Nonparametric Individual-Tree Growth Models for the Northern Forest

Neal F. Maker and John D. Foppert[1]

Abstract.—Tree growth modeling is important for forest managers, but is difficult in structurally and compositionally diverse forests where growth depends on complex ecological relationships. Many of those relationships are not well studied, and traditional parametric models struggle to adequately account for them. We describe a nonparametric machine learning approach to modeling, which uses data to determine the methodology, accommodates high dimensional data, and could be well suited to capturing complicated, poorly understood interactions between predictors. We compare this approach to the traditional parametric approach by constructing diameter increment models for the United States northern forest region. In this case, the approach led us to build a gradient boosting machine that decreased the root-mean-squared error by 6.5 percent over the traditional, nonlinear least squares model, and which was especially accurate in predicting the growth of trees in less common situations, such as especially large trees and trees in forests with below- and above-average basal areas. Analysis of the model structure showed that its predictions are driven by multiple complex interactions. We explore several of them and discuss the potential of the nonparametric machine learning approach for predictive tree growth modeling and for studying ecological processes underlying tree growth.

INTRODUCTION

Forest managers depend on tree growth models to evaluate management approaches and establish forest valuations, but modeling has proven difficult in mixed-species and multi-aged forests where growth is driven by particularly complex ecological relationships. For example, recent research has demonstrated that the species identity of neighboring trees can influence growth, and that the effects of species mixing depend on the species involved (Canham et al. 2004, Fien et al. 2019, Pretzsch 2017). In many cases, trees in mixed-species stands have higher potential growth rates than those in mono-specific stands because of positive interactions between species (Pretzsch and Zenner 2017). Yet very few growth models account for species mixing, preventing managers from accounting for mixing effects in forest valuations and inhibiting efforts to take advantage of mixing effects through silvicultural interventions. Stand structural diversity is similarly absent from most operational growth models but has been shown to relate to tree growth in complicated ways (Ali 2019). Measures of structural diversity are positively correlated to growth in some settings (e.g., Dănescu et al. 2016, Lei et al. 2009) and negatively correlated in others (e.g., Edgar and Burk 2001, Liang et al. 2005), depending on the particular species and their development. These are complicated relationships that have not been studied in most forest types (O'Hara 2014, Pretzsch et al. 2017).

One approach to modeling growth in complex forests has been to shift analytical attention from stand to neighborhood dynamics (Canham and Uriarte 2006). This approach conceptualizes a forest stand as a collection of distinct patches of land (neighborhoods) that are typically 0.01 to 0.1 hectares in size and are assumed to be developing independently from one another (Bugmann

[1] Consulting Foresters (NFM and JDF), Pekin Branch Forestry, 1324 West County Road, Calais, VT 05648; Assistant Professor of Silviculture and Economics (JDF), Paul Smith's College. NFM is corresponding author: to contact, call 802-229-9757 or email neal@pekinbranch.com.

2001). Emphasis on neighborhood dynamics grew out of ecological research demonstrating that plant-to-plant interactions are primarily local, neighbors' effects attenuate with distance, and beyond a certain distance, plants have no detectable effect on one another (Stoll and Weiner 2000). To simplify analyses and data collection, neighborhood-scale models can ignore the specific locations of trees within each neighborhood, thereby offering a compromise between high resolution, difficult to apply distance-dependent models, and low resolution, easy to apply stand-scale models (Bugmann 2001). Historically, neighborhood-scale models have been used to model long-term system-level dynamics (e.g., Forsyth et al. 2015), often under the moniker "gap models," and the majority of research attention has been directed toward representing regeneration, recruitment, and mortality based on an understanding of ecological processes (Weiskittel et al. 2011). These "mechanistic" models rely on causal relationships and ecological theories and are well suited to studying ecosystem changes (Cuddington et al. 2013), but they are often less accurate than empirical models in predicting overstory tree growth, more difficult to initialize, and incapable of capturing ecological processes that have not been well studied (Taylor et al. 2009), limiting their application for practitioners.

There are also many statistical forest growth models built for operational decision-making (Weiskittel et al. 2011). These are generally parametric models constructed with traditional regression-based techniques, which predict the growth of individual trees and employ stand-scale competition metrics (Weiskittel et al. 2011). These models could be adapted to a neighborhood framework by incorporating neighborhood-scale competition metrics in place of stand-scale metrics, potentially improving overstory growth predictions and making neighborhood-scale models easier to initialize. Still, existing parametric models fail to account for many complex and poorly understood effects, which could limit their accuracy usefulness in mixed-species, multi-aged forests.

In this paper, we explore the potential for integrating nonparametric models into the neighborhood-scale model framework, using a machine learning approach. Nonparametric models are especially good at uncovering complex and poorly understood relationships in data (Breiman 2001). They use data to determine their own structures and make minimal assumptions about the processes that generated the data. This is in contrast to the traditional parametric approach, in which the data are assumed to have come from a stochastic model and previous research guides the selection of a model form and the predictors to be used (Breiman 2001). Nonparametric models are trained using machine learning, which is a branch of artificial intelligence that uses algorithms to allow computers to "learn" from data. Breiman (2001) points out three advantages of machine learning over traditional modeling, which he refers to as the three important "lessons" of machine learning: *First*, machine learning recognizes that many different models can do a good job explaining the same phenomenon and takes an experimental approach to model selection; *second*, machine learning does not shy away from complex model formulations when they can improve predictions; and *third,* machine learning accommodates high dimensional data, seeking relationships among many interrelated variables. For these reasons, it has been successful in modeling complex processes with interrelated predictors and in situations where extensive or high-dimensional data are available.

Nonparametric models do have drawbacks, though. Because they are entirely data-driven, they do not ensure biologically robust behavior, which could be problematic when using models for long-term simulations. They also cannot extrapolate well beyond the limits of their training data and tend to be more computationally demanding and more difficult to interpret.

