

RESEARCH ARTICLE

Diversity and Distributions WILEY

Unpacking the 'black box': Improving ecological interpretation of regression-based models

Anantha Prasad¹  | Matthew Peters¹ | Stephen Matthews^{1,2} | Louis Iverson¹

¹Northern Research Station, USDA Forest Service, Delaware, Ohio, USA

²School of Environment and Natural Resources, The Ohio State University, Columbus, Ohio, USA

Correspondence

Anantha Prasad, Northern Research Station, USDA Forest Service, 359 Main Road, Delaware, OH 43015, USA.
Email: anantha.prasad@usda.gov

Editor: Marion Pfeifer

Abstract

Aim: Many tree species distribution models use black-box machine learning techniques that often neglect interpretative aspects and instead focus mainly on maximizing predictive accuracy. In this study, we outline an interpretative modelling framework to gain better ecological insights while mapping abundance patterns of six North American species.

Location: Continental United States and Canada.

Methods: We develop an innovative procedure using regression trees by stabilizing variance, and mapping dominant rules which we term 'optimized regression tree bagging for interpretation and mapping' (ORTBIM). We apply this technique to understand ecological features influencing the abundance patterns of three eastern (*Pinus strobus*, *Acer saccharum* and *Quercus montana*), and three western (*Picea engelmannii*, *Pinus ponderosa* and *Pseudotsuga menziesii*) tree species in North America. For these species, we assess and map the dominant climate–terrain interactions that partly determine abundance patterns in the eastern and western regions. In the process, we examine the role of varying responses and scales and explore finer-scale species climate–terrain niches and non-linear relationships.

Results: Our study emphasizes the prominent role of elevation and heat–moisture variables in the west and the greater importance of seasonal precipitation and seasonal temperature in the east. The abundance patterns under future climate (SSP5–8.5) show climate–terrain habitats shifting northward and westward into Canada and Alaska for the eastern species, and predominantly north-westward for the western species.

Conclusion: Our interpretative modelling framework can be used to gain a more comprehensive understanding of the abundance patterns across the full species range, formulate better predictive models and facilitate improved management practices under climate change.

KEYWORDS

bagged regression tree rules, climate change, ecological interpretation, interpretative modelling framework, species distribution models, tree abundance modelling

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. *Diversity and Distributions* published by John Wiley & Sons Ltd.

1 | INTRODUCTION

The evolutionary history of temperate North American vegetation has been shaped by various global-scale climatic and biogeographic forces, the accompanying disturbance regimes and their interactions. These forces have resulted in the current distribution of extant species we see today, and sets the stage for species potential changes under contemporary anthropogenic climate and land use changes (Schoonmaker & Foster, 1991). Over deep time, geological forces have shaped the continents and the terrain, resulting in biogeographic constraints to migration pathways and niches of various tree species (Barton et al., 2003). These geological forces created mountain ranges and other features that not only altered migration routes but also greatly influenced regional climates, including glacial events that shaped the vegetation of North America for both Gymnosperms (starting ~350mya) and Angiosperms (starting ~120mya). The changing landforms, especially in the last 120 Mya when Angiosperms and Gymnosperms competed amidst climatic swings, and the accompanying large-scale biotic and abiotic disturbances, have shaped the migration pathways and niches where tree species have subsequently colonized (Graham, 1999; Williams et al., 2004). The climate–terrain interactions of present-day abundance patterns, therefore, are the predominant outcome of this rich eco-evolutionary legacy across their entire range (Silander, 2001), while regional and local variations are the result of idiosyncratic responses due to the myriad of land use, edaphic and biotic interactions. Such deep history underscores the challenge of niche modelling using modern machine learning techniques, which only captures a portion of the eco-evolutionary dynamics in multivariate space–time and deemphasizes interpretation and explainability in favour of maximizing predictive accuracy. The residual deviance in the species distribution models – the unexplained patterns of present-day vegetation, therefore, opens further avenues of inquiry into the rich historical legacies of tree species.

In this study, we explore and analyse the macroscale distribution of six North American species derived from their current abundance patterns, and their interaction with climate and terrain throughout their range. Our focus is on the interpretative aspects of modern machine learning techniques rather than the maximization of predictive accuracy.

1.1 | Modelling rationale

Modern species distribution modelling began roughly in the last decades of the 20th century, with the development of reliable climatic interpolation methods (Booth, 1990, 2016). A shift away from traditional parametric methods to non-parametric machine learning techniques (especially those based on decision trees) was also evident during this period (Iverson & Prasad, 1998; Verbyla, 1987). In addition to decision trees, semi-parametric data-driven functional models like generalized additive models (GAM) that account for non-linear relationships between response and explanatory variables

became popular (Guisan et al., 2002). The attractiveness of decision tree (DT) models stems from their superior ability to handle multivariate non-linearity, including all possible interactions, in a hierarchical tree, with the important macroscale splits (spatially 'broad') forming earlier in the tree and the more regional effects in the lower branches (Breiman et al., 1984; Iverson & Prasad, 1998). Another important aspect is the interpretability of the decision trees, which enables one to understand complex relationships intuitively. While the DT approach shines in intelligibility, interpretability and the handling of interactions, they generally do not provide robust predictions (Dwyer & Holte, 2007) because they suffer from higher tree variance even with small data perturbations, and a tendency to overfit. To improve DT predictions, ensemble methods are used – for example, bagging, Random Forests and boosted trees (Prasad, 2018; Prasad et al., 2006). But these ensemble methods are typically black-box techniques designed to maximize prediction accuracy (Kuhn & Johnson, 2013). The explainable features of these black-box models are mainly limited to variable importance scores and approximate partial dependency plots. However, sacrificing interpretability for prediction with black-box models can be costly (Chakraborty et al., 2021; Rudin, 2019), and efforts are underway to find a good compromise between interpretability and prediction (Linardatos et al., 2020; Lucas, 2020; Ryo et al., 2021).

1.2 | Our approach

Our earlier efforts in modelling suitable habitats and colonization potentials used multi-model ensemble and hybrid techniques optimized for predictive accuracy which were mostly black-box techniques (Peters et al., 2019; Prasad et al., 2020). In this paper, we focus on 'ecological interpretation', rather than predictive performance, as our main goal by examining more closely the variance inherent in ecological phenomena instead of trying to minimize it for achieving better prediction. By focusing on ecological rules governing continuous abundance patterns across the entire range of the species, we shed light on the non-linear interactions that are typical of vegetation–environment interactions. We construct a procedure we term 'optimized regression tree bagging for interpretation and mapping' (ORTBIM) using specialized machine learning techniques for extracting dominant patterns and rules geared towards interpretation and mapping. By formulating a procedure based on examining the variance of singular and bagged regression trees (Breiman, 1996), we boost confidence in ecological interpretation. The ORTBIM procedure consists of (a) selecting ecologically meaningful climate and terrain variables; (b) using a bootstrapping technique called recursive feature elimination to pick the most relevant variables from this set; (c) using hyper-parameter tuning to achieve optimally pruned tree; (d) using frequency distribution of species abundance to choose an abundance threshold for selecting dominant rules; (e) using bagging trees (bootstrap resampling) to quantify variances; and (f) to combine and map dominant rules (henceforth referred to as *dominant bagged rules*) under current conditions, and to examine trends under a future climate scenario.

Using this novel procedure, we explore patterns of tree abundance for six representative tree species in the eastern and western regions of the North American continent (north of Mexico), derived from Forest Inventory and Analysis (FIA) data for the conterminous United States and projected across the entire continental United States and Canada. The goal is to understand and explore dominant tree distribution patterns under current conditions and a future climate at varying resolutions and therefore lay the foundation for improved modelling of regional and local trends. In the process, we seek to answer the following questions:

- What is the role of elevation and terrain in the distribution of the species across the east and west?
- What can we understand from interactions between elevation/terrain and climate, especially between the eastern and western tree species?
- How sensitive are the rules to changes in scale (10 and 20 km²)?
- Does the type of response (relative abundance vs. absolute abundance) have a large effect on the results?
- Are there patterns in finer-scale plot-level FIA data with the explanatory variables that shed light on niche occupation patterns when contrasting eastern and western species?
- How do the trends in the mapped dominant bagged rules of current and future abundances inform improved management of tree species in the face of changing climate?

By examining the above questions for the six representative species in a nuanced interpretative framework, we set the stage for building a better prediction modelling framework in North America.

2 | DATA AND METHODS

2.1 | The abundance responses

We use the USDA Forest Service Forest Inventory and Analysis (FIA; <https://apps.fs.usda.gov/fia/datamart/datamart.html>) plot-level data to derive relative importance value (RIV, also relative abundance) and absolute importance value (AIV, also absolute abundance) for three eastern species – eastern white pine (*Pinus strobus*), sugar maple (*Acer saccharum*) and chestnut oak (*Quercus montana*); and three western species – Engelmann spruce (*Picea engelmannii*), ponderosa pine (*Pinus ponderosa*) and Douglas fir (*Pseudotsuga menziesii*). The RIV index is calculated as:

$$\text{RIV}(x) = 50 \times \frac{\text{BA}(x)}{\sum \text{BA}(\text{all species in plot})} + 50 \times \frac{\text{NS}(x)}{\sum \text{NS}(\text{all species in plot})}$$

where x is the species in the plot, BA is the basal area and NS is the number of stems (summed for overstorey and understorey trees). In monotypic stands, the RIV would reach a maximum of 100%.

The absolute abundance (AIV) is calculated as:

$$\text{AIV}(x) = \text{SumDia}(x) + \text{NS}(x)$$

where x is the species in the plot, SumDia is the sum of diameter of the species in the plot and NS is the number of stems (summed for overstorey and understorey trees). AIV is normalized by dividing it by the maximum value of AIV for the species in the plot and multiplying it by 100 to get values between 0 and 100. AIV is an absolute measure of both the abundance and dominance of the species within the plot without standardizing them to the other trees in the plot. Because the FIA plot locations are unequally sampled across the United States, we aggregated relative importance values averaged over grid cells at (a) 20×20 km (20k) and (b) 10×10 km (10k). We excluded plots that indicated artificial regeneration, ensuring that only natural stands were aggregated for reasons that will be explained in Section 4.

The range and spread of the six selected species capture the spatial trends of many of the other species in the conterminous United States, except those that are quite rare or widely scattered in distribution. We used the ORTBIM procedure (explained in detail below) to model the responses for RIV and AIV for these six species at the two spatial resolutions. The abundances were calculated from the accurate and consistent annual FIA data confined to the conterminous United States, but the explanatory variables spanned the continental United States and Canada.

2.2 | The explanatory variables

Because our domain of inquiry included continental United States and Canada, the explanatory variables (also referred to as ‘features’) were confined to climate, elevation and terrain variables (Table 1). We did not use soil variables for two reasons: (a) the soil data quality and metrics are inconsistent at continental scales and (b) they have been shown to be generally less relevant at these scales, becoming more important only at regional and local scales (Dyer & Hutchinson, 2019; Mathys et al., 2014). We grouped the variables into six categories according to climate and terrain characteristics to assist in summarizing the results (Table 1).

2.3 | Elevation and terrain

The elevation and terrain data (Table 1) were derived originally from the USGS GMTED2010 DEM data at 500m resolution for the North American continent (Global Multi-resolution Terrain Elevation Data, 2010). The terrain variables were derived from the DEM data using R (using terrain in *terra* package; Hijmans, 2022; R Core Team, 2022) and the topographic wetness index was derived using SAGA GIS (RSAGA package in R; Brenning et al., 2018). Because our domain included the whole of the continental United States and Canada, computational bottlenecks were avoided by choosing the relatively coarse 500m DEM data for this macroscale analysis. The elevation and terrain variables were aggregated (using averaging) to 10- and 20-km grids.

TABLE 1 Climate, elevation and terrain (21) variables selected as ecologically meaningful (and their six groupings) and used in the ORTBIM procedure.

