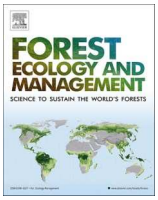
Forest Ecology and Management 572 (2024) 122334

Contents lists available at ScienceDirect

Forest Ecology and Management

journal homepage: www.elsevier.com/locate/foreco



Nature vs. nurture: Drivers of site productivity in loblolly pine (*Pinus taeda* L.) forests in the southeastern US

Vicent A. Ribas-Costa [a,b,*], Aitor Gastón [a], Sean A. Bloszies [b], Jesse D. Henderson [c], Andrew Trlica [b], David R. Carter [d], Rafael Rubilar [e,f], Timothy J. Albaugh [d], Rachel L. Cook [b]

[a] Centro para la Conservación de la Biodiversidad y el Desarrollo Sostenible (CBDS), ETSI Montes, Forestal y del Medio Natural, Universidad Politécnica de Madrid, Calle José Antonio Novais 10, Madrid 28040, Spain
[b] Department of Forestry and Environmental Resources, NC State University, Raleigh, NC 27695, USA
[c] USDA Forest Service Southern Research Station, Research Triangle Park, Raleigh, NC 27709, USA
[d] Department of Forest Resources and Environmental Conservation, Virginia Polytechnic Institute and State University, 228 Cheatham Hall, Blacksburg, VA 24061, USA
[e] Departamento de Silvicultura, Facultad de Ciencias Forestales, Universidad de Concepción, Victoria 631, Casilla 160-C, Concepción, Chile
[f] Centro Nacional de Excelencia para la Industria de la Madera (CENAMAD) - ANID BASAL FB210015, Pontificia Universidad Católica de Chile, Santiago, Chile

ARTICLE INFO

Keywords:
Site index
Soils
Geology
Physiographic province
Management
Climate
LiDAR
Random forest

ABSTRACT

Forest productivity is one of the most important aspects of forest management, landscape planning, and climate change assessment. However, although there are multiple elements known to affect productivity, most of them rely on the "nature" of the edaphic, climatic, and geographic conditions, and only some specific aspects can be modified through forest management or "nurture". Through evaluation of site resource availability and an understanding of the main drivers of productivity, management can present solutions to overcome site resource limitations to productivity. Therefore, understanding the implications of a specific management regime requires understanding what drives productivity across large spatial extents and among different management regimes. In this study, we used data from over 1 million hectares of industrial forestland, covering over 6000 different soils and several management regimes of *Pinus taeda* L. plantations, as well as plot-based data from the Forest Inventory and Analysis (FIA) program, facilitating a comparison of planted and natural *Pinus taeda* stands. Combined with US Geological Survey LiDAR data, we computed site index and generated wall-to-wall productivity maps for planted *Pinus taeda* stands in the southeastern US, as well as point-based site index estimates for the FIA dataset. We modeled site index using a random forest algorithm considering edaphic, geologic, and physiographic province information based on the Forest Productivity Cooperative "SPOT" system, and also included climate and management history data. Our model predicted site index with an R^2 of 0.701 and RMSE of 1.41 m on the industrial data and R^2 of 0.417 and RMSE of 1.84 m for the FIA data. We found that year of establishment of the forest, physiographic province, and geology, were the most important drivers of site index. The soil classification modifier indicating root restrictions were the most important soil-specific variable. Additionally, we found an average increase in site index of 3.05 m since the 1950s for all FIA data, and an average increase of 4.73 m for all industrial data since the 1970s. For the latest period analyzed (2000–2012), average site index in planted FIA plots was 1.2 m higher than naturally regenerated FIA plots, and site index in all industrial forestland had a site index almost 3 m greater than planted FIA plots. Overall, we believe this work sets the foundation for better understanding of forest productivity and highlights the importance of intensive silviculture to improve productivity, and as an additional tool to achieve the economic, environmental, and social objectives.

1. Introduction

Loblolly pine (*Pinus taeda* L.) is the most planted and intensively managed species in the southeastern US (Restrepo et al., 2019), occupying over 14 million hectares and making up 71 % of planted timberland in the US. The southeastern US produces more timber than any

* Corresponding author at: Centro para la Conservación de la Biodiversidad y el Desarrollo Sostenible (CBDS), ETSI Montes, Forestal y del Medio Natural, Universidad Politécnica de Madrid, Calle José Antonio Novais 10, Madrid 28040, Spain
E-mail addresses: vribasc@ncsu.edu, va.ribas@upm.es, vicentribas11@gmail.com (V.A. Ribas-Costa).

https://doi.org/10.1016/j.foreco.2024.122334
Received 22 July 2024; Received in revised form 9 October 2024; Accepted 10 October 2024
Available online 18 October 2024

other country in the world and generates over 26 % of wood products nationally (Johnston et al., 2022; Oswalt et al., 2019). In the southeastern US, these forests have multiple levels of management, ranging from natural regeneration with essentially no management to plantations with high levels of inputs (intensive forest management), such as chemical vegetation control, fertilization, thinning (Gyawali and Burkhart, 2015), and the use of genetically improved seedlings (McKeand, 2015). Specifically, loblolly pine forests have a long history of management improvements (Fox et al., 2007b). At present, the most intensive forest management can be characterized by interventions that optimize genotypes and silvicultural treatments based on site-specific characteristics to ameliorate resource limitations and to optimize value (Allen et al., 2005). While some environmental conditions (such as climate, soil texture, or geology) that affect forest growth cannot be changed (Clutter et al., 1992; Koirala et al., 2021), there are other variables that forest managers can address (soil drainage, nutrient deficiency) with management regimes (Carter and Foster, 2006; Isabel et al., 2019).

In plantation forestry, growth or productivity is typically the primary objective to optimize (Gyawali and Burkhart, 2015). Sites will respond differently to management in terms of potential productivity due to their inherent resource limitations and potential productivity (Fox et al., 2007a). In forest management, both in naturally regenerated stands and plantations, site productivity is defined as the forest production that can be realized at a certain site with a given management regime and environmental conditions (Skovsgaard and Vanclay, 2008), and it is normally described in terms of site index (SI), an expression that relates the stand's dominant height at a given stand age. The relative independence of SI from stand density (Weiskittel et al., 2011), and the relative ease of data collection compared to other site productivity indicators (Koirala et al., 2021), have made SI one of the most common forest productivity indicators. The ability to map current productivity and estimate potential future productivity is key for forest management planning (Hennigar et al., 2017). Furthermore, with models of site productivity, it is possible to predict where additional inputs will provide a growth response, which forest managers need to inform their decision-making process (Cook et al., 2024).

A second major consideration is the optimization of carbon management in forests (Susaeta et al., 2016). Intensively managed loblolly pine plantations can contribute to timber and biomass supply, but also play a crucial role in greenhouse gas emission reduction goals (Zhao et al., 2023). Different forest management techniques in even-aged loblolly pine plantations can produce differences in productivity over time for timber purposes (Fox et al., 2007a). These findings are transferable to carbon sequestration (Aspinwall et al., 2012; Clay et al., 2019; Zhao et al., 2023; Puls et al., 2024). If management can optimize current and potential productivity, forest carbon sequestration can also be optimized.

To this end, it is important to evaluate the drivers or limitations of tree growth either through empirical field trials or observational study designs. Normally, loblolly pine productivity models rely on measured tree data, which is then converted to SI or to another estimator (e.g., Hennigar et al., 2017; Koirala et al., 2021). These data can come from multiple sources (Allen and Burkhart, 2015) such as ground-measured forest inventory datasets, national forest inventories (for instance in the US the Forest Inventory and Analysis - FIA), or, most recently, from Light Detection and Ranging (LiDAR) datasets. The major benefit of the remotely sensed methods over ground data is the potential to cover larger areas at a lower cost than traditional on-the-ground methods (Lefsky et al., 2005). However, the combination of both approaches can aid the understanding of the relationships between forest productivity and its main drivers, as well as strengthen the power and accuracy of model predictions.

Multiple previous attempts at modeling loblolly pine site productivity using environmental and management data have focused on specific aspects such as water balance variables (Koirala et al., 2021), soil and physiographic properties (Sabatia and Burkhart, 2014; Subedi and Fox, 2016), management (Jokela et al., 2004; Zhao et al., 2016), genetic improvement (McKeand, 2015), and even CO_2 fertilization (Burkhart et al., 2018). However, these previous studies used plot-based local productivity measures, either from a specific location within the southeastern US, or based on limited data. We take a more comprehensive approach incorporating soils, geology, climate, and management in a large-scale, range-wide study that encompasses most established productivity drivers in a single analysis. We compare data from natural regeneration to the most intensive industrial management throughout the southeastern US and combine LiDAR-derived wall-to--wall productivity information with plot-based data.

Accordingly, the specific objectives of the study are to:

1. Characterize the change in SI over time by management regime (natural or planted regeneration, low-input silviculture to high-input silviculture).
2. Model SI using edaphic, geologic, climatic, and management data, to predict SI across the southeastern US, and specifically:
3. Explain which factors drive forest productivity, and how site and management factors interact.
4. Evaluate how forest management affects SI, and how that effect has varied over time.

2. Materials and methods

In this study, we used climate, edaphic, geology, and management data (Table 1) to model and predict SI in two different datasets and compare the obtained results. SI was obtained from Forest Inventory and Analysis (FIA) plot-based data and industrial plantation land using United States Geological Survey (USGS) LiDAR data to obtain SI raster maps. We then applied a random forest model to (1) predict SI using the previously mentioned variables, and (2) explain their importance and interaction level, assessing management intensity over time.

2.1. Climatic, edaphic, geologic, and physiographic province data

The climatic data was obtained from the 30-year normal (daily values averaged over the 1991–2020 period) baseline datasets from the PRISM Climate Group (https://www.prism.oregonstate.edu/normals/) at the Northwest Alliance for Computational Science and Engineering. We downloaded and used the 800-meter resolution rasters of precipitation (mm), maximum temperature (°C), and maximum and minimum vapor pressure deficit (kPa). We used the Forest Productivity Co-operative's (FPC's) Site Productivity Optimization for Trees (SPOT) classification system (Cook et al., 2024) as the data source for soil, geology, and physiographic province information. Incorporating these soils and climate data into both FIA and Industrial datasets resulted in 14 variables, whose names, ranges, and units are provided in Table 1.

2.2. Loblolly pine forest inventory and analysis data

The Forest Inventory and Analysis (FIA) Database includes tree height, site condition, and species identification in naturally regenerated and planted loblolly pine and was queried with exact coordinates across the southeastern US. Plot data in the field were collected following FIA field inventory standards (Reams et al., 2005), and SI was computed as an average of the per-tree SI values applying the Diéguez-Aranda et al. (2006) model using a 25-year base age. We intersected plot locations with the SPOT classification system layer (Cook et al., 2024) to provide information on soil conditions and underlying geology. We intersected the climate variables listed in Table 1 and SI with FIA plot locations and then averaged those variables for all the FIA plots with loblolly pine trees within the same SPOT code and origin type (planted or natural). We excluded SPOT codes represented by fewer than three FIA plots. There was a total of 1220 plots with a mean of 17.6 plots and a median of

Table 1
Variables used to model of site index based on management regime, soils, geology, and climate. For more details about edaphic and geologic data, refer to Cook et al., (2024). IND, Industrial dataset; FIA, Forest Inventory and Analysis dataset.