Still, it appears that nonparametric techniques could be helpful for modeling tree growth in mixed-species, multi-aged forests, since large amounts of forest inventory data have been

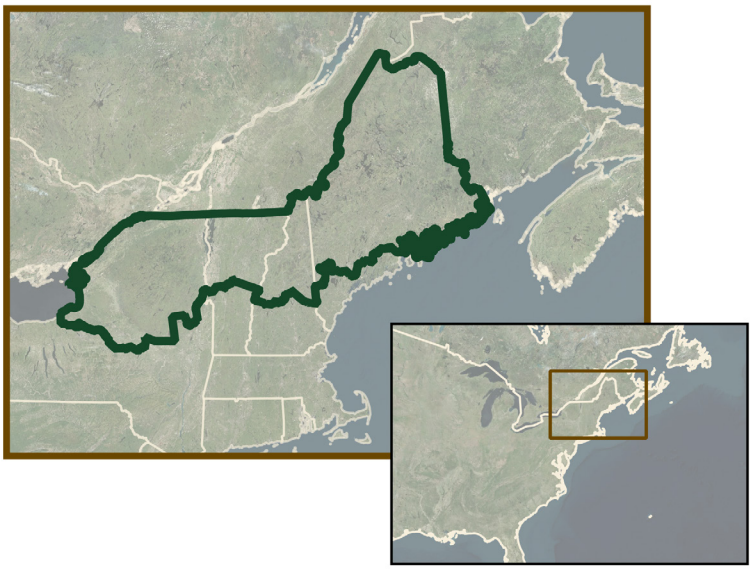

Figure 1.—Northern forest region.

collected in many places over extended time periods (Kershaw et al. 2017), which could provide sufficient information for nonparametric models and alleviate problems with extrapolation. Forest ecologists have begun using nonparametric models regularly for predicting nonlinear ecosystem behaviors (Liu et al. 2018), but their application to forest growth modeling has mostly been limited to tree mortality (e.g., Fan et al. 2006, Guan and Gertner 1991, King et al. 2000), which has been especially hard to predict using traditional parametric techniques.

We illustrate the nonparametric approach by constructing a neighborhood-scale, tree diameter increment model, using extensive inventory data from the ecologically complex United States northern forest region. We compare the nonparametric model to a second, parametric model that we parameterized to the same dataset to get a sense of its efficacy, and we explore a few of the unexpected relationships that it uses to guide its predictions. Finally, we discuss the potential for nonparametric techniques in predicting and studying tree growth in complex forests.

STUDY AREA

The "northern forest" covers 36 counties in the New York Adirondacks and northern Vermont, New Hampshire, and Maine (Fig. 1). It is located between the temperate forests of the eastern United States and the boreal forests of Canada, and contains characteristics of both biomes, giving it a diverse mixture of tree species and site conditions. Trees in the training data (described in the next section) belong to 28 different species groups and are located in 17 different forest types and 14 different landscape positions. The most common species groups (by number of stems) are fir (*Abies balsamea* (L.) Mill.), spruce (mostly *Picea rubens* Sarg. and *Picea mariana* (Mill.) Britton, Sterns & Poggenburg), soft maple (*Acer rubrum* L.), hard maple (*Acer saccharum* Marshall), cedar (*Thuja occidentalis* L.), beech (*Fagus grandifolia* Ehrh.), yellow birch (*Betula alleghaniensis* Britt.), paper birch (*Betula papyrifera* Marshall), hemlock (*Tsuga canadensis* (L.) Carrière), and eastern white pine (*Pinus strobus* L.). Elevations in the region range from 0 to 1,917 meters above sea level.

The region is well suited to studying the complexities of tree growth because its forests tend to be complex. Most are naturally regenerated, second-growth forests with varied management histories. In the training data, some 68 percent of plots contain three or more species growing together, while the spread of within-plot tree diameters averages 35 cm.

DATA

Data were obtained from the USDA Forest Service's Forest Inventory and Analysis (FIA) program (https://apps.fs.usda.gov/fia/datamart/datamart.html, downloaded December 6, 2020) from remeasured plots in the northern forest region. We used data from the current national standard plot design, which was adopted in the mid-1990s (Woudenberg et al. 2010). These plots are systematically located and evenly distributed across the region. Each plot is made up of four 168-m^2 subplots with a 13.5-m^2 microplot inside each subplot. Trees greater than 12.7 cm diameter at breast height (d.b.h.) are measured on subplots, and trees between 2.54 and 12.7 cm d.b.h. are measured on microplots.

FIA data relate to numerous tree-, plot-, and stand-level attributes, and are collected through ground sampling and remote sensing. Our objective was to build a model for operational decision-making, so we focused on attributes that can be measured reasonably easily by practitioners. They include d.b.h., species, compacted crown ratio (CR), crown class, tree class (a measure of general quality), forest type, stocking level, landscape (physiographic) position, site class, slope, aspect, elevation, latitude, and longitude. We also used existing data to calculate a number of new attributes, including subplot basal area (BA), tree overtopping basal area (BAL), and species-specific subplot BA and tree BAL. To keep the data and resulting model relevant to practitioners, we grouped tree species into species groups and FIA forest types into more general forest types. Our groupings were designed to match common inventory protocols that we have observed. Finally, we removed trees from the analysis that died during the remeasurement period and removed all plots with missing data. The final dataset includes 399,358 observations from 11,286 different plots. Remeasurement intervals on the plots ranged from 2.9 to 7.8 years and averaged 5.1 years.

Twenty percent of the plots were randomly selected, and their data were set aside for testing final models, to account for over-fitting and allow a fair comparison between different modeling approaches. This splitting was based on plots rather than individual trees to make sure that the testing data are truly independent. The data from the remaining 80 percent of plots were used for exploratory analysis, to compare different modeling methodologies, and to train the final models.

All of the data acquisition and processing, and the analysis that follows, was done using the programming software R, version 4.0.1 (R Core Team 2020). Much of the general data manipulation and visualization was carried out with R Tidyverse packages (Wickham et al. 2019); models were trained, tuned, and compared using Kuhn et al.'s (2020) "caret" framework. Detailed methods are available at https://github.com/nealmaker/tree-growth-nf.

