New Phytologist

Research

Machine learning predicts large scale declines in native plant phylogenetic diversity

Daniel S. Park[1], Charles G. Willis[2], Zhenxiang Xi[3], John T. Kartesz[4], Charles C. Davis[1] and Steven Worthington[5]

[1]Department of Organismic and Evolutionary Biology and Harvard University Herbaria, Harvard University, Cambridge, MA 02138, USA; [2]Department of Biology Teaching and Learning, University of Minnesota, Minneapolis, MN 55108, USA; [3]Key Laboratory of Bio-Resource and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu 610065, China; [4]Biota of North America Program, 9319 Bracken Lane, Chapel Hill, NC 27516, USA; [5]Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138, USA

Authors for correspondence:
Daniel S. Park
Tel: +1 617 496 0515
Email: danielpark@fas.harvard.edu

Steven Worthington
Tel: +1 917 680 2608
Email: sworthington@iq.harvard.edu

Received: *6 September 2019*
Accepted: *12 April 2020*

New Phytologist (2020)
doi: 10.1111/nph.16621

Key words: artificial intelligence, biodiversity, climate change, machine learning, phylogenetic diversity, vascular plants.

Summary

- Though substantial effort has gone into predicting how global climate change will impact biodiversity patterns, the scarcity of taxon-specific information has hampered the efficacy of these endeavors. Further, most studies analyzing spatiotemporal patterns of biodiversity focus narrowly on species richness.
- We apply machine learning approaches to a comprehensive vascular plant database for the United States and generate predictive models of regional plant taxonomic and phylogenetic diversity in response to a wide range of environmental variables.
- We demonstrate differences in predicted patterns and potential drivers of native vs nonnative biodiversity. In particular, native phylogenetic diversity is likely to decrease over the next half century despite increases in species richness. We also identify that patterns of taxonomic diversity can be incongruent with those of phylogenetic diversity.
- The combination of macro-environmental factors that determine diversity likely varies at continental scales; thus, as climate change alters the combinations of these factors across the landscape, the collective effect on regional diversity will also vary. Our study represents one of the most comprehensive examinations of plant diversity patterns to date and demonstrates that our ability to predict future diversity may benefit tremendously from the application of machine learning.

Introduction

Climate change poses one of the greatest threats to biodiversity in the Anthropocene (Williams *et al.*, 2007; Tilman *et al.*, 2017; Vellend *et al.*, 2017). In a matter of decades, large portions of the globe and its inhabitants will experience climates not seen in the present or recent past (Parmesan, 2006). This will lead to local and regional species turnover and changes in species diversity via a combination of adaptation, dispersal, and extinction (Peterson *et al.*, 2002). Identifying effective conservation strategies depends on reliable, spatially explicit predictions of the effects of climate change on biodiversity (Mokany & Ferrier, 2011).

Though substantial efforts have been made to predict how biodiversity patterns will be altered in response to climate change (Bakkenes *et al.*, 2002; McKenney *et al.*, 2007; Lima-Ribeiro *et al.*, 2017), the majority of such studies focus on species richness, mostly ignoring higher taxonomic levels and phylogenetic relatedness despite increasing understanding of its functional importance (Vellend *et al.*, 2017; Daru *et al.*, 2019). Also, these efforts generally do not consider nonnative biodiversity separately, despite evidence that biodiversity maintenance mechanisms differ between native and nonnative dominated communities (Wilsey *et al.*, 2009).

The most commonly applied method for predicting patterns of diversity is to model individual species' habit preferences by linking species' presences (and absences) to environmental conditions, and forecast these species distribution models (SDMs) onto future climate scenarios (Elith & Leathwick, 2009; Zhang *et al.*, 2017). The inferred species ranges are then summed, or multiple models are combined to produce ensemble forecasts of overall diversity (Thuiller *et al.*, 2005; Algar *et al.*, 2009). This approach builds upon the hypothesis that species richness may indicate the sum of the effects on individual species' environmental tolerances (Boucher-Lalonde *et al.*, 2013). However, though recent advances in generating and mobilizing biodiversity data have improved our general knowledge of species' ranges at large scales (e.g. country or state level; Kartesz, 2015; Meineke *et al.*, 2019; Hedrick *et al.*, 2020), accurate fine-scale occurrence data necessary for SDMs (i.e. point coordinates) are still lacking for most species, and available data are affected by a wide range of gaps,

© 2020 The Authors
New Phytologist © 2020 New Phytologist Trust

biases and uncertainties (Meyer *et al.*, 2016; Park & Davis, 2017; Daru *et al.*, 2018). Furthermore, it has been suggested that individual species do not track the richness–climate relationship that accounts for regional variation in species diversity (Algar *et al.*, 2009; Boucher-Lalonde *et al.*, 2013). Finally, this approach does not account for the carrying capacity of the environment. It is hypothesized that the number of taxa that can tolerate the environmental conditions in any given location is generally much greater than the number that actually occur there (Cornell, 1985; Currie, 1991; Cornell & Karlson, 1996).

