Journal of Forestry (2025) 123:653–674
https://doi.org/10.1007/s44392-025-00044-x

ORIGINAL PAPER

Quantifying Stand-level Carbon Using an Alternative Method: A Case Study of National Forests and Parks in Southern Appalachian Mountains

Sheng-I Yang[1] · Thomas J. Brandeis[2] · Tara Skiba[3]

Received: 24 September 2024 / Accepted: 29 May 2025 / Published online: 24 July 2025

Abstract

As vital components of the global carbon cycle, forest ecosystems play an important role in sustainable forest management. With the new data collected and recent development of carbon equations in the region, there is a need to update the essential quantitative information about forest carbon for monitoring these vulnerable populations and making appropriate management decisions. This study aims to quantify stand-level aboveground live carbon using an alternative method among different national forests and parks in the southern Appalachian region. Hybrid models with linear mixed model and random forests were built using forest type as a random effect with common stand variables as predictors. Long-term data collected by the USDA Forest Service, Forest Inventory and Analysis (FIA) program were used in analyses. Results show that the average aboveground live carbon per hectare varies among study sites, ranging from 58.2 to 91.5 Mg C/ha. Monongahela National Forest and Great Smoky Mountains National Park have higher average carbon per hectare than other sites. Within a site, forest carbon per hectare varies among different forest type groups. The hybrid model can provide satisfactory predictions of total carbon at the stand level, even for data not used in model fitting. As parametric models remain the most popular quantitative methods in forestry practice, the proposed method retains the interpretability and portability of the parametric models, but also improves prediction accuracy by leveraging the strengths of the machine learning algorithm.

Keywords Random forest · Mixed models · Great Smoky Mountains National Park · Mixed species forests · Forest type · Machine learning · Publically owned forest

Study Implications Mixed-species forests in the southern Appalachian mountains sequester atmospheric carbon, as well as provide timber products, wildlife habitats and ecosystem services to the surrounding communities. Data collected from eight national forests and two national parks with 45 distinct forest types were analyzed. This study demonstrates the use of a hybrid model to predict stand-level forest carbon across national forests and parks in the region. The hybrid model combines the advantages of the classic parametric method and the machine learning algorithm. The findings of this study provide an alternative method for providing quantitative information to manage the vulnerable ecosystems in the region.

Extended author information available on the last page of the article

Introduction

As vital components of the global carbon cycle, forest ecosystems play an important role in sustainable forest management. Mixed-species forests in the southern Appalachian mountains sequester atmospheric carbon, as well as provide timber products, wildlife habitats and ecosystem services to the surrounding communities. While forest land in the eastern US is primarily privately owned, the publicly-owned forests of the southern Appalachians are of particular importance. They are relatively large, generally contiguous areas of forest with minimal anthropogenic disturbance and management activity when compared to the surrounding private forest lands. Most of the upper elevations in the southern Appalachians and the unique vegetation communities found there are in national forests or parks (Bailey 2016; USDA Forest Service 2023; Jenkins 2007). The health and sustainability of mixed-species forests communities has received increased concern due to threats to some of the ecologically important species found there, such as the montane spruces and firs (*Picea rubens* and *Abies fraser)*, eastern hemlocks (*Tsuga canadensis*), and oaks (*Quercus* spp.) observed in the region (e.g., Koo et al. 2011; Ellison et al. 2018). Koo et al. (2011) describe the losses of upper elevation spruce-fir forests to a combination of air pollution and an introduced pest, the balsam wooly adelgid (*Adelges piceae*). Eastern hemlock, a foundational species throughout much of the Appalachians, continue to decline in abundance despite decades of attempts to control the spread of an introduced pest, the hemlock wooly adelgid (*Adelges tsugae*) (Ellison et al. 2018). Keyser et al. (2015) indicated changes in the composition and structure of oak-dominated forests have been unprecedented and rapid, caused by extensive logging, wildfires, exotic insects, diseases, and invasive plants. These changes in species composition and structure along with expected changes due to climate change affect carbon storage and sequestration in these southern Appalachian forests. With the new data collected and recent development of carbon equations in the region, there is a need to update the essential quantitative information about forest carbon for monitoring these vulnerable populations and making appropriate management decisions.

Studies of mixed-species forests can frequently be more challenging because of species richness. Data grouping based on similar characteristics (e.g., taxonomic rank, functional traits, tree size) is a commonly-used strategy to build statistical models for species rich populations due to a small sample size for diverse species (e.g., Vanclay 1991; Phillips et al. 2002; Zhang et al. 2022). Mixed-effects models have been proposed as an alternative quantitative solution as they do not require data grouping (e.g., Russel et at. 2014, Yang et al. 2022, Kuehne et al. 2020). A mixed-effects model consists of two components: fixed and random effects, making it advantageous to handle data with a large number of groups. Estimated with all data together, the model's fixed effects provide the overall mean response for the target variable of interest, while its random effects provide the adjusted response for each group (i.e., group-specific response). With the development of computationally intensive algorithms, machine learning algorithms have been integrated in mixed modeling. For example, machine learning is used to estimate fixed effects as it is more advantageous to capture the complex relationships between covariates for high dimensional data than conventional parametric approaches. Yang et al. (2022) employed mixed-effects random forests, incorporating species as a random effect, to analyze height-diameter relationships for Caribbean species. The algorithm performed better than original random forests and parametric models in terms of prediction accuracy (Yang et al. 2022). However, machine learning does not provide estimated coefficients of fixed effects like parametric models, limiting the explicit interpretation of relationships between response and predictor variables. Rather than estimating fixed effects, another use of machine learning is to capture the remaining variability after a mixed model is fitted, denoted as a hybrid model (e.g., Cornelio et al. 2022; de Mattos Neto et al. 2022; Jiang et al. 2024; de Lima Duarte et al. 2024). Specifically, a mixed model is first built, and then a machine learning algorithm is trained using the residuals computed from the fitted mixed model as a response variable. The final predictions are made by combining the predicted values from both mixed model and the predicted residuals from the machine learning algorithm (a more detailed description of the model building procedure is given in the method section). This approach can offer more interpretable fixed effects from a mixed model and leverage machine learning's capability to handle nonlinear and complex data.

The objective of this study was to quantify stand-level forest carbon for eight national forests and two national parks in the southern Appalachian region. For a given national forest or park, a hybrid model composed of a linear mixed model and random forests were built with forest type as a random effect. Stand-level carbon prediction models can provide a quick estimate of total forest carbon without the need to identify the species of every sample tree and use of species-specific carbon equations, which could save time and reduce cost in data collection and processing. Long-term data collected by the USDA Forest Service, Forest Inventory and Analysis (FIA) program were used in analyses. This work serves as a practical example of applying the methodology to continuously monitor forest resources for the vulnerable ecosystems in the region.

Materials and Methodology

Data

Study Area

The study area covers eight national forests (Chattahoochee-Oconee, Cherokee, Nantahala, Pisgah, George Washington-Jefferson, Daniel Boone, Monongahela, and Wayne) and two national parks (Great Smoky Mountains, Shenandoah), which encompass a total of 2,755,856 ha. These national parks and forests are primarily situated along the Tennessee-North Carolina and West Virginia-Virginia borders, and extend into adjacent parts of Ohio, Kentucky, and as far south as central Georgia. Among these areas, the George Washington-Jefferson National Forest covers the largest physical area of all study sites (1,800,269 ha), while Shenandoah National Park is the smallest study area with only 79,900 ha (USDA Forest Service 2023).

The southeast region where the targeted forests and parks are located falls within the temperate humid domain of Bailey's ecoregions, largely within the Eastern Broadleaf Forest, Central Appalachian Broadleaf Forest, and Southeastern Mixed Forest provinces (Bailey 2016). The study sites also share several physiographic similarities, like mountain ranges that have elevations up to 2,037 m (Mount Mitchell, which is found in the North Carolina State Park named after it, and 2,025 m for Kuwohi (old name: Clingmans Dome) in the Great Smoky Mountain National Park), surrounded by the rolling forested hills and rocky bluffs that have become identifiable characteristics of southern Appalachia. The exception to this is Wayne National Forest, as it is located in the Appalachian foothills and does not contain the same mountainous terrain as the other areas. While these geographical similarities exist, each study site also has its own unique attributes. For example, the Great Smoky Mountains National Park ranks as the most biodiverse park in the national park system where over 1,300 native plant species have been identified (Jenkins 2007). Much of the ecological diversity in this area is due to the temperate rainforest climate of the southeast. This kind of climate consists of long, wet winters and shorter, dry yet foggy summers which makes water resources plentiful, so that Nantahala National Forest is the second wettest region in the continental US after the Pacific Northwest (Spencer and Lewis 2017).