METHODS

Model Development

We used all 18 aforementioned predictors to train our nonparametric diameter growth model. Remeasurement intervals varied in the data, and we annualized diameter growth rates to make them comparable. To our knowledge, no model-agnostic, nonlinear interpolation methods exist for annualizing remeasurement data, so we assumed linear growth through the remeasurement period and trained models on average annual d.b.h. growth rates.

We used a number of pre-processing steps on predictors so they could be widely used across methods and to increase model accuracy (Kuhn and Johnson 2013). Nominal categorical predictors were dummy-encoded; numeric predictors were centered, scaled, and normalized using Yeo-Johnson transformations (Yeo and Johnson 2000); and highly correlated predictors were filtered.

While parametric models prevent over-fitting by using simplified, predetermined model forms that cannot be bent out of shape to fit the noise in the training data, nonparametric models are more flexible, and over-fitting must be deliberately addressed (Hastie et al. 2017). We did this with 10-fold cross validation and model tuning, using the *recipes* framework developed by Kuhn and Wickham (2020). Pre-processing steps were applied independently to each training in the cross validation, to prevent information from the reserved fold from biasing the estimated prediction error.

We spot-checked 18 different algorithms representing a wide variety of modeling methods to find the most promising ones. We were most concerned with accuracy, which we measured using root-mean-squared error (RMSE), but we also compared prediction speed and model object size (megabytes) to account for how easily models could be incorporated into growth simulators. We used a random subset of the training data and parallel processing to speed the process. Nonlinear methods like model trees, random forests, gradient boosting machines, and support vector machines were the most accurate in this particular case, while regularized models were less accurate, but smaller and faster at making predictions. We chose a gradient boosting machine as the best compromise, because it offered high accuracy with relatively fast prediction speeds and was reasonably small compared to many of the other nonlinear methods.

Having chosen the best algorithm for the application, we reparameterized the model using all of the training data and a more organized tuning process described by Boehmke and Greenwell (2020, section 12.3.3). The final tuned model has a shrinkage rate of 0.5 and uses 14,000 regression trees, with each tree having a minimum terminal node size of 15 observations and a maximum tree depth of two branches. See Hastie et al. (2017, pp. 337–388) for a more detailed description of gradient boosting trees and their tuning parameters.

A Parametric Model for Comparison

As a point of comparison, we also built a nonlinear least squares model of individual tree diameter increment, based on a similar model that Weiskittel et al. (2016) built for the New York Adirondacks. We experimented with nonlinear mixed-effects modeling, treating plots as random, but found no improvement over nonlinear least squares modeling. Both approaches have been used successfully to model individual tree growth in the Northeast (Kuehne et al. 2020). Our model used the form:

$$ADI = exp(b_1 + b_2 \cdot ln(DBH) + b_3 \cdot DBH + b_4 \cdot BAL + b_5 \cdot \sqrt{BA} + b_6 \cdot SC)$$

where *ADI* is the annual diameter increment ($cm \cdot yr^{-1}$), *DBH* is diameter at breast height (cm), *BAL* is overtopping basal area ($m^2 \cdot ha^{-1}$), *BA* is plot basal area ($m^2 \cdot ha^{-1}$), and *SC* is site class. These predictors have been well studied in the northeastern United States in relation to diameter increment (Teck and Hilt 1991, Westfall 2006), their use in predicting diameter increment has a strong conceptual basis (Weiskittel 2011), and they were found to be important predictors for the FIA data. The equation was fit to the training data for each species group to obtain species-specific coefficients.

We used McDill and Amateis' (1993) nonlinear "two-step" interpolation method to account for the fact that different plots in the training data had different remeasurement intervals. The method assumes that the diameter growth rate over the period varies according to the model form. Nonlinear interpolation has been shown to be more accurate than assuming a linear growth rate over the remeasurement period (Cao 2010, McDill and Amateis 1993), or using the remeasurement interval as a predictor (Juma et al. 2014).

Model Assessment

We compared the nonparametric model to the parametric model based on its ability to accurately predict d.b.h. growth in the independent testing data. The models predict annualized growth increments, while growth was actually observed over multi-year remeasurement intervals. To predict whole-interval growth (ΔDBH) comparable to observed growth, we applied the models iteratively to each observation, predicting 1 year's growth at a time and updating the DBH, BA and BAL before predicting the next year's growth. We assumed a linear change in BA and BAL over each remeasurement period (after Weiskittel et al. 2016). Overall accuracy was assessed using RMSE and mean absolute error (MAE).

We also used several model-agnostic interpretation tools to explore the structure of the nonparametric model. Nonparametric models like the gradient boosting machine can be very complicated, and instead of trying to fully describe the model's structure, we chose a couple of examples of important growth effects to illustrate the interpretive process. To start, we calculated permutation feature importance (Fisher et al. 2019) to see which predictors contribute the most to the models' predictions. Then we used Friedman's H-statistic to search for potential interactive effects (Friedman and Popescu 2008) and visualized several of them with partial dependence plots (Friedman 2001). We implemented all three techniques using Molnar et al.'s (2018) *iml* package for R.

RESULTS

The mean observed ΔDBH over all species and tree sizes was 1.27 cm, for remeasurement intervals that averaged 5.1 years. The distribution of observed ΔDBH was skewed right, with 90 percent of observations falling between 0 and 3 cm. Our nonparametric gradient boosting machine was more accurate in predicting ΔDBH than the parametric nonlinear least squares model, reducing the RMSE by 6.5 percent and the MAE by 7.9 percent (Table 1). These predictive gains came at the cost of speed and size, with the nonparametric model taking 27 times as long to make predictions and using almost 44,000 times the space in memory.