Features	Explanation	Group
AHM	Annual heat moisture index, calculated as (MAT+10)/(MAP/1000)	Heat-moisture
SHM	Summer heat-moisture index, calculated as MWT/ (MSP/1000)	Heat-moisture
CMD	Hargreave's climatic moisture index	Heat-moisture
CMI	Hogg's climate moisture index (mm)	Heat-moisture
NFFD	Number of frost-free days	Growing season
DD5	Degree days above 5°C (growing degree days)	Growing season
Tave_wt	Winter (December to February) mean temperature (°C)	Seasonal temperature
Tave_sp	Spring (March to May) mean temperature (°C)	Seasonal temperature
Tave_sm	Summer (June to August) mean temperature (°C)	Seasonal temperature
EMT	Extreme minimum temperature over 30 years	Seasonal temperature
Tave_at	Autumn (September to November) mean temperature (°C)	Seasonal temperature
PPT_wt	Winter (December to February) precipitation (mm)	Seasonal precipitation
PPT_sp	Spring (March to May) precipitation (mm)	Seasonal precipitation
PPT_sm	Summer (June to August) precipitation (mm)	Seasonal precipitation
PAS	Precipitation as snow (mm)	Seasonal precipitation
PPT_at	Autumn (September to November) precipitation (mm)	Seasonal precipitation
elv	Elevation in m	Elevation
twi	Topographic wetness index	Terrain
ttri	Topographic roughness index	Terrain
tslope	Slope angle	Terrain
taspect	Aspect	Terrain

Note: The climate variables are from the AdaptWest database (AdaptWest 2021) for current and SSP5-8.5 scenarios.

2.4 | Climate data and the ORTBIM procedure

We obtained climate data from the AdaptWest database (AdaptWest Project, 2021; Mahony et al., 2022; Wang et al., 2016) for climate variables for the current (1991–2020) and future (2071–2100) time periods. We used data for future climate scenario (CMIP6) associated with shared socioeconomic pathways (SSPs) to examine the degree to which habitat suitability differed under these alternative

pathways. SSPs are socioeconomic pathways that use quantitative and qualitative elements pertaining to society, energy, economy and environment (O'Neill et al., 2016). We chose the worst-case SSP5-8.5 (roughly corresponding to CMIP5's RCP-8.5) to highlight the extreme trends in our interpretative analysis. The SSP5-8.5 projection is based on an ensemble of 13 general circulation models – ACCESS-ESM1-5, BCC-CSM2-MR, CNRM-ESM2-1, CanESM5, EC-Earth3, GFDL-ESM4, GISS-E2-1-G, INM-CM5-0, IPSL-CM6A-LR, MIROC6, MPI-ESM1-2-HR, MRI-ESM2-0 and UKESM1-0-LL (CMIP6 AOGCMs – AdaptWest Project, 2021). For further details about these 13 models, refer to https://wcrp-cmip.github.io/CMIP6_CVs/docs/CMIP6_source_id.html. The original data, at 1 km resolution, were aggregated to 10- and 20-km grids to match the abundances at these scales.

To stabilize the variance of regression trees to map dominant rules based on the abundance distribution of tree species, we used the ORTBIM procedure (as outlined previously). The ORTBIM procedure is most suited to exploring spatial distributions of continuous responses although it can be modified to adapt to different types of responses. A complete explanation of this procedure is listed below.

1. From the 33 climatic variables (AdaptWest Project, 2021), we selected a subset of 16 variables (Table 1) that we found to be most ecologically relevant based on our experience from previous modelling efforts (Iverson et al., 2008, 2019; Iverson & Prasad, 1998). The five elevation and terrain variables, along with the 16 climatic variables (we favoured seasonal temperature/precipitation and heat-moisture ratio variables over annual ones – Table 1) were used in the first stage of the modelling procedure.
2. Recursive feature elimination (explained in the next section) was used to select the most relevant of the 21 variables (based on bagging trees), using the *rfe* and *caret* packages in R (Kuhn, 2008, 2022). We did not let the variables chosen by *rfe* to exceed 12 to sharpen interpretation and minimize noise in the dominant rules.
3. The pruning parameter for the regression tree (*rpart* package in R; Therneau & Atkinson, 2022) was optimized via the *caret* package using bootstrapping and cross-validation and using the variables selected by *rfe* for that species. Variable selection via *rfe* and optimized pruning via *rpart* (using 'anova' method to split since the abundance response is continuous) together resulted in an optimally pruned regression model specific to the tree species.
4. The abundance threshold for defining dominance was based on the frequency distribution of abundance. We chose the values of abundance above the first quartile as dominant for the species in order to exclude 'noise' in the lower values of abundance (below the first quartile). This allowed the rules to vary depending on the species' abundance distribution.
5. After screening for the best algorithm to obtain rules, we decided on using a singular regression tree (see more in Section 4) chosen by the optimal pruning parameter and the abundance threshold to examine the main rules (limited to a maximum of six rules for practicality) and key ecological variables.

6. We performed bagging with 30 trees (using the *ipred* package in R; Peters & Hothorn, 2021) with the optimal regression tree chosen in step 3, and the abundance threshold chosen in step 4, to (a) examine rule stability via deviances explained (which would vary based on the species and the scale of analysis) and to (b) map each of the individual bagged rules.
7. We averaged the 30 individual bagged maps that represented the dominant rules (in step 6) for current conditions and the SSP5-8.5 scenario. By combining bagged maps via averaging, we incorporate variability among the correlated bagged trees, which is essential for capturing dominant trends.

By using species-specific rules determined by quartiles of the abundance distribution, and further averaging the 30 bagged trees, we focus on the main ecological trends in the distribution while minimizing noisy signals in the other branches. To explore dominant abundance patterns, we examine the rules derived from key ecological drivers based on a singular regression tree (step 5). The mapped products, therefore, incorporate variability among correlated bagged trees, and the main rules from singular regression tree help explain the key drivers. It should be noted that the bagged rules derived from current climate were assumed to be applicable in the future timeframe (2071–2100) because it is generally agreed that there will be an adaptation lag and most species are unlikely to adapt quickly to rapid climate change (Etterson et al., 2020; Fréjaville et al., 2020; Gray & Hamann, 2013). We emphasize that the ORTBIM procedure is an interpretative modelling framework developed to gain better insights into response–covariate interactions and provides an interpretative framework to assist in building better predictive models – it is not optimized to provide the best prediction.

2.5 | Recursive feature elimination

Recursive feature elimination (*rfe* function in *caret* package in R; Kuhn & Johnson, 2013; Kuhn, 2022) uses a backward selection procedure to find the optimal number of explanatory variables (features). After building a model with all features, *rfe* evaluates the importance of each feature and iteratively removes the least important ones based on a model evaluation metric (root mean squared error, RMSE). Our procedure consisted of using repeated resampling with replacement ('bagging' option in *rfe*) for selecting the variables (with five cross-validations and three repeats). We did not select the 'Random Forest' option in *rfe* because we had already preselected ecologically meaningful variables and randomizing the predictors would result in decorrelated DTs, which are optimized for prediction, not interpretation.

2.6 | Analysis

Based on the ORTBIM procedure, we extracted the deviances explained for the six species at the two resolutions (20 and 10k) for both RIV and AIV responses. We also compiled the deviances

explained for the bagged statistics (minimum, maximum, median, mean and standard deviation) for the 30 trees – to compare performance (Table 2). The dominant rules chosen by the ORTBIM procedure (singular tree, step 5) at 10 and 20k resolutions for the six species were extracted for RIV, and the deviances explained by the explanatory variables were summed at the tree nodes and summarized (Table 3). The combined bagged rules (step 7 in the ORTBIM procedure) were mapped for the current and future climate (SSP5-8.5) at 20k (Figures 3–8) and 10k (Figures S1 and S2). Because FIA-based abundances (Figure 2) provide the best plot-based estimate of basal area available, the ORTBIM procedure created higher fidelity rules which were applied to the whole continent.

It is noteworthy that the diversity of tree species in a single FIA plot is low for many western species, elevating the relative importance to higher values. Because of this pattern, we wanted to test whether our choice of response, relative abundance (RIV), was superior to absolute abundance (AIV) which was not affected by this artefact (since AIV does not consider other species in the plot).

2.7 | FIA plot-level analysis

We also evaluated the correspondence between our macroscale analysis and FIA plot-level responses to the explanatory variables, and tested whether plot data offered additional insights. We, therefore, calculated FIA plot-level abundances and extracted elevation and climatic features at 1 km resolution (to match the approximate extent of the fuzzed FIA plot) to understand the following responses.

2.7.1 | RIV versus elevation

The response of FIA plot abundances to elevation was plotted using loess smoothing for the six species (Figure 9). We also extracted the most important explanatory variables based on the modelled results to examine if these responses were accentuated at different elevation range classes. We, therefore, split the elevation into six range classes (–400–500; 500–1000; 1000–1500; 1500–2000; 2000–2500 and 2500–5000m) and created facet/trellis plots to examine the relationships (see Figures S3–S8). Negative elevations were included in the range classes because there are areas in the continent that are below sea level.

2.7.2 | Niche contour plots

To illustrate the relationship of FIA plot abundances with climatic and topographic drivers, we mapped FIA plot-based RIV against mean annual temperature (MAT) and mean total annual precipitation (MAP) using interpolation (*akima* package in R; Akima, 1991; Akima & Gebhardt, 2021). To highlight the role of topography in the niche space, elevation contours were draped onto the 3D plots of the six species (Figure 10).

TABLE 2 The deviances explained and the bagged statistics for 20km (20k) and 10km (10k) resolutions for relative abundance (RIV) and absolute abundance (AIV).

		RIV		AIV	
		20k	10k	20k	10k
Engelmann Spruce (5482)	Optimal Pruned Deviance explained (%)	60	50	51	41
	Bagged stats (Min/Max/Median)	60/68/66	47/53/51	51/63/59	41/50/45
	Bagged stats (Mean/SD)	65/2.1	50/1.4	58/3.1	44/2.2
Ponderosa Pine (9759)	Optimal Pruned Deviance explained (%)	38	29	44	26
	Bagged stats (Min/Max/Median)	41/52/47	30/36/32	42/55/47	26/35/29
	Bagged stats (Mean/SD)	47/2.7	32/1.3	47/2.2	29/1.9
Douglas Fir (15344)	Optimal Pruned Deviance explained (%)	54	40	50	45
	Bagged stats (Min/Max/Median)	53/60/57	39/44/41	51/61/56	45/52/48
	Bagged stats (Mean/SD)	57/1.8	42/1.3	56/2.4	48/1.4
Eastern white pine (7386)	Optimal Pruned Deviance explained (%)	31	18	22	16
	Bagged stats (Min/Max/Median)	34/42/37	18/26/21	24/36/29	14/24/20
	Bagged stats (Mean/SD)	37/2.4	21/1.7	29/3.1	20/2.2
Sugar maple (14781)	Optimal Pruned Deviance explained (%)	41	29	41	27
	Bagged stats (Min/Max/Median)	42/48/44	25/30/28	39/47/43	26/32/29
	Bagged stats (Mean/SD)	44/1.6	28/1.2	43/2.1	29/1.9
Chestnut oak (5724)	Optimal Pruned Deviance explained (%)	53	36	48	33
	Bagged stats (Min/Max/Median)	53/60/56	37/42/40	50/58/53	36/41/39
	Bagged stats (Mean/SD)	56/1.7	40/1.4	53/1.9	38/1.6

Note: The number in parenthesis after the species name is the number of FIA plots with the species present.

3 | RESULTS

The deviances explained by regression trees (pruned singular and bagged) among the six species for both RIV and AIV, along with summary statistics (Table 2), show that RIV generally outperformed AIV. This is true for all species except ponderosa pine at 20k, and all species except Douglas fir at 10k (Table 2). Also, for the bagged trees, the standard deviation (SD) is almost always higher with AIV compared to RIV, making RIV less variable and hence a more stable metric for modelling the response. Based on these results, we decided to confine our further analysis to RIV, which also has the added advantage of incorporating a portion of the interspecific interaction by way of including all species in the plot. AIV, however, can be a useful measure for some species when the interspecific component is not needed, and for some species for which the deviance explained is significantly higher.

Focusing only on the RIV response, the quartile distributions (minimum, maximum and median), as well as the mean and standard deviations of the bagged trees, portray the *explained deviance* statistics displayed by the six species (Table 2). Because both the singular tree and bagged trees use optimally pruned regression trees derived from the ORTBIM procedure, the range (maximum minus minimum) in the explained deviance is due to bootstrapping in the bagged trees. Using singular tree with the full training data is essential to unravel the main features influencing the distribution.