Here, for the first time to our knowledge, we employ machine learning – the practice of using algorithms to parse data, learn from it, and then make predictions – to predict the regional biodiversity of counties in the contiguous United States (hereafter referred to as the US) in response to a wide range of climatic, geographic, and edaphic variables, as opposed to individual species (Ferrier & Guisan, 2006; Sommer *et al.*, 2010). Models built using machine learning are able to incorporate complex, high-dimensional, correlated data, and account for nonlinear relationships, rendering them ideal for modeling complicated patterns of biodiversity (Olden *et al.*, 2008; Kelling *et al.*, 2009). The 'top-down' modeling approach we employ can be applied in situations where insufficient data are available for modeling the distributions of individual species (Mokany *et al.*, 2010; Mokany & Ferrier, 2011), and builds on the theory that there exists a direct link between species richness and climate, which imposes limits on overall richness, regardless of individual species identities (Algar *et al.*, 2009; Boucher-Lalonde *et al.*, 2013). Indeed, strong climate–richness relationships have been identified in a number of studies across a wide range of taxa, including plants, invertebrates, and vertebrates (H-Acevedo & Currie, 2003; Hawkins *et al.*, 2003, 2011; Kreft & Jetz, 2007; Park & Razafindratsima, 2018).

We train models of total, native, and nonnative (introduced) plant diversity on an exceptionally robust and well-curated dataset of the US vascular flora. We then project these models into the near future under seven climate change scenarios to determine how spatial patterns of biodiversity may shift over time. Though this approach assumes that the current diversity of taxa is at or near carrying capacity, and that the processes that generated and maintain this diversity can respond over the timespan of the prediction interval, it can generate baseline estimates of how the diversity in an area may change in the future. The ecological and evolutionary responses of individual taxa will ultimately determine whether and when predictions are met, but such models can provide useful benchmarks for immediate climate change mitigation and biodiversity management. In addition to metrics of taxonomic richness, we assess metrics of phylogenetic diversity, which take into account the shared evolutionary history of species in a region. It has been suggested that environmental variation can affect the phylogenetic diversity and structure of communities (Kerkhoff *et al.*, 2014; Kamilar *et al.*, 2015; Park & Razafindratsima, 2018), and plant phylogenetic diversity in particular, has been linked to ecosystem productivity (Cadotte *et al.*, 2008; Flynn *et al.*, 2011; Srivastava *et al.*, 2012), ecosystem stability (Cadotte *et al.*, 2012), and animal diversity (Dinnage *et al.*, 2012; Park & Razafindratsima, 2018). Despite its importance for conservation, however, relatively few studies have addressed the impact of climate change on phylogenetic diversity on large scales (Zhang *et al.*, 2015, 2017; González-Orozco *et al.*, 2016).

Machine learning models employ complex and opaque algorithms that often render it difficult to ascertain the effects of individual predictor variables and their importance. Indeed, there is no general consensus on the best way to compute – or even define – variable importance in such predictive models (Grömping, 2009). Therefore, we additionally explored the effects of a subset of environmental variables that have been hypothesized to drive plant diversity using traditional modeling methods (Holdridge, 1947; Parker, 1963; Stephenson, 1990; Pigott & Pigott, 1993; Francis & Currie, 2003; Venevsky & Veneskaia, 2003). Our combined analyses represent one of the most comprehensive examinations of plant diversity patterns in the US and highlight significant differences among patterns and drivers of native and nonnative plant diversity.

Materials and Methods

Species occurrence data

Species richness and nativity data on vascular plants were derived from the Biota of North America Program's (BONAP; http://www.bonap.org/) North American Plant Atlas (NAPA; Kartesz, 2015), representing 19 039 taxa from 227 families. The dataset is available as presence/absence data for 3067 counties in the US, excluding Alaska and Hawaii. BONAP's NAPA represents the first comprehensive attempt to provide state- and county-level distribution maps of all vascular plant taxa in the US, and integrates county records, derived from herbaria, museums and other plant repositories, coupled with monographic and revisionary literature, and other selected bibliographic references into arguably the most complete floristic treatment of a large region. The vast majority of the nearly 6000 000 county records of the BONAP's database are verified by taxonomic and floristic specialists. Nativity status is derived from historical floristic accounts, taxonomic literature, and plant repository vouchers from multiple institutions across North America. Though counties and their equivalents are not standard area units, they often represent finer geographic and climatic units than those used in many similarly large-scale studies (e.g. 1° cells), and mean county climate has been shown to be a reasonable proxy for point climate when point occurrence data are not available (Park & Davis, 2017).