Both models are relatively unbiased overall, but the nonparametric model is less biased than the parametric model in predicting the growth of especially slow- and fast-growing trees (Fig. 2) and in predicting growth in trees with less common attributes, such as large d.b.h. (Fig. 3a) and small or large CR (Fig. 3b), and in less common settings, such as below- or above-average BA (Fig. 3c) and BAL (Fig. 3d).

Table 1. —Performance metrics for the parametric nonlinear least squares model (NLS) and the nonparametric gradient boosting machine (GBM). Mean error is the mean of observed responses minus the mean of predictions. Speed is the time to make 1,000 predictions on our computer and is relative.

Metric	NLS	GBM
RMSE (cm)	0.364	0.341
MAE (cm)	0.269	0.248
Mean Error (cm)	0.005	0.003
Speed (sec)	0.07	1.97
Size (MB)	0.003	131

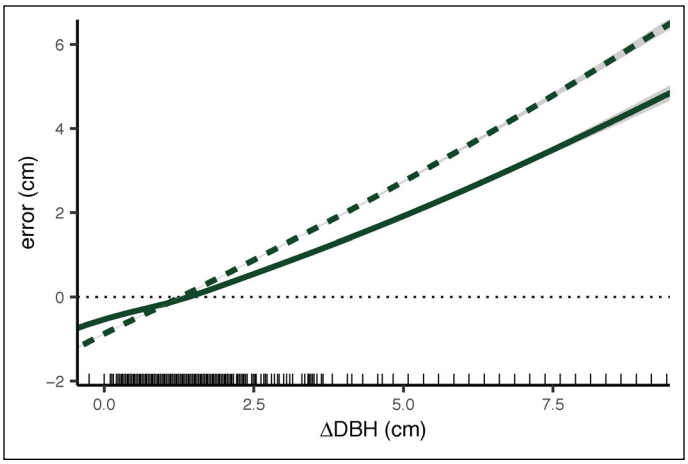

Figure 2.—Prediction error trends (observed minus predicted) based on observed ΔDBH, for the gradient boosting machine (solid line) and nonlinear least squares model (dashed line). Trends are drawn with generalized additive models and shaded regions show 95 percent confidence intervals. The rug plot on the x axis shows the ΔDBH distribution, with a higher rug density indicating a greater number of observations.

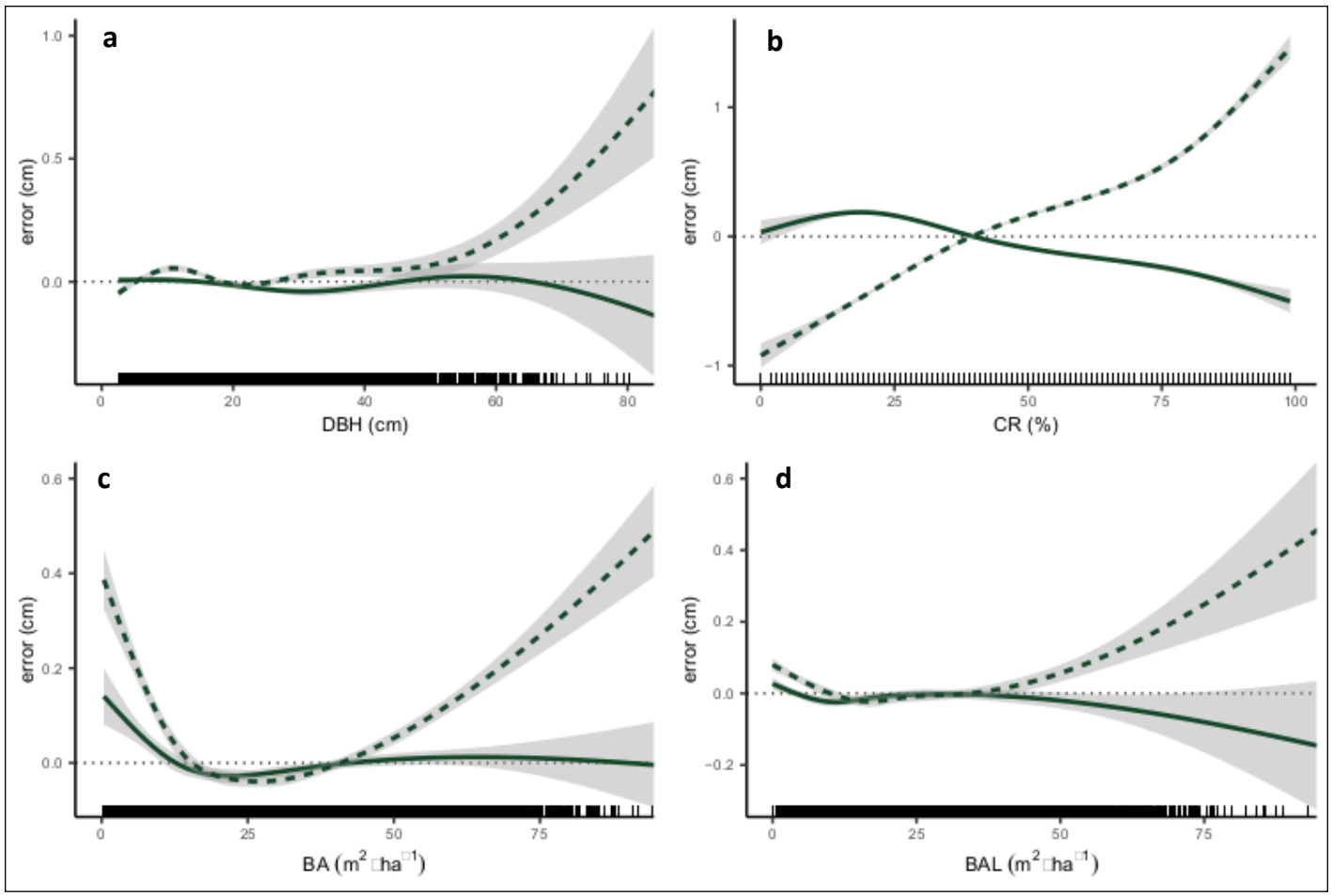

Figure 3.—Prediction error trends (observed minus predicted) based on (a) DBH, (b) CR, (c) BA, and (d) BAL, for the gradient boosting machine (solid line) and nonlinear least squares model (dashed line). Trends are drawn with generalized additive models and shaded regions show 95 percent confidence intervals. Rug plots on the x axes show predictors' distributions, with higher rug densities indicating more observations.