The deviance explained varied among the species, with Engelmann spruce, followed by Douglas fir having the highest values (Table 2). The lower deviance explained by eastern white pine (with only 31% even at 20k) may reflect that it is a highly adaptable species that has naturalized in many parts of the eastern United States, and hence more problematic to model.

The deviances explained by the explanatory variables were summed along with the variable defining the topmost split in the regression tree (Table 3). The climate/terrain grouping categories (Table 1) were used to summarize the deviances explained by these variables (Table 3) and plot their total importance for the western and the eastern species at 20 and 10k resolutions (Figure 1). At both 20 and 10k, for the western species, elevation is the most dominant variable explaining the bulk of the deviance – it is noteworthy that for the eastern species, elevation does not figure at 20k and is marginal at 10k. For Engelmann spruce, elevation is the most important variable at both scales, explaining 44.5% and 39.1% at 20 and 10k, respectively (Table 3), while dwarfing the other variables. For ponderosa pine, elevation explains 9.6% of the deviance at 20k and 6.4% at 10k; for Douglas fir, elevation explains 12% at 20k and was not important at 10k. For both ponderosa pine and Douglas fir, heat-moisture variables (AHM, CMI and CMD – see Table 3) become quite important at both scales. However, at 20k, the influence of seasonal precipitation and seasonal temperature was quite important for eastern species, followed by growing season variables. Heat-moisture indices are more important in the west and the role of

TABLE 3 Deviances explained by the explanatory variables at 20km (20k) and 10km (10k) resolutions.

	20k		10k			20k		10k	
	Var	Dev	Var	Dev		Var	Dev	Var	Dev
Engelmann spruce	elv*	44.5	elv*	39.1	Eastern white pine	PPT_sm*	7.1	Tave_sm	4.5
	Tave_at	4.4	CMD	3.5		DD5	6.9	PPT_sp*	3
	CMD	3.1	SHM	2.3		PAS	4.6	PAS	3
	PPT_sm	2.7	PAS	1.9		Tave_sm	3.4	PPT_sm	3
	CMI	2.4	Tave_at	1.8		ttri	2.8	PPT_at	1.3
	SHM	1.9	CMI	1		Tave_sp	2.4	PPT_wt	1.2
	PAS	0.9	ttri	0.7		PPT_wt	1.7	CMD	1.1
	tslope	0.3				SHM	1.1	elv	1.1
	Total	60.2	Total	50.3		PPT_at	0.5		
				Total	30.5			18.2	
Ponderosa pine	elv*	9.6	AHM	10.3	Sugar maple	PPT_sp*	14.4	AHM	6.8
	AHM	7.8	elv*	6.4		Tave_sm	7	Tave_at*	5.5
	Tave_wt	5.9	Tave_wt	4.5		CMI	6.6	elv	4.1
	ttri	4.8	ttri	4.2		NFFD	5.7	Tave_wt	3.8
	tslope	4.3	Tave_at	1.1		PPT_sm	2.7	EMT	3.7
	Tave_at	1.4	twi	0.9		PPT_at	1.9	PPT_wt	1.4
	Tave_wt	1.4	PPT_at	0.7		PAS	1.2	NFFD	0.9
	Tave_sm	1.1	Tave_sp	0.6		EMT	0.6	PAS	0.9
	PPT_sm	0.9	Tave_sm	0.5				CMI	0.7
	DD5	0.8						DD5	0.7
	Total	38		29.2		Total	40.1		
Douglas fir	CMI	20.8	AHM	13	Chestnut oak	PPT_sp*	14.9	Tave_at	11.7
	elv*	11.6	CMD*	8.4		Tave_sm	13.5	tslope	8.3
	CMD	11.6	PAS	7.7		ttri	12.4	PPT_sm*	4.3
	PPT_sp	4.9	PPT_at	5.7		PPT_sm	4.8	ttri	4.2
	PPT_sm	4	PPT_sp	3.8		EMT	2.8	EMT	2.2
	NFFD	0.7	Tave_sm	0.9		PPT_wt	2.7	Tave_sp	1.9
			Tave_wt	0.5		DD5	1.5	NFFD	1.5
	Total	53.6		40		NFFD	0.7	Tave_sm	0.8
								twi	0.6
						SHM	0.3		
				Total	53.3			35.8	

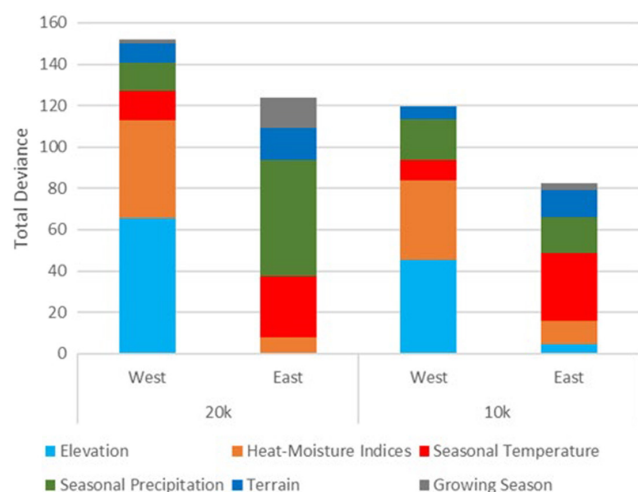


FIGURE 1 Comparison of the total deviances explained by the four dominant rules in the six explanatory groups (Table 1) – for the eastern and western species at 20km (20k) and 10km (10k) resolutions. The y-axis represents the total summed per cent deviances explained by each category.

The current distribution maps (Figure 2) illustrate the current FIA baseline conditions for the six species. They can be used to compare dominant modelled trends from bagged trees which use the FIA data from conterminous United States to map suitable habitats across the entire continental United States and Canada. The results for the six species described below pertain mainly to the 20k maps (Figures 3–8). Because 20k performance was higher (in terms of deviance explained – Table 2), the interpretation of the macroscale patterns is more robust compared to 10k (Figures S1–S6). However, interesting and revealing contrasts with the 10k distributions are discussed.

3.2 | Engelmann spruce

By comparing the current distribution of Engelmann spruce (Figure 2a) to dominant bagged rules at 20k (Figure 3a), we see the latter captures most of the current distribution in the conterminous United States and also some outlined by Little's range in Canada (Little, 1971). The deviances explained by the rules as discussed in the previous section and Table 2 clearly show the

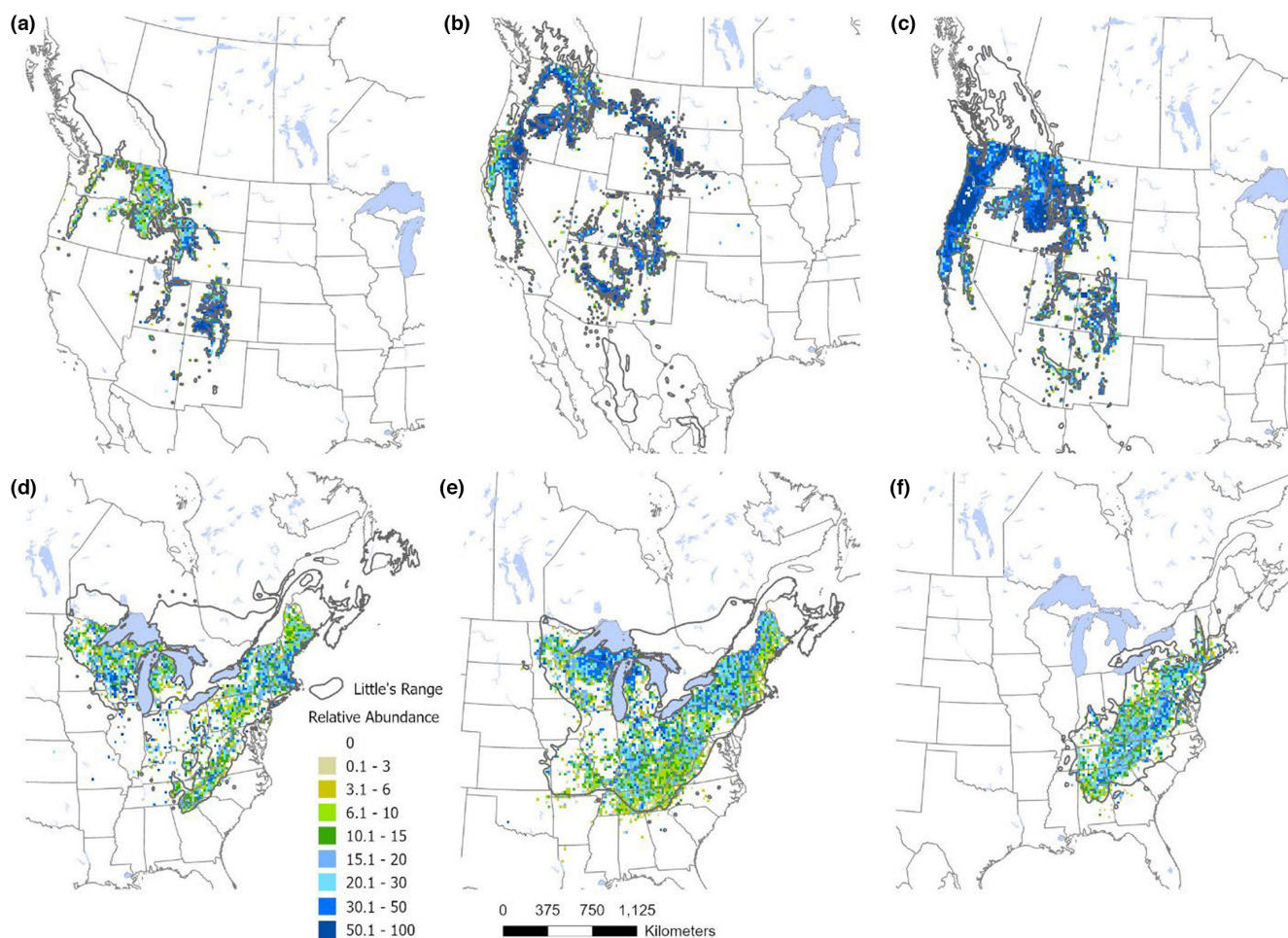


FIGURE 2 The current distribution of the six species using FIA for the six species. (a) Engelmann spruce; (b) Ponderosa pine; (c) Douglas fir; (d) Eastern white pine; (e) Sugar maple and (f) Chestnut oak. The dark-grey outline is Little's range map.

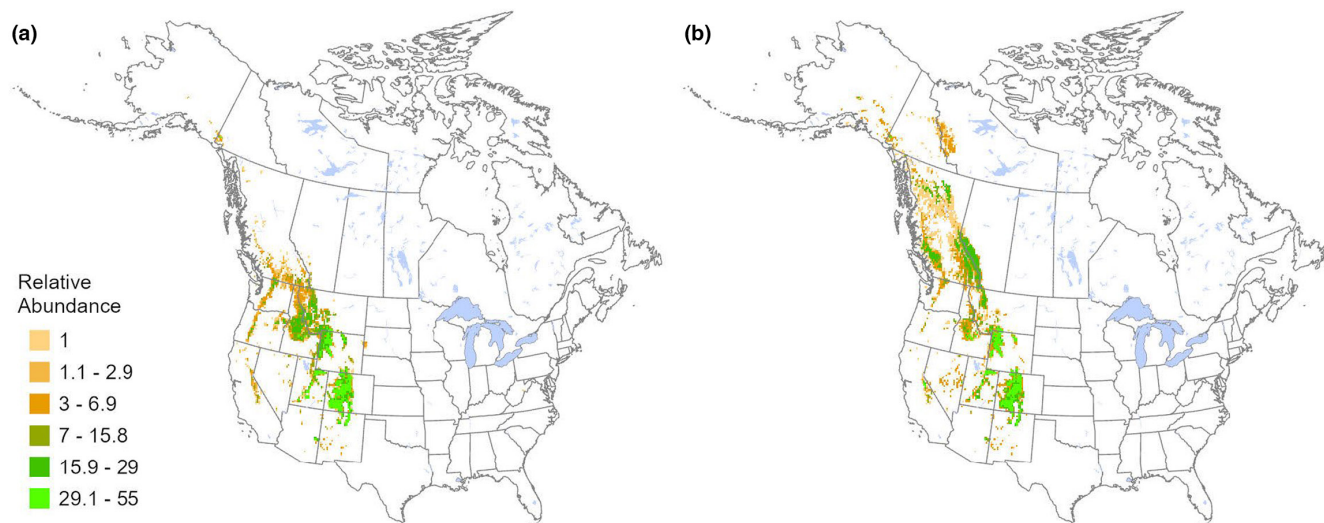


FIGURE 3 The bagged dominant rules mapped at 20k for Engelmann spruce (*Picea engelmannii*) for current (a) and future climate SSP5-8.5 (b). The legend pertains to per cent relative abundance derived from the quantile distribution of the species.