Variable name	Variable code	Variable type	Number of levels	Variable group	Unit	Range (max-min)	Variable description
Establishment year / Average establishment year	EstbYr / AvgEstbYr	Continuous		Manage-ment regime	Year	IND.: 2012–1973 FIA: 2006–1917	Year the plantation was established for Industrial dataset and the average establishment year among plots for FIA dataset.
Management	Mgmt	Categorical	IND.: 8 FIA: 2	Manage-ment regime			Management regime (Table 1 for Industrial dataset or origin planted/natural for FIA dataset)
Maximum temperature	TMax	Continuous		Climate	°C	IND.: 27.5–19.6 FIA: 26.8–19.4	30-year normal maximum temperature average
Precipitation	Ppt	Continuous		Climate	mm	IND: 1756–1086 FIA: 1682–1126	30-year normal total precipitation average
Minimum vapor pressure deficit	VPDMin	Continuous		Climate	hPa	IND: 1.5–0.07 FIA: 1.2–0.4	30-year normal minimum vapor pressure deficit
Maximum vapor pressure deficit	VPDMax	Continuous		Climate	hPa	IND: 20.1–12.6 FIA: 19.7–11.5	30-year normal maximum vapor pressure deficit
Major soil group	MajorSoil	Categorical	IND.: 7 FIA: 7	Soil			Primarily soil texture
Depth to clay	DepthClay	Categorical	IND.: 7 FIA: 7	Soil			Soil depth to increase in clay horizon
Drainage	Drainage	Categorical	IND.: 7 FIA: 7	Soil			Based on rate of water removal
Nature of surface	NatSurface	Categorical	IND.: 6 FIA: 5	Soil			Soil surface modifiers
Nature of subsoil	NatSubsoil	Categorical	IND.: 7 FIA: 5	Soil			Nature of soil subsurface modifiers, describes minerology
Additional Limitation or Resources	AddLimRes	Categorical	IND.: 9 FIA: 6	Soil			E.g., root restrictions, ponding, alkaline, salt affected
Geocode	Geocode	Categorical	IND.: 33 FIA: 27	Geophys			Geology, geologic formations, and coastal plain terraces
Physiographic Province	PhysioPro	Categorical	IND.: 10 FIA: 10	Geophys			Grouped Major Land Resource Area

5 plots per county within the southeastern US (both inside and outside the native range of loblolly pine) (Fig. 1).

2.3. Loblolly pine industrial data

To get a thorough representation of the plantation industrial-owned loblolly pine land, eleven timber management companies shared their planted loblolly pine data, including stand boundaries, year of establishment, and past silvicultural treatments. We standardized the management data into an eight-level categorical variable (Table 2) based on the records of thinning, mid-rotation chemical and/or mid-rotation fertilization application. It is important to note that the category "unknown" includes the stands without silvicultural information, yet these can be either real non-treated stands or stands with missing silvicultural information (as records may be lost in the transfer of land). We included the establishment year of the plantation to track the effect of silviculture, genetics, and environmental changes over time.

To obtain wall-to-wall SI information for the loblolly pine industrial plantations, we followed the workflow presented in Ribas-Costa et al. (2024) and used LiDAR data from the United States Geological Survey (USGS) nationwide acquisitions. The USGS provides three levels of LiDAR data qualities (QL) in the 3D Elevation Program (3DEP), ranging in aggregate nominal pulse density from $\geq$0.5 pulses m^{-2} in QL3 to $\geq$8 pulses m^{-2} in QL1. The most common quality was QL2 with an aggregate nominal pulse density of $\geq$2 and $\leq$ 8 pulses m^{-2}. For this study, we used LiDAR acquisitions in QL2 or QL1. The specific flights used for this study occurred between 2011 and 2022 (Fig. 2). These data were accessible on the National Map App (https://apps.nationalmap.gov/lidar-explorer/#/) and can be streamed through the Amazon Web Services (AWS) Public Dataset project USGS 3DEP LiDAR (https://registry.opendata.aws/usgs-lidar/). Approximately 5 TB of LiDAR were downloaded from the AWS cloud repository and processed on a local machine.

Given that the approach presented in Ribas-Costa et al. (2024) for estimating SI from LiDAR data was built for plantations between 9 and 43 years old, we excluded all the stands that were not in that age range at the time of USGS flight. Ordinary processing steps for the LiDAR data (i.e., removing the outliers, ground classification, and point cloud normalization) were applied, followed by the application of the dominant height model (Eq. (1)) to the point cloud at a 20-m pixel basis:

$$H_{dom} = 2.186 + 0.8989 * PCT95 \quad (1)$$

Where *PCT*95 represents the 95th percentile of the height above ground distribution for all returns > 1 m above ground. To calculate the age of the stand at the time of the flight, we used the time of the USGS flight and the year of establishment, and computed age based on a 1st of July cut-off to account for real growing seasons (Ribas-Costa et al., 2024). Finally, we applied the Diéguez-Aranda et al. (2006) SI model. Fig. 3 represents a stepwise summary of this process.

Once the SI raster was completed, we intersected the stand boundaries with the SPOT soils classification map, similar to the FIA analysis. We computed zonal statistics for each SPOT code within each stand by applying the ArcGIS Pro 3.10 function *Zonal Statistics as Table* to add to the previous intersected layer the values of raster-based SI within the target zones (SPOT codes within stand IDs). Climate variables were also added to the dataset via zonal statistics by SPOT code polygon within

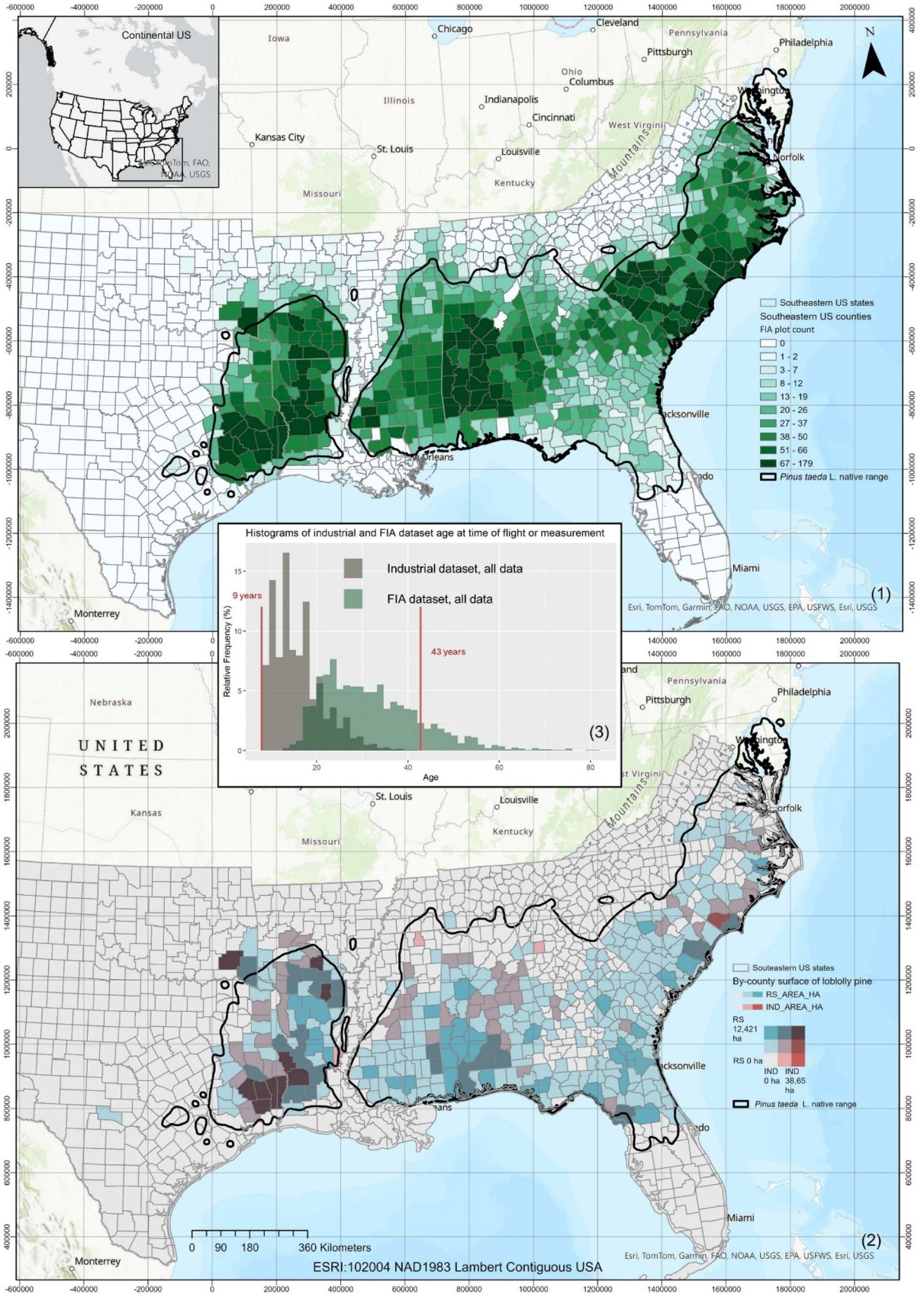


Fig. 1. Datasets used for this study. (1) represents a per-county number of FIA plots used in the analysis, and (2) represents the area of industrial land utilized in our study (IND_AREA_HA), against the remotely-sensed area of loblolly pine (RS_AREA_HA)) from Thomas et al. (2021). The dark black line represents the native range of loblolly pine (Little, 1971). (3) represents the histograms of the ages of the plantations used in the study from each dataset (industrial and FIA).

Table 2
Management intensity categorical variable based on the known silvicultural treatments per stand.

Number of treatments	Number of combinations	Possible variables	Variable description
3	1	"chem+fert+thin"	Stands with at least one mid-rotation chemical release, one thinning, and one fertilization treatment.
2	3	"chem+fert", "chem+thin", "fert+thin"	Stands with at least two treatments, any combination, among mid-rotation chemical release, thinning, and/or fertilization.
1	3	"chem", "fert", "thin"	Stands with at least one treatment, such as mid-rotation chemical release, thinning, or fertilization.
0	1	"unknown"	Stands with no treatment records.

stand ID. Finally, management information at a stand level was added to the dataset. Prior to modeling steps, we excluded all observations with fewer than 10 pixels of SI values (in other words, we did not use any combination of SPOT code within stand ID that represented less than 4000 m^2). At this stage, the dataset had a total of N = 107,844 observations and corresponded to a total of approximately 1 million hectares of land located in the same general region as the remotely sensed distribution of managed loblolly pine plantations across the southeastern US (Thomas et al., 2021; Fig. 1).

2.4. Statistical methods

2.4.1. SPOT code presence and site index evolution over time

To characterize the evolution of SI over time, specifically with regard to management regime, we first needed to limit our analysis to SPOT codes shared in both datasets (1194 total), given that the number unique SPOT codes represented in each data type varied (Industrial unique codes = 6275; FIA unique codes = 1542). By analyzing sites in only shared SPOT codes, we compared the observed SI values across the two data sets and avoided the potential biases associated with some soils (either more or less productive) only being represented in one dataset. We then organized this subset of the data in the following manner:

A. "IND high": Industrial dataset with documentation of receiving at least one chemical application in mid-rotation, one fertilization treatment, and one thinning.
B. "IND low": Industrial dataset with at least one documented management procedure (chemical application, fertilization, or thinning).
C. "IND unknown": Industrial where management is unknown.
D. "FIA planted": FIA dataset where origin is planted.
E. "FIA natural": FIA dataset where origin is natural regeneration.