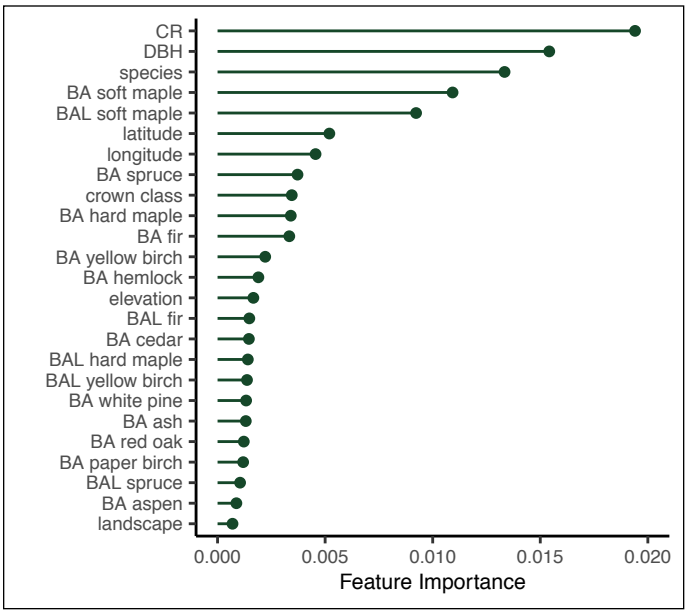

Figure 4.—Permutation feature importance of the 25 most important predictors. BA and BAL are decomposed by species.

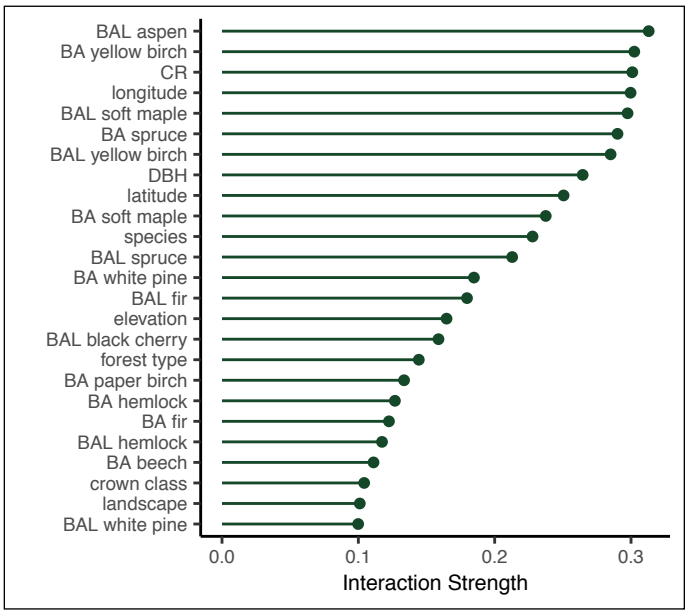

Figure 5.—Interaction strengths of the 25 predictors demonstrating the most interaction, calculated using Friedman's H-statistic. BA and BAL are decomposed by species.

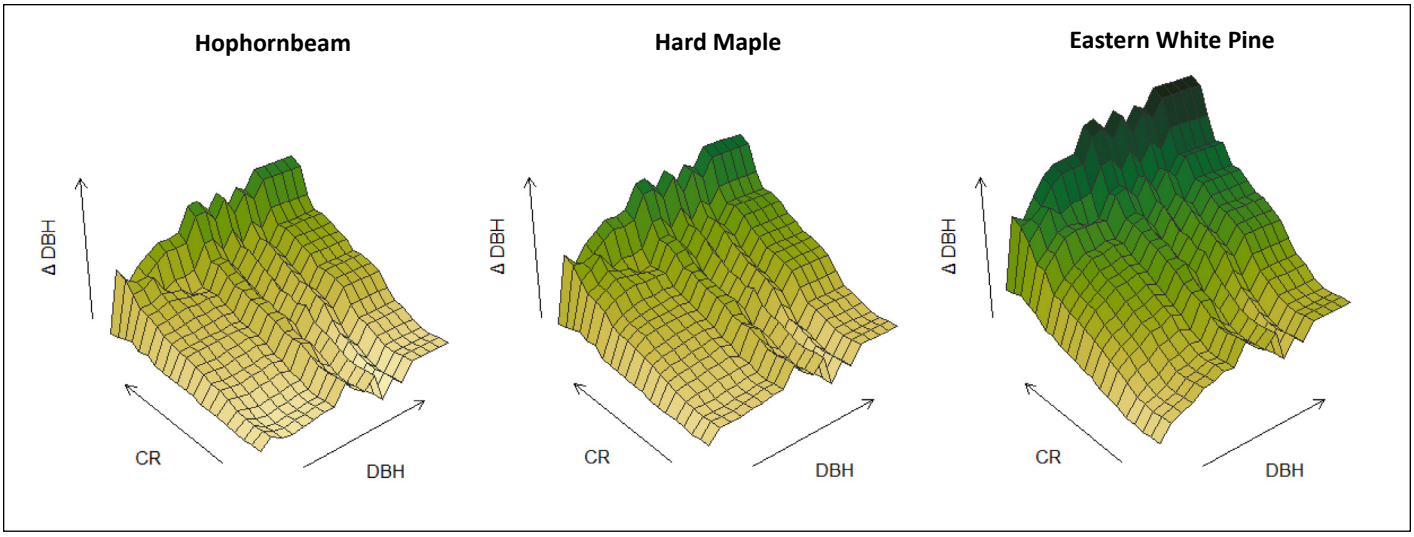

Figure 6.—Relationship between DBH, CR, and predicted ΔDBH for three individual species groups, using partial dependence plotting. The shape of the response curve for each predictor changes depending on the other predictors, indicating that the predictors interact with one another.