Forest Type

According to forest typing algorithm used by FIA (Burrill et al. 2024), a total of 45 distinct forest types are present in the study area. Thirty-three different forest types were identified in Cherokee and George Washington-Jefferson National Forests, followed by Chattahoochee-Oconee National Forest and Great Smoky Mountains National Park with 32 forest types each (Table 1). The study sites are primarily dominated by oak (*Quercus* spp.) forests along with a few pine-dominated (*Pinus* spp.) forest types (Table 1).

FIA Long-term Permanent Plots and Tree Data

Data from the permanent national forest inventory plots installed and maintained by the the FIA program were used in this study. These data were extracted by querying the FIA database for forested plots that fell within the targeted national forests and national parks. The spatial distribution of the 10 national forests and national parks is shown in Fig. 1. In each permanent plot, four 7.3-m (24-ft) fixed-radius subplots were installed with four 2.1-m (6.8-ft) fixed-radius microplots nested (see USDA Forest Service 2023 for details). Trees measured on a subplot need to have diameter at breast height (DBH) greater than or equal to 12.7 cm (5 in.), whereas those below 12.7-cm (5-in.) but greater than 2.54-cm (1 in.) DBH were measured on a microplot. At each sampling point, the area (the area of the permanent plot) was classified based on site conditions, such as land use, forest type, stand size, regeneration status, tree density, stand origin, ownership group, and disturbance history. If there was more than one distinct condition, the plot

Table 1 Number of observations (Obs) and forest types (FT) among 11 different forest type groups represented in eight national forests (NF) and two national parks (NP) in the southern Appalachian Mountains

Study site	Asp/Bir		Elm/Ash/Cw		Exo Hw		Lob/SLP		Map/Bch/Bir		Oak/Gum/Cyp		Oak/Hic		Oak/Pine		Other Hw		Spr/Fir		WP/RP/JP		Total	
	Obs	FT	Obs	FT	Obs	FT	Obs	FT	Obs	FT	Obs	FT	Obs	FT	Obs	FT	Obs	FT	Obs	FT	Obs	FT	Obs	FT
Chattahoochee-Oconee NF	0	0	7	2	0	0	152	5	1	1	0	0	1143	15	278	5	14	1	0	0	73	3	1668	32
Cherokee NF	0	0	0	0	0	0	143	5	73	3	0	0	881	15	145	5	17	1	1	1	66	3	1326	33
Nantahala NF	0	0	0	0	0	0	13	5	18	2	0	0	417	13	32	4	10	1	0	0	17	2	507	27
Pisgah NF	2	1	0	0	0	0	6	2	11	1	0	0	368	12	35	4	23	1	4	1	15	2	464	24
George Washington-Jefferson NF	0	0	0	0	1	1	106	3	81	4	2	1	1980	15	230	4	18	1	2	1	60	3	2480	33
Daniel Boone NF	0	0	3	1	0	0	7	3	32	3	0	0	503	14	16	4	0	0	0	0	7	1	568	26
Monongahela NF	2	1	2	1	0	0	4	3	595	4	0	0	501	13	18	3	28	1	51	1	20	3	1221	30
Wayne NF	8	1	25	3	0	0	6	1	29	2	0	0	268	13	17	2	1	1	0	0	16	1	370	24
Great Smoky Mountains NP	0	0	1	1	0	0	16	5	57	3	1	1	219	12	23	4	51	1	9	1	7	3	384	31
Shenandoah NP	0	0	1	1	0	0	0	0	0	0	0	0	156	11	1	1	3	1	0	0	0	0	161	14

Forest type is used in model fitting. Due to space constraints, only a summary of forest type groups is included here. A table detailing all forest types is given in Table S1 in the supplementary material. Eleven forest type groups include: aspen/birch (Asp/Bir), elm/ash/cottonwood (Elm/Ash/Cw), exotic hardwoods (Exo Hw), Loblolly/shortleaf pine (Lob/SLP), Maple/beech/birch (Map/Bch/Bir), Oak/gum/cypress (Oak/Gum/Cyp), Oak/hickory (Oak/Hic), Oak/pine, Other hardwoods (Other Hw), Spruce/fir (Spr/Fir), and White/red/jack pine (WP/RP/JP). Each forest type group includes one or more forest types

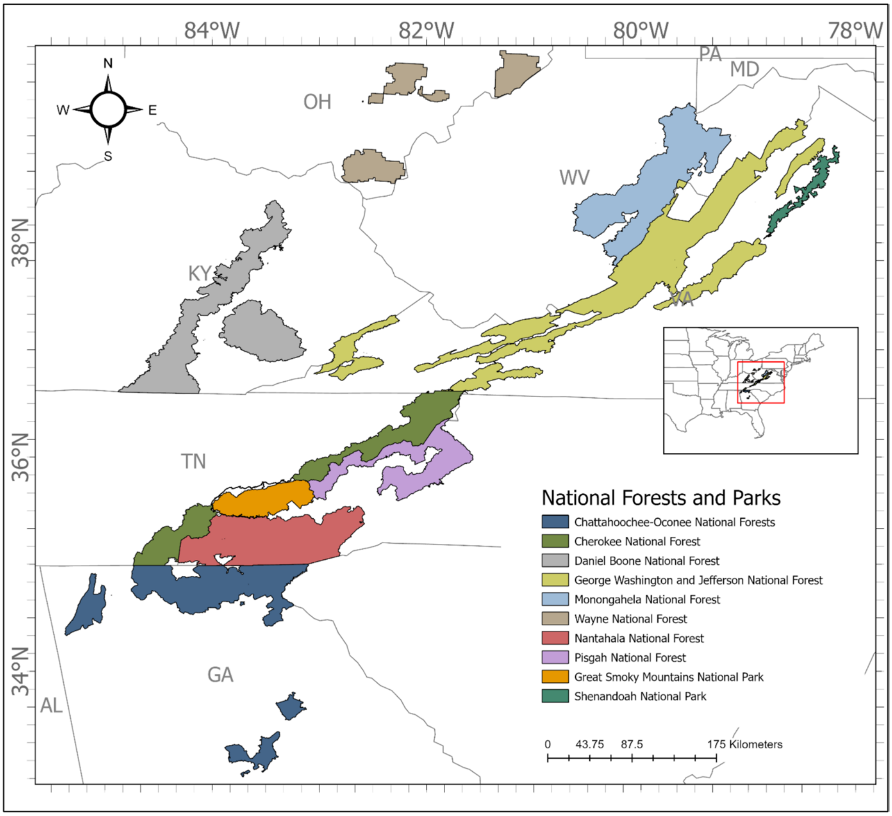


Fig. 1 Spatial distribution of the study area

would be mapped by condition class even though the boundaries cut through subplots or microplots. Each permanent plot was remeasured every 5–7 years.

Tree data collected in the study area were queried from the FIA database. Individual trees were excluded if they met any of the following criteria: (1) they were classified as rotten cull (indicating severe damage and decay which would affect the tree's form and growth; tree class code = 4 in the FIA database); (2) they were located on non-stocked lands (areas which were considered forest land but were still in an early or arrested stage of reforestation; forest type code = 999 in the FIA database); and (3) their height or diameter measurements were missing. Tree data were then used to compute stand-level variables given in the following section.

Model Variables

To better capture the site condition at each installation, a forested condition class on a plot, denoted as condition (or plot condition), was treated as a basic sampling unit. All variables were calculated from live trees on a "condition class" rather than a plot. In other words, each plot's condition was treated as an individual stand of trees. For a given sampling unit, repeated measurements were denoted as observations. For example, plot condition A was remeasured in 2002 and 2009. Plot condition A is a sampling unit while measurements taken in 2002 and 2009 are two separate observations for a given sampling unit. Observations with a large diameter (over 40 cm) but low carbon stocking (50–100 Mg C/ha) that deviated from the general trend of mean diameter and carbon per ha were considered potential outliers and manually removed. A total number of 9,175 observations were used in analyses. Details descriptions of all variables used are given below.