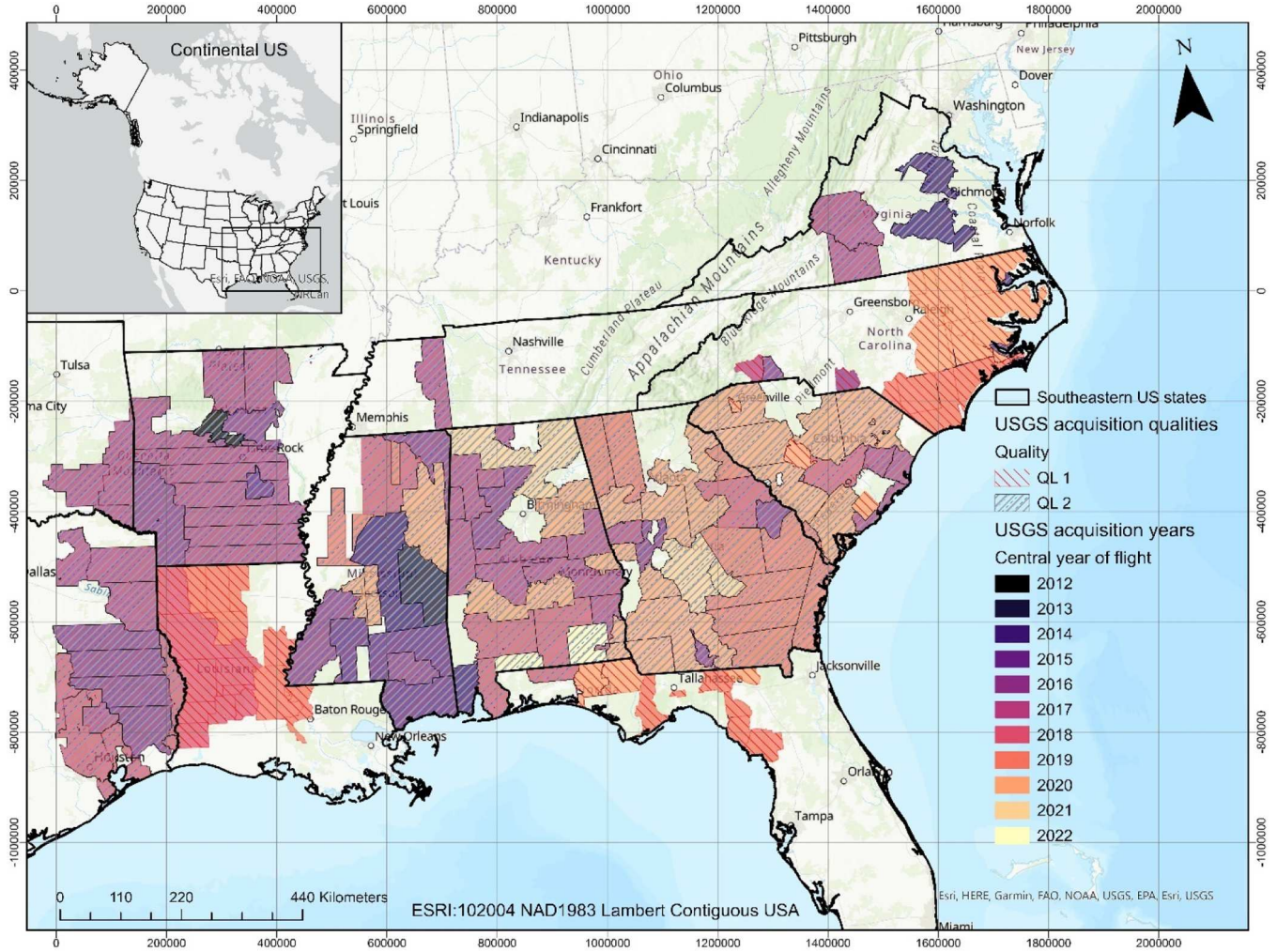


Fig. 2. USGS acquisitions used for this study. The color represents acquisition date and the texture represents acquisition quality (QL1, QL2).

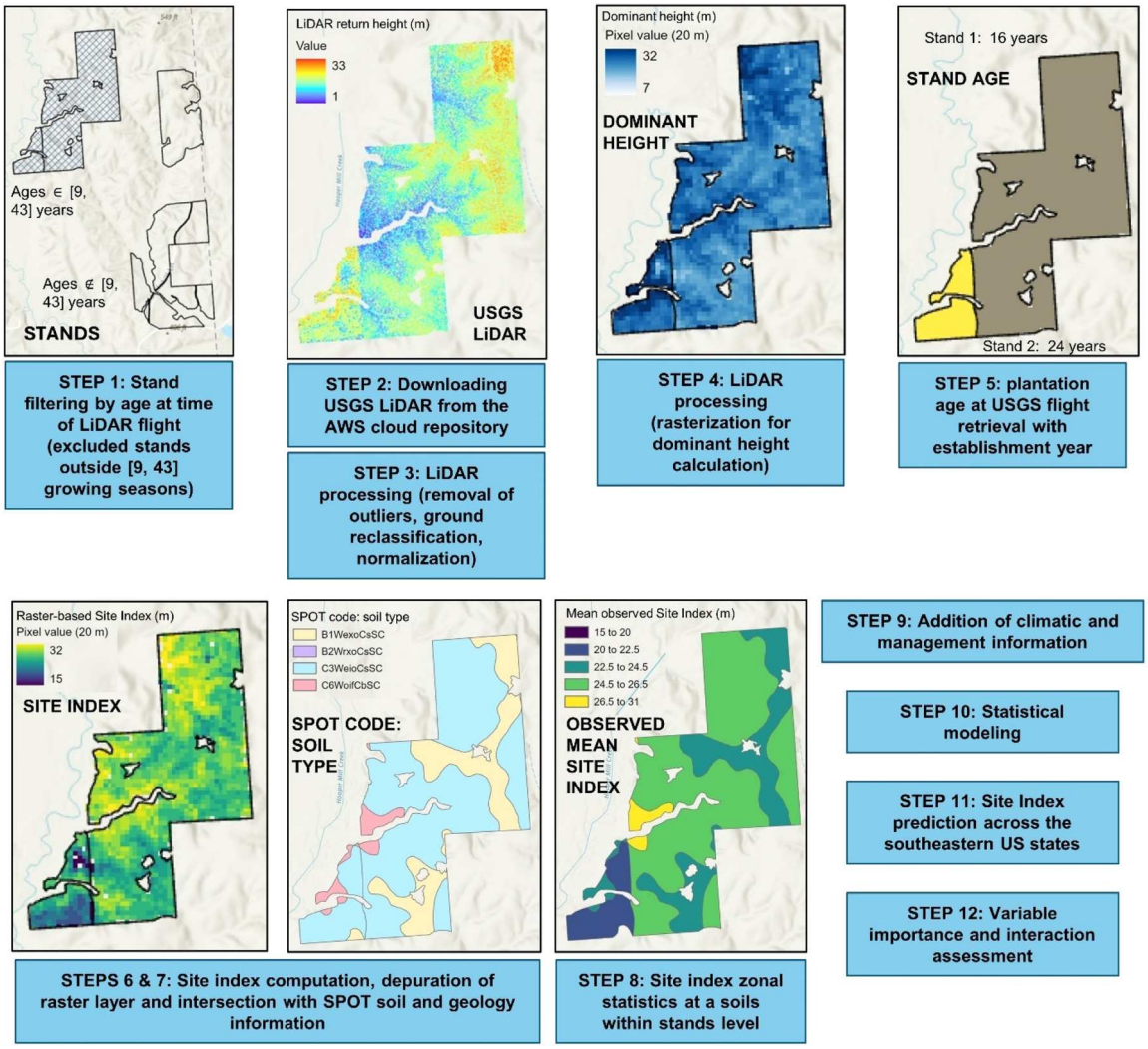


Fig. 3. Stepwise summary of the process performed in the Industrial dataset to obtain SI average measures for each SPOT code within each stand ID from USGS LiDAR data, showing all the necessary steps prior statistical modeling using soils, geology, climate, and silvicultural information. AWS stands for Amazon Web Services.

After that, we evaluated the effect of establishment year on SI for each of the groups by applying linear regressions and evaluating model fit and comparing slope coefficients.

2.4.2. *Site index random forest modeling*

After analyzing the general trends over time, we employed a non-parametric random forest (Breiman et al., 2001; RF) model to (1) assess the relative importance of each variable and (2) create a predictive model for SI in each data set following a similar approach to other studies (e.g., Hennigar et al., 2016; Pahlavan-Rad et al., 2020; Cook et al., 2024). Due to the nature of the algorithm, RF cannot extrapolate beyond the underlying data (Hennigar et al., 2016; Jeong et al., 2016). However, despite the inherent limitations of categorical data (i.e., any new level of any of the variables blocks the performance of the model) it had valuable benefits and abilities: RF was able to tolerate autocorrelation and high dimensional data, handle interactions between variables, identify informative inputs using a permutation-based RF variable importance index, and capture complex phenomena while revealing non-linear relationships (Antoniadis et al., 2021; Cheng et al., 2020).

Prior to modeling, we excluded the levels of any categorical variable that had less than 15 observations, as well as observations with unrepresentative conditions, such as plantations in soils labeled as dumps or water layers. The ready-to-use Industrial dataset had 107,331 observations (industrial stand IDs with one SPOT code), whereas the FIA dataset had 1985 observations (FIA plots with one SPOT code). We then implemented the random forest algorithm via the "ranger" R package (Wright and Ziegler, 2017), performing a hyperparameter tuning optimization via the "tuneRanger" R package (Probst et al., 2019). To prevent the bias associated with categorical variables with many levels or in favor of high-frequency levels within categorical variables, we built trees without resampling, utilizing the permutation option for the variable importance assessment and the mean response encoding option as in Hastie et al. (2009) (Loecher, 2022; Probst et al., 2019; Nembrini et al., 2018; Altmann et al., 2010). The out-of-bag (OOB) coefficient of determination, root mean square error, and relative root mean square error (RMSE and RRMSE, Eqs. (2) and (3), respectively) were then used to assess the predictive performance of the model.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (2)$$

$$RRMSE = \frac{RMSE}{\bar{y}} \quad (3)$$

Where y_i is the observed SI, $\hat{y}_i$ is the predicted SI i, and $\bar{y}$ is the mean of the observed SI. The final proposed model integrated management regime, soil information, geocode, physiographic province, and climatic information (Eq. (4), in simplified notation for RF modeling).

$$Site\ Index \sim YEAR + Mgmt + TMax + Ppt + VPDMin + VPDMax + MajorSoil + DepthClay + Drainage + NatSurface + NatSubsoil + AddLimRes + Geocode + PhyisioPro \quad (4)$$

The parameter *YEAR* reflects either *EstbYr* in the Industrial dataset or *AvgEstbYr* in the FIA dataset. Additionally, the parameter *Mgmt* indicates the management level for Industrial dataset or origin for FIA dataset. Then, *TMax* represents the maximum temperature; *Ppt* the average total yearly precipitation; *VPDMin* the minimum vapor pressure deficit; *VPDMax* the maximum vapor pressure deficit; *MajorSoil* the major soil group (texture) from the SPOT code; *DepthClay* the soil depth to increase in clay horizon, *Drainage* the drainage level; *NatSurface* the soil surface modifiers, *NatSubsoil* the soil subsurface modifiers, *AddLimRes* any additional limitations or resources, *Geocode* the geology, geologic formations, and coastal plain terraces; and *PhysioPro* the grouped major land resource area. Furthermore, to prevent overfitting, we limited the hyperparameters, fixing the number of features per tree to four and the minimum node size to five (Mentch and Zhou, 2020). To validate the model's performance and stability, we ran a 5-fold cross-validation procedure and analyzed the stability of the OOB R^2 and RMSE across folds.

2.4.3. Evaluation of the effect of covariates in site index modeling

To assess the importance of the variables, we trained a random forest model using soils, geology, climate, management, and year of establishment (Eq. (4)), ranking the variables' importance after ten subsequent runs, presenting a position and dispersion measure of those repetitions. Furthermore, to visualize the relation between the different features and the SI, we produced partial dependence plots (PDP) (Friedman, 2001), individual conditional expectation (ICE) curves (Goldstein et al., 2015), and accumulated local effects (ALE) plots (Apley and Zhu, 2020). PDPs assess the mean predicted value response variable on the y-axis against the marginal distribution of the studied feature in the x-axis, while ICE plots represent the individual effect for each observation, typically used in conjunction with PDPs. ALE plots assess the local effect of a feature, considering only a subpopulation within a specific range of the studied feature (Molnar et al., 2018). As such, ALE plots avoid extrapolation and are more computationally efficient (Apley and Zhu, 2020); however, the local effects displayed are only applicable to the specific subpopulation for which they were calculated (Loef et al., 2022), making them challenging to interpret.

In this work, we used PDPs, ICE curves, and ALE plot to examine the main effects of a subset of features. Following the methodology of previous studies (e.g., Loef et al., 2022) we employed PDPs to gain a general understanding of the effect size attributed to each feature. Concurrently, ALE plots were utilized to determine whether the behaviors observed in PDPs could be attributed to incorrect extrapolation. importance of variables and their interactions were evaluated using "iml" (Molnar et al., 2018) R package. To explore variable interactions, we applied Friedman's H-statistic (Friedman and Popescu, 2008), which measures the intensity of an interaction based on the combined variation of two input features. The H-statistic for a feature assess the strength of interaction with other features and then computes an average, yielding a value that indicates the interaction strength of each feature. This metric is derived from the proportion of the variance of the two-dimensional partial dependence function that cannot be explained by the sum of the two one-dimensional partial dependence functions (Friedman and Popescu, 2008).