Biodiversity assessment

Our phylogenetic dataset for the North American flora was assembled using the program Phlawd (Smith *et al.*, 2009). We harvested sequence data from GenBank Release 205.0 based on our entire species list, targeting 12 commonly used molecular loci (plastid: *atpB*, *atpB-rbcL*, *matK*, *ndhF*, *rbcL*, *rps4* and *trnL-trnF*; mitochondrial: *atp1*, *atpA*, *matR* and *rps3*; nuclear: ITS). Species names were cross-checked against potential synonyms listed in

© 2020 The Authors
New Phytologist © 2020 New Phytologist Trust

GenBank. DNA sequences of each locus were aligned separately using MAFFT v.7.220 (Katoh & Standley, 2013) and then concatenated together using PHYUTILITY v.2.4 (Smith & Dunn, 2008). We were able to retrieve DNA sequences for 10 147 species, and the final concatenated matrix contained 23 022 sites (percentage of gaps and missing data: 88.7%). Maximum likelihood (ML) phylogenies including 100 bootstrap replicates with replacement were constructed using EXAML v.3.0.1 (Kozlov *et al.*, 2015) with a general time reversible (GTR) + Γ model specified for each locus. These phylogenies were then dated using the penalized likelihood approach as implemented in TREEPL v.3.26.2013 (Smith & O'Meara, 2012). The smoothing parameter was determined using the random subsample and replicate cross-validation (RSRCV) approach. Thirty-three fossils described in detail by Bell *et al.* (2010) and five age constraints used by Jiao *et al.* (2011) were adopted as calibration points. The resulting trees are available in a Zenodo repository (Park *et al.*, 2020). Taxonomic richness was calculated as the sum of taxa at each level (i.e. species, genera, and families) occurring in each county and region based on the presence–absence matrices compiled as described in the Species occurrence data section above. Phylogenetic diversity (PD), mean phylogenetic distance (MPD) and their standardized effect sizes (PD_S, MPD_S) were calculated using the package PHYLOMEASURES v.2.1 (Tsirogiannis & Sandel, 2016) in R v.3.4.1 (R Core Team, 2017). Standardized effect sizes account for effects of species richness and are calculated as: (observed value − expected value)/standard deviation of the expected value. Expected values of PD and MPD were calculated from a null distribution of 1000 random assemblages of species drawn without replacement from the species pool of US taxa. Therefore, positive values of PD_S and MPD_S indicate phylogenetic overdispersion, whereas negative values indicate clustering, relative to random assemblages of taxa. These metrics were calculated for the following: all taxa regardless of native status (^{T}), native taxa (^{N}), and nonnative (introduced) taxa (^{I}), across all phylogenies. Molecular data for vascular plant species is still lacking; thus, our phylogenies do not include all taxa present in the US. However, our phylogenies represent one of the most comprehensive phylogenetic reconstructions of the North American flora to date, and the standard error of each metric derived from the phylogenetic bootstrap replicates are presented in Supporting Information Fig. S1. Our downstream modeling efforts account for differences in the proportion of taxa in each county represented on the phylogeny.

Environmental data

Environmental data comprised climatic, edaphic, and geographic data collected at 2.5 arc-minute resolution. Climatic data included the 19 bioclimatic variables available in the WorldClim database v.1.4 (Hijmans *et al.*, 2005). Elevation data for each county was derived from the USGS GMTED2010 dataset (Danielson & Gesch, 2011). Edaphic data, including fraction soil clay content, fraction soil gravel content, fraction soil sand content, percentage organic content in soil, soil pH, soil salinity, fraction soil silt content, cation exchange capacity (CEC), and $CaSO_4$ concentration were derived from the Harmonized World Soil Database v.1.21 (Fischer *et al.*, 2008). For predictive models, we included the mean, minimum, maximum, and standard deviation of all bioclimatic variables, soil variables, and elevation at the county level. Geographic variables included county area, presence of coast, and glaciation status during the last glacial maximum inferred from the United States Geological Survey database (Haj *et al.*, 2018).

To examine future patterns of plant biodiversity, predictive models were projected onto Hadley Centre Global Environment Model v.2 (HadGEM2-ES (HE); Collins *et al.*, 2011) climate predictions for 2050 and 2070 across four representative concentration pathways (RCPs; 2.6, 4.5, 6.0 and 8.5) as used in the Fifth Assessment Intergovernmental Panel on Climate Change report (IPCC, 2014). Among these, RCP 4.5 reflects a somewhat optimistic scenario where the goals of the Paris Climate Agreement are assumed to be met. Thus, to focus on climate change scenarios based upon RCP 4.5 in more depth, we incorporated future climate predictions from the following additional Coupled Model Intercomparison Project Phase 5 (CMIP5) models: ACCESS1.0 (AC), GFDL-ESM2G (GD) and GISS-E2-R (GS). These projections were also derived from the WorldClim database (Hijmans *et al.*, 2005).

Statistical methods

Machine learning can either be 'supervised', where a response variable (e.g. species richness) is observed and therefore some ground-truth is known about the relationship between predictors and response, or 'unsupervised', where no response variable exists. Supervised learning, as employed here, represents the gold-standard for producing accurate out-of-sample predictions of response variables and allows us to consider an unprecedented number of environmental variables potentially linked to biodiversity patterns. Compared to more traditional SDMs, machine learning SDMs typically have higher predictive performance and greater flexibility to incorporate complex nonlinear effects, interaction effects, and noisy high dimensional data. Our analysis goals were twofold: to predict future biodiversity, in terms of taxonomic richness and phylogenetic diversity, under various climatic scenarios; and to identify explanators of contemporary biodiversity. To achieve these goals, we built two different statistical models: predictive and explanatory.