Total carbon per unit area (Mg C/ha) at the condition level was used as a response variable, while stand variables were used as predictor variables in the models. Specifically, the aboveground carbon of each individual live tree was estimated with the updated carbon and biomass equations implemented by FIA (Westfall et al., 2023). Then, the estimated carbon (c, kg) of all trees on a sampling unit were used to compute the total carbon per unit area (C, Mg C/ha), which can be written as:

$$C = \frac{\sum c * tph}{1000} \tag{1}$$

where tph is the number of trees per hectare represented by a sampled tree, and 1,000 is the conversion factor from kg to Mg. In addition, six stand characteristics used as predictors were computed for every sampling unit (i.e., plot condition), including mean diameter at breast height (D, cm), exponential D (Exp. D), mean total tree height (H, m), mean h–d ratio (H–D ratio[1]), number of trees per hectare (TPH), and basal area per hectare (BA, m^2/ha). A summary of stand variables is given in Table 2.

Statistical Models

A hybrid model is composed of two components: a mixed model and a random forest algorithm, which were built sequentially. First, a linear mixed-effects model was built with all predictors as fixed-effects variables and forest type as a random-effect variable. Namely,

$$C = \sum x\beta + b + \epsilon \tag{2}$$

where $\sum x\beta$ is a linear combination of predictors (x) and their corresponding coefficients (β), b is a forest-type-specific random effect, and ϵ is the error term. In this study, we assumed that b follows a normal distribution with mean zero and forest-type-specific variance σ^2_{fort} $\left(i.e.,\ b \sim N\left(0, \sigma^2_{fort}\right)\right)$, and ϵ follows a normal distribution with mean zero and variance $\sigma^2\left(i.e., \epsilon \sim N\left(0, \sigma^2\right)\right)$.

Secondly, a random forest model was trained with the same predictors and the residuals (e_c, observed C – predicted C) calculated from the mixed model as the response variable. The final prediction of total carbon $\left(\hat{C}_{final}\right)$ is made by combining

[1] H–D ratio was computed as mean(h*100/d) where h and d are total tree height (m) and dbh (cm) of each sampled tree, respectively.

the predicted C ($\widehat{C}$) from the linear mixed model and the predicted residual $\left(\widehat{e}_c\right)$ from random forests, which can be written as:

$$\widehat{C}_{final} = \widehat{C} + \widehat{e}_C \quad (3)$$

The "lmer" function was used for fitting the mixed models, and the "ranger" function was used for training random forests in R (Bates et. al 2024; Wright and Ziegler 2017). For a given national forest (or park), the hyperparameters: (1) the number of predictors randomly selected from the total predictors at each split (mtry), and (2) the minimum number of observations that a leaf node can contain (min.node.size), were tuned using a grid search and a five-fold cross validation using "trainControl", "expand.grid" and "train" functions (Kuhn et al. 2023). According to various studies (e.g., Probst et al 2018; Bartz et al 2022), the grid searches for mtry values between 2 and 10 in increments of 2, while min.node.size is set to 1, 5, and 10. The other hyperparameters were set to their default values in the ranger function.

Model Building and Evaluation

To quantify the uncertainty of the model performance, the hybrid model (linear mixed model + random forests) was built using 1,000 cluster bootstrap samples. In cluster bootstrap, resampling was done at the sampling unit level to account for the potential

Table 2 Summary statistics of stand variables among eight national forests (NF) and two national parks (NP) in the southern Appalachian mountains. Stand variables include mean diameter at breast height (D, cm), exponential D (Exp. D), mean total tree height (H, m), mean h–d ratio (H–D ratio), number of trees per hectare (TPH), and basal area per hectare (BA, m^2/ha). Means are given followed by standard deviations in parentheses

Study site	D	Exp. D	H	H–D ratio	TPH	BA
Chattahoochee-Oconee NF	23.2 (5.2)	4.48e + 14 (5.48e + 15)	18.2 (3.5)	90.6 (13.4)	1021.6 (757.2)	23.6 (11.7)
Cherokee NF	22 (6.1)	1.98e + 14 (4.37e + 15)	17.2 (4.5)	91.5 (15.8)	1130.4 (905)	22.4 (12.1)
Nantahala NF	23.2 (5.7)	1.84e + 14 (1.44e + 15)	17.6 (4.1)	89.4 (14.5)	1088.5 (890.9)	23.3 (10.9)
Pisgah NF	23.5 (6.3)	5.71e + 14 (5.57e + 15)	17.5 (4.5)	86.8 (16.2)	1060.8 (902.4)	24.6 (12.9)
George Washington-Jefferson NF	22.6 (5.1)	4.98e + 14 (7.20e + 15)	16.1 (3.8)	82.8 (15.6)	1103.9 (800.9)	23.4 (10.4)
Daniel Boone NF	22 (5.2)	1.26e + 13 (8.91e + 13)	18 (3.8)	97.1 (14.3)	1179.6 (985.9)	20.2 (10.5)
Monongahela NF	25.8 (5.3)	1.20e + 15 (9.46e + 15)	18.2 (3.8)	78.9 (13)	721 (754.7)	27.4 (11.7)
Wayne NF	24.5 (5.3)	5.66e + 14 (5.34e + 15)	18.5 (3.7)	84.3 (12.9)	520.6 (567.9)	16.9 (9.4)
Great Smoky Mountains NP	25 (5.3)	2.28e + 15 (1.64e + 16)	18.8 (4.2)	86.4 (13.2)	1002 (709.6)	27.6 (12.7)
Shenandoah NP	23.5 (5.6)	5.60e + 13 (2.31e + 14)	15.3 (3.8)	78 (10.5)	1074.1 (618.8)	24.2 (8.8)

temporal autocorrelation among observations (i.e., repeated measurements) on a given sampling unit (i.e., plot condition). If a plot condition is sampled, all repeated measurements on a plot condition are selected and grouped into a training dataset, denoted as in-bag data. The unselected sampling units are then used as a test dataset, denoted as out-of-bag data. At each bootstrap round, the importance score of each predictor was calculated, and evaluation statistics (see Section "Evaluation statistics") were computed using both in-bag and out-of-bag data. The medians of 1,000 bootstrap estimates were used to represent the importance score and evaluation statistics for a given national forest or national park. The upper and lower bounds of 95% confidence intervals of evaluation statistics were computed by the 2.5% and 97.5% quantiles of 1,000 bootstrap samples. For a given study site, one out-of-bag sample was used to plot the pairs of observations and predictions. A 1:1 diagonal line and a LOWESS smoothing curve with its 95% confidence interval were added to assess the model's prediction accuracy. LOWESS smoothing curve with its 95% confidence interval were created using the geom_smooth function in the R package ggplot2 (Wickham et al. 2024).

The importance score was used to evaluate the contribution of model prediction accuracy from a predictor. The greater score indicates the predictor is more important. In this study, a permutation approach was used to calculate the importance score, which is the average differences of prediction accuracy between the reference and permutated-predictor trees. The reference tree is defined as the tree grown with the original set of predictor and response variables. Then, a new dataset is constructed by randomly shuffling the values of a predictor and keeping pairs of all other predictors and response unchanged. The permutated-predictor tree is built using the new dataset, and the prediction accuracy of the permutated-predictor tree is computed as the reference tree. Detailed descriptions of the approach can be found in Wright and Ziegler (2017). The importance score was provided by random forests using the "ranger" function in R (Wright and Ziegler 2017).

Evaluation Statistics

The prediction accuracy from the hybrid model was evaluated using mean bias (MB) and root-mean-square error (RMSE), which were computed as:

$$MB = \frac{\sum e}{n} \tag{4}$$

$$RMSE = \sqrt{\frac{\sum e^2}{n}} \tag{5}$$

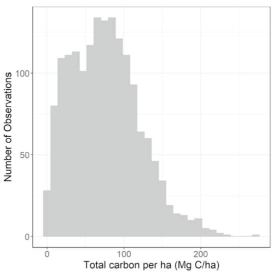
where $\overline{C}$ is the mean of C, n is total number of observations, and e is prediction residual, which was defined as:

$$e = C - \widehat{C}_{final} \tag{6}$$

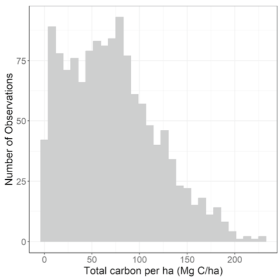
where all variables have been defined as above. R script for statistical analyses is available upon request.