3. Results

3.1. Site index evolution over time for the common soils

For the unique SPOT codes that were common in the industrial and the FIA datasets (n = 1194), we found a statistically significant linear increase in SI over time for all sub-datasets (in all cases, p-value < 0.0001; Fig. 4). Interestingly, the slope coefficient was very similar for the naturally regenerated (0.07 m year^{-1}) and planted (0.09 m year^{-1}) stands in the FIA dataset but was significantly higher in the Industrial dataset at any silvicultural level (0.26 – 0.32 m year^{-1}), indicating that the average rate of SI increase per year was higher in the Industrial dataset. The lowest slope for SI over time was found in the naturally regenerated FIA plots, while the highest was found in the Industrial dataset with high silvicultural management (chemical application, fertilization, and thinning). In the FIA dataset, the higher increase per year of establishment was found in planted origin stands. In the Industrial dataset, high silviculture management stands had the highest increase per year. However, in all cases, the low R^2 suggested that there were other factors influencing productivity besides establishment year.

3.2. Site index modeling and variable effect evaluation

For the Industrial dataset, the predictive performance of the RF model was high: when running a 5-fold cross-validation, R^2 and RMSE outcomes were stable, with average and standard deviation values of 0.701 ± 0.003 for R^2 and 1.41 m ± 0.004 m for RMSE. In relative terms, the industrial model exhibited an RRMSE of 5.98 %. When performing the 5-fold cross-validation on the FIA dataset, the results indicated both a lower goodness fit compared to the Industrial dataset, and a lower stability across the folds. The reported average and standard deviation values were 0.417 ± 0.030 for R^2 and 1.84 m ±0.047 m for RMSE, which translated into a 9.21 % RRMSE. Both the Industrial and FIA data showed the general trend of underestimation of higher SI values and overestimation of low SI values (Fig. 5).

When constructing the 1-way PDPs for the variables in the Industrial dataset (see Fig. 6 and Supplementary 1), it became evident which variables exert a stronger influence on the model. Specifically, establishment year, physiographic province, geocode, maximum temperature, and precipitation had stronger effects on the predictions than other variables. These variables showed a similar influence in the FIA dataset (Fig. 6 and Supplementary 2), except that management level (planted or natural origin) had a much stronger effect on the predictions than the management level in the Industrial dataset. Additionally, the ALE plots for all variables (see Supplementary 2 and 4) followed a similar trend to the PDPs, providing further evidence that no problematic bias was present.

As an example of specific levels of the most influential variables in the Industrial dataset, the most beneficial physiographic province for site productivity was found to be the Mississippi Valley Loess Plain (LP), which added up to 1.2 m of additional site index (Fig. 6, Supplementary 2). The second most beneficial physiographic province was the Atlantic Coastal Plain Flatwoods (AF), with a boost in 0.8 m of additional site index (Fig. 6, Supplementary 2). The Atlantic Coastal Plain Flatwoods are generally poorly drained but very responsive to nutrient inputs.

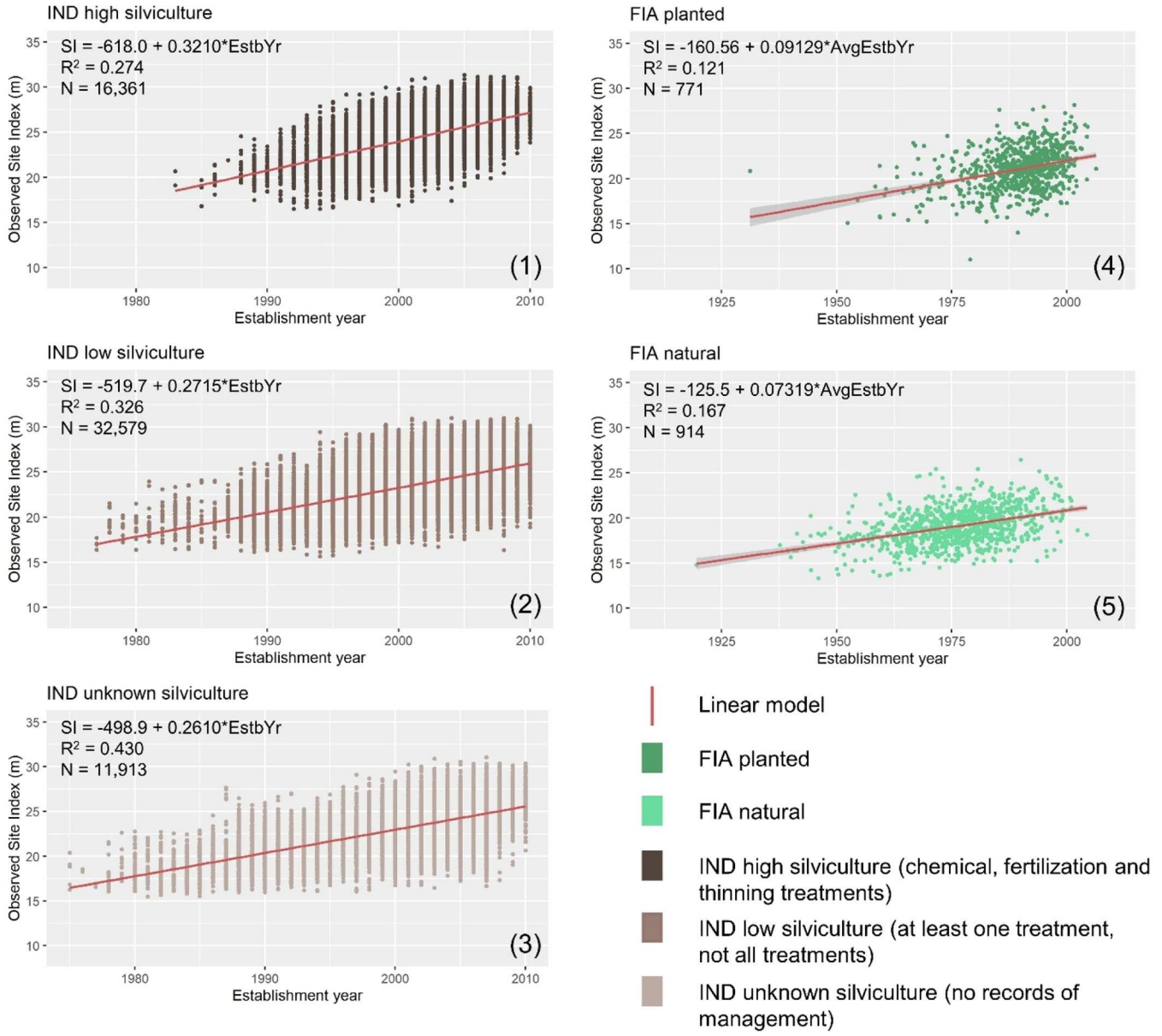


Fig. 4. Evolution of SI over time for SPOT codes common in both the Industrial and FIA datasets. Regression parameters were generated with all the observations within each data subset, for (1) Industrial with at least one chemical application, fertilization, and thinning, (2) Industrial with at least one treatment, but not all three, (3) industrial with unknown management level, (4) FIA planted, and (5) FIA natural. The shade in the linear regression lines represents the 95 % confidence intervals, yet in the Industrial dataset, it is so close to the regression line that it cannot be seen at the presented scale. IND, Industrial dataset; FIA, Forest Inventory and Analysis dataset.

When looking at soil depth to argillic layer, we found deeper increases in clay content indicated lower productivity potential. This result is to be expected as clay content provides a reservoir of nutrient and water holding capacity. Deep sands are well known to have lower productivity potential. The FIA dataset followed the same patterns, but with a smaller productivity increase.

The two most influential climatic variables, annual precipitation and maximum temperature, showed average values of approximately 1400 mm of rainfall and maximum temperature of about 24 °C represented the central conditions where neither a boost nor a decrease in productivity was expected (Fig. 6, Supplementary 2 and Supplementary 4), both for Industrial and FIA datasets and across all the climatical range that exists in the study region. Higher maximum temperature or higher precipitation seemed to boost forest production about 0.5 m in SI, regardless of the dataset. However, it also appeared that when annual precipitation exceeded 1600 mm or the maximum annual temperature rose above 26°C, productivity tended to decline.

In terms of feature importance, establishment year clearly outranked all the other variables in both datasets, generating an increase of RMSE of about 1.8 m for both datasets. The next most important features in the Industrial dataset were physiographic province, maximum temperature, geology, and precipitation (Figure 7.1), whereas in the FIA dataset management, physiographic province, precipitation, and geology were the most important ones (Figure 7.2). Across both datasets, the most important feature within the group of soil variables was additional limitations or resources, and the least important was nature of surface, also in both datasets. Finally, in terms of overall interaction strength physiographic province, precipitation, and year of establishment showed the greatest degree of interaction with other features (Fig. 8), both for the Industrial and FIA datasets.

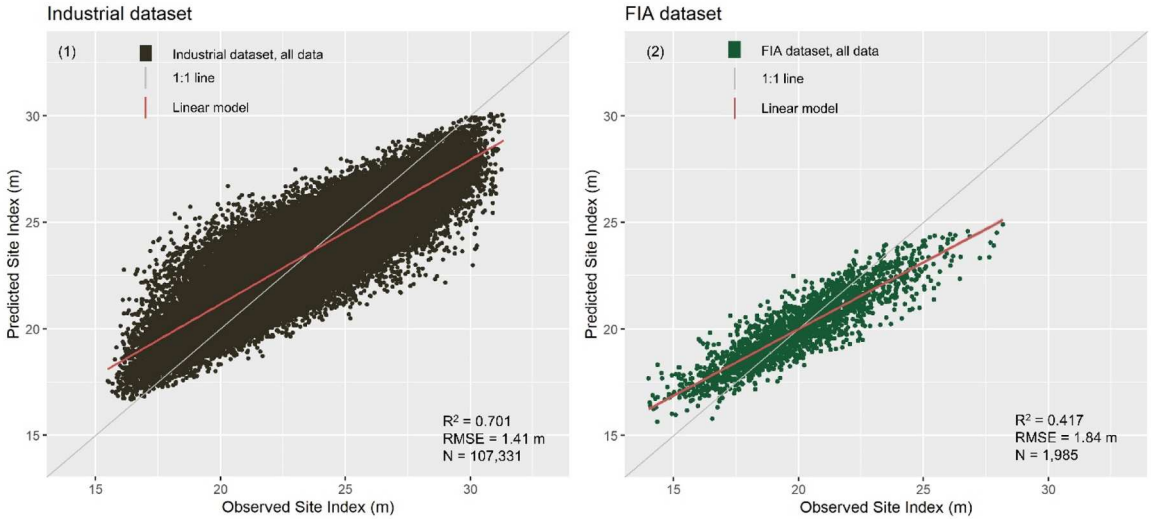


Fig. 5. Scatterplot of observed vs. predicted SI values for the random forest model for Industrial dataset (1) and for the FIA dataset (2). In both cases, R^2 and root mean square error (RMSE) are the outcomes of a 5-fold cross-validation procedure. The red line (linear model) demonstrates the deviation of the predictions is greater at the extreme high and low values. The gray line is the 1:1 line.