For predictive models, we first cleaned the data by imputing missing predictor data (< 0.5% of observations for 6 out of 125 variables) using bagging (bootstrap aggregation: a method where models are trained on bootstrap resamples of the original data and then averaged to obtain results with less variance than individual models) and transforming some response variables to natural logarithms if this improved predictive accuracy. For model selection and validation, we partitioned data into 80% (training) and 20% (test) subsets. Model specific parameters were tuned by fitting models over a grid of parameter values and selecting the combination of values that minimized predictive error. To select the best performing model out of 13 candidate models (which

© 2020 The Authors
New Phytologist © 2020 New Phytologist Trust

were selected as a representative sample of different modeling strategies/algorithms from a larger set of 80+ machine learning models for continuous outcomes available in the R language), while preventing over-fitting, we used *k*-fold cross validation (CV; 10-folds, with 10 repeats) on the training data. The CV was nonspatial, because the out-of-sample predictions generated from the models are for new temporal units, but the same spatial units as the training data. For each response variable and combination of model specific parameter values, CV involved partitioning the training data into *k*-folds, then iteratively fitting each candidate model *k* times to $k-1$ folds of the data, until all folds had been excluded from model estimation (Fig. S2). Performance was assessed by predicting response values for the excluded out-of-sample data folds and comparing them to observed response values from the same folds. The *k*-fold partitioning step was repeated 10 times to provide 100 estimates of performance for each candidate model, response variable, and set of model specific parameters (Fig. 1). Model predictive accuracy was compared using the coefficient of determination (R^2) and root mean squared error (RMSE). The model with the highest average R^2 and lowest average RMSE was selected as the best predictive model, which was in this case a Cubist regression tree model. For this final model, we performed external validation using the test dataset and report R^2 and RMSE as out-of-sample predictive accuracy measures (Table S1). Predictive models were fitted in R v.3.4.1 (R Core Team, 2017) using the package CARET v.6.0-77 (Khun, 2018).

To explore the effects of specific environmental factors, we examined the relationship between contemporary taxonomic richness/diversity patterns and a subset of variables used in the predictive models using linear mixed effects models (explanatory models). These models shared a common specification. To account for state-level heterogeneity in the reporting of data, random intercepts were grouped by state. To ameliorate the confounding effects of spatial autocorrelation, we included a residual autocovariate (RAC) term. The spatial range and functional form (linear inverse distance or quadratic inverse distance) of autocorrelation differed for each outcome and was determined by optimizing these parameters on a variogram. To facilitate comparisons of effect size magnitude, all focal explanatory variables were standardized so that a one-unit change was equivalent to a change of one standard deviation. Response variables were not standardized, but some were transformed to the natural logarithm scale when this improved residual diagnostics. Models included explanatory climatic, geographic, and edaphic variables that are representative of major, independent axes of environmental variation across the US and are thus hypothesized to influence taxonomic richness/phylogenetic diversity. These variables included: mean annual temperature (BIO1), annual temperature range (BIO7), mean temperature of the wettest quarter (BIO8), annual precipitation (BIO12), precipitation seasonality (BIO15), mean elevation, standard deviation of elevation, fraction soil clay content, fraction soil gravel content, fraction soil sand content, percentage organic content in soil, soil pH, soil