Results

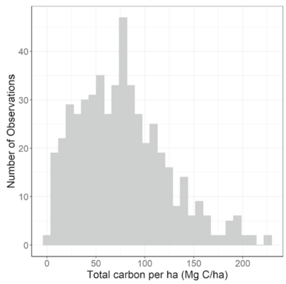
As shown in Fig. 2, total carbon per unit area exhibits right-skewed distributions across all study sites. The average carbon per hectare varies among study sites, ranging from 58.2 to 91.5 Mg C/ha. Monongahela National Forest (91.5 Mg C/ha) and Great Smoky Mountains National Park (92.2 Mg C/ha) have higher average carbon per hectare than other sites (58.2–81.0 Mg C/ha) (Table 3). In general, national forests share similar shapes of distributions where observations concentrate below 200 Mg C/ha of forest carbon (Fig. 2). Seven out of eight national forests have a peak between 50 and 100 Mg C/ha. The Cherokee and Pisgah National Forests show a relatively even distribution between 0 and 100 Mg C/ha, while the Monongahela National Forest has the highest count around 100 Mg C/ha. Both national parks share similar patterns with



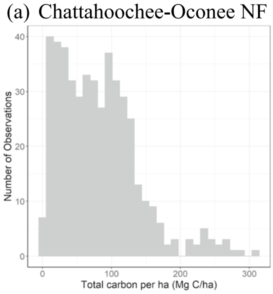
(a) Chattahoochee-Oconee NF



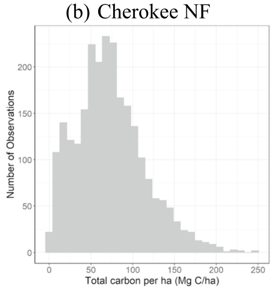
(b) Cherokee NF



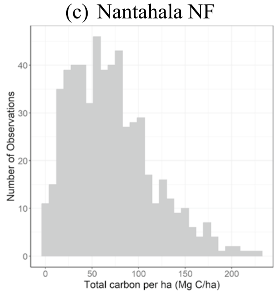
(c) Nantahala NF



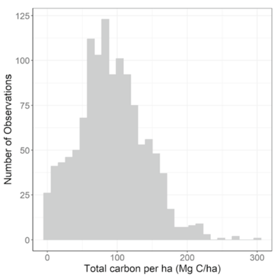
(d) Pisgah NF



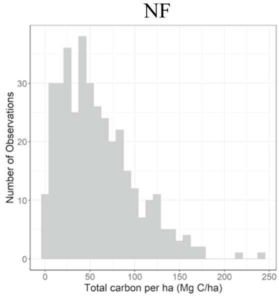
(e) George Washington-Jefferson NF



(f) Daniel Boone NF



(g) Monongahela NF



(h) Wayne NF

Fig. 2 Distribution of total carbon per hectare (Mg C/ha) among eight national forests (NF) in the southern Appalachian mountains

Table 3 Average total carbon per hectare (Mg C/ha) in eight national forests (NF) and two national parks (NP) both overall and by forest type group. Means are given followed by standard deviations in parentheses

Study site	Total	Asp/Bir	Elm/Ash/Cw	Exo Hw	Lob/SLP	Map/Bch/Bir	Oak/Gum/Cyp	Oak/Hic	Oak/Pine	Other Hw	Spr/Fir	WP/RP/JP
Chattahoochee-Oconee NF	76 (44.6)	0	60.4 (36.8)	0	55.4 (42.4)	116.8 (0.0)	0	79 (43.8)	71.4 (41.5)	81.6 (44.7)	0	87.8 (59.6)
Cherokee NF	70.9 (46.7)	0	0	0	37.9 (30.6)	85.6 (47.5)	0	76.7 (47.7)	58.5 (38.7)	87.7 (51.6)	7 (0)	73.2 (40.4)
Nantahala NF	76.9 (45.4)	0	0	0	50.6 (44.3)	61.6 (39.9)	0	80.6 (45.6)	57.5 (38.0)	67.9 (50.3)	0	64.7 (39.3)
Pisgah NF	81 (57.1)	4.2 (3.4)	0	0	32.7 (11.5)	47.6 (42.4)	0	88.3 (59.4)	64.2 (30.7)	50.2 (35.9)	90.1 (16.9)	41.7 (35.9)
George Washington-Jefferson NF	72.9 (41.3)	0	0	1.8 (0)	39.4 (22.1)	94.2 (45.6)	57.9 (6.8)	76.8 (41.5)	55.6 (29.0)	19.5 (16.8)	110.3 (12)	58.1 (33.3)
Daniel Boone NF	70.3 (44.1)	0	14.9 (1.3)	0	27.7 (22.6)	68 (48.7)	0	71.7 (44)	50.9 (28.0)	0	0	91.2 (41.8)
Monongahela NF	91.5 (47.3)	52 (30.8)	3.2 (0.4)	0	50.3 (10)	96.2 (46.2)	0	91.3 (48.9)	67.6 (18.2)	77.1 (51)	62.9 (35.8)	95.8 (39.3)
Wayne NF	58.2 (41)	78.3 (34.4)	56.4 (32)	0	24.5 (17.6)	54.3 (49.8)	0	61.2 (41.8)	50 (28.8)	2.3 (0)	0	32.7 (24)
Great Smoky Mountains NP	91.1 (47.7)	0	72.1 (0)	0	50.6 (34.6)	85.8 (46.3)	16.4 (0)	97.3 (49.2)	79.1 (32.5)	81.2 (43.1)	135.5 (45.4)	98.5 (25.6)
Shenandoah NP	76.4 (37.8)	0	1.4 (0)	0	0	0	0	76.6 (37.6)	46.7 (0)	101.3 (16.3)	0	0

Forest type is used in model fitting. Due to space constraints, only a summary of forest type groups is included here. A table detailing all forest types is given in Table S2 in the supplementary material. Eleven forest types include: aspen/birch (Asp/Bir), elm/ash/cottonwood (Elm/Ash/Cw), exotic hardwoods (Exo Hw), Loblolly/shortleaf pine (Lob/SLP), Maple/beech/birch (Map/Bch/Bir), Oak/gum/cypress (Oak/Gum/Cyp), Oak/hickory (Oak/Hic), Oak/pine, Other hardwoods (Other Hw), Spruce/fir (Spr/Fir), and White/red/jack pine (WP/RP/JP)

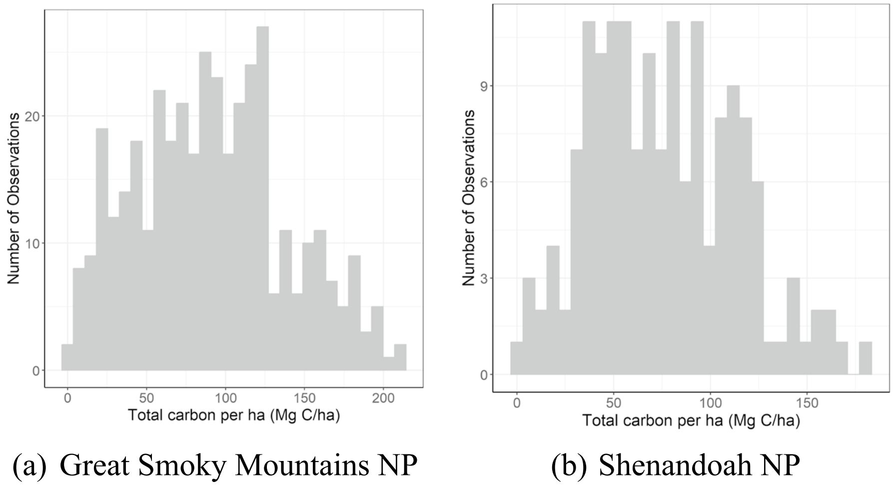


Fig. 3 Distribution of total carbon per hectare (Mg C/ha) among two national parks (NP) in the southern Appalachian mountains

increasing counts to 120 Mg C/ha, followed by a decline up to 200 Mg C/ha (Fig. 3). When examining forest carbon among forest type groups, the highest carbon per hectare varies across study sites (Table 3). For example, aspen/birch forests have the greatest carbon stocking at the Wayne National Forest, whereas the spruce/fir group is the forest type with the highest carbon stocking in the Great Smoky Mountains National Park. Oak/hickory and oak/pine forest types are present in all study sites. For a given study site, oak/hickory forests (61.2–97.3 Mg C/ha) have greater total carbon per hectare than oak/pine forests (46.7–79.1 Mg C/ha). The highest carbon stocking for both forest type groups was found in the Great Smoky Mountains National Park (97.3 Mg C/ha for oak/hickory forests and 79.1 Mg C/ha for oak/pine forests) (Table 3).