Subplot BA was the most important predictor in the nonparametric model (with a feature importance score of 0.035), followed by CR (0.019), BAL (0.019), DBH (0.015), and species (0.013). Figure 4 shows the relative importance of the 25 most important predictors with species-specific BAs and BALs listed separately.

The predictors with the highest overall interaction strengths in the nonparametric model are shown in Figure 5. Calculations of specific interaction strengths between predictors with high overall strengths revealed several important interactions that influenced the model. The most important is a three-way interaction between species, d.b.h., and CR. Figure 6 illustrates the interaction using three of the 28 species groups. Generally, the effect of CR is amplified at larger d.b.h., the effect of d.b.h. is amplified at larger CR, and both the CR and d.b.h. effects are larger in some species (e.g., eastern white pine) than in others (e.g., hophornbeam). A second interaction

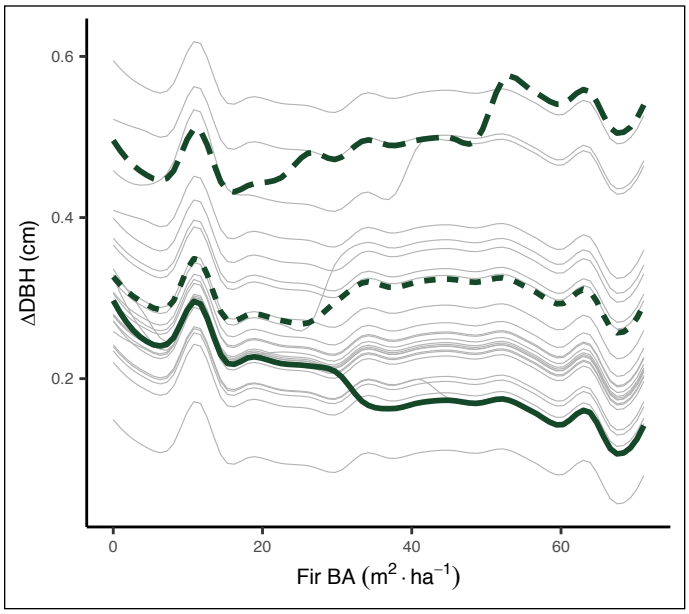

Figure 7.—Relationship between fir BA and predicted ΔDBH for individual species, using partial dependence plotting. Three species have been highlighted: aspen (dashed line on top) demonstrates increased growth at higher fir BA, hemlock (dotted line in middle) demonstrates growth poorly correlated to fir BA, and yellow birch (solid line on bottom) demonstrates reduced growth at higher fir BA.

occurs between fir BA and species (Fig. 7). For most species, the model predicts reduced d.b.h. growth on plots with higher fir BA, but fir BA does not affect predicted growth for other species like hemlock. Aspen actually experiences increased predicted growth rates at higher fir BA. These are two of the more important interactions driving the nonparametric model, but they are not the only ones. Interaction strengths suggest additional interactions related to latitude and longitude, and to a number of different species-specific BAs and BALs.

DISCUSSION

We did not observe large gains from the nonparametric model overall or for average trees in common situations, but the nonparametric model is more accurate for trees in abnormal situations, such as for larger diameter trees, trees with above- or below-average crown sizes, and trees growing with above- or below-average competition. Part of that advantage likely comes from the nonparametric model's use of numerous predictors. While parametric techniques focus on the most important predictors to build a robust and parsimonious model, the nonparametric approach also makes use of less important predictors whose simple effects might add to the predictive ability when they are aggregated.

The nonparametric model also accounts for some complex interactions that the parametric model does not capture. We illustrated the three-way interaction we found between species, d.b.h., and CR, and the uneven effect of fir BA on different species as examples. Interaction strengths pointed to several more potential interactions, and there could be many more among less important predictors, which would contribute little to overall accuracy, but could be important in specific situations. Parametric techniques could be used to model interactions like these, but only if they were understood ahead of time and deliberately included.

All three of Breiman's (2001) "lessons" probably came into play in uncovering these interactions: *First,* the nonparametric approach did not limit model selection to match a preconceived conception of growth dynamics, and was able to find a particularly good model algorithm that had not been tried in this context before; *second,* the nonparametric approach did not impose limitations on model complexity, and employed a complicated model more capable of representing interactions; and *third,* the nonparametric approach embraced the high dimensional data, which

contained information about interactions that might have been missed otherwise. It is important to note that these benefits depend on nonparametric models following the data more or less blindly. Cross validation and other resampling methods can prevent over-fitting, but any biases in the data will lead to inaccurate models. This is true of parametric models as well but is probably more acute for nonparametric models.

When extensive, unbiased data are available, the complex, flexible models that machine learning can produce may be especially good at uncovering the subtleties of neighborhood-scale growth dynamics without requiring expensive inventory or parameterization. In this way they could be a boon to practitioners working in diverse and complex forests. But several questions remain about their effectiveness for predictive modeling. First, can nonparametric models accurately predict forest growth over longer time periods? Parametric models are somewhat more mechanistic in that their structures are often informed by biological relationships, which could give them an edge in longer-term simulations. Suitable long-term reinventory data are needed to determine if nonparametric models' more empirical forms disadvantage them. And second, are nonparametric models' relatively subtle gains in predictive ability worth their cost in increased computing power and complexity? Small gains in describing growth variability could be quite valuable to managers of diverse hardwood forests, where some trees are worth much more than others and the nuances of growth are more important. However, this information will probably be less valuable to managers of simpler forests geared toward fiber production.