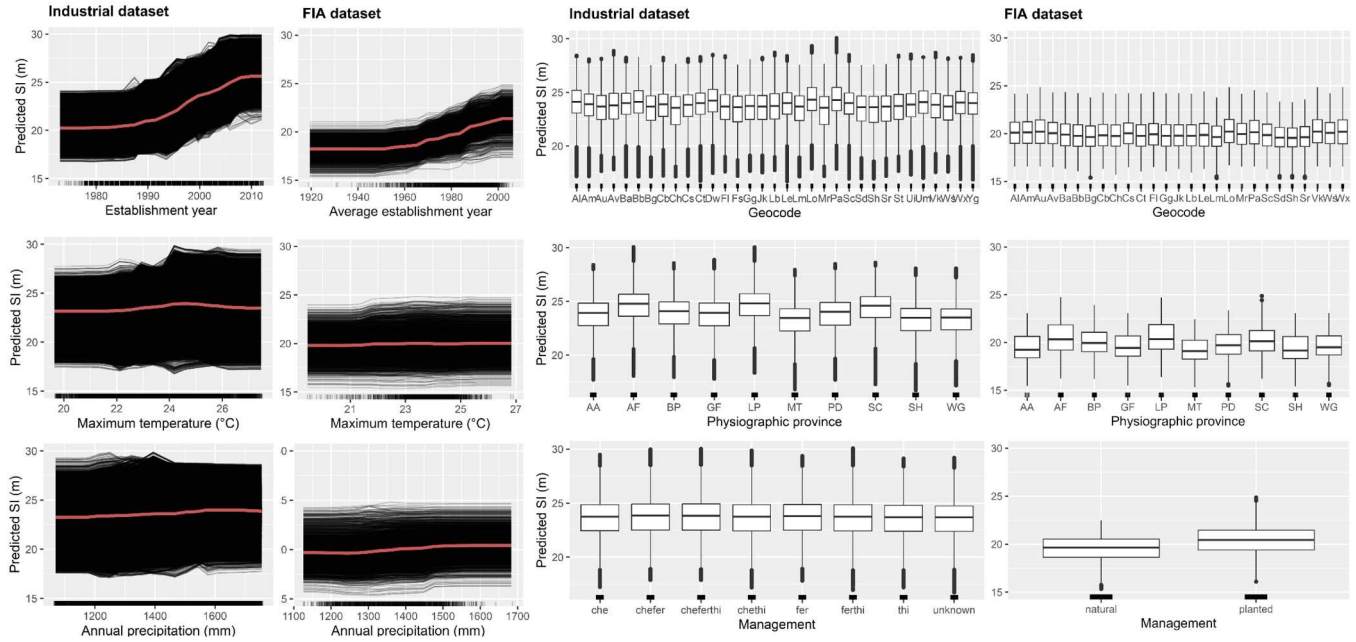


Fig. 6. 1-way partial dependence plots and individual conditional expectation curves for establishment year, maximum temperature, annual precipitation, geocode, physiographic province, and management, for industrial and FIA datasets. In the box and whisker plots of the predicted SI values, the box represents the interquartile range (IQR), this is the 25–75 % interval of the predictions, the line represents the median, and the whiskers represent the median ± 1.5 IQR. Observations above that limit are marked as outliers. A higher oscillation of the box plots can be understood as a higher influence of that variable on the overall SI model.

3.3. Effect of management on site index over time

From the early 1980s until the 2000s, management regime had a more pronounced effect than in the subsequent years. This trend is evident in the 2-way partial dependence plot (see Fig. 9), where the management effect gradually diminished in its influence on the average SI prediction over time. When analyzing the FIA dataset, we notice a similar trend. However, the difference between natural and planted site indices remains relatively constant across time compared to the Industrial dataset (see Table 3). Also, the predicted SI values in the FIA dataset are notably lower than those in the Industrial dataset. Interestingly, both datasets showed a plateau in SI growth in the latest data. Around the year 2000, the response curve began to level off in the FIA data, while in the Industrial dataset, this effect becomes noticeable around 2008 (Fig. 9).

When looking at the differences across management levels (summarized in FIA planted, FIA natural, Industrial high silviculture, Industrial low silviculture, and industrial unknown silviculture) for each of the decades (Table 3), the largest difference in consecutive management levels was achieved in 2000–2012 between FIA planted and industrial unknown silviculture, with almost 3 m of difference in the SI predictions. The greatest absolute difference occurred also in the same decade between FIA natural and industrial high silviculture, with greater than 4 m of predicted SI difference. Since the 1950s, there has been an average increase of 3.05 m for FIA plots (for both natural and planted origin), and an increase of 4.73 m for industrial stands (for all

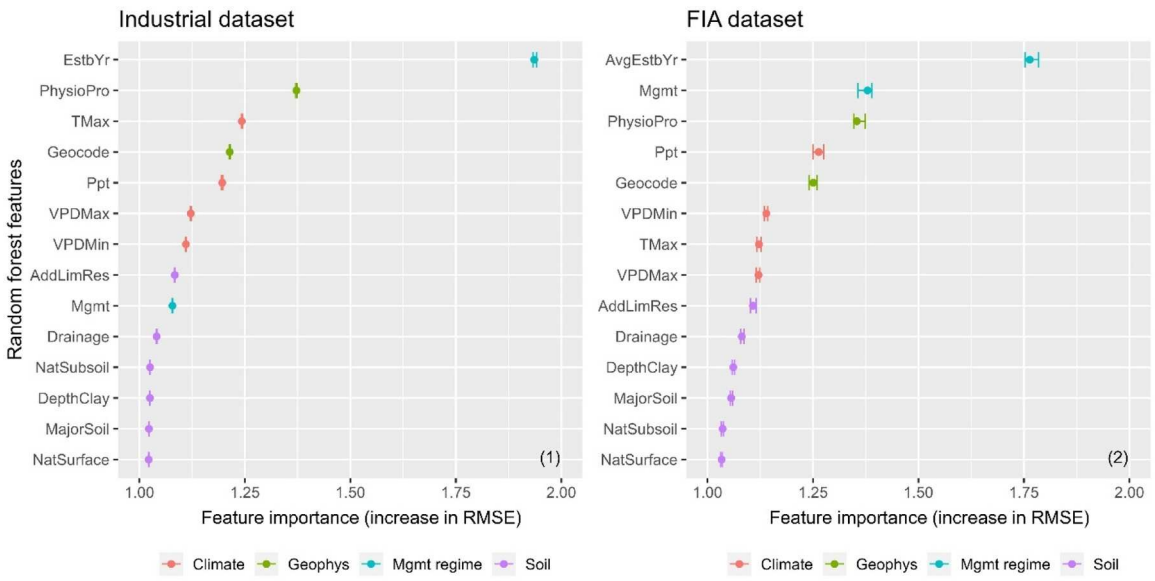


Fig. 7. Feature importance ranking for a 10-run RF model for both industrial (1) and FIA (2) datasets. The features are colorized by type (climate, geology and physiographic province, management, or soil). The dot represents the mean value, and the interval represents the 5th and 95th confidence interval (alpha = 0.05). The RMSE increase is reported in absolute difference between the original model and the model with the shuffled feature.

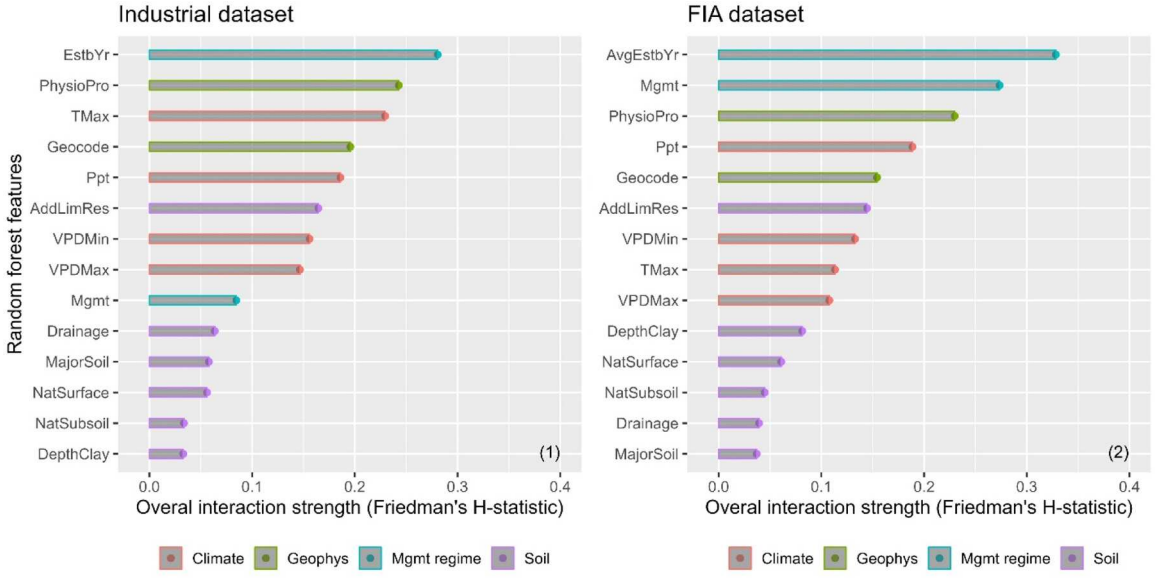


Fig. 8. Variable overall interaction strength, based on Friedman's H-statistic, for both industrial (1) and FIA (2) datasets. The H-statistic varies between 0 (no interaction) to 1 (full interaction).

management levels) since the 1970s. When looking at the latest data (2000–2012), we observed predicted SI in FIA planted stands was 1.2 m higher than naturally regenerated FIA stands, although planted FIA stands were still 2.9 m lower than predicted SI in industrial stands, averaged across management levels.

4. Discussion

4.1. Site index evolution over time and management influence

The year of establishment was the most important variable when modeling SI. This finding is consistent with previous studies, which have demonstrated that various factors have contributed to increased productivity in loblolly pine plantations over time. These factors include improved and more efficient management (Jokela et al., 2004; Fox et al., 2007b; Zhao et al., 2016), carryover effects of fertilization (Everett and Palm-Leis, 2009), atmospheric nitrogen deposition, enhanced chemical weed control (Subedi et al., 2019), changes in climate, and ambient CO_2 concentration (Burkhart et al., 2018; Aguilos et al., 2021), as well as enhanced productivity genetics resulting from traditional tree breeding efforts (McKeand, 2015; McKeand et al., 2020). Interestingly, even naturally regenerated stands present in the FIA dataset, exhibited a similar trend of higher SI values for more recent stands. This trend suggests that environmental factors such as CO_2 fertilization (when nitrogen is not limiting) (Huang et al., 2007), nitrogen deposition, temperature increases, and perhaps even genetically-improved pollen from planted loblolly pines and/or the slow recovery of soils from past agricultural use have contributed to increased productivity over time, independently of management practices (Van Lear et al., 2004; Davis et al., 2022).

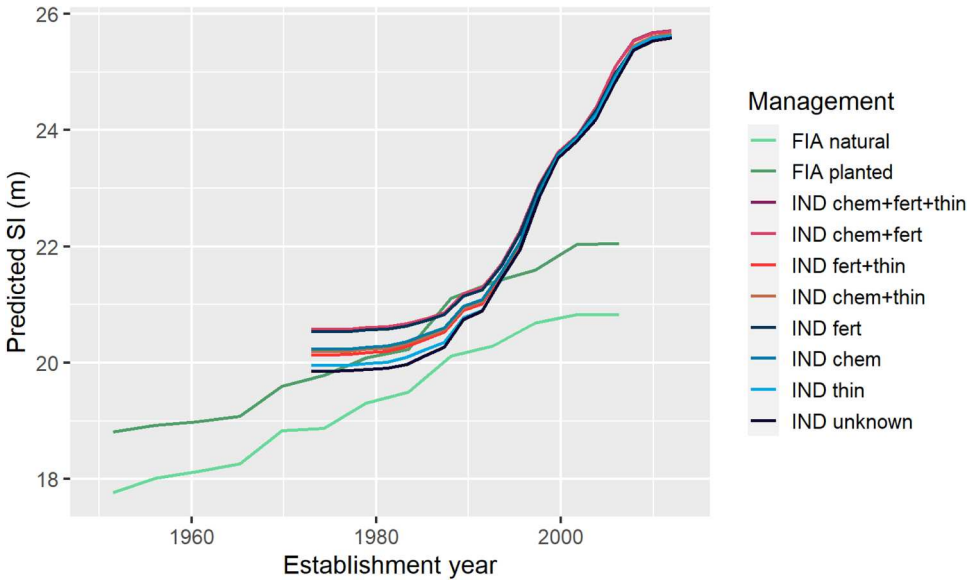


Fig. 9. Evolution of the effect of management level on the predicted site index (SI) depending on the establishment year. Values extracted from the 2-way partial dependence of establishment year and management type, for both the industrial (IND) and FIA datasets.

Table 3
Differences in the predictions for Forest Inventory and Analysis (FIA) natural and planted, and Industrial (IND). Only combination of consecutive levels of management are shown. IND high includes the treatments with at least one chemical mid-rotation release, fertilization application, and thinning; IND low includes the treatments with at least one silvicultural treatment, and IND unknown includes all the observations with no records of treatments. Data was obtained from the partial dependence plots of management and establishment year and averaged by management levels and decades.