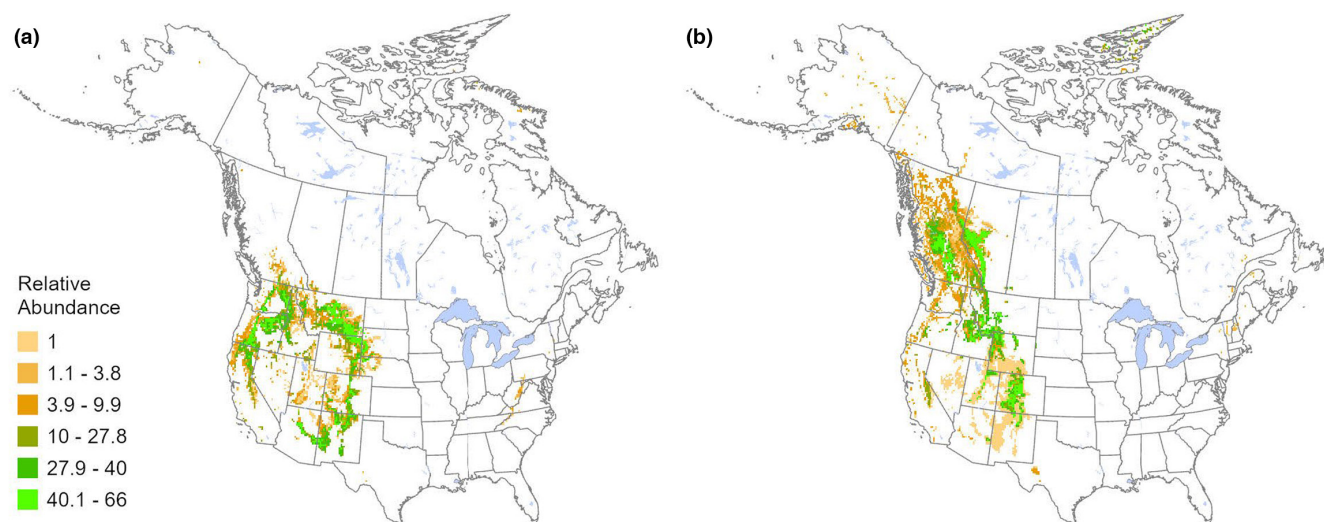


FIGURE 4 The bagged dominant rules mapped at 20k for Ponderosa pine (*Pinus ponderosa*) for current (a) and future climate SSP5-8.5 (b). The legend pertains to per cent relative abundance derived from the quantile distribution of the species.

dominance of elevation. Under SSP5-8.5 future climate (Figure 3b), there is further northern extension to Canada. The patterns at 10k (Figure S1A.a,A.b) for current and future climate are similar to that of 20k. On the whole, Engelmann spruce shows expected behaviour (upslope and northward trajectory) and does not reveal any surprising trends at both 20 and 10k – indicating the strong elevational trends that influence its current distribution.

3.3 | Ponderosa pine

The dominant bagged model (Figure 4a) shows good correspondence with the current distribution, capturing even the extension into Canada shown by Little's range map (Figure 2b). There is a pronounced extension of the range into Canada under SSP5-8.5,

showing some low-suitability areas even in Alaska. There is lowering of habitat suitability that is quite pronounced in the margins of the distribution along with some low suitability spots appearing in the more southern Appalachian regions under the future scenario. At 10k (Figure S1B.a,B.b), Ponderosa pine follows a similar trend compared to 20k, not showing any major departures.

3.4 | Douglas fir

There is good match of the dominant bagged model (Figure 5a) with the current distribution (Figure 2c), with the former capturing some of the areas in Little's range map in Canada. Notice that Douglas fir shows suitable habitats in the east because it was previously planted there. Because we excluded planted plots, these are likely sporadic

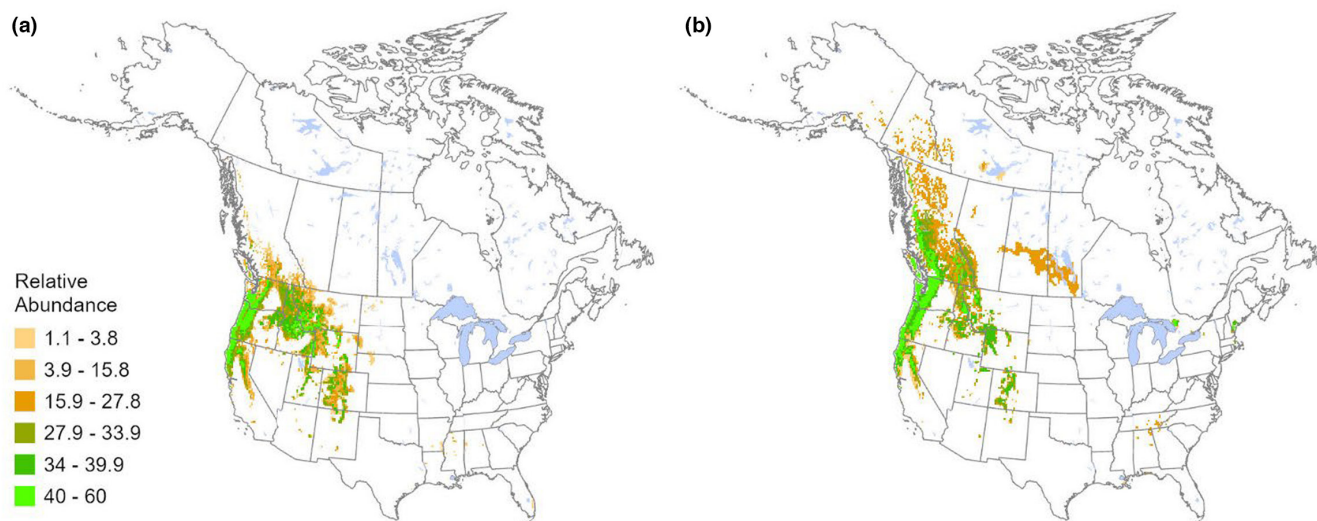


FIGURE 5 The bagged dominant rules mapped at 20k for Douglas fir (*Pseudotsuga menziesii*) for current (a) and future SSP5-8.5 (b). The legend pertains to per cent relative abundance derived from the quantile distribution of the species.

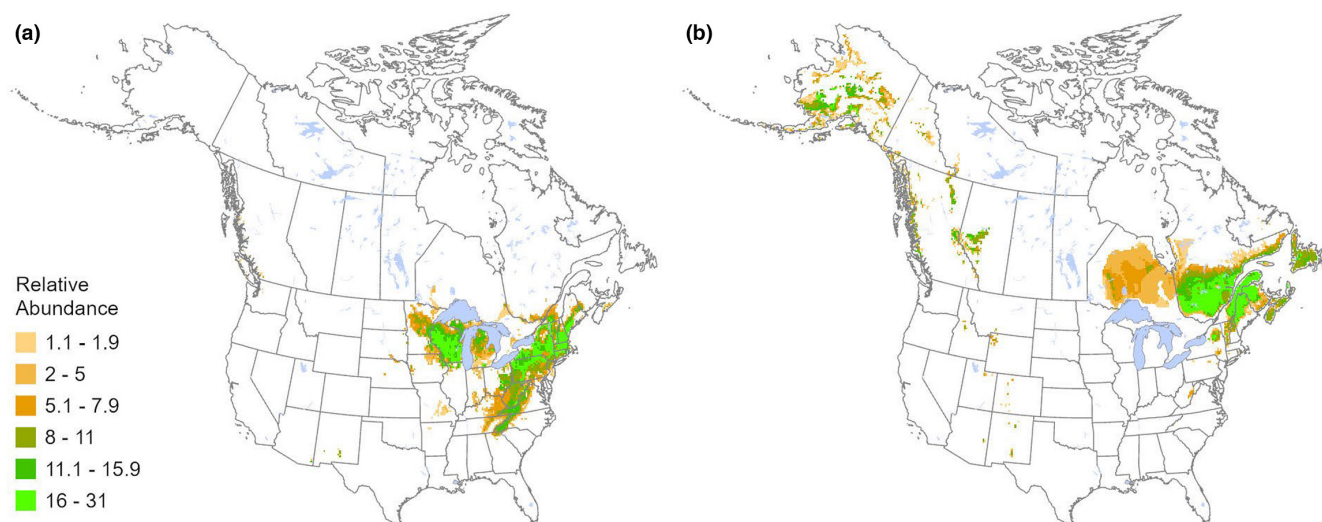


FIGURE 6 The bagged dominant rules mapped at 20k for eastern white pine (*Pinus strobus*) for current (a) and future climate SSP5-8.5 (b). The legend pertains to per cent relative abundance derived from the quantile distribution of the species.

naturalized populations that escaped plantings. Under SSP5-8.5 (Figure 5b), there is an extension of suitable habitats north-west into Canada, into Yukon. Lower-quality habitats appear in mid-western Canada, along with some locations of higher suitable habitats appearing in Ontario and the north-eastern United States. Under the future scenario, degradation and thinning of the range are visible in the eastern portions of the Rockies. At 10k (Figure S1C.a,C.b), the eastern regions gain more habitats with a substantial portion of Ontario acquiring higher-quality habitats, along with some regions in the north-eastern and southern Appalachians. Because heat-moisture variables figure strongly at both 20 and 10k, but elevation is prominent only at 20k (Table 3), it appears that 10k is showing expansion to the east (without elevation constraining the acquisition of suitable habitats in the east). The thinning of habitats in the eastern

portions of its current range is visible under the future scenario at both 20 and 10k.

3.5 | Eastern white pine

Eastern white pine captures the current distribution of white pine (Figure 2d) fairly well, however, only 31% of the deviance is explained even at 20k (Table 2). The dominant trends under SSP5-8.5 (Figure 6b) show that the main shift is towards the northeast although there is a substantial presence in Alberta, Canada and to the northwest into Alaska (Figure 6b). Most of the dominant suitable habitats tend to be lost in the United States, except in the north-eastern states. At 10k, which only explains 18% of the deviance

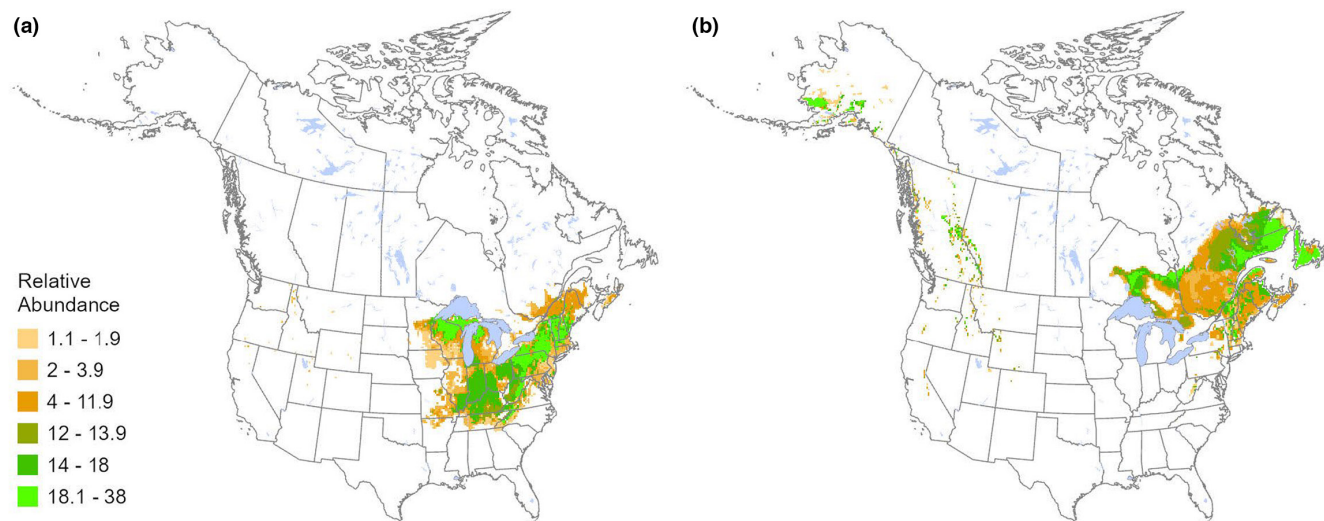


FIGURE 7 The bagged dominant rules mapped at 20k for sugar maple (*Acer saccharum*) for current (a) and future climate SSP5-8.5 (b). The legend pertains to per cent relative abundance derived from the quantile distribution of the species.