In general, both linear mixed model and hybrid model can provide unbiased predictions of total carbon for all study sites based on the 95% confidence intervals of MB for both in-bag and out-of-bag data (Table 4). Regarding the precision level, the hybrid model produced lower RMSE than the linear mixed model for all sites. For out-of-bag samples, RMSE ranges from 7.2 to 13.7 Mg C/ha for the hybrid model, compared to 7.7 to 15.5 Mg C/ha for the linear mixed model. The greatest precision was found for the Shenandoah National Park (7.2 Mg C/ha) and George Washington-Jefferson National Forest (7.5 Mg C/ha) whereas the Nantahala (13.7 Mg C/ha) and Pisgah (12.0 Mg C/ha) National Forests have a lower precision level than other study sites. The accuracy level is higher when predicting in-bag samples than out-of-bag samples, which is expected as in-bag observations were used in model building. As shown in Figs. 4 and 5, except for high C-stocking observations (100–250 Mg C/ha), the 1:1 diagonal lines lie within the 95% confidence intervals of the smoothing curves. This implies that observed and predicted total carbon values are similar, particularly in George Washington-Jefferson National Forest. In addition, the level of importance of predictor varies among study sites. In general, diameter-based predictors (D, Exp. D and BA) are more important than other variables, especially D (mean diameter at

Table 4 Evaluation statistics of mean bias (MB) and root mean square error (RMSE) for eight national forests (NF) and two national parks (NP) in the southern Appalachian mountains. The evaluation statistics were computed from the medians of 1,000 bootstrap samples using the in-bag data (IB, training data) and out-of-bag data (OOB, validation data)

	MB				RMSE			
Linear mixed model	IB	95% CI	OOB	95% CI	IB	95% CI	OOB	95% CI
Chattahoochee-Oconee NF	0.0	0.0–0.0	0.0	−1.5–1.5	9.3	8.7–9.8	10.0	9.0–11.0
Cherokee NF	0.0	0.0–0.0	0.0	−2.0–2.2	10.2	9.4–10.9	11.2	9.9–22.6
Nantahala NF	0.0	0.0–0.0	0.0	−3.6–3.5	10.7	8.0–12.1	11.6	8.5–15.1
Pisgah NF	0.0	0.0–0.0	0.0	−5.1–5.6	12.4	9.6–13.8	15.5	12.6–27.7
George Washington-Jefferson NF	0.0	0.0–0.0	0.0	−1.1–1.1	8.5	8.1–8.8	9.0	8.3–9.7
Daniel Boone NF	0.0	0.0–0.0	0.0	−2.8–3.3	9.3	8.2–10.2	11.1	9.1–13.0
Monongahela NF	0.0	0.0–0.0	0.0	−2.0–2.4	12.0	11.2–12.7	12.9	11.7–14.3
Wayne NF	0.0	0.0–0.0	0.0	−2.7–3.5	9.0	7.9–9.9	10.9	9.1–13.7
Great Smoky Mountains NP	0.0	0.0–0.0	0.0	−3.4–3.7	9.4	7.9–10.3	11.4	9.5–13.5
Shenandoah NP	0.0	0.0–0.0	0.0	−3.8–4.3	5.7	4.6–6.6	7.7	6.1–10.4
Hybrid model								
Chattahoochee-Oconee NF	0.0	0.0–0.1	0.2	−1.2–1.4	3.2	3.0–3.4	8.8	7.8–9.7
Cherokee NF	0.0	0.0–0.1	0.2	−1.7–2.0	3.4	3.2–3.6	9.5	8.4–21.5
Nantahala NF	0.1	0.0–0.1	0.2	−3.3–3.9	3.5	2.5–4.3	12.0	8.1–15.0
Pisgah NF	0.1	0.0–0.2	0.7	−3.8–5.7	3.8	2.9–4.3	13.7	10.3–27.4
George Washington-Jefferson NF	0.0	0.0–0.0	0.1	−0.8–1.0	3.0	2.9–3.2	7.5	6.8–8.3
Daniel Boone NF	0.1	0.0–0.1	0.5	−2.4–3.4	4.3	3.8–4.7	10.2	8.3–12.2
Monongahela NF	0.0	0.0–0.1	0.4	−1.6–2.2	4.7	4.3–4.9	11.2	10.0–12.4
Wayne NF	0.0	−0.1–0.1	0.6	−2.3–3.4	3.5	3.1–3.9	10.0	8.0–12.9
Great Smoky Mountains NP	0.1	0.0–0.2	0.5	−2.7–4.1	3.6	2.9–4.0	10.8	8.7–13.1
Shenandoah NP	0.0	−0.1–0.1	−0.1	−3.3–3.8	2.4	2.0–2.9	7.2	5.7–10.2

breast height). BA ranks the top predictor for five of ten sites, while H–D ratio ranks the least important variable for five out of eight national forests (Table 5). For two national parks, TPH has the least importance score among all six predictors. The estimated coefficients of fixed and random effects for all study sites are given in Table 6.

Discussion

Comparison of Forest Carbon Among Different Study Sites

Smith et al. (2019) examined the aboveground live carbon per hectare across various federal lands in the US (e.g., USDA Forest Service (Forest Service), National

Fig. 4 Observed and predicted total carbon per hectare (Mg C/ha) for out-of-bag samples for Chattahoochee-Oconee, Cherokee, Nantahala, Pisgah, George Washington-Jefferson and Daniel Boone National Forests (NF). The blue curves represent the LOWES smoothing curve (solid blue) and its 95% confidence interval (dashed blue)

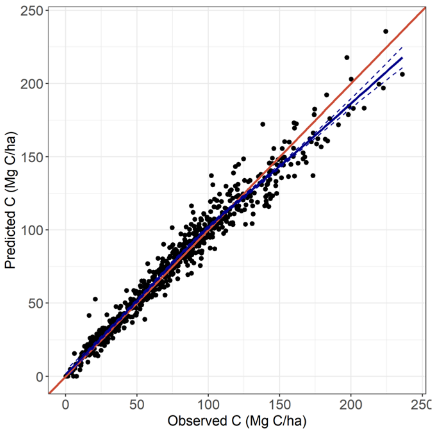
Park Service, Bureau of Land Management) It was reported that Forest Service lands have an average of 73.1 Mg C/ha, which is higher than the 63.1 Mg C/ha on National Park Service lands in the southern US (Smith et al. 2019). The density of forest carbon at the study sites (70.3–91.5 Mg C/ha, except for Wayne National Forest at 58.2 Mg C/ha) is close to or greater than the average estimates reported by Smith et al. (2019), particularly for national parks (76.4–91.1 Mg C/ha). The differences may be due to: (1) Smith et al. (2019) covering a broader geographical area than this study within the southern US, as our study sites represent only a portion of the lands owned by the Forest Service and National Park Service in the region, and (2) the study sites being located in some of the highest carbon stocking areas in the southern US (Wilson et al. 2013).

Hoover and Smith (2021) further broke down the estimates of aboveground live carbon by forest type groups and regions. Their estimates for loblolly/shortleaf pine forests (57.3 Mg C/ha) are higher than those in this study (24.5–55.4 Mg C/ha). In contrast, our estimates for oak/hickory (71.7–97.3 Mg C/ha, except for Wayne National Forest at 61.3 Mg C/ha) forests exceed the estimate (69.8 Mg C/ha) provided by Hoover and Smith (2021). The differences of carbon estimates between the two studies reflect the different distribution of forest types within the region. Hoover and Smith (2021) covered the entire southeastern US, from the coastal plain to the Piedmont, without distinguishing smaller geographical areas. Consequently, their higher average carbon stocking may be due to the inclusion of planted pine forests, which are more prevalent in coastal areas than in the southern Appalachians. In contrast, our study sites are primarily dominated by oaks, hickories, and other hardwood species, which store more carbon per hectare here than similar forest types in other areas of the southern US.