Decade / Mgmt.	Number of observations in each dataset					Differences in the average predictions for consecutive management levels and extreme range (m)				
	N FIA planted	N FIA natural	N IND high silviculture	N IND low silviculture	N IND unknown silviculture	FIA planted - FIA natural	IND unknown - FIA planted	IND low - IND unknown	IND high - IND low	IND high – FIA natural
1950–1960	13	160	-	-	-	0.98	-	-	-	-
1960–1970	51	438	-	-	-	0.81	-	-	-	-
1970–1980	156	792	3	71	165	0.85	−0.07	0.41	0.29	1.48
1980–1990	640	667	188	2362	2791	0.87	−0.46	0.36	0.24	1.00
1990–2000	847	211	10,769	21,572	6173	1.00	0.64	0.14	0.12	1.90
2000–2012	114	15	19,801	39,192	13,645	1.21	2.85	0.08	0.08	4.23
Accumulated difference	882	1076	28,141	58,204	20,986	5.72	2.95	1.00	0.74	8.61

Establishment origin (natural or planted) in the FIA dataset and silvicultural intensity in the Industrial dataset were also found to have a significant effect on SI over time. Specifically, management had a greater effect on SI in the early decades (1980 – 1995), compared to 1995 onwards. This observation may be linked to the critical role of management practices in an earlier era of plantation forestry, when genetic improvements were still in early development (McKeand et al., 2020), and other environmental factors were not as pronounced as they appear to be now (Burkhart et al., 2018). Before 1980, the lower management level in industrial plantations more closely resembled natural forests, while the current "baseline" silviculture in the southeastern US represents a higher level of management than the best silvicultural practices of the 1980s (Fox et al., 2007a, 2007b). Additionally, at present at least half of the loblolly pine forestland in the southeastern US is under some kind of silvicultural management (Oswalt et al., 2019; Thomas et al., 2021), while simultaneously, timber industries are strategically pursuing the acquisition of land with a high return on investment, potentially resulting in a higher concentration of more productive sites within the Industrial dataset. However, when comparing the same soils in both datasets, SI values for FIA stands remain stagnant, highlighting a missed opportunity for enhancement. These differences could potentially be attributed to the inherent differences in management practices between the two types of land rather than the differences in the sites themselves. By isolating the effects of management improvement over time, the remaining contribution of management practices became minimal. In other words, when the slope of the SI curve (see Fig. 9) was higher, the different management level curves converge. This trend is also slightly visible in the FIA data, where the productivity curves for FIA-planted and naturally regenerated forests briefly converge in the 1990s, likely due to genetic advancements enhancing both forest types.

It is also worth noting that SI gain plateaued in recent establishment years beginning around 2008 for the Industrial dataset. We believe that three simultaneous components would influence this phenomenon: economic restrictions, limitation in development and deployment of new methods, and environmental changes. Regarding the first two, and while SI has not yet reached the empirical, physiological limit of 32 m at 25 years (Zhao et al., 2016), it may have reached a point of diminishing return on investment for additional inputs. At the same time, there was a significant spike in fertilizer prices and the crash of the housing market in 2008 dramatically reducing stumpage prices, particularly for sawtimber, which subsequently reduced fertilizer application (Albaugh et al., 2019). Two key processes happening at that time could be driving the observed behavior: (1) intensive management practices became less common, as they were seen as a "luxury" treatment, leading to a clearer

differentiation between managed and unmanaged sites; and (2) forest owners extended their rotation periods, reducing the introduction of newer genetic material, which had previously driven productivity improvements. This caused the curve to flatten and therefore highlighted management's residual effects again. Another possible reason for this reduction in productivity could be the shift in focus within tree breeding. Initially centered on productivity, tree breeding efforts later shifted toward optimizing tree straightness for higher quality sawtimber (Aspinwall et al., 2013; McKeand, 2019; McKeand et al., 2020). It is also likely that the physiological limit in SI for loblolly pine can only be achieved on a fraction of sites in the southeastern US. In terms of FIA data, this plateauing starts about 10 years before that of the industrial data, which may be an indicator of the potential change that will be observed for industrial land (although given that the management approaches are different among FIA and Industrial datasets, FIA stands are not useful for making predictions on industrial land).

Other limiting factors of loblolly pine productivity are the pests or diseases, such as the Nantucket pine tip moth (*Rhyacionia frustrana* Comstock) or the southern pine beetle (*Dendroctonus frontalis* Zimmermann). While these factors have been shown to cause decreases in productivity, it has also been demonstrated that active management can mitigate these negative effects (Nowak and Berisford, 2000; Carter and Foster, 2006; Asaro et al., 2017). Returning to the third component previously mentioned, the changing environment in temperature, rainfall, CO2 concentration, and nitrogen deposition, potentially contributesd to the limitation in productivity. For instance, a 1°C drop in U.S. temperatures in 2008 likely contributed to the plateau in productivity, yet while temperatures rose by nearly 2°C a decade later (Climate Change Knowledge Portal, 2024), more data is needed to assess how recent environmental changes will influence future trends.

Finally, another data-related restriction must be noted. Until the 1990s, the structure of forest industry companies shifted from a vertically integrated ownership (i.e., the companies owned and managed their own lands and processed their timber in their mills) to the alternate Real Estate Investment Trust (REITs) and Timber Investment Management Organization (TIMOs) models. Vertically integrated companies owned land for a long time and would have had relatively good (accurate) long-term records of what was done on a given area. The change in ownership structure created an increased churn where land would be bought and sold regularly resulting in a problematic transfer of the information about what had been done on a given piece of land. The result is that all the land would end up looking the same in the sense of recorded management inputs because the record keeping has become more haphazard over time.

4.2. Site index modeling and variable effect evaluation

The predictive performance achieved in this work was in line with previous studies (e.g., Cook et al., 2024), where approximately 70 % of the variability in the Industrial dataset and 40 % in the FIA dataset were explained. The underperformance of the FIA model when compared to the Industrial one can potentially be attributed to higher data variability and smaller sample size. While "planted" management in the FIA dataset may indicate an intention to participate in timber markets, the Industrial data was expected to represent a narrow, highly productive, subset of planted forests (i.e., a more "controlled" and consistent management environment). Additionally, the different nature of the type of observation between the FIA and Industrial datasets might be causing some of these differences too. For example, an observation in the Industrial dataset represents one single age, yet an observation within the FIA dataset corresponds to the average SI and age of all the FIA plots that fell within the same soil type. Therefore, because of the nature of the data collection and collation, there is more inherent variability in the FIA dataset, and the predictive performance cannot be directly compared among the models.

However, these models outperformed other models built with similar edaphic and climatic data: Sabatia and Burkhart (2014) reported a predictive ability (R^2) between 0.3 and 0.6 using annual precipitation, soil depth, water availability, a growing season dryness index, and elevation. When modeling height growth, Cohrs (2022) reported correlations between 0.3 and 0.4 using only soil information. Other studies that include much more detail of edaphic properties (e.g., nitrogen, phosphorus or potassium concentrations, pH or percentage of clay), have reported correlations up to 0.7 using parametric modeling techniques for SI (Subedi and Fox, 2016). Other research utilized water deficit, excess water, and available water to model dominant height, reporting prediction errors of < 1 m, which then was used to model SI in the Western Gulf of the US (Koirala et al., 2021). Jiang et al. (2015) reported an adjusted R^2 of 0.6 and a RMSE of 4.5 m when modeling conifer SI using climate and soils. It needs to be mentioned that, except Cook et al. (2024), previous models did not include establishment year as a covariate, and therefore the strong effect of steadily improving SI that we found was potentially missed. On the other hand, the sample size of the underlying data in all studies was smaller than the one presented in this study, potentially masking the strength of this effect.

Regarding variable importance, we observed a trend: the larger the area represented by a variable, the greater its impact on the model. Physiographic province, representing the broadest areas, emerged as a significant driver, followed by geology, which represents edaphic and climatic properties within each geological region. The observed benefits or restrictions different physiographic provinces or geologies provide was consistent with the reported Cook et al. (2024) when referring to management implications of each variable, at least in the southeastern US context. Other studies have found geology and physiographic province to be important drivers when modeling forest productivity in loblolly pine (Amateis et al., 2006; Everett and Thorp, 2008) and other forest environments (Hennigar et al., 2017; Moore et al., 2022). Parent material (signified by Geocode) is one of the soil formation factors that influences the inherent soil nutrient status and physical properties. Precipitation and maximum temperature were other important variables in our model and similar to Sabatia and Burkhart (2014), VanderSchaaf and Prisley (2006), and Jiang et al. (2015). However, our use of 30-year averages for climatic variables could be affecting the correlation between observed productivity and the climate variables because the data did not correspond to that exact period of growth (Bryars et al., 2013). Research with more detailed climatic data suggests that monthly averages of precipitation and temperature do affect loblolly pine productivity (VanderSchaaf and Prisley, 2006; Davis et al., 2022). Furthermore, vapor pressure deficit, another climatic variable ranked highly, has increased over the past decades and could negatively impact future productivity (Ficklin and Novick, 2017). Other studies indicate that limitations related to water availability (atmospheric or in the soil) or water excess can impact forest productivity (Koirala et al., 2021).

Once the variability from the large-scale factors is accounted for, then soils play a larger role in explaining variability of productivity (Jiang et al., 2015; Hennigar et al., 2017). We found the additional limitations or resources modifier was the most important soil variable, followed by drainage class. Additional limitations and resources primarily include variables for root restrictions which determine soil volume. We believe our large sample size influenced our identification of this variable as an important one because our soils were not biased toward those selected for fertilization field trials. Soil properties were ranked last in the model variable importance, which was also found by Cook et al. (2024) using the same SPOT code system to evaluate SI in a network of experimental stands in southeastern US loblolly pine. However, they found that major soil group (primarily texture), nature of subsoil, and drainage class were the most influential soil properties. In another study, Sabatia and Burkhart (2014) found that soil depth was influential when predicting SI from biophysical parameters. The differences in productivity due to soils is potentially more complex as soil factors may depend on and interact with management and landform. For example, Everett and Thorp (2008) found drainage class to be influential

when assessing site quality in the lower coastal plain, yet this effect can be partially overcome with modern site preparation techniques such as bedding (Fox et al., 2005). Different soil factors can be limiting in different contexts. The effect of soils is potentially more important at a smaller scale when assessing potential responses to silvicultural treatments than within broader physiographic or geological regions.

Regarding variable interactions, we found that year of establishment, geology, physiographic province, and some of the climatic variables (mainly, maximum temperature and precipitation), as well as origin of the plantation for FIA dataset, were the most strongly interacting of the variables of the model. However, we did not analyze which of the other variables they interacted the most with, as other studies did. For instance, Cohrs (2022) also found that physiographic province was one of the most interactive variables when modelling growth following random forest techniques, which especially interacted with runoff potential. Hennigar et al. (2017), when modeling a biomass-based site productivity estimate, found that interactions among terrain properties (e.g., slope and depth to water table) were important to distinguish productive and unproductive soils, something that could align with what in our study could be inferred to the additional limitations and resources variable – drainage pair of variables. Regarding climatic variables, Jiang et al. (2015), when modeling SI for different species in the eastern US, found that the interaction between precipitation and temperature was significant, something that aligns with our results.

4.3. Limitations and further directions

For the Industrial dataset, the results obtained in this work rely on the methodology proposed by Ribas-Costa et al. (2024), and therefore some of the limitations presented in that study apply to this work too. These include (1) the restrictions based on the availability of updated age records, which can have errors (le Maire et al., 2011); (2) differences due to variable USGS LiDAR acquisition quality, and (3) the increase in uncertainty when translating modeled dominant height to modeled SI. Related to the first of these conditions, this study is based on observational data, so all conclusions derived from it should be supported by past or future research. Furthermore, unlike other studies (e.g., Davis et al., 2022), where they were able to attribute to CO_2 increase and other climate change-derived variables a specific percentage of effect on the productivity increase, our study encapsulates all those effects in the establishment year variable. It is also worth noting the way climate data is accounted for in each dataset (FIA vs. Industrial): whereas in the latter dataset, climate data was added at a soil-within-stand level, in the former dataset climate data had to be averaged by SPOT code.