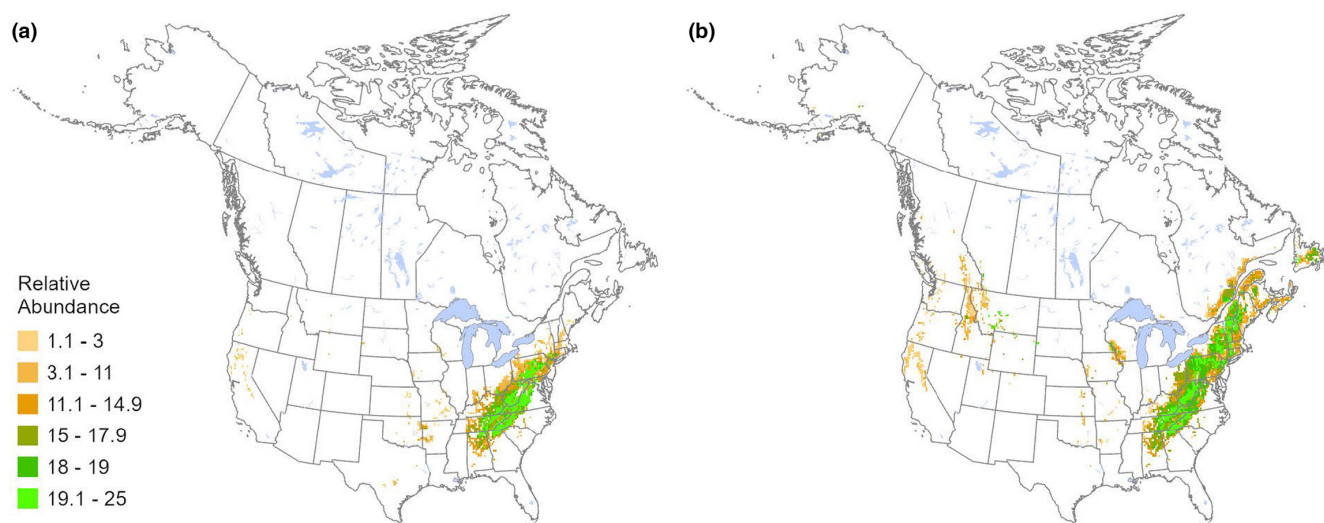


FIGURE 8 The bagged dominant rules mapped at 20k for chestnut oak (*Quercus montana*) for current (a) and future climate SSP5-8.5 (b). The legend pertains to per cent relative abundance derived from the quantile distribution of the species.

(Table 2), the trend towards the west is apparent under SSP5-8.5 (Figure S2A.a,A.b), with lower-quality habitats.

3.6 | Sugar maple

Sugar maple's dominant trends are modelled fairly well (Figure 7a) compared to the current distribution (Figure 2e), although Canada's portion of Little's range is missing. A substantial portion of sugar maple's dominant habitat moves north-eastward under SSP5-8.5 (Figure 7b). Scattered suitable habitats appear in the west and Alaska. The quality of future habitats also tends to be lower in several regions. At 10k, the bagged current (Figure S2B.a,B.b) does not capture the

dominant trends of the current distribution (Figure 2e) well. However, the trends in the east are somewhat similar, although SSP5-8.5 shows high-quality habitats in the east much further north into Canada. However, it should be noted that there is a substantial drop in the deviance explained for sugar maple from 20k (41%) to 10k (29%) (Table 2).

3.7 | Chestnut oak

The dominant abundance pattern of chestnut oak (Figure 2f) is captured quite well (Figure 8a), although the model shows some lower-quality habitats further west even under current conditions. Under SSP5-8.5, there is a strong north-eastern extension. The suitable

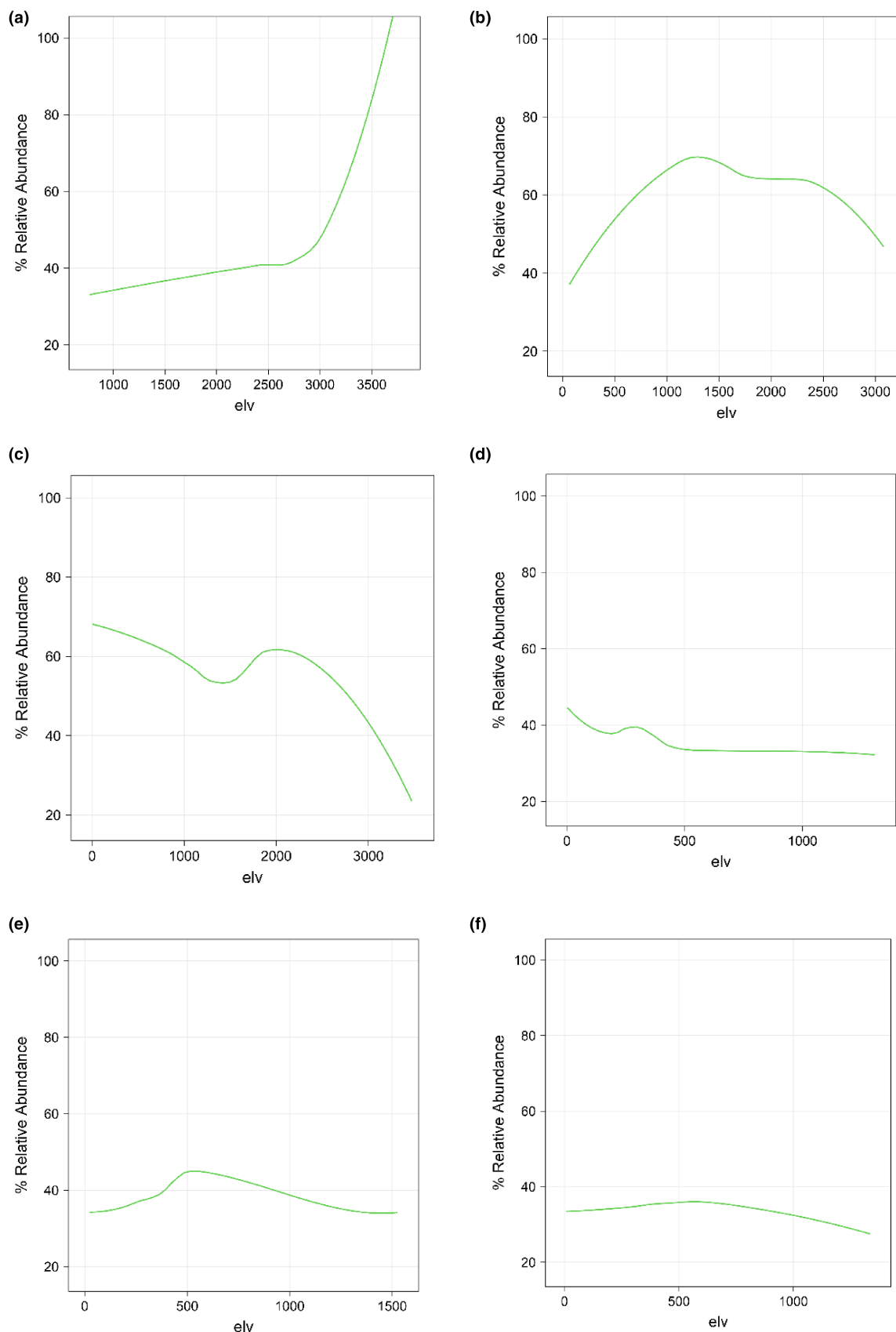


FIGURE 9 The overall response of elevation (x-axis) versus FIA plot-based relative abundance (y-axis) for the six species – (a) Engelmann spruce; (b) Ponderosa pine; (c) Douglas fir; (d) Eastern white pine; (e) Sugar maple and (f) Chestnut oak.

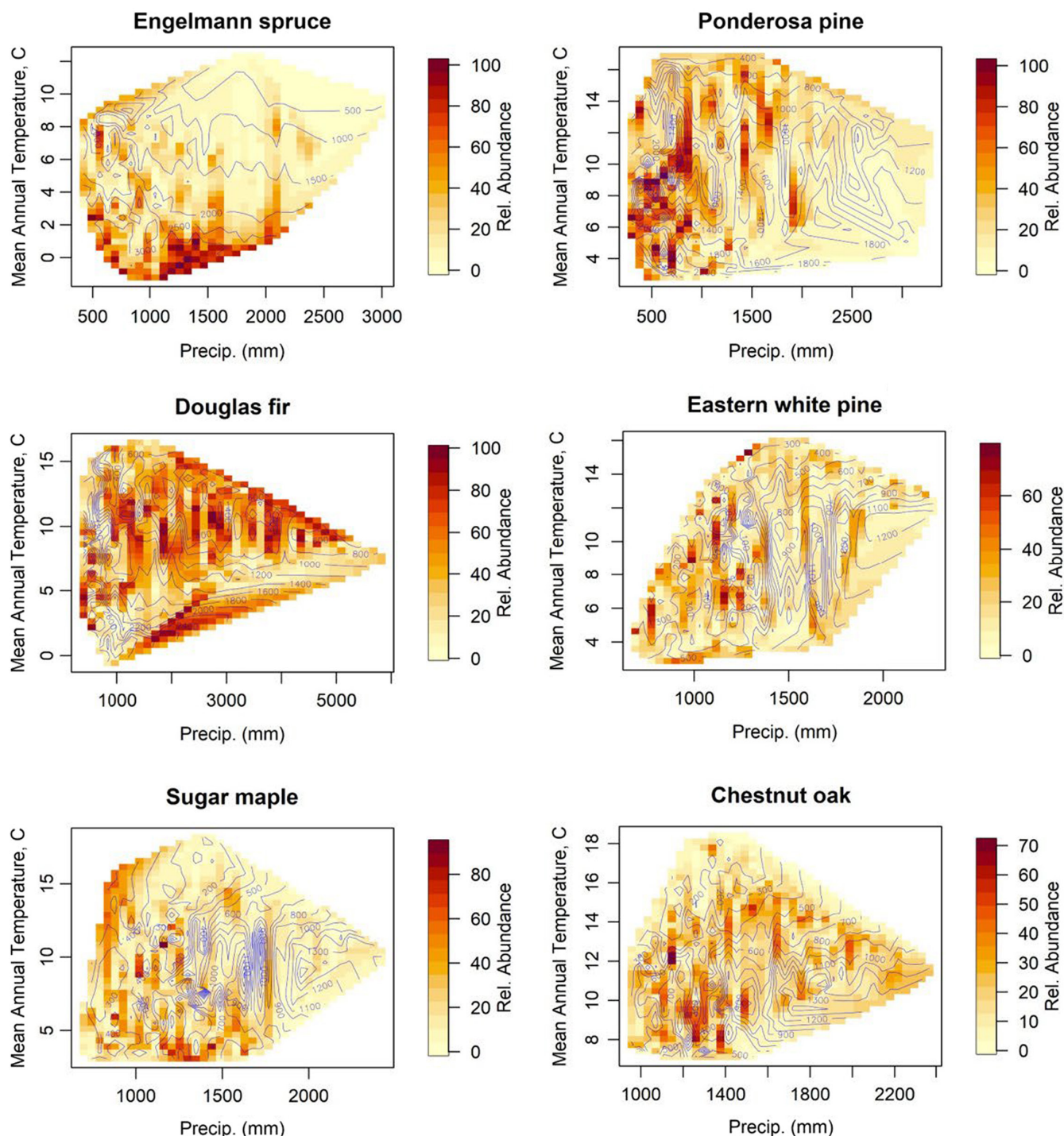


FIGURE 10 The niche plots depicting interpolated relative abundance plotted against mean annual temperature (C) and total annual precipitation (mm). The contours depict elevation.

habitats in the west are also more prominent under SSP5-8.5 (Figure 8b). Compared to the other eastern species (eastern white pine and sugar maple), the quality of the future habitats is not impacted much. At 10k (Figure S2C.a,C.b), the pattern is somewhat similar, although there is substantial thinning and degradation of habitats under SSP5-8.5. It should be noted that the drop in the deviance explained is quite steep with only 36% explained at 10k compared to 53% at 20k (Table 1).

