
Duke University
Durham, North Carolina 27708
USA
E-mail: schliepe@missouri.edu
alan@stat.duke.edu
jimclark@duke.edu
bradley.tomasek@duke.edu