Regardless of their usefulness for predictive modeling, it does seem clear that nonparametric techniques are very valuable for studying tree growth in complex forests. Existing forest inventory data contain a wealth of information for machine learning techniques to mine, and the rapidly expanding new field of interpretable machine learning is devoted to studying machine learning models to understand the relationships they glean from data (Molnar 2020). Variable importance, interaction scores, and partial dependence plots are a few of the better-established techniques, but there are many more, and they are growing more sophisticated as the field matures. Models that used to be considered black boxes are now becoming powerful microscopes, capable of finding complicated relationships between multiple variables.

LITERATURE CITED

Ali, A. 2019. **Forest stand structure and functioning: current knowledge and future challenges.** Ecological Indicators. 98: 665–677. https://doi.org/10.1016/j.ecolind.2018.11.017.

Boehmke, B.; Greenwell, B. 2020. **Hands-on machine learning with R.** https://bradleyboehmke.github.io/HOML/.

Breiman, L. 2001. **Statistical modeling: the two cultures (with comments and a rejoinder by the author).** Statistical Science. 16 (3): 199–231. https://doi.org/10.1214/ss/1009213726.

Bugmann, H. 2001. **A review of forest gap models.** Climatic Change. 51(3–4): 259–305. https://doi.org/10.1023/A:1012525626267.

Canham, C.D.; LePage, P.T.; Coates, K.D. 2004. **A neighborhood analysis of canopy tree competition: effects of shading versus crowding.** Canadian Journal of Forest Research. 34(4): 778–787. https://doi.org/10.1139/x03-232.

Canham, C.D.; Uriarte, M. 2006. **Analysis of neighborhood dynamics of forest ecosystems using likelihood methods and modeling.** Ecological Applications. 16(1): 62–73. https://doi.org/10.1890/04-0657.

Cao, Q.V. 2010. **Adjustments of individual-tree survival and diameter-growth equations to match whole-stand attributes.** In: Stanturf, J.A., ed. Proceedings of the 14th biennial southern silvicultural research conference. Gen. Tech. Rep. SRS–121. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southern Research Station. 121: 369–373. https://www.fs.usda.gov/treesearch/pubs/35885.

Cuddington, K.; Fortin, M.J.; Gerber, L.R.; Hastings, A.; Liebhold, A.; OConnor, M.; Ray, C. 2013. **Process-based models are required to manage ecological systems in a changing world.** Ecosphere. 4(2): 1–12. https://doi.org/10.1890/ES12-00178.1.

Dănescu, A.; Albrecht, A.T.; Bauhus, J. 2016. **Structural diversity promotes productivity of mixed, uneven-aged forests in southwestern Germany.** Oecologia. 182(2): 319–333. https://doi.org/10.1007/s00442-016-3623-4.

Edgar, C.; Burk, T. 2001. **Productivity of aspen forests in northeastern Minnesota, U.S.A., as related to stand composition and canopy structure.** Canadian Journal of Forest Research. 31(6): 1019–1029. https://doi.org/10.1139/cjfr-31-6-1019.

Fan, Z.; Kabrick, J.; Shifley, S. 2006. **Classification and regression tree based survival analysis in oak-dominated forests of Missouri's Ozark highlands.** Canadian Journal of Forest Research. 36: 1740–1748. https://doi.org/10.1139/x06-068.

Fien, E.K.; Fraver, S.; Teets, A.; Weiskittel, A.R.; Hollinger, D.Y. 2019. **Drivers of individual tree growth and mortality in an uneven-aged, mixed-species conifer forest.** Forest Ecology and Management. 449: 117446. https://doi.org/10.1016/j.foreco.2019.06.043.

Fisher, A.; Rudin, C.; Dominici, F. 2019. **All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously.** Journal of Machine Learning Research. 20(177): 1–81.

Forsyth, D.M.; Wilson, D.J.; Easdale, T.A.; Kunstler, G.; Canham, C.D.; Ruscoe, W.A.; Wright, E.F. et al. 2015. **Century-scale effects of invasive deer and rodents on the dynamics of forests growing on soils of contrasting fertility.** Ecological Monographs. 85(2): 157–180. https://doi.org/10.1890/14-0389.1.

Friedman, J.H. 2001. **Greedy function approximation: a gradient boosting machine.** The Annals of Statistics. 29(5): 1189–1232.

Friedman, J.H.; Popescu, B.E. 2008. **Predictive learning via rule ensembles.** The Annals of Applied Statistics. 2(3): 916–954. https://doi.org/10.1214/07-AOAS148.

Guan, B.; Gertner, G. 1991. **Modeling red pine tree survival with an artificial neural network.** Forest Science. 37(November): 1429–1440. https://doi.org/10.1093/forestscience/37.5.1429.

Hastie, T.; Tibshirani, R.; Friedman, J. 2017. **The elements of statistical learning: data mining, inference, and prediction.** 2nd ed. Berlin, Germany: Springer-Verlag. 767 p. https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12_toc.pdf.

Juma, R.; Pukkala, T.; Sergio de Miguel; Muchiri, M. 2014. **Evaluation of different approaches to individual tree growth and survival modeling using data collected at irregular intervals – a case study for *Pinus patula* in Kenya.** Forest Ecosystems. 1(August): 14. https://doi.org/10.1186/PREACCEPT-9097767721272407.

Kershaw, J.A.; Weiskittel, A.R.; Lavigne, M.B.; McGarrigle, E. 2017. **An imputation/copula-based stochastic individual tree growth model for mixed species Acadian forests: a case study using the Nova Scotia permanent sample plot network.** Forest Ecosystems. 4(1): 15. https://doi.org/10.1186/s40663-017-0102-2.

King, S.L.; Bennett, K.P.; List, S. 2000. **Modeling noncatastrophic individual tree mortality using logistic regression, neural networks, and support vector methods.** Computers and Electronics in Agriculture. 27(1): 401–406. https://doi.org/10.1016/S0168-1699(00)00106-X.

Kuehne, C.; Russell, M.B.; Weiskittel, A.R.; Kershaw, J.A. 2020. **Comparing strategies for representing individual-tree secondary growth in mixed-species stands in the Acadian forest region.** Forest Ecology and Management. 459(March): 117823. https://doi.org/10.1016/j.foreco.2019.117823.