3.8 | Dominant rules from singular tree

Table S1a-f shows the dominant rules for the six species derived from the optimally pruned singular tree for 20 and 10k resolutions (Step 5, ORTBIM procedure). While rules from singular tree exhibit variation and should not be used for prediction, they are highly valuable for interpreting when due consideration is given to the per cent deviance explained. They show which explanatory variables, along

with their interactions, are influencing the dominant distribution under current and future (since the rules would not vary much due to adaptational lag) conditions. These rules should be interpreted in relation to maps of bagged dominant rules (Figures 3–8). The variations exhibited by these rules between 20 and 10k can be scrutinized for greater insight into the nuances that drive the distributions under these two scales, especially when examining the climate–terrain niche space of the species, greatly aiding interpretation, which we discuss below.

3.9 | Plot-level responses

The FIA plots can highlight finer-scale responses compared to the aggregated response models at 20 and 10k. The plot-level relative abundance response to elevation (Figure 9) and the most important explanatory variable (Table 3, and Figures S3–S8) were plotted as smooth loess response curves (family='gaussian', span=2/3, degree=2) to highlight the non-linear relationships and examine the elevation class that is most similar to the overall trend. Loess plots are based on the idea that any function can be well approximated in a small neighbourhood by a low-order polynomial and that simple models can be fit to data easily.

The relationship across the entire range of elevation for the six species (Figure 9a–f) clearly shows the important role elevation has for the western species. RIV for Engelmann spruce increases exponentially starting around 2500m; for Ponderosa pine, RIV increases up to ~1300m and then levels off up to ~2300m and then starts decreasing. Douglas fir shows an interesting non-linear pattern where RIV decreases up to ~1200m, levels off and increases up to 1800m, levels off and starts to decrease exponentially from around 2100m. Compared to the western species, the response of RIV to elevation for the eastern species (Figure 9d–f) is much less pronounced.

FIA plot-level data trends also show finer-scale relationships to the important variables in the model. When the most important variable (one that explains the most deviance, Table 3) at 20k is plotted against RIV by splitting elevation into classes for each species (Figures S3–S8), we can observe the elevation class in which the response curve resembles the overall response the most, and how it responds differently at various elevation classes. It should be noted that loess curves represent approximate trends because they simulate the finer FIA plot data and are likely to fit some noise with the signal – however, the overall trends provide insights into the distributions of the species at finer scales. For Engelmann spruce, the overall trend with abundance for the most important variable CMD is clearly apparent at 1000–1500m elevation class and also somewhat similar at 2000–2500m (Figure S3). For Ponderosa pine (Figure S4), the relationship of AHM with RIV is visible between 1500 and 2000m. For Douglas fir, the relationship of RIV with CMI (Figure S5) at different elevation classes shows that this trend is visible at several classes (–400–500, 1000–1500 and 1500–2000m). For the eastern species, the curves are much flatter – eastern white

pine's (PPT_sm) pattern is visible under both –400–500m and 500–1000m classes (Figure S6). For sugar maple (PPT_sp), the overall response is most apparent at –400–500m and 500–1000m classes (Figure S7). The abundance response of chestnut oak to spring precipitation (PPT_sp) decreases gradually, reflected both at –400–500m class as well as at 500–1000m (Figure S8).

3.10 | Climatic niche–elevation contour plots

The species' niche in climate space casts light on the climatic range the species occupies. To plot this, we mapped interpolated RIV (from FIA plot data) against averaged total annual precipitation, mm (MAP – x-axis) and mean annual temperature, C (MAT – y-axis) and draped elevation contours (Figure 10). These niche–elevation contour plots are a succinct way of summarizing the differences between the eastern and western tree species. For example, we can see that, in general, the range of precipitation (x-axis) is larger in the western species compared to the eastern species. The temperature range (y-axis) is somewhat higher for the eastern species, especially when compared to Engelmann spruce which prefers much cooler temperatures. Bands of higher relative abundances tend to be narrower for Engelmann spruce which prefers lower temperatures, and mostly 1200–2000mm of rainfall. Ponderosa pine has a narrower range of precipitation where it performs well, but a much broader temperature range. The eastern species are generally climatically more tolerant (with the caveat that they occupy a smaller range of precipitation compared to the western species) and occupy broader niches compared to their western counterparts – although Douglas fir in the west fares well across its entire climatic range. When one examines the elevation contours, the contrast between the eastern and western species gets sharper. Engelmann spruce, for example, shows large abundance at higher elevations. The spacing of the contours reflects terrain heterogeneity – with Engelmann spruce preferring higher elevations over rugged terrain, whereas terrain heterogeneity becomes more important for Ponderosa pine which prefers relatively lower elevations, but dissected terrain. For Douglas fir, the pattern is more difficult to discern as it does well across its entire climatic niche. For the eastern species, chestnut oak, which mainly spans the Appalachian ranges and their foothills, the importance of terrain heterogeneity is more obvious, preferring more rugged terrain compared to eastern white pine and sugar maple. These interactions between climate and terrain are captured well in the dominant rules (Table S1a–f), as well as the optimally pruned bagged maps (Figures 3–8).

4 | DISCUSSION

4.1 | Background

In this study, we have emphasized the exploratory and interpretive aspects of macroscale analysis by focusing on ecologically dominant patterns that are influencing the abundance distribution of six major

species in the United States and Canada – and comparing important features of these distributions in the eastern and western regions. This analysis helps lay an interpretative framework for conducting better predictive analysis by capturing the main ecological trends in species abundance distribution. Within this context, some variance in the rules can be tolerated because many ecological questions typically involve inherently variable biotic and abiotic feedbacks in multivariate space, and the target species themselves are adaptable and influence their niches (Laland et al., 2016).

Initially, when formulating the ORTBIM procedure, we explored Rulefit, Sirrus, Cubist and interpretML (Bénard et al., 2021; Chang et al., 2021; Friedman & Popescu, 2008; Kuhn & Quinlan, 2023) as alternatives to a singular decision tree – but these algorithms did not yield results that were interpretable and mappable using our spatial datasets. The singular regression tree derived from an optimal pruning parameter proved to be the best option. Singular trees can overfit and exhibit variation for prediction, but can still be valuable for interpretation, especially when combined with bagging to examine variability. In addition, the rules derived from a singular tree are useful for mapping the predictors to understand which predictors and their interactions are driving the distribution geographically (Iverson & Prasad, 1998; Prasad, 2015). It should be emphasized that, in spite of some variability in the rules due to the singular regression tree, it is the dominant pattern in the distribution of habitat suitability and not the prediction of abundance values that is the focus of this analysis. However, the ORTBIM procedure stabilizes the variability in the rules to a minimal set of ecologically relevant features, while still incorporating correlated variability via bagging. This interpretative exploration is urgently needed because the entire current species range and beyond could be the domain of disturbance under climate change, especially with the northern latitudes warming faster than the southern ones (Walsh, 2014). By specifically exploring the dominant patterns of abundance by focusing on quartiles greater than the first, and comparing these macroscale patterns with plot-level inventory data, we reveal dynamic aspects of species distribution that are often hidden by black-box predictive methods.

An assumption and potential limitation inherent in our approach is that the realized niche reflected by FIA-based abundance data is sufficient for the ORTBIM procedure. Several studies have pointed out that many tree species can colonize and survive in areas well beyond their current climatic profile (e.g. see Booth, 2017). The assumption of using the realized niche to extract rules was a compromise that was needed in our interpretative approach for regression-based modelling of relative abundance. Gathering binary presence/absence data from other sources may have captured a somewhat broader species range to assess climatic requirements beyond the realized niche (Booth, 2016; McKenney et al., 2007; Zhang et al., 2017). But this would not have yielded the insights that only continuous abundance data using the rigorous sampling procedure of FIA can provide. Based on our experience with modelling continuous data, we are confident that our approach is robust enough to extract the principal rules from the core of the distribution, which are well represented by the realized niche. To explore this further,

we created climatic envelopes of species (using climatic variables like mean annual temperature, extreme minimum temperature, annual heat moisture index, etc.) based on their realized niche. The results showed that the envelopes were far greater than the actual species distribution, lending support to our assumption that the rules derived from the realized niche will not limit the climatic potential of the species. Also, when the bagged rules are mapped under SSP5-8.5 (Figures 3–8), we see that several eastern species extend their potential range into northern Canada and Alaska, providing further confidence about the climatic plasticity inherent in the rules derived from the realized niche.

Another potential limitation in our approach is that we use only 'natural' forested plots which do not capture the slightly wider niche provided by including plantations. We did this for two reasons: (a) only commercial species have plantations – which would cause data asymmetry among species; and (b) the relative abundance of non-natural stands would be misleadingly high due to the elimination of competition for forest management objectives – which will adversely affect model performance. However, we include all forested occurrences in FIA plots when aggregating to 20 and 10k by rounding RIV values of abundance greater than 0, however tiny, as 1.

A further limitation of our approach is that the current distribution is derived only from the conterminous US FIA data, which constrains the models to be conservative when mapping bagged rules across Alaska and Canada. We refrained from using Canada's photo-plot data because they stem from a different protocol and their combination with FIA data would have complicated the interpretative aspects of our analysis. However, this omission is justifiable for the species chosen because the FIA-based distribution (which is based on rigorous, systematic sampling) represents the species' core climate–terrain range sufficiently for the bagged rules to map reliably across the entire domain of inquiry for interpretative purposes.

4.2 | East versus West

Our earlier attempts to model and project tree species' suitable habitats (Iverson et al., 2008, 2019) have been confined to the eastern United States, although we have modelled a subset of the eastern species across most of their range (Prasad et al., 2020). We have so far not attempted to model the species in the western regions of United States and Canada using our habitat suitability and colonization likelihood models – an active area of biotic and abiotic change (Stanke et al., 2021). While the six species we have studied here are a tiny sample of the many found in the east and west, they represent wide-ranging species characteristic of these two regions. Importantly, we can be confident that the interpretations using our ORTBIM procedure can extract features that are relevant to modelling multiple species across their entire range using a more predictive ensemble modelling framework. Crucially, the unexplained deviance opens avenues to probe other historical, biogeographic and disturbance factors that have influenced

current distributions, and will continue to shape future responses to rapid climate change.

From our results, the predominant difference between the western and eastern species, as expected, is the role of elevation, which has a major influence on our examined species in the west at both 20 and 10k (Figure 1, Table 3). Correspondingly, the seasonal precipitation variables are important in the east at 20k and less so at 10k. Terrain-related features appear to be more important for chestnut oak which spans the Appalachian in the east (Table 3), most likely because its abundance response varies with topographic heterogeneity (partly captured by *ttri* variable), and not primarily high elevation ranges more common in the west.