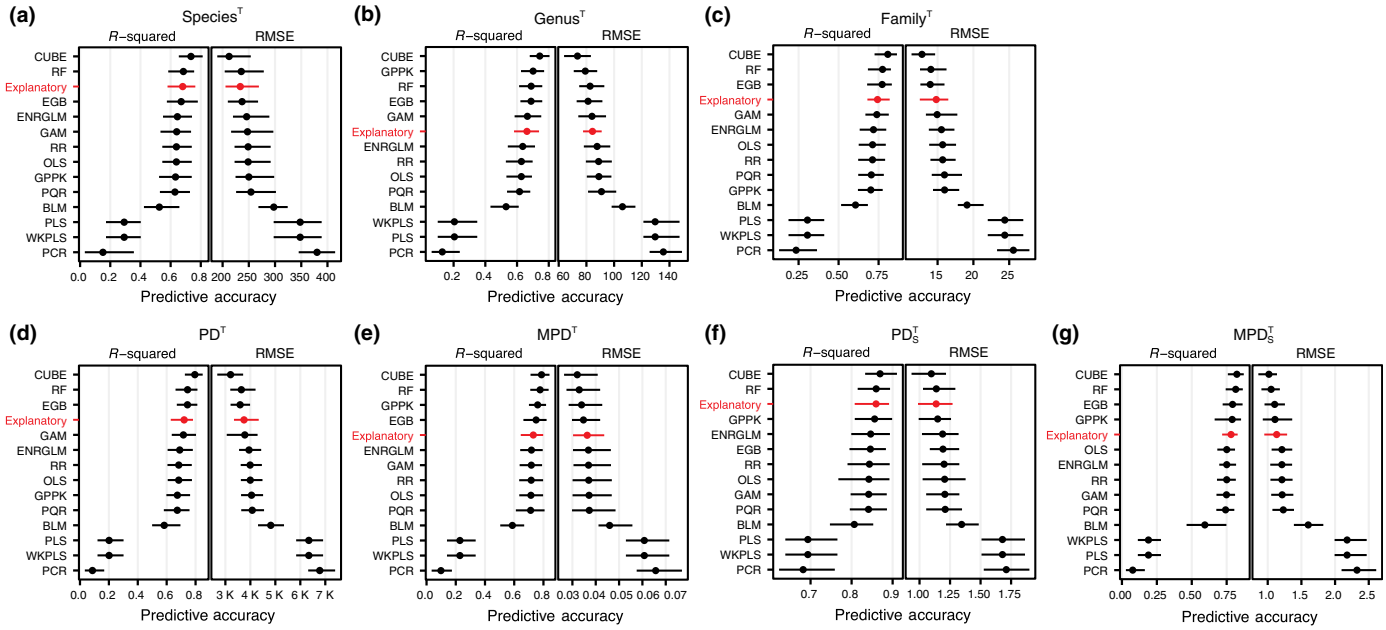

Fig. 1 Model selection for seven total taxonomic and phylogenetic diversity responses using cross-validation (10-fold, 10-repeats) on the training sets (80%). Diversity responses include total (a) species richness (Species^T), (b) genus richness (Genus^T), (c) family richness (Family^T), (d) phylogenetic diversity, (e) mean phylogenetic diversity, (f) standardized effect size of phylogenetic diversity (PD_S^T) and (g) standardized effect size of mean phylogenetic diversity (MPD_S^T). Thirteen predictive models, plus one explanatory model, are shown, ranked by R^2 and RMSE. Points are averages, while error bars denote minima and maxima, over the folds and repeats. BLM, boosted linear model; CUBE, Cubist model; EGB, extreme gradient boosting; ENRGLM, elastic-net regularized GLM; explanatory, explanatory linear mixed effects model; GAM, generalized additive model; GPPK, Gaussian process with polynomial kernel; OLS, general linear model; PCR, principal component regression; PLS, partial least squares; PQR, penalized quantile regression; RF, random forest; RR, ridge regression; WKPLS, wide kernel partial least squares.

© 2020 The Authors
New Phytologist © 2020 New Phytologist Trust

salinity, county area, presence of coast and glaciation history. Both point and interval (95% confidence) estimates are reported. Explanatory models were fitted in R v.3.4.1 (R Core Team, 2017) using the package LME4 v.1.1-14 (Bates *et al.*, 2014). Replication code and data are available in a Zenodo repository (Park *et al.*, 2020).

Results

Predictive model results

Of 13 candidate models, we determined that a Cubist regression tree model yielded the most accurate predictions for all response variables, including species, genus, and family richness, PD, MPD, PD_S and MPD_S (Fig. 1). Cubist models are rule-based and fit separate linear models at each node of a decision tree. An ensemble procedure is used to combine many decision trees into one omnibus model. As with other ensemble methods (e.g. random forest, stochastic gradient boosting), combining multiple trees improves model stability and performance. Our Cubist models yielded an out-of-sample predictive accuracy between 88% and 70% (Table S1). Predicted patterns of biodiversity were highly correlated across all general circulation models (GCMs) and emission scenarios, differing only in severity, with more severe emission scenarios eliciting larger predicted changes (Table S2; Figs S3, S4). Therefore, we present predictions based on RCP4.5, which reflects a scenario in which the goals of the Paris Climate Agreement are met. As the predicted responses of several diversity metrics were correlated (Fig. S5), we focus on our results for species richness and MPD below.

Predicted changes to taxonomic diversity were highly variable across counties (Figs 2, S6, S7). Overall, taxonomic diversity was predicted to increase in the majority of US counties, but less so in desert areas of the southwest. In particular, average gains in species richness were 9% by 2050 and 14% by 2070, but overall increases in nonnative taxa were predicted to be much greater than native taxa. Native MPD (MPD^N) was expected to decrease in over 80% of the counties examined. By contrast, nonnative MPD (MPD^I) was predicted to increase overall. Along these lines, standardized native MPD (MPD_S^N) was predicted to decrease by 81% by 2070, on average, while standardized nonnative MPD (MPD_S^I) was predicted to decrease by only 11% (Fig. S6). In general, predicted patterns of total biodiversity mirrored those predicted for native taxa (Figs S3, S7).

Contrasting biogeographical patterns emerged across different aspects of plant diversity when predicted changes in plant diversity were examined across ecoregions following their current-day extents (Figs 3, S8, S9). In general, phylogenetic diversity loss was greater in ecoregions east of the Rocky Mountains (Great Plains (GP), eastern temperate forests (ETF), tropical wet forests (TWF), northern forests (NF)), whereas taxonomic gains tended to be greatest in northern forest ecoregions (marine west coast forest (MWCF), northwestern forested mountains (NFM), NF). Northern forests were predicted to gain the most diversity in terms of species richness, but not MPD. On the other hand, southern semi-arid highlands (SSH) were predicted to lose native taxonomic diversity while becoming more phylogenetically diverse, especially in terms of MPD^I. Along these lines, most regions were predicted to lose native phylogenetic diversity while gaining nonnative phylogenetic diversity. For instance, TWF were predicted to lose the most phylogenetic diversity on average, in terms of both total MPD and MPD^N. However, this region was simultaneously predicted to gain the most diversity in terms of MPD^I. When native and nonnative plant diversity were modeled and predicted together (total diversity), the resulting biogeographical patterns mirrored those of predicted native diversity in general (Fig. S8).