We saw a general trend in the underestimation of high SI values and overestimation of the low SI values, a phenomenon known as regression to the mean. This effect is a very common outcome when modeling ecological processes (Mazalla and Diekmann, 2022), and was also found in other studies modeling SI in loblolly pine (Sabatia and Burkhart, 2014). Another potential source of error is the fact that the method for estimating SI from dominant height used in this work, the Diéguez-Aranda et al. (2006) model, was developed based on older stands. Potentially, the use of a more recent SI model, that includes more recent data and data from more intensively managed pine plantations, could help reduce the uncertainty in both the models and their outcomes for more recent stands. It is also worth noting that the year of establishment in the industrial data comes from plantation records that start counting from age 0. However, in FIA data, the year of establishment is computed as a function of the age at measurement, obtained from bole cores at taken at breast height, and is therefore an estimate subject to measurement error. There is also an intrinsic difference between the two estimates, as FIA age estimates do not consider the time that the tree required to reach the breast height, which will be several years.

A better understanding of current forest productivity opens several opportunities for land managers, foresters, or other stakeholders to improve their estimations and models of both timber volume and carbon stock. The proposed model will allow for (1) a spatially continuous map of SI based on average management (i.e., year of establishment and silvicultural practices) and (2) a spatially explicit SI calculator so that each forest owner can correct the previous estimation with their own management data. Our models could be used to assess potential productivity, by comparing the observed productivity with the modeled one given specific condition. Furthermore, given the variables included in our models, especially maximum temperature and average precipitation, future research can examine climate change scenarios, addressing how changes in current temperature, precipitation, and water availability affect the expected productivity of the forests. Although we can project future productivity trends with increasing temperatures and changes in rainfall, its combined effect in current forests will be less predictable when past thresholds are surpassed (although that entangles the risks related to temporal extrapolation). Other stochastic events may interfere with predictions such as heat waves, droughts, excess rainfall, or hurricanes.

5. Conclusion

In this study, we explored the key drivers of the evolution of site index across different forest management levels. Our finding consistently highlights the year of establishment as a major factor influencing forest productivity. Specifically, it has been responsible for an increase in SI of more than 5 m in industrial plantations since the 1970s and more than 3 m in the FIA dataset since the 1950s. Moreover, both datasets reveal that broad factors, such as physiographic provinces or geologic regions, which share similar edaphic and climatic conditions, significantly impact forest productivity at the landscape scale. Climate variables, particularly precipitation and maximum temperature, emerge as strong drivers of forest productivity. Within soil properties, soil volume was the most influential factor. Improving genetic material via traditional tree breeding techniques and implementing better silvicultural practices is a crucial solution when addressing climate change and carbon sequestration and meeting global wood fiber demands.

CRediT authorship contribution statement

Vicent A. Ribas-Costa: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Aitor Gastón:** Writing – review & editing, Validation, Methodology, Formal analysis. **Sean A. Bloszies:** Methodology, Investigation. **Jesse D. Henderson:** Writing – review & editing, Resources, Methodology, Investigation. **Andrew Trlica:** Writing – review & editing, Methodology. **David R. Carter:** Writing – review & editing, Methodology. **Rafael Rubilar:** Writing – review & editing, Methodology. **Timothy J. Albaugh:** Writing – review & editing, Methodology. **Rachel L. Cook:** Writing – review & editing, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Data curation, Conceptualization.

Declaration of Competing Interest

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.

Acknowledgements

We gratefully acknowledge the continued support of the Forest Productivity Cooperative. This work would not have been possible without the generous contribution of data from Forest Productivity Cooperative members BTG Pactual, Campbell Global, LLC, Forest Investment Associates, Jordan Lumber & Supply, Manulife Investment Management, Roseburg Resources Co., and others. Vicent A. Ribas-Costa

was funded by the Doctoral INPhINIT Retaining fellowship n. º LCF/BQ/DR22/11950023 from 'la Caixa' Foundation (ID 100010434). Other funding sources include the program ANID BASAL (FB210015) and NSF I/UCRC Center for Advanced Forestry Systems (Award #1916552). We also receive support from the North Carolina State University Department of Forestry and Environmental Resources, Virginia Tech Departement of Forest Resources and Environmental Conservation, the Facultad de Ciencias Forestales, Universidad de Concepcion, and the Departamento de Ciencias Florestais, Universidade Federal de Lavras.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.foreco.2024.122334.

Data availability

USGS LiDAR is openly available, as well as climate data; FIA data is managed by USFS at USDA, and the SPOT code is a proprietary system owned by the FPC.

References

Aguilos, M., Sun, G., Noormets, A., Domec, J., McNulty, S., Gavazzi, M., Prajapati, P., Minick, K.J., Mitra, B., King, J., 2021. Ecosystem Productivity and Evapotranspiration Are Tightly Coupled in Loblolly Pine (Pinus taeda L.) Plantations along the Coastal Plain of the Southeastern U.S. Forests 12 12 (8), 1123. https://doi.org/10.3390/f12081123.

Albaugh, T.J., Fox, T.R., Cook, R.L., Raymond, J.E., Rubilar, R.A., Campoe, O.C., 2019. Forest Fertilizer Applications in the Southeastern United States from 1969 to 2016. For. Sci. 65, 355–362. https://doi.org/10.1093/forsci/fxy058.

Allen, L.H., Fox, T.R., Campbell, R.G., 2005. What is Ahead for Intensive Pine Plantation Silviculture in the South? South. J. Appl. For. 29, 62–69. https://doi.org/10.1093/sjaf/29.2.62.

Allen, M.G., Burkhart, H.E., 2015. A comparison of alternative data sources for modeling site index in loblolly pine plantations. Can. J. For. Res. 45 (8), 1026–1033. https://doi.org/10.1139/cjfr-2014-0346.

Altmann, A., Toloşi, L., Sander, O., Lengauer, T., 2010. Data and text mining Permutation importance: a corrected feature importance measure. Bioinformatics 26, 1340–1347. https://doi.org/10.1093/bioinformatics/btq134.

Amateis, R.L., Prisley, S.P., Burkhart, H.E., Liu, J., 2006. The Effect of Physiographic Region and Geographic Locale on Predicting the Dominant Height and Basal Area of Loblolly Pine Plantations. South. J. Appl. For. 30, 147–153. https://doi.org/10.1093/sjaf/30.3.147.

Antoniadis, A., Lambert-Lacroix, S., Poggi, J.M., 2021. Random forests for global sensitivity analysis: A selective review. Reliab. Eng. Syst. Saf. 206, 107312. https://doi.org/10.1016/J.RESS.2020.107312.

Apley, D.W., Zhu, J., 2020. Visualizing the effects of predictor variables in black box supervised learning models. J. R. Stat. Soc. Ser. B: Stat. Metho. 82, 1059–1086. https://doi.org/10.1111/rssb.12377.

Asaro, C., Nowak, J.T., Elledge, A., 2017. Why have southern pine beetle outbreaks declined in the southeastern U.S. with the expansion of intensive pine silviculture? A brief review of hypotheses. For. Ecol. Manag. 391, 338–348. https://doi.org/10.1016/j.foreco.2017.01.035.

Aspinwall, M.J., McKeand, S.E., King, J.S., 2012. Carbon Sequestration from 40 Years of Planting Genetically Improved Loblolly Pine across the Southeast United States. For. Sci. 58, 446–456. https://doi.org/10.5849/forsci.11-058.

Aspinwall, M.J., King, J.S., McKeand, S.E., 2013. Productivity differences among loblolly pine genotypes are independent of individual-tree biomass partitioning and growth efficiency. Trees 27, 533–545. https://doi.org/10.1007/s00468-012-0806-4.

Breiman, L., Last, M., Rice, J., 2001. Random forests: Finding quasars. Statistical Challenges in Astronomy. Springer, New York, New York, NY, pp. 243–254.

Bryars, C., Maier, C., Zhao, D., Kane, M., Borders, B., Will, R., Teskey, R., 2013. Fixed physiological parameters in the 3-PG model produced accurate estimates of loblolly pine growth on sites in different geographic regions. For. Ecol. Manag. 289, 501–514. https://doi.org/10.1016/j.foreco.2012.09.031.

Burkhart, H.E., Brooks, E.B., Dinon-Aldridge, H., Sabatia, C.O., Gyawali, N., Wynne, R. H., Thomas, V.A., 2018. Regional Simulations of Loblolly Pine Productivity with CO2 Enrichment and Changing Climate Scenarios. For. Sci. 64, 349–357. https://doi.org/10.1093/forsci/fxy008.

Carter, M.C., Foster, C.D., 2006. Milestones and millstones: A retrospective on 50 years of research to improve productivity in loblolly pine plantations. For. Ecol. Manag. 227, 137–144. https://doi.org/10.1016/j.foreco.2006.02.014.

Climate Change Knowledge Portal 2024. Observed Annual Mean Surface Air Temperature of the US for the 1901 – 2022 period. ⟨https://climateknowledgeportal.worldbank.org/country/united-states/climate-data-historical⟩ [Accessed: 27/09/2024].

Cheng, L., De Vos, J., Zhao, P., Yang, M., Witlox, F., 2020. Examining non-linear built environment effects on elderly's walking: A random forest approach. Transp. Res. D. Transp. Environ. 88, 102552. https://doi.org/10.1016/J.TRD.2020.102552.

Clay, L., Motallebi, M., Song, B., 2019. An Analysis of Common Forest Management Practices for Carbon Sequestration in South Carolina. Forests 10 (11), 949. https://doi.org/10.3390/f10110949.

Clutter, J.L., Fortson, J.C., Pienaar, L.V., Brister, G.H., Bailey, R.L., 1992. Timber Management: A Quantitative Approach (Reprint. ed.). John Wiley & Sons Inc, New York.

Cohrs, C.W., 2022. Optimizing pine plantation management via geospatial data science. PhD dissertation, North Carolina State University. ⟨https://www.researchgate.net/publication/362112608⟩ [Accessed 01/31/2024].

Cook, R., Fox, T.R., Allen, H.L., Cohrs, C.W., Ribas-Costa, V., Trlica, A., Ricker, M., Carter, D.R., Rubilar, R., Campoe, O., Albaugh, T.J., Kleto, P., O'Brien, E., McEachern, K., 2024. Forest soil classification for intensive pine plantation management: "Site Productivity Optimization for Trees" system. For. Ecol. Manag. 556, 121732. https://doi.org/10.1016/j.foreco.2024.121732.

Davis, E.C., Sohngen, B., Lewis, D.J., 2022. The effect of carbon fertilization on naturally regenerated and planted US forests. Nat. Commun. 13. https://doi.org/10.1038/s41467-022-33196-x.

Diéguez-Aranda, U., Burkhart, H., Amateis, R., 2006. Dynamic site model for loblolly pine (*Pinus taeda* L.) plantations in the United States. For. Sci. 52, 262–272. https://doi.org/10.1093/forsci/fxz050.

Everett, C.J., Palm-Leis, H., 2009. Availability of residual phosphorus fertilizer for loblolly pine. For. Ecol. Manag. 258, 2207–2213. https://doi.org/10.1016/j.foreco.2008.11.029.

Everett, C.J., Thorp, J.H., 2008. Site quality evaluation of loblolly pine on the South Carolina Lower Coastal Plain, USA. J. For. Res. 19, 187–192. https://doi.org/10.1007/s11676-008-0033-4.

Ficklin, D.L., Novick, K.A., 2017. Historic and projected changes in vapor pressure deficit suggest a continental-scale drying of the United States atmosphere. J. Geophys. Res. Atmospheres 122, 2061–2079. https://doi.org/10.1002/2016JD025855.

Fox, T.R., Kyle, K.H., Andrews, L.J., Aust, W.M., Burger, J.A., Hansen, G.H., 2005. Long-Term Effects of Drainage, Bedding, and Fertilization on Growth of Loblolly Pine (Pinus taeda L.) in the Coastal Plain of Virginia. South. J. Appl. For. 29, 205–214. https://doi.org/10.1093/sjaf/29.4.205.

Fox, T.R., Jokela, E.J., Allen, H.L., 2007b. The Development of Pine Plantation Silviculture in the Southern United States. J. For. 105 (7), 337–347. https://doi.org/10.1093/jof/105.7.337.