4.3 | Scale issues and residual deviance

One of our goals was to understand the role of scale, and the type of response on the outcome. It is clear that relative abundance, as compared to absolute abundance, produces a stronger signal even in the west (except for Ponderosa pine), in spite of lesser diversity at the plot level. Relative abundance has been the response of choice all along in the east, as presented in the Climate Change Tree Atlas (Iverson et al., 2008; Iverson & Prasad, 1998; Peters et al., 2019). The domination of plots by single species, much more predominant in the west compared to the east, still yields good interpretive models and is most likely to be true for predictive models also – boosting our confidence in this response metric as we extend our models to the west.

Ecological questions should be framed based on the scale of the analysis – a macroscale perspective allows us to understand emergent patterns across the full range of the species, and subsequently, help probe more regional and local phenomena based on this broad understanding. Within this macroscale perspective, the deviances explained in our study vary with resolution – higher at 20k compared to 10k for all the species (Table 2). This systematic pattern is likely because aggregation at coarser resolutions brings out the emergent properties of the response (in this case relative abundance) to match more closely with the scale of climatic and terrain variables. This relationship breaks down with finer resolutions because there are sharper variances between the abundance and climate-related patterns compared to the coarser resolutions (e.g. a 10k aggregated pixel is more likely to have fewer FIA plots), which cannot be explained by the explanatory variables. Therefore, the trade-off for modelling at finer resolutions is to sacrifice interpretability and predictability and add significant computational costs – which is not attractive for macroscale investigation of patterns across entire ranges for multiple species.

While climate and terrain remain the main macroscale factors influencing species range (Pederson et al., 2015; Toledo et al., 2011), other ecologically relevant factors like disturbances and soils can become important at finer scales. The upper branches of the regression tree (especially the top split) are based on more regional features defining the distribution, with the lower branch

splits based more on local features (Iverson & Prasad, 1998). Thus, it is possible that the inclusion of soil variables could have better explained some of the nuances of the distribution. But from our past experience with modelling in the eastern United States, there is often a scale-based confounding of climate–soil interactions, and furthermore, even including soil variables is insufficient to compensate for the lower deviances explained at finer resolutions. Additionally, many aspects of the species' historical biogeographic and eco-evolutionary legacies, as well as past and current disturbances become crucial at finer resolutions, and are unknown or incomplete to incorporate into models. Therefore, for macroscale analysis across the entire range of the species, the coarser 20 km cells based on climate–terrain interactions are better able to capture a greater portion of the emergent properties and are more likely to reflect dominant trends across species ranges. It is noteworthy that for some species, 10k in spite of explaining less deviance can also provide some important insights when contrasted with 20k, as each scale tells a different story. When investigating regional patterns, a hybrid approach can be helpful by combining 20k with 10k resolutions (Peters et al., 2019).

4.4 | Management issues

As noted previously, the deviance not explained by the models also becomes a rich source of inquiry into the historical biogeography of past and current disturbances. The species' legacies and acquired evolutionary traits can give us vital information about how they are likely to respond to disturbance currently and in the future. These can be systematically tabulated for each species as modifying factors (Matthews et al., 2011), and many of the disturbance layers (land-use change, fire impacts, contemporary climate change, etc.) can be made spatially explicit to provide additional insights into the management of the species as post-model filters at specific locations. In conjunction with model interpretations and predictions, they can also provide insights into landscape features that could act as refugia that can facilitate or deter species migrations (Roberts & Hamann, 2015). For example, we know from our analysis that elevation is the most influential variable for some of the western species (Figure 1, Table 3). We also know that elevation is interacting with many seasonal and heat–moisture variables in the west (rules in Table S1a–f) which are likely to have a major influence as the climate warms, creating both refugia as well as barriers, depending on the topographic features. The maps of dominant rules (Figures 3–8) can provide spatial context adding information to where these forces are likely to operate in the landscape, enabling further investigations into these regions along with disturbance regimes.

The habitat suitability distribution projected by the dominant bagged rules (at 20k) is often wider in geographic scope compared to the current FIA-based distribution – and can appear in unexpected places throughout continental United States–Canada (as we see in all the species except Engelmann spruce, which is more clearly defined by elevation), and this pattern is more likely under future climates. For

example, the eastern species we tested had suitable habitats appear in the north-western regions of the continent under future climates (Figures 6–8). Similarly, some western species appeared in suitable spots in the east, for example, Ponderosa pine (along central and southern Appalachians under SSP5-8.5 – Figure 4) and Douglas fir in Ontario and Maine (Figure 5). The eastern species are influenced by a mix of factors (Figure 1) with seasonal precipitation patterns having a larger influence compared to others. However, these are likely to interact with other features to influence abundances which are captured by mapping the dominant bagged rules (Figures 3–8). We see strong tendencies for eastern species to shift their habitats north-eastward to Canada, and also to the western parts all the way to Alaska. Comparatively, there is less tendency for the western species to move eastward, the dominant pattern being north-westward into Canada. The fact that we do not currently find the species in these suitable habitats is related to climatic lags, climatic disequilibrium and the ability of the species to migrate to suitable habitats amidst other interspecific and disturbance interactions (Davis, 1989; Hoek, 2001; Svenning & Sandel, 2013). The potential of changing habitat suitability, however, has major implications for management in both the United States and Canada with increasing need for greater research and management collaboration as the climate warms faster in the northern latitudes compared to the south (Walsh, 2014). Many of the species show both habitat degradation and thinning in the southern regions under future climates; this tendency is more pronounced for eastern species (Figures 6–8). The patterns at 10k (Figures S1 and S2) generally mirror that of 20k, although there are some exaggerated patterns (especially for the eastern species) that could be related to lesser deviance explained compared to 20k.

It is noteworthy that tree species have always been opportunistic, migrating in time and space based on niche availability (with strong interactions with climate and topography) and their own acquired evolutionary traits. Therefore, the near future will be no different with niches appearing in unexpected places but with human influence acting both as a facilitator and a barrier to the occupation of these niches.

4.5 | Plot-level insights

While the bagged rules provide information on the important variables that are influencing the distribution via dominant rules, these relationships operate at coarse scales. The FIA plot-level analysis of trends, however, can provide a finer-scale perspective when overlaid on the dominant trends. Understanding the correspondence between coarser-scale trends and plot-level trends can help management because they provide important insights based on matching and contrasting patterns at these two scales. While the relative importance of elevation is reinforced for the western species (Figure 9), by examining the abundance response of other important variables with elevation classes, we can also uncover at what range(s) of elevation the overall pattern is likely to be manifesting (Figures S3–S8). For example, for ponderosa pine, the overall pattern

is most evident at elevation range of 1500–2000m (Figure S4), but for Douglas fir, the overall pattern is partly reflected in several classes (Figure S5). This strong correspondence between elevation and abundance breaks down in the eastern species (Figures S6–S8).

A succinct way of summarizing climatic and elevational differences between the eastern and western species is via the niche-elevation contour plots (Figure 10). In addition to the elevational suitability within the climatic range, it illustrates the suitability of terrain heterogeneity – for example, the Engelmann spruce prefers higher elevations compared to the higher preference for more dissected terrain for ponderosa pine.

Both these types of plots (Figures S3–S8 and Figure 10) reveal finer plot-level non-linear relationships that are harder to decipher from the maps (Figures 3–8) and the rules (Table S1a–f). These, together with the macroscale analysis, reveal important insights into where to plant and at what ranges the response is likely to be most pronounced – helping with assisted migration plans of selected species. This interpretative framework can also be extended to explore how variables other than climate-terrain can influence species-specific rules and distribution – for example, to investigate the role of trait-fitness relationships, etc. However, the procedure has to be modified by carefully considering how these variables scale with respect to the response variable.

In conclusion, our ORTBIM procedure provides a robust framework to address ecologically meaningful interpretation of the dominant rules influencing the abundance patterns of tree species and provides insights into how climatic-terrain habitats can rearrange under future climates. In addition, it offers a meaningful template to examine the role of scale in interpreting ecological patterns – and differentiates the important variables influencing the distribution in the eastern and western parts of the continent. Unexplained deviance can be used as a feature to examine historical biogeographical and disturbance components of current distributions, offering guidelines on how they may change under future climates. Our procedure, which is appropriate for examining macroscale distribution of species across their entire range (together with finer-scale FIA plot-based analysis of non-linear relationships), can be applied wherever there is sufficient inventory-based abundance data for tree species. Where only presence/absence data are available, a modification of the procedure is needed to yield reliable results. We believe our procedure is a useful addition to the mostly predictive black-box modelling frameworks that are predominant today.

ACKNOWLEDGEMENTS

We are grateful to the field crews of the Forest Inventory Analysis programme in the United States for gathering data that made this paper possible. We would like to thank Andy Gougherty, Brice Hanberry and Erik Lilleskov for their helpful reviews of the manuscript. And to the three anonymous reviewers whose comments greatly improved the manuscript. Disclaimer: The use of trade, firm or corporation names in this publication is for the information and convenience of the reader. Such use does not constitute an official endorsement or approval by the U.S. Department of Agriculture or the Forest Service of any product or service.

CONFLICT OF INTEREST STATEMENT

We have no competing interests.

PEER REVIEW

The peer review history for this article is available at <https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/ddi.13707>.

DATA AVAILABILITY STATEMENT

For peer review process: <https://datadryad.org/stash/share/sxGhOoMEsfAeAexYVQXTH98rczPLVYW13I46fQn4HIE>. Can be accessed via: <https://doi.org/10.5061/dryad.d7wm37q46> (20MB). (1) Unzip the downloaded file to a folder, (2) View the README.md and (3) The required files with instructions are in rulepc_ManuscriptData.zip.