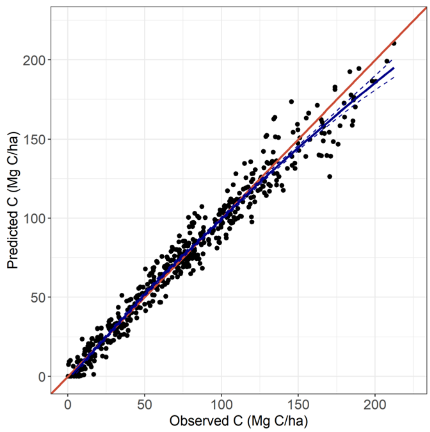
Evaluation of Using the Hybrid Model

As the results show, the hybrid model can reliably quantify forest carbon at the stand level, even for data not used in model building. By adding the random forest model along with the linear mixed model, the model provides unbiased predictions as well as the precision level improved for all study sites (Table 4). The model predictability is consistent across all study sites. For several study sites, it is noted that the level of prediction accuracy reduces as carbon stocking increases (e.g., Cherokee, Wayne, Monongahela) (Figs. 4 and 5). It may be because the number of high stocking plots is relatively low on the study sites. The models were not able to "fully learn" from the data to make more accurate predictions compared to lower stocking plots.

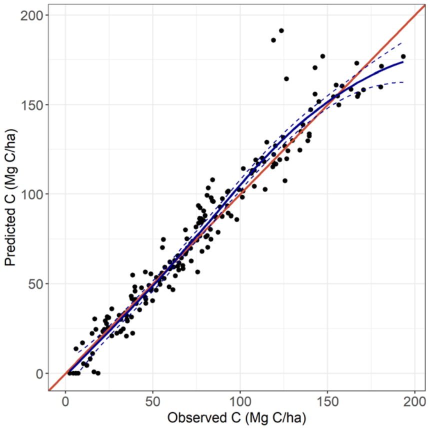
To simulate a common setting in forest practice, only stand variables were used in model building. All variables used in this work can be derived from common measurements recorded in a typical forest inventory (e.g., DBH, height, and number of



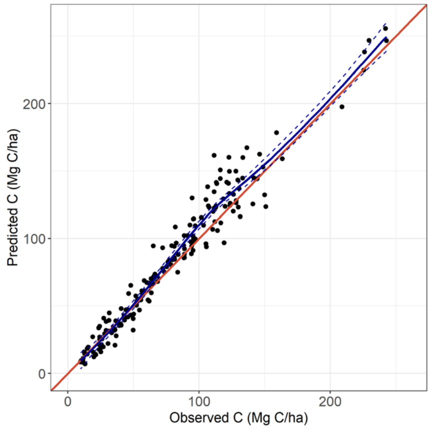
(a) Chattahoochee-Oconee NF



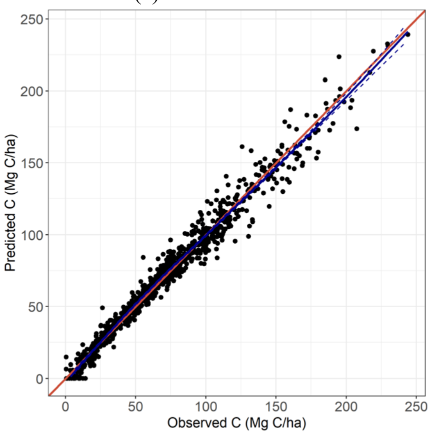
(b) Cherokee NF



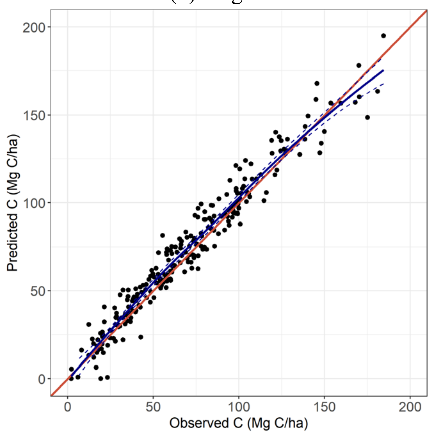
(c) Nantahala NF



(d) Pisgah



(e) George Washington-Jefferson NF



(f) Daniel Boone NF

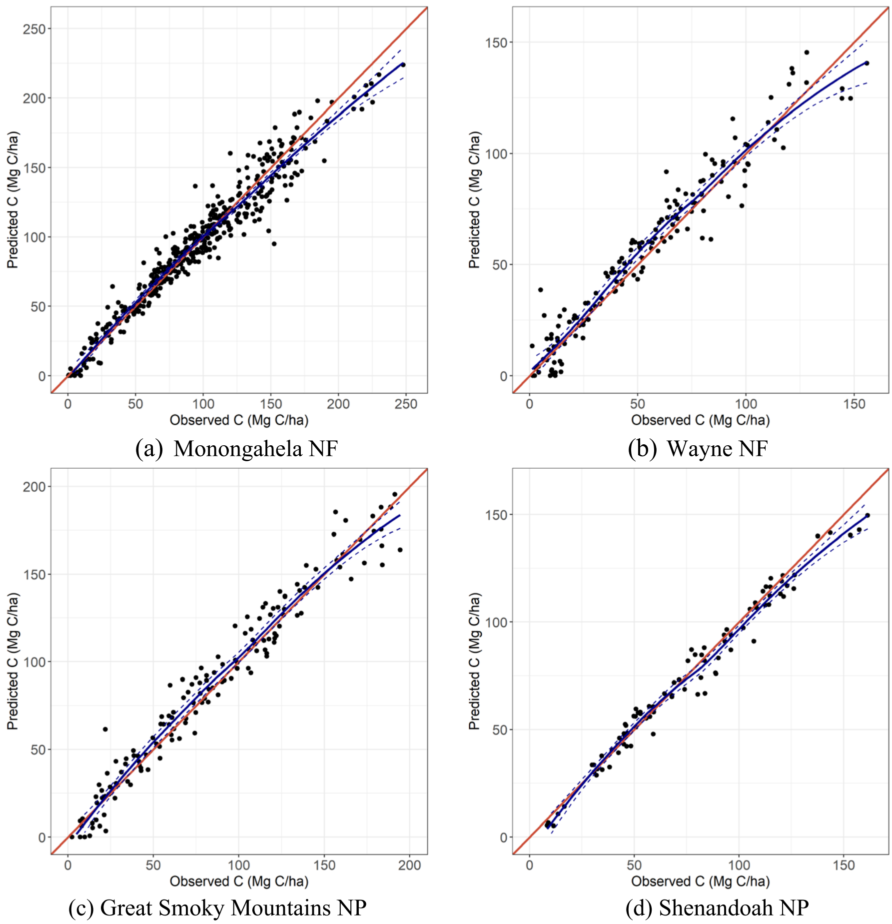


Fig. 5 Observed and predicted total carbon per hectare (Mg C/ha) for out-of-bag samples for Monongahela and Wayne National Forests (NF) and Great Smoky Mountains and Shenandoah National Parks (NP) in the southern Appalachian mountains. The blue curves represent the LOWES smoothing curve (solid blue) and its 95% confidence interval (dashed blue)

trees). Similar results were reported by Khan et al (2020) where the model with basal area per hectare, mean tree height and wood density yielded more accurate predictions than other candidates for Bangladeshi forests. In our preliminary analysis, a few additional variables, such as slope, aspect, and species diversity index, were examined, but they did not appreciably improve prediction accuracy. Thus, these variables were not included in the formal analysis to maintain the parsimony of the model. In addition, obtaining additional variables may require extra effort and cost, which may not be practical in routine forest inventories. Besides, deriving certain variables (e.g., climatic variables) may require the location information of ground sample plots to pair with remote sensing data. As the actual plot GPS coordinates are confidential in FIA, general users will not be able to access this information via public database.