Explanatory model results

As with our future predictions, current patterns of plant diversity vary greatly across space, metric, and native status (Fig. S10). The climatic factors that consistently explained taxonomic diversity were annual temperature range (BIO7) and precipitation seasonality (BIO15) (Tables S3–S5; Figs S11, S12). Species richness was lower in areas with greater seasonal variation in temperature and precipitation, and the magnitude of these climatic effects was greater on native diversity than nonnative diversity (Fig. 4; Tables S4, S5). On average, with a one SD unit increase in annual temperature variation, native and nonnative species richness decreases by 12.0% and 8.2%, respectively. Similarly, for a one SD unit increase in annual precipitation variation, native and nonnative species decline by 12.1% and 2.5%, respectively. Similarly, the effects of edaphic variables and presence of coast were directionally consistent across native and nonnative species richness, but of greater magnitude in the case of native diversity. Of these geophysical factors, proximity to coast had the greatest impact on diversity, with coastal counties having significantly higher relative diversity than inland counties.

As with species richness, both native and nonnative phylogenetic diversity was lower in areas with greater seasonal variation in temperature (BIO7) and precipitation (BIO15), and those without a coast (Figs 4, S12, S13). However, some environmental associations differed significantly between native and nonnative phylogenetic diversity (Tables S4, S5). For instance, MPD^N decreased 5.7% per unit increase in mean annual temperature (BIO1), while MPD^I increased by 3.0%. Higher mean elevation was associated with higher MPD^N but did not have a significant effect on MPD^I. Other geophysical factors had undetectable ($P > 0.05$) and/or weak (< 1%) associations with native and nonnative phylogenetic diversity. Similar results were found when examining the standardized versions of these metrics (Figs S12, S13).

Discussion

Mitigating the effects of climate change on Earth's biodiversity requires the means to accurately predict future biodiversity change and understand factors that influence its distribution and maintenance. Harnessing the power of machine learning, we generated large-scale predictive models of multiple facets of native and nonnative biodiversity, using climatic, geographic, and

© 2020 The Authors
New Phytologist © 2020 New Phytologist Trust

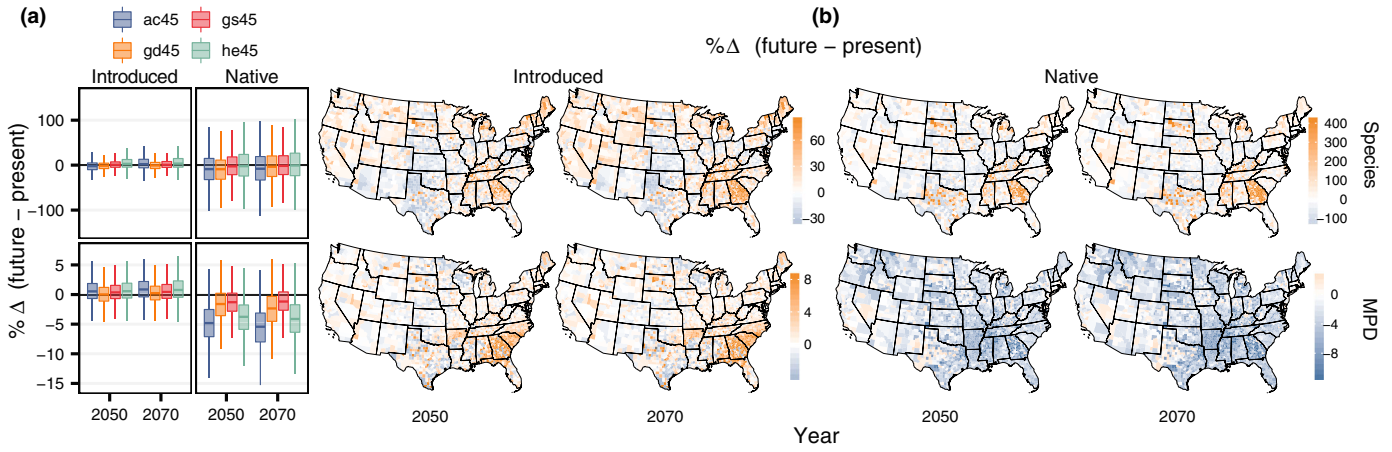

Fig. 2 Predicted changes in native vs nonnative introduced species richness and mean phylogenetic diversity in 2050 and 2070 relative to current values under RCP4.5. (a) Boxplots depict the percentage difference between predicted future values for 2050/2070 and present day observed values, for each of four different general circulation models (ACCESS1-0 (ac), GISS-E2-R (gs), GFDL-ESM2G (gd) and HadGEM2-ES (he)) and two response variables (species, MPD). (b) Choropleth maps illustrate the average percentage difference over four different general circulation models (ACCESS1-0 (ac), GISS-E2-R (gs), GFDL-ESM2G (gd), and HadGEM2-ES (he)) between predicted future values for 2050 and 2070 and present day observed values, for each of two response variables (species, MPD). The color gradient indicates increases (orange) and decreases (blue) in each metric.