ORCID

Anantha Prasad  <https://orcid.org/0000-0002-4645-6260>

REFERENCES

- AdaptWest Project. (2021). *Gridded current and projected climate data for North America at 1 km resolution, generated using the ClimateNA v7.01 software*. <https://adaptwest.databasin.org/pages/adaptwest-climatena/>
- Akima, H. (1991). A method of univariate interpolation that has the accuracy of a third-degree polynomial. *ACM Transactions on Mathematical Software*, 17(3), 341–366. <https://doi.org/10.1145/114697.116810>
- Akima, H., & Gebhardt, A. (2021). *akima: Interpolation of irregularly and regularly spaced data*. R package version 0.6-2.3. <https://CRAN.R-project.org/package=akima>
- Barton, K. E., Howell, D. G., & Vigil, J. F. (2003). *The North America tapestry of time and terrain*. USGS Geologic Investigations Series map I-2781. <https://pubs.usgs.gov/imap/i2781/>
- Bénard, C., Biau, G., Da Veiga, S., & Scornet, E. (2021). SIRUS: Stable and interpretable rule set for classification. *Electronic Journal of Statistics*, 15(1), 1–79. <https://doi.org/10.1214/20-EJS1792>
- Booth, T. H. (1990). Mapping regions climatically suitable for particular tree species at the global scale. *Forest Ecology and Management*, 36(1), 47–60. [https://doi.org/10.1016/0378-1127\(90\)90063-H](https://doi.org/10.1016/0378-1127(90)90063-H)
- Booth, T. H. (2016). Estimating potential range and hence climatic adaptability in selected tree species. *Forest Ecology and Management*, 366, 175–183. <https://doi.org/10.1016/j.foreco.2016.02.009>
- Booth, T. H. (2017). Assessing species climatic requirements beyond the realized niche: Some lessons mainly from tree species distribution modelling. *Climatic Change*, 145(3–4), 259–271. <https://doi.org/10.1007/s10584-017-2107-9>
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140. <https://doi.org/10.1007/BF00058655>
- Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Wadsworth.
- Brenning, A., Bangs, D., & Becker, M. (2018). RSAGA: SAGA geoprocessing and Terrain analysis. R package version 1.3.0. <https://CRAN.R-project.org/package=RSAGA>
- Chakraborty, D., Başağaoğlu, H., & Winterle, J. (2021). Interpretable vs. noninterpretable machine learning models for data-driven hydro-climatological process modeling. *Expert Systems with Applications*, 170, 114498. <https://doi.org/10.1016/j.eswa.2020.114498>
- Chang, C.-H., Tan, S., Lengerich, B., Goldenberg, A., & Caruana, R. (2021). *How interpretable and trustworthy are GAMs?* ArXiv:2006.06466 [Cs, Stat]. <http://arxiv.org/abs/2006.06466>
- Davis, M. B. (1989). Lags in vegetation response to greenhouse warming. *Climatic Change*, 15(1–2), 75–82. <https://doi.org/10.1007/BF00138846>
- Dwyer, K., & Holte, R. (2007). Decision tree instability and active learning. In J. N. Kok, J. Koronacki, R. L. de Mantaras, S. Matwin, D. Mladenić, & A. Skowron (Eds.), *Machine learning: ECML 2007* (Vol. 4701, pp. 128–139). Springer.
- Dyer, J. M., & Hutchinson, T. F. (2019). Topography and soils-based mapping reveals fine-scale compositional shifts over two centuries within a central Appalachian landscape. *Forest Ecology and Management*, 433, 33–42. <https://doi.org/10.1016/j.foreco.2018.10.052>
- Etterson, J. R., Cornett, M. W., White, M. A., & Kavajecz, L. C. (2020). Assisted migration across fixed seed zones detects adaptation lags in two major North American tree species. *Ecological Applications*, 30(5), e02092. <https://doi.org/10.1002/eap.2092>
- Fréjaville, T., Vizcaíno-Palomar, N., Fady, B., Kremer, A., & Benito Garzón, M. (2020). Range margin populations show high climate adaptation lags in European trees. *Global Change Biology*, 26(2), 484–495. <https://doi.org/10.1111/gcb.14881>
- Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3), 1–41. <https://doi.org/10.1214/07-AOAS148>
- Global Multi-resolution Terrain Elevation Data. (2010). (GMTED2010). <https://doi.org/10.5066/F7J38R2N>
- Graham, A. (1999). *Late Cretaceous and Cenozoic history of North American vegetation: North of Mexico*. Oxford University Press.
- Gray, L. K., & Hamann, A. (2013). Tracking suitable habitat for tree populations under climate change in western North America. *Climatic Change*, 117(1–2), 289–303. <https://doi.org/10.1007/s10584-012-0548-8>
- Guisan, A., Edwards, T. C., & Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: Setting the scene. *Ecological Modelling*, 157(2–3), 89–100. [https://doi.org/10.1016/S0304-3800\(02\)00204-1](https://doi.org/10.1016/S0304-3800(02)00204-1)
- Hijmans, R. J. (2022). *terra: Spatial data analysis*. R package version 1.5-21. <https://CRAN.R-project.org/package=terra>
- Hoek, W. Z. (2001). Vegetation response to the ~14.7 and ~11.5 ka cal. BP climate transitions: Is vegetation lagging climate? *Global and Planetary Change*, 30(1–2), 103–115. [https://doi.org/10.1016/S0921-8181\(01\)00081-9](https://doi.org/10.1016/S0921-8181(01)00081-9)
- Iverson, L., Peters, M., Prasad, A., & Matthews, S. (2019). Analysis of climate change impacts on tree species of the eastern US: Results of DISTRIB-II modeling. *Forests*, 10(4), 302. <https://doi.org/10.3390/f10040302>
- Iverson, L. R., Prasad, A. M., Matthews, S. N., & Peters, M. (2008). Estimating potential habitat for 134 eastern US tree species under six climate scenarios. *Forest Ecology and Management*, 254(3), 390–406. <https://doi.org/10.1016/j.foreco.2007.07.023>
- Iverson, L. R., & Prasad, A. M. (1998). Predicting abundance of 80 tree species following climate change in the eastern United States. *Ecological Monographs*, 68(4), 21.
- Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5), 1–26.
- Kuhn, M. (2022). *caret: Classification and regression training*. R package version 6.0-91. <https://CRAN.R-project.org/package=caret>
- Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.
- Kuhn, M., & Quinlan, R. (2023). *Cubist: Rule- and instance-based regression modeling*. <https://topepo.github.io/Cubist/>, <https://github.com/topepo/Cubist/>
- Laland, K., Matthews, B., & Feldman, M. W. (2016). An introduction to niche construction theory. *Evolutionary Ecology*, 30(2), 191–202. <https://doi.org/10.1007/s10682-016-9821-z>
- Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A review of machine learning interpretability methods. *Entropy*, 23(1), 18. <https://doi.org/10.3390/e23010018>

- Little, E. L. (1971). *Atlas of United States trees. Conifers and important hardwoods*. Miscellaneous Publication 1146, vol. 1. U.S. Department of Agriculture, Forest Service.
- Lucas, T. C. D. (2020). A translucent box: Interpretable machine learning in ecology. *Ecological Monographs*, 90(4), e01422. <https://doi.org/10.1002/ecm.1422>
- Mahony, C. R., Wang, T., Hamann, A., & Cannon, A. J. (2022). A global climate model ensemble for downscaled monthly climate normals over North America. *International Journal of Climatology*, 42(11), 5871–5891. <https://doi.org/10.1002/joc.7566>
- Mathys, A., Coops, N. C., & Waring, R. H. (2014). Soil water availability effects on the distribution of 20 tree species in western North America. *Forest Ecology and Management*, 313, 144–152. <https://doi.org/10.1016/j.foreco.2013.11.005>
- Matthews, S. N., Iverson, L. R., Prasad, A. M., Peters, M. P., & Rodewald, P. G. (2011). Modifying climate change habitat models using tree species-specific assessments of model uncertainty and life history-factors. *Forest Ecology and Management*, 262(8), 1460–1472. <https://doi.org/10.1016/j.foreco.2011.06.047>
- McKenney, D. W., Pedlar, J. H., Lawrence, K., Campbell, K., & Hutchinson, M. F. (2007). Potential impacts of climate change on the distribution of North American trees. *Bioscience*, 57(11), 939–948. <https://doi.org/10.1641/B571106>
- O'Neill, B. C., Tebaldi, C., van Vuuren, D. P., Eyring, V., Friedlingstein, P., Hurtt, G., Knutti, R., Kriegler, E., Lamarque, J.-F., Lowe, J., Meehl, G. A., Moss, R., Riahi, K., & Sanderson, B. M. (2016). The scenario model intercomparison project (ScenarioMIP) for CMIP6. *Geoscientific Model Development*, 9(9), 3461–3482. <https://doi.org/10.5194/gmd-9-3461-2016>
- Pederson, N., D'Amato, A. W., Dyer, J. M., Foster, D. R., Goldblum, D., Hart, J. L., Hessler, A. E., Iverson, L. R., Jackson, S. T., Martin-Benito, D., McCarthy, B. C., McEwan, R. W., Mladenoff, D. J., Parker, A. J., Shuman, B., & Williams, J. W. (2015). Climate remains an important driver of post-European vegetation change in the eastern United States. *Global Change Biology*, 21(6), 2105–2110. <https://doi.org/10.1111/gcb.12779>
- Peters, A., & Hothorn, T. (2021). *ipred: Improved predictors*. R package version 0.9-12. <https://CRAN.R-project.org/package=ipred>
- Peters, M. P., Iverson, L. R., Prasad, A. M., & Matthews, S. N. (2019). Utilizing the density of inventory samples to define a hybrid lattice for species distribution models: DISTRIB-II for 135 eastern U.S. trees. *Ecology and Evolution*, 9, 8876–8899. <https://doi.org/10.1002/ece3.5445>
- Prasad, A., Pedlar, J., Peters, M., McKenney, D., Iverson, L., Matthews, S., & Adams, B. (2020). Combining US and Canadian forest inventories to assess habitat suitability and migration potential of 25 tree species under climate change. *Diversity and Distributions*, 26, 1142–1159. <https://doi.org/10.1111/ddi.13078>
- Prasad, A. M. (2015). Macroscale intraspecific variation and environmental heterogeneity: Analysis of cold and warm zone abundance, mortality, and regeneration distributions of four eastern US tree species. *Ecology and Evolution*, 5(21), 5033–5048. <https://doi.org/10.1002/ece3.1752>
- Prasad, A. M. (2018). Machine learning for macroscale ecological niche modeling—A multi-model, multi-response ensemble technique for tree species management under climate change. In G. R. Humphries, D. R. Magness, & F. Huettmann (Eds.), *Machine learning for ecology and sustainable natural resource management*. Springer International Publishing. <https://doi.org/10.1007/978-3-319-96978-7>
- Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems*, 9(2), 181–199. <https://doi.org/10.1007/s10021-005-0054-1>
- R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org>
- Roberts, D. R., & Hamann, A. (2015). Glacial refugia and modern genetic diversity of 22 western North American tree species. *Proceedings of the Royal Society B: Biological Sciences*, 282(1804), 20142903. <https://doi.org/10.1098/rspb.2014.2903>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
- Ryo, M., Angelov, B., Mammola, S., Kass, J. M., Benito, B. M., & Hartig, F. (2021). Explainable artificial intelligence enhances the ecological interpretability of black-box species distribution models. *Ecography*, 44(2), 199–205. <https://doi.org/10.1111/ecog.05360>
- Schoonmaker, P. K., & Foster, D. R. (1991). Some implications of paleoecology for contemporary ecology. *Botanical Review*, 57(3), 204–245.
- Silander, J. A. (2001). Temperate forests. In *Encyclopedia of biodiversity* (pp. 607–626). Elsevier.
- Stanke, H., Finley, A. O., Domke, G. M., Weed, A. S., & MacFarlane, D. W. (2021). Over half of western United States' most abundant tree species in decline. *Nature Communications*, 12(1), 451. <https://doi.org/10.1038/s41467-020-20678-z>
- Svenning, J. C., & Sandel, B. (2013). Disequilibrium vegetation dynamics under future climate change. *American Journal of Botany*, 100(7), 1266–1286. <https://doi.org/10.3732/ajb.1200469>
- Therneau, T., & Atkinson, B. (2022). *rpart: Recursive partitioning and regression trees*. R package version 4.1.16. <https://CRAN.R-project.org/package=rpart>
- Toledo, M., Poorter, L., Peña-Claros, M., Alarcón, A., Balcázar, J., Leaño, C., Licona, J. C., Llanque, O., Vroomans, V., Zuidema, P., & Bongers, F. (2011). Climate is a stronger driver of tree and forest growth rates than soil and disturbance: Climate drives tree and forest growth. *Journal of Ecology*, 99(1), 254–264. <https://doi.org/10.1111/j.1365-2745.2010.01741.x>
- Verbyla, D. L. (1987). Classification trees: A new discrimination tool. *Canadian Journal of Forest Research*, 17, 1150–1152.
- Walsh, J. E. (2014). Intensified warming of the Arctic: Causes and impacts on middle latitudes. *Global and Planetary Change*, 117, 52–63. <https://doi.org/10.1016/j.gloplacha.2014.03.003>
- Wang, T., Hamann, A., Spittlehouse, D., & Carroll, C. (2016). Locally downscaled and spatially customizable climate data for historical and future periods for North America. *PLoS One*, 11(6), e0156720. <https://doi.org/10.1371/journal.pone.0156720>
- Williams, J. W., Shuman, B. N., Webb, T., Bartlein, P. J., & Leduc, P. L. (2004). Late-Quaternary vegetation dynamics in North America: Scaling from taxa to biomes. *Ecological Monographs*, 74(2), 309–334. <https://doi.org/10.1890/02-4045>
- Zhang, J., Nielsen, S. E., Chen, Y., Georges, D., Qin, Y., Wang, S.-S., Svenning, J.-C., & Thuiller, W. (2017). Extinction risk of North American seed plants elevated by climate and land-use change. *Journal of Applied Ecology*, 54(1), 303–312. <https://doi.org/10.1111/1365-2664.12701>

BIOSKETCH

Anantha Prasad's work involves designing ensemble and simulation models to map habitat distributions and colonization potentials of tree species under current and future climate conditions. His interests span eco-evolutionary aspects of macroscale tree species distributions and understanding within-species variations of populations to climate change.

Author contributions: Anantha Prasad conceived and designed the research, did data analysis and interpretation and wrote the manuscript. Matt Peters was responsible for figures and providing major support. Steve Matthews and Louis Iverson provided additional support to complete the manuscript. All authors contributed to and reviewed the drafts, and gave final approval for publication.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Prasad, A., Peters, M., Matthews, S., & Iverson, L. (2023). Unpacking the 'black box': Improving ecological interpretation of regression-based models. *Diversity and Distributions*, 29, 926–945. <https://doi.org/10.1111/ddi.13707>