Table 5 Importance score of all predictors for eight national forests (NF) and two national parks (NP) in the southern Appalachian mountains

	Chattahoochee-Oconee NF		Cherokee NF		Nantahala NF		Pisgah NF		George Washington-Jefferson NF	
Variable	Imp. score	Rank	Imp. score	Rank	Imp. score	Rank	Imp. score	Rank	Imp. score	Rank
D	37.7	2	36.6	3	39.6	1	39.2	2	31.0	1
Exp. D	37.6	3	35.9	4	39.5	2	38.2	3	31.0	1
H	24.1	6	45.9	1	24.9	4	36.9	4	30.3	3
H–D ratio	29.9	5	28.3	6	20.6	6	23.3	6	24.5	5
TPH	36.0	4	30.1	5	25.7	3	27.7	5	21.8	6
BA	39.1	1	37.9	2	22.7	5	77.3	1	27.6	4
	Daniel Boone NF		Monongahela NF		Wayne NF		Great Smoky Mountains NP		Shenandoah NP	
Variable	Imp. score	Rank	Imp. score	Rank	Imp. score	Rank	Imp. score	Rank	Imp. score	Rank
D	25.6	2	64.3	1	18.5	1	19.1	3	5.1	2
Exp. D	25.1	3	63.8	2	18.2	2	18.9	5	5.1	2
H	14.4	5	49.7	3	7.5	6	19.0	4	3.4	4
H–D ratio	10.9	6	34.4	6	9.8	5	21.5	2	2.6	5
TPH	18.9	4	49.1	4	18.0	3	14.7	6	2.3	6
BA	25.8	1	41.7	5	16.7	4	34.2	1	5.6	1

Table 6 Estimated coefficients of fixed and random effects for eight national forests (NF) and two national parks (NP) in the southern Appalachian mountains

Study site	Intercept	D	Exp. D	H	H–D ratio	TPH	BA	σ_{fort}	σ
Chattahoochee-Oconee NF	−118.12	3.48	4.82*10^−18	−2.22	0.85	−0.010	3.59	5.35	9.41
Cherokee NF	−84.20	2.62	2.06*10^−17	−1.59	0.61	−0.009	3.52	5.46	10.37
Nantahala NF	−96.06	2.31	−5.93*10^−16	−1.65	0.78	−0.010	3.80	4.76	10.99
Pisgah NF	−145.24	5.05	−2.91*10^−16	−4.04	1.08	−0.010	3.86	9.17	12.82
George Washington-Jefferson NF	−107.04	2.53	−6.36*10^−17	−1.03	0.77	−0.009	3.54	6.37	8.57
Daniel Boone NF	−96.16	2.49	4.59*10^−15	−1.68	0.68	−0.010	4.21	7.04	9.61
Monongahela NF	−117.29	2.70	−1.78*10^−16	−1.33	0.88	−0.006	3.51	9.13	12.23
Wayne NF	−100.03	3.03	2.64*10^−16	−2.26	0.76	−0.011	3.78	7.45	9.47
Great Smoky Mountains NP	−103.60	2.18	−3.84*10^−17	−0.61	0.70	−0.008	3.56	7.06	9.90
Shenandoah NP	−91.31	2.49	3.48*10^−15	−1.74	0.81	−0.010	3.42	3.74	6.06

However, several studies have shown that adding environmental or auxiliary variables (e.g., climate, topography, etc.) to the model can enhance prediction accuracy, particularly for mixed-species forests (e.g., Aguirre et al. 2021, Liepiņš et al. 2022; Yang et al. 2022). For example, Martonne aridity index derived from annual precipitation and mean temperature has a positive correlation with stand dry weight for natural pine forests in the Spanish Iberian Peninsula (Aguirre et al. 2021). Liepiņš et al (2022) found that incorporating species-related metrics into the model provided better predictions of stand-level biomass for hemiboreal forests in Latvia. Yang et al (2022) showed that model prediction accuracy was improved with the inclusion of climatic and topographic variables for Caribbean trees. Since the random forest algorithm is well-suited for handling high-dimensional data without requiring predefined relationships among variables, the model used in this study can be extended by incorporating additional predictors, making it useful for future follow-up studies.

To quantify forest carbon at the stand level, one approach is to estimate individual tree carbon using species-specific equations and then aggregate the estimated carbon of all sample trees to the entire stand. Another approach, as demonstrated in this study, is the direct use of stand-level models, which provide a quick estimate of total forest carbon based on forest type and stand-level predictors. While obtaining these predictors may require tree-level data, the selected variables do not depend on species identification. As a result, identifying the species of every sample tree is unnecessary, reducing time and costs in data collection and processing, particularly in species-rich ecosystems. Future studies may assess the model's predictive accuracy using metrics derived from remote sensing data, such as LiDAR and satellite imagery, potentially further lowering field data collection costs.

As indicated previously, this study focuses on forest carbon on federal lands to address a research gap in the region. The proposed methodology can be applied to other ownership lands (e.g., privately-owned lands) in the region. In addition, total carbon is predicted using both estimated fixed and random effects if forest types are included in model training. However, for a forest type not used in model training, only fixed effects are applied for prediction in the validation data. To predict total carbon for a new forest type, it is recommended to retrain the model by incorporating new observations. Since random forest is an ensemble algorithm, it may produce poor predictions outside the range of the training dataset (i.e., extrapolation), as noted by Yang and Burkhart (2020). This limitation may also explain the poor prediction performance for high-carbon stocking plots in this study. While this study employed a linear mixed model as a working example, more complex models (e.g., nonlinear mixed models) or alternative machine learning algorithms could potentially enhance prediction accuracy. Lastly, we did not examine temporal changes or the effects of different management practices on carbon storage across forest types in national forests and parks, which is suggested as a direction for future research.

Conclusion

In summary, the average aboveground live carbon per hectare varies among national forests and parks in the southern Appalachian mountains, ranging from 58.2 to 91.5 Mg C/ha. Monongahela National Forest and Great Smoky Mountains National Park have higher average carbon per hectare than other sites. For a given study site, forest carbon per hectare varies among different forest type groups. This study provides a working example of using a hybrid model with a combination of linear mixed model and random forest algorithm for stand-level, forest carbon predictions. In a typical carbon inventory, stand-level carbon is estimated by aggregating individual tree carbon computed from species-specific equations. However, in the species-rich ecosystem, identifying the species of every sample tree can be challenging and time consuming. The model developed in this study, which incorporates forest type and common stand variables (e.g., mean DBH, mean total tree height, trees per hectare, and basal area per hectare), offers an alternative method for stand-level carbon assessment without relying on species identification. As parametric models remain the most popular quantitative methods in forestry practice, the proposed method retains the interpretability and portability of the parametric models, but also improves prediction accuracy by leveraging the strengths of the machine learning algorithm. With the increasing availability of programming tools and publicly accessible databases, implementing such models in forestry is expected to become more feasible. Further research is warranted to comprehensively evaluate this method and its limitations under diverse site conditions.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s44392-025-00044-x.

Acknowledgements Support from USDA Forest Service (project # 21-CR-11330145-070) and Warnell School of Forestry and Natural Resources is greatly appreciated. We are grateful for the insightful comments from the Editors and four anonymous reviewers.

Author's Contribution Conceptualization – Yang, Brandeis, Skiba; Data curation – Brandeis; Formal analysis – Yang; Methodology – Yang, Brandeis; Writing—original draft – Yang; Writing—review & editing – Brandeis, Skiba.

Data Availability The FIA data can be found at https://research.fs.usda.gov/products/dataandtools/tools/fia-datamart.

Declarations

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

References

Aguirre, A., M. del Río, R. Ruiz-Peinado, and S. Condés. 2021. Stand-level biomass models for predicting C stock for the main Spanish pine species. *Forest Ecosystems* 8:29. https://doi.org/10.1186/s40663-021-00308-w.

Bailey, R. G. 2016. Bailey's ecoregions and subregions of the United States, Puerto Rico, and the U.S. Virgin Islands. *Fort Collins, CO: USDA Forest Service Research Data Archive*. https://doi.org/10.2737/RDS-2016-0003. Accessed 6 Sept 2024.

Bartz, E., T., Bartz-Beielstein, M., Zaefferer and O. Mersmann. 2022. Hyperparameter Tuning for Machine and Deep Learning with R. Springer. 323.

Bates, D., Maechler M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B. Scheipl, F., Grothendieck. G., Green, P., Fox, J., Bauer, A., Krivitsky, P. N., Tanaka, E. and M. Jagan. 2024. R Package "lme4" - Linear Mixed-Effects Models using 'Eigen' and S4. 129.

Burrill, E. A., G. A. Christensen, B. L. Conkling, A. M. DiTommaso, L. C. Lepine, C. J. Perry, S. A. Pugh, J. A. Turner, and D. M. Walker. 2024. The Forest Inventory and Analysis Database: FIADB User Guides-Volume Database Description (version 9.2). *US Department of Agriculture Forest Service* 1042. https://research.fs.usda.gov/understory/forest-inventory-and-analysis-database-user-guide-nfi. Accessed 26 Aug 2024.