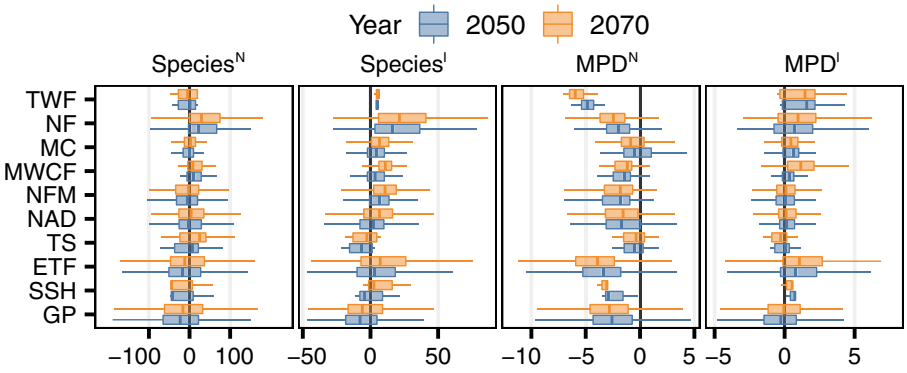

Fig. 3 Predicted changes in native vs nonnative introduced plant diversity and community structure in 2050 and 2070 for RCP4.5 by political divisions and ecoregions (https://www.epa.gov/eco-research/ecoregions-north-america). Boxplots depict the average percentage difference over four different general circulation models (ACCESS1-0 (ac), GISS-E2-R (gs), GFDL-ESM2G (gd) and HadGEM2-ES (he)) between predicted future values for 2050/2070 and present day observed values, for each of two response variables (species, MPD), grouped by eco-regions. ETF refers to eastern temperate forests; GP, Great Plains; MC, Mediterranean California; MWCF, marine west coast forest; NAD, North American deserts; NF, northern forests; NFM, northwestern forested mountains; SSH, southern semiarid highlands; TS, temperate sierras; and TWF, tropical wet forests. Native status is denoted in superscript for each diversity metric, where ^{N} indicates native, and ^{I} nonnative introduced.

edaphic data. We also identify key climate and geophysical factors that may influence patterns of biodiversity and demonstrate that the importance of these factors differed between native and nonnative taxa. Our models also took into account environmental heterogeneity, which despite being known to be an important driver of biodiversity (Ricklefs, 1977, 2004) is frequently ignored (Cramer & Verboom, 2017). Thus, the modeling approaches applied here provide a comprehensive examination of potential changes in plant biodiversity. Ultimately, the delay between environmental changes and colonization events (colonization lags), speciation events (speciation rate), and extirpations (extinction debt) will determine whether our projections are met within the modeled timeframe. As it may take longer amounts of time for plant diversity and the composition of biological communities to adjust to changing environments (Hector *et al.*, 1999), our projections can be considered estimates of the steepness of the gradient across which plant diversity is predicted to change over time (Currie, 2001). Thus, these projections can serve as a baseline for assessing and managing the future distribution of plant diversity in the face of climate change.

Disparate responses to climate and homogenization of diversity

Patterns of predicted change varied across the different facets of biodiversity. While considerable variation existed between counties, taxonomic diversity was predicted to increase on average in the US. Predicted changes among metrics of taxonomic diversity were positively correlated regardless of nativity (Fig. S5). These projections support previous studies predicting an overall increase