Kuhn, M.; Johnson, K. 2013. **Applied predictive modeling.** New York, NY: Springer. https://doi.org/10.1007/978-1-4614-6849-3.

Kuhn, M.; Wickham, H. 2020. **Recipes: preprocessing tools to create design matrices** (version 0.1.15). https://CRAN.R-project.org/package=recipes.

Kuhn, M.; Wing, J.; Weston, S.; Williams, A.; Keefer, C.; Engelhardt, A.; Cooper, T. [et al.]. 2020. **Caret: classification and regression training** (version 6.0-86). https://CRAN.R-project.org/package=caret.

Lei, X.; Wang, W.; Peng, C. 2009. **Relationships between stand growth and structural diversity in spruce-dominated forests in New Brunswick, Canada.** Canadian Journal of Forest Research. 39(October): 1835–1847. https://doi.org/10.1139/X09-089.

Liang, J.; Buongiorno, J.; Monserud, R.A. 2005. **Growth and yield of all-aged Douglas-fir – western hemlock forest stands: a matrix model with stand diversity effects.** Canadian Journal of Forest Research. 35: 14. https://doi.org/10.1139/X05-137.

Liu, Z.; Peng, C.; Timothy, W.; Candau, J.N.; Desrochers, A.; Kneeshaw, D. 2018. **Application of machine-learning methods in forest ecology: recent progress and future challenges.** Environmental Reviews. 26(4): 339–350. https://doi.org/10.1139/er-2018-0034.

McDill, M.E.; Amateis, R.L. 1993. **Fitting discrete-time dynamic models having any time interval.** Forest Science. 39(3): 499–519.

Molnar, C. 2020. **Interpretable machine learning.** https://christophm.github.io/interpretable-ml-book/.

Molnar, C.; Casalicchio, G.; Bischl, B. 2018. **Iml: An R package for interpretable machine learning.** Journal of Open Source Software. 3(26): 786. https://doi.org/10.21105/joss.00786.

O'Hara, K. 2014. **Multiaged silviculture: managing for complex forest stand structures.** Oxford, UK: Oxford University Press. 240 p. https://doi.org/10.1093/acprof:oso/9780198703068.001.0001.

Pretzsch, H. 2017. **Individual tree structure and growth in mixed compared with monospecific stands.** In: Pretzsch, H.; Forrester, D.I.; Bauhus, J., eds. Mixed-species forests: Ecology and Management. Berlin, Germany: Springer. 271–336. https://doi.org/10.1007/978-3-662-54553-9_6 .

Pretzsch, H.; Forrester, D.I.; Bauhus, J., eds. 2017. **Mixed-species forests: ecology and management.** Berlin, Germany: Springer. 653 p.

Pretzsch, H.; Zenner, E.K. 2017. **Toward managing mixed-species stands: from parametrization to prescription.** Forest Ecosystems. 4(1): 19. https://doi.org/10.1186/s40663-017-0105-z.

R Core Team. 2020. **R: A language and environment for statistical computing.** Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org/.

Stoll, P.; Weiner, J. 2000. **A neighborhood view of interactions among individual plants.** In: Dieckmann, U.; Law, R.; Metz, J.A., eds. The geometry of ecological interactions: simplifying spatial complexity, 1st ed. Cambridge, England: Cambridge University Press. 580 p. https://doi.org/10.1017/CBO9780511525537.003.

Taylor, A.R.; Han Y. H.C.; VanDamme, L. 2009. **A review of forest succession models and their suitability for forest management planning.** Forest Science. 55(1): 23–36.

Teck, R.M.; Hilt, D.E. 1991. **Individual tree-diameter growth model for the northeastern United States.** Res. Pap. NE-649. Radnor, PA: US. Department of Agriculture, Forest Service, Northeastern Forest Experiment Station. 11 p. https://doi.org/10.2737/NE-RP-649.

Weiskittel, A.R.; Hann, D.W.; Kershaw, J.A., Jr.; Vanclay, J.K., eds. 2011. **Forest growth and yield modeling.** Hoboken, NJ: Wiley. 430 p.

Weiskittel, A.R.; Kuehne, C.; McTague, J.P.; Oppenheimer, M. 2016. **Development and evaluation of an individual tree growth and yield model for the mixed species forest of the Adirondacks region of New York, USA.** Forest Ecosystems. 3(1): 26. https://doi.org/10.1186/s40663-016-0086-3.

Westfall, J.A. 2006. **Predicting past and future diameter growth for trees in the northeastern United States.** Canadian Journal of Forest Research. 36: 1551–1562. https://doi.org/10.1139/x06-045.

Wickham, H.; Averick, M.; Bryan, J.; Chang, W.; McGowan, L.D. [et al.]. 2019. **Welcome to the Tidyverse.** Journal of Open Source Software. 4(43): 1686. https://doi.org/10.21105/joss.01686.

Woudenberg, S.W.; Conkling, B.L.; O'Connell, B.M.; LaPoint, E.B.; Turner, J.A.; Waddell, K.L. 2010. **The Forest Inventory and Analysis database: database description and users manual, version 4.0 for Phase 2.** Gen. Tech. Rep. RMRS-GTR-245. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 336 p. https://doi.org/10.2737/RMRS-GTR-245.

Yeo, I.K.; Johnson, R.A. 2000. **A new family of power transformations to improve normality or symmetry.** Biometrika. 87(4): 954–959. https://doi.org/10.1093/biomet/87.4.954.

CITATION: Maker, Neal F.; Foppert, John D. 2023. **Nonparametric Individual-Tree Growth Models for the Northern Forest.** In: Kern, Christel C.; Dickinson, Yvette L., eds. Proceedings of the first biennial Northern Hardwood Conference 2021: Bridging science and management for the future. Gen. Tech. Rep. NRS-P-211. Madison, WI: U.S. Department of Agriculture, Forest Service, Northern Research Station: 83–95. https://doi.org/10.2737/NRS-GTR-P-211-paper19.