Cornelio, J., S. Mohd Razak, Y. Cho, H. Liu, R. Vaidya, and J. Behnam. 2022. Residual learning to integrate neural network and physics-based models for improved production prediction in unconventional teservoirs. *SPE Joural* 22:3328–3350. https://doi.org/10.2118/210559-PA.

de Lima Duarte, F. C., P. S. G. de Mattos Neto, and P. R. A. Firmino. 2024. A hybrid recursive direct system for multi-step mortality rate forecasting. *Journal of Supercomputing* 80:18430–18463. https://doi.org/10.1007/s11227-024-06182-x.

de MattosNeto, P. S. G., G. D. C. Cavalcanti, and O. D. S. de Santos Júnior. 2022. Hybrid systems using residual modeling for sea surface temperature forecasting. *Scientific Reports* 12:487. https://doi.org/10.1038/s41598-021-04238-z.

Ellison, A. M., D. A. Orwig, M. C. Fitzpatrick, and E. L. Preisser. 2018. The past, present, and future of the hemlock woolly adelgid (*Adelges tsugae*) and its ecological interactions with eastern hemlock (*Tsuga canadensis*) forests. *Insects* 9:172.

Henson, M. B., O. D. Kolawole, and O. Ayeni. 2014. Bankfull curves for the temperate rainforests in the Southern Appalachian Mountains of Western North Carolina. *Cinq Continents* 4 (9): 5–15.

Hoover, C. M., and J. E. Smith. 2021. Current aboveground live tree carbon stocks and annual net change in forests of conterminous United States. *Carbon Balance and Management* 16:17. https://doi.org/10.1186/s13021-021-00179-2.

Jenkins, M. A. 2007. Vegetation communities of Great Smoky Mountains National Park. *Southeastern Naturalist* 6 (2): 35–56.

Jiang, P., C. Wang, J. Fan, T. Ji, L. Mu, X. Lu, and J. Zhu. 2024. Hybrid residual modelling of biomass pyrolysis. *Chemical Engineering Science* 293:120096.

Keyser, P. D., T. Fearer, and C. A. Harper. 2015. *Managing oak forests in the eastern United States*, 289. Taylor & Francis Group: CRC Press.

Khan, M. N. I., M. R. Islam, A. Rahman, M. S. Azad, A. S. Mollick, M. Kamruzzaman, M. N. Sadath, S. Feroz, M. G. Rakkibu, and A. Knohl. 2020. Allometric relationships of stand level carbon stocks to basal area, tree height and wood density of nine tree species in Bangladesh. *Global Ecology and Conservation* 22:e01025.

Koo, K., B. C. Patten, and I. F. Creed. 2011. *Picea rubens* growth at high versus low elevations in the Great Smoky Mountains National Park: Evaluation by systems modeling. *Canadian Journal of Forest Research* 41 (5): 945–962. https://doi.org/10.1139/x10-243.

Kuehne, C., M. B. Russell, A. R. Weiskittel, and J. A. Kershaw. 2020. Comparing strategies for representing individual-tree secondary growth in mixed-species stands in the Acadian Forest region. *Forest Ecology and Management* 459:117823.

Kuhn, M., J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B., Kenkel, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang, C. Candan, T. Hunt. 2023. R Package 'caret' - Classification and Regression Training. 224.

Liepiņš, J., A. Lazdiņš, S. Kalēja, and K. Liepiņš. 2022. Species composition affects the accuracy of stand-level biomass models in hemiboreal forests. *Land* 11 (7): 1108. https://doi.org/10.3390/land11071108.

Phillips, P. D., I. Yasman, T. E. Brash, and P. R. van Gardingen. 2002. Grouping tree species for analysis of forest data in Kalimantan (Indonesian Borneo). *Forest Ecology and Management* 157 (1–3): 205–216.

Probst, P., M. Wright, and A. Boulesteix. 2018. Hyperparameters and Tuning Strategies for Random Forest. arXiv:1804.03515

Russell, M., A. R. Weiskittel, and J. A. Kershaw. 2014. Comparing strategies for modeling individual-tree height and height-to-crown base increment in mixed-species Acadian forests of northeastern North America. *European Journal of Forestry Research* 133:1121–1135.

Smith, J. E., G. M. Domke, M. C. Nichols, and B. F. Walters. 2019. Carbon stocks and stock change on federal forest lands of the United States. *Ecosphere* 10 (3): e02637. https://doi.org/10.1002/ecs2.2637.

Spencer, M., and A. C. B. J. Lewis. 2017. *Nantahala National Forest: a history*. History Press Library Editions.

USDA Forest Service. 2023. Land areas of the national forest system. FS-383. Washington, DC: *USDA Forest Service*. 273. https://www.fs.usda.gov/land/staff/lar/LA2023/LAR_Book_FY2023.pdf. Accessed 27 Aug 2024.

USDA Forest Service. 2023. Forest inventory and analysis national core field guide for the nationwide forest inventory. Version 9.3. Washington, DC. 539 p. USDA Forest Service, Forest Inventory and Analysis, 201 14th Street, Washington, DC 20250. https://research.fs.usda.gov/understory/nationwide-forest-inventory-field-guide. Accessed 26 Aug 2024.

Vanclay, J. K. 1991. Aggregating tree species to develop diameter increment equations for tropical rainforests. *Forest Ecology and Management* 42:143–168.

Westfall, J. A., Coulston, J. W., Gray, A. N., Shaw, J. D., Radtke, P. J., Walker, D. M., Weiskittel, A. R., MacFarlane, D. W., Affleck, D. L., Zhao, D., Temesgen, H., Poudel, K. P., Frank, J. M., Prisley, S. P., Wang, Y., Meador, A. J. S., Auty, D., & G. M., Domke. 2023. *A national-scale tree volume, biomass, and carbon modeling system for the United States*. https://doi.org/10.2737/wo-gtr-104Wright

Wickham, H., Chang, W., Henry, L., Pedersen, T. L., Takahashi, K., Wilke, C.. Woo, K., Yutani, H. Dunnington, D., van den Brand, T., and P. B. C. Posit. 2024. R Package "ggplot2" - Create Elegant Data Visualisations Using the Grammar of Graphics. 354.

Wilson, B. T., C. W. Woodall, and D. M. Griffith. 2013. Imputing forest carbon stock estimates from inventory plots to a nationally continuous coverage. *Carbon Balance and Management* 8:1. https://doi.org/10.1186/1750-0680-8-1.

Wright, M. N., and A. Ziegler. 2017. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software* 77 (1): 1–17.

Yang, S., and H. E. Burkhart. 2020. Robustness of parametric and nonparametric fitting procedures of tree-stem taper with alternative definitions for validation data. *Journal of Forestry* 118 (6): 576–583.

Yang, S., T. J. Brandeis, E. H. Helmer, M. P. Oatham, T. Heartsill-Scalley, and H. Marcano-Vega. 2022. Characterizing height-diameter relationships for Caribbean trees using mixed-effects random forest algorithm. *Forest Ecology and Management* 524:120507.

Zhang, J., R. L. Bras, M. Longo, and T. HeartsillScalley. 2022. The impact of hurricane disturbances on a tropical forest: implementing a palm plant functional type and hurricane disturbance module in ED2-HuDi V1. 0. *Geoscientific Model Development* 15 (13): 5107–5126. https://doi.org/10.5194/gmd-15-5107-2022.

Ziegler, M. N. A., and I. R. König. 2016. Do little interactions get lost in dark random forests? *BMC Bioinformatics* 17:145.

Authors and Affiliations

Sheng-I Yang[1] · Thomas J. Brandeis[2] · Tara Skiba[3]

✉ Sheng-I Yang
syang23@uga.edu

Thomas J. Brandeis
thomas.j.brandeis@usda.gov

Tara Skiba
tskiba@vols.utk.edu

1 Warnell School of Forestry and Natural Resources, University of Georgia, 180 E Green Street, Athens, GA 30602, USA

2 USDA Forest Service, Southern Research Station, 2730 Cherokee Farm Way, Suite 101, Knoxville, TN 37920, USA

3 School of Natural Resources, University of Tennessee, 2431 Joe Johnson Dr, Suite 401, Knoxville, TN 37996, USA