Fox, T.R., Allen, H.L., Albaugh, T.J., Rubilar, R., Carlson, C.A., 2007a. Tree Nutrition and Forest Fertilization of Pine Plantations in the Southern United States. South. J. Appl. Fortry. 31 (1), 5–11. https://doi.org/10.1093/sjaf/31.1.5.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat.: 1189-1232.

Friedman, J.H., Popescu, B.E., 2008. Predictive learning via rule ensembles. Ann. Appl. Stat. 2 (3), 916–954. https://doi.org/10.1214/07-AOAS148.

Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E., 2015. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. J. Comp. Graph. Stat. 24 (1), 44–65. https://doi.org/10.1080/10618600.2014.907095.

Gyawali, N., Burkhart, H.E., 2015. General response functions to silvicultural treatments in loblolly pine plantations. Can. J. For. Res. 45 (3), 252–265. https://doi.org/10.1139/cjfr-2014-0172.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Springer Series in Statistics. Springer.

Hennigar, C., Weiskittel, A., Allen, H.L., Maclean, D.A., 2017. Development and evaluation of a biomass increment-based index for site productivity. Can. J. For. Res. 47, 400–410. https://doi.org/10.1139/CJFR-2016-0330.

Huang, J., Bergeron, Y., Denneler, B., Berninger, F., Tardif, J., 2007. Response of Forest Trees to Increased Atmospheric CO2. Crit. Rev. Plant Sci. 26, 265. https://doi.org/10.1080/07352680701626978.

Isabel, N., Holliday, J.A., Aitken, S.N., 2019. Forest genomics: Advancing climate adaptation, forest health, productivity, and conservation. Evolut. Appl. 13, 1–241. https://doi.org/10.1111/eva.12902.

Jeong, J.H., Resop, J.P., Mueller, N.D., Fleisher, D.H., Yun, K., Butler, E.E., Timlin, D.J., Shim, K.-M., Gerber, J.S., Reddy, V.R., Kim, S.-H., 2016. Random Forests for Global and Regional Crop Yield Predictions. Plos One 11 (2), e0156571. https://doi.org/10.1371/journal.pone.0156571.

Jiang, H., Radtke, P.J., Weiskittel, A.R., Coulston, J.W., Guertin, P.J., 2015. Climate- and soil-based models of site productivity in eastern US tree species. Can. J. For. Res. 45 (3), 325–342. https://doi.org/10.1139/cjfr-2014-0054.

Johnston, C.M.T., Guo, J., Prestemon, J.P., 2022. RPA forest products market data for U. S. RPA Regions and the world, historical (1990-2015), and projected (2020-2070) using the Forest Resource Outlook Model (FOROM). https://doi.org/10.2737/RDS-2022-0073.

Jokela, E.J., Dougherty, P.M., Martin, T.A., 2004. Production dynamics of intensively managed loblolly pine stands in the southern United States: a synthesis of seven long-term experiments. For. Ecol. Manag. 192, 117–130. https://doi.org/10.1016/j.foreco.2004.01.007.

Koirala, A., Montes, C.R., Bullock, B.P., 2021. Modeling dominant height using stand and water balance variables for loblolly pine in the Western Gulf, US. For. Ecol. Manag. 479, 118610. https://doi.org/10.1016/j.foreco.2020.118610.

le Maire, G., Marsden, C., Nouvellon, Y., Grinand, C., Hakamada, R., Stape, J., Laclau, J., 2011. MODIS NDVI time-series allow the monitoring of Eucalyptus plantation biomass. Remote Sens. Environ. 115, 2613–2625. https://doi.org/10.1016/j.rse.2011.05.017.

Lefsky, M.A., Turner, D.P., Guzy, M., Cohen, W.B., 2005. Combining lidar estimates of aboveground biomass and Landsat estimates of stand age for spatially extensive validation of modeled forest productivity. Remote Sens. Environ. 95, 549–558. https://doi.org/10.1016/j.rse.2005.01.004.

Little, E.L., Jr., 1971. Atlas of United States trees. Volume 1. Conifers and important hardwoods. Miscellaneous Publication 1146. Washington, DC: U.S. Department of Agriculture, Forest Service. 9 p., illus. [313 maps, folio].

Loecher, M., 2022. Unbiased variable importance for random forests. Commun. Stat. Theory Methods 51, 1413–1425. https://doi.org/10.1080/03610926.2020.1764042.

Loef, B., Wong, A., Janssen, N.A.H., Strak, M., Hoekstra, J., Picavet, H.S., Boshuizen, H.C. H., Verschuren, W.M.M., Herber, G.M., 2022. Using random forest to identify longitudinal predictors of health in a 30-year cohort study. Sctfic. Rep. 12, 10372. https://doi.org/10.1038/s41598-022-14632-w.

Mazalla, L., Diekmann, M., 2022. Regression to the mean in vegetation science. J. Veg. Sci. 33. https://doi.org/10.1111/jvs.13117.

McKeand, S.E., 2015. The success of tree breeding in the southern US. BioResources 10, 1–2.

McKeand, S.E., 2019. The evolution of a seedling market for genetically improved loblolly pine in the Southern United States. J. For. 117, 293–301. https://doi.org/10.1093/jofore/fvz006.

McKeand, S.E., Payn, K.G., Heine, A.J., Abt, R.C., 2020. Economic Significance of Continued Improvement of Loblolly Pine Genetics and Its Efficient Deployment to Landowners in the Southern United States. J. For. 119, 62–72. https://doi.org/10.1093/jofore/fvaa044.

Mentch, L., Zhou, S., 2020. Randomization as Regularization: A Degrees of Freedom Explanation for Random Forest Success. J. Mach. Learn. Res. 21 (171), 1–36. ⟨https://jmlr.org/papers/v21/19-905.html⟩.

Molnar, C., Bischl, B., Casalicchio, G., 2018. iml: An R package for Interpretable Machine Learning. J. Opn. Srce. Soft. 3 (26), 786. https://doi.org/10.21105/joss.00786.

Moore, J.A., Kimsey, M.J., Garrison-Johnston, M., Shaw, T.M., Mika, P., Poolakkal, J., 2022. Geologic Soil Parent Material Influence on Forest Surface Soil Chemical Characteristics in the Inland Northwest, USA. Forests 13 (9), 1363. https://doi.org/10.3390/f13091363.

Nembrini, S., König, I.R., Wright, M.N., 2018. The revival of the Gini importance? Bioinformatics 34, 3711–3718. https://doi.org/10.1093/bioinformatics/bty373.

Nowak, J.T., Berisford, C.W., 2000. Effects of Intensive Forest Management Practices on Insect Infestation Levels and Loblolly Pine Growth. J. Econ. Entomol. 93, 336–341. ⟨https://www.doi.orb/10.1603/0022-0493-93.2.336⟩.

Oswalt, S.N., Smith, W.B., Miles, P.D., Pugh, S.A., 2019. Forest resources of the United States, 2017: A technical document supporting the Forest Service 2020 RPA Assessment (No. WO-GTR-97). U.S. Department of Agriculture, Forest Service, Washington, DC. https://doi.org/10.2737/WO-GTR-97.

Pahlavan-Rad, M.R., Dahmardeh, K., Hadizadeh, M., Keykha, G., Mohammadnia, N., Gangali, M., Keikha, M., Davatgar, N., Brungard, C., 2020. Prediction of soil water infiltration using multiple linear regression and random forest in a dry flood plain, eastern Iran. Catena 194, 104715. https://doi.org/10.1016/j.catena.2020.104715.

Probst, P., Wright, M.N., Boulesteix, A., 2019. Hyperparameters and tuning strategies for random forest. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 9, e1301. https://doi.org/10.1002/widm.1301.

Puls, S.J., Cook, R.L., Baker, J.S., Rakestraw, J.L., Trlica, A., 2024. Modeling wood product carbon flows in southern us pine plantations: implications for carbon storage. Carbon Balance Manag. 19, 8. https://doi.org/10.1186/s13021-024-00254-4.

Reams, A.G., Smith, W.D., Hansen, M.H., Bechtold, W.A., Roesch, F.A., Moisen, G.G., 2005. The Forest Inventory and Analysis Sampling Frame". In: Bechtold, W.A., Patterson, P.L. (Eds.), The Enhanced Forest Inventory and Analysis Program – national Sampling Design and Estimation Procedures. USDA Forest Service, Southern Research Station.

Restrepo, H.I., Bullock, B.P., Montes, C.R., 2019. Growth and yield drivers of loblolly pine in the southeastern U.S.: A meta-analysis. For. Ecol. Manag. 435, 205–218. https://doi.org/10.1016/j.foreco.2018.12.00.

Ribas-Costa, V.A., Gastón-González, A., Cook, R.L., 2024. Modeling dominant height with USGS 3DEP LiDAR to determine site index in even-aged loblolly pine (Pinus taeda L.) plantations in the southeastern US. For.: Int. J. For. Res., cpae034 https://doi.org/10.1093/forestry/cpae034.

Sabatia, C.O., Burkhart, H.E., 2014. Predicting site index of plantation loblolly pine from biophysical variables. For. Ecol. Manag. 326, 142–156. https://doi.org/10.1016/j.foreco.2014.04.019.

Skovsgaard, J.P., Vanclay, J.K., 2008. Forest site productivity: a review of the evolution of dendrometric concepts for even-aged stands. For.: Int. J. For. Res. 81, 13–31. https://doi.org/10.1093/forestry/cpm041.

Subedi, P., Jokela, E.J., Vogel, J.G., Martin, T.A., 2019. Sustained productivity of intensively managed loblolly pine plantations: Persistence of fertilization and weed control effects across rotations. For. Ecol. Manag. 446, 38–53. https://doi.org/10.1016/j.foreco.2019.05.025.

Subedi, S., Fox, T.R., 2016. Predicting Loblolly Pine Site Index from Soil Properties Using Partial Least-Squares Regression. For. Sci. 62, 449–456. https://doi.org/10.5849/forsci.15-127.

Susaeta, A., Adams, D.C., Carter, D.R., Gonzalez-Benecke, C., Dwivedi, P., 2016. Technical, allocative, and total profit efficiency of loblolly pine forests under changing climatic conditions. For. Policy Econ. 72, 106–114. https://doi.org/10.1016/j.forpol.2016.06.021.

Thomas, V.A., Wynne, R.H., Kauffman, J., McCurdy, W., Brooks, E.B., Thomas, R.Q., Rakestraw, J., 2021. Mapping thins to identify active forest management in southern pine plantations using Landsat time series stacks. Remote Sens. Environ. 252, 112127. https://doi.org/10.1016/j.rse.2020.112127.

Van Lear, D.H., Harper, R.A., Kapeluck, P.R., Carroll, W.D., 2004. History of Piedmont Forests: Implications For Current Pine Management. Gen. Tech. Rep. SRS-71. U.S. Department of Agriculture, Forest Service, Southern Research Station, Asheville, NC, pp. 127–131.

VanderSchaaf, C.L., Prisley, S.P., 2006. In: Prisley, S., Bettinger, P., Hung, I.-K., Kushla, J. (Eds.), Factors affecting site productivity of loblolly pine plantations across the southeastern United States. Proceedings of the 5th Southern Forestry and Natural Resources GIS Conference, June 12-14, 2006, Asheville, NC. Warnell School of Forestry and Natural Resources, University of Georgia, Athens, GA.

Weiskittel, A.R., Hann, D.W., Kershaw, J.A., Vanclay, J.K., 2011. Forest site evaluation in forest growth and yield modeling. John Wiley & Sons, Ltd, West Sussex, UK.

Wright, M.N., Ziegler, A., 2017. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J. Stat. Soft. 77 (1), 1–17. https://doi.org/10.18637/jss.v077.i01.

Zhao, D., Bullock, B.P., Wang, M., 2023. Long-term dynamics of aboveground carbon stocks in managed loblolly pine plantations in the southeast United States. For. Ecol. Manag. 546, 121384. https://doi.org/10.1016/j.foreco.2023.121384.

Zhao, D., Kane, M., Teskey, R., Fox, T.R., Albaugh, T.J., Allen, H.L., Rubilar, R., 2016. Maximum response of loblolly pine plantations to silvicultural management in the southern United States. Forest Ecol. Manag. 375, 105–111. https://doi.org/10.1016/j.foreco.2016.05.035.