© 2020 The Authors
New Phytologist © 2020 New Phytologist Trust

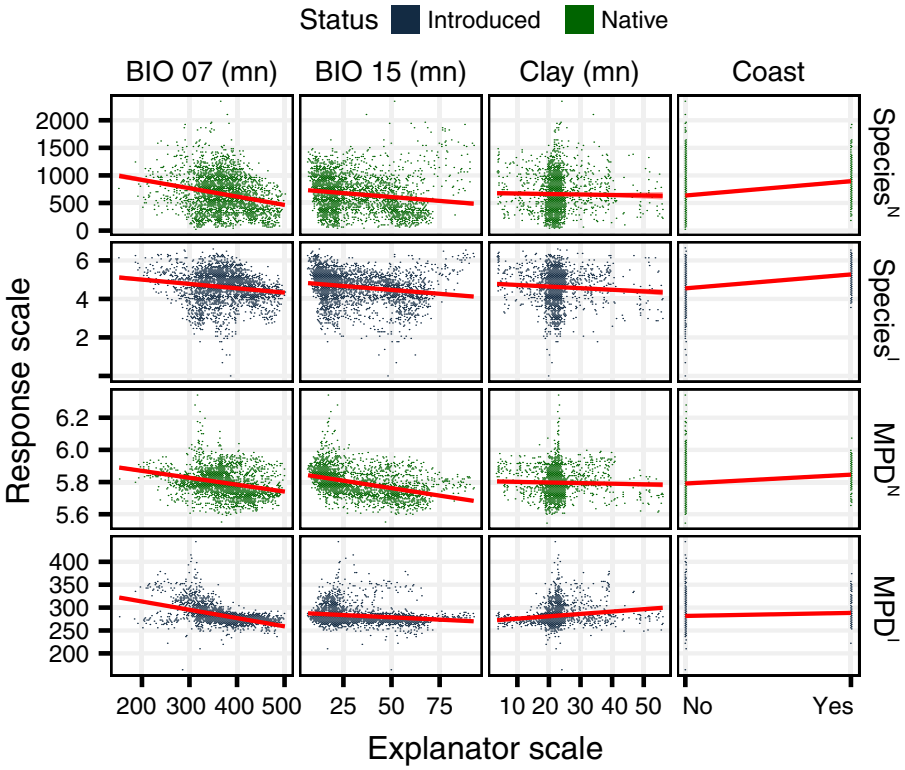

Fig. 4 Present day observed relationships between four climatic explanator variables (BIO 07 (mean), BIO 15 (mean), clay (mean) and coast (binary)) and two response variables (species, mean phylogenetic distance (MPD)).

of plant species richness in the US (Currie, 2001; Iverson & Prasad, 2001; Sommer *et al.*, 2010). However, increases in family diversity were predicted to be much smaller than those for species and genus, and under certain emission scenarios, family diversity was predicted to decrease slightly on average in 2050. This suggests that increases in plant biodiversity are likely to occur at lower taxonomic levels, or shallower phylogenetic nodes. Along these lines, our models predicted an overall decrease in phylogenetic diversity, especially in terms of native MPD. On the other hand, PD is inherently correlated with species richness, as greater numbers of species almost always correspond to greater summed branch lengths on a phylogeny (Venail *et al.*, 2015). While our predictions reflect this relationship, percent increases in PD were much smaller compared to those of species and genera, again indicating that increases in diversity may primarily occur within shallower nodes. This is also supported by the fact that phylogenetic clustering (MPD_S) was predicted to increase across the vast majority of the counties examined. This is especially alarming, as the loss of phylogenetic diversity is the loss of biodiversity *per se*, and may affect ecosystem function and stability negatively (Cadotte *et al.*, 2008, 2012; Staab *et al.*, 2016; Knapp *et al.*, 2017; Park & Razafindratsima, 2018). Furthermore, predicted changes in plant diversity were negatively correlated with the current amount of diversity present across all metrics examined (Table S2). Larger increases were predicted in counties with lower levels of diversity, whereas smaller increases and larger losses were predicted for counties with higher levels of standing diversity. We find that this has a homogenizing effect, leading to an overall decrease in the variation of plant diversity across the landscape, with few exceptions (Table S6). While it is possible that this pattern could be influenced by sampling bias, where counties with conditions conducive to high levels of plant diversity have been subject to undersampling, similar results have previously been reported (Sommer *et al.*, 2010). Together, our results suggest that while more counties will gain an increased capacity for taxonomic diversity, this gain will mostly support the proliferation of closely related native or nonnative species and relatively few lineages.

At the ecoregion scale, gains in taxonomic richness were greatest in northern forest ecoregions (MWCF, NFM, NF). In these regions, numerous taxa, including nonnative invasive species, have been limited by (seasonal) extreme cold and ice cover, and shorter growth periods (Sakai & Weiser, 1973; Grodowitz *et al.*, 1991; Owens & Madsen, 1995; Ayres & Lombardero, 2000; Owens *et al.*, 2004). However, the northeastern and northwestern US are experiencing disproportionately high amounts of climate change (Wuebbles *et al.*, 2017). The resulting relaxation of such thermal constraints is likely to increase taxonomic diversity (Sommer *et al.*, 2010). Loss of phylogenetic diversity was generally predicted to be greater in ecoregions east of the Rocky Mountains (GP, ETF, TWF, NF), suggesting that changes in climate may select for certain evolutionary lineages in these regions. In particular, TWF were predicted to lose the most phylogenetic diversity on average, suggesting that warm-adapted tropical lineages in these areas may be at greater risk. These patterns highlight how changes in different facets of biodiversity are not necessarily linked, and that regional capacities for biodiversity

© 2020 The Authors
New Phytologist © 2020 New Phytologist Trust

may shift in unexpected ways. Indeed, regional changes in climate, which can be highly spatially heterogeneous, are more relevant in the context of ecological response to climatic change, compared to global or continental trends (Walther *et al.*, 2002). Further, although many studies have predicted that species in the US will have to migrate northward with global warming (Morse *et al.*, 1993; Iverson & Prasad, 1998; McKenney *et al.*, 2007; Zhang *et al.*, 2017), our results suggest that that northern regions will not necessarily increase in their taxonomic and phylogenetic carrying capacity. Thus, mitigating the effects of climate change will require region-specific strategies, as well as approaches specific to the biodiversity facet (e.g. species richness, phylogenetic diversity) of management interest.

Decreases in native diversity coupled with increases in nonnative diversity

Compared to their native counterparts, many nonnative species have broad climatic tolerances and large geographic ranges, short generation times, rapid growth, high fecundity, strong dispersal ability, and independence from (specific) mutualists, all of which may affect their responses to climate change (Pyšek *et al.*, 1995; Rejmánek & Richardson, 1996; Goodwin *et al.*, 1999; Qian & Ricklefs, 2006; Bradley *et al.*, 2010; Park & Potter, 2015). Thus, we might expect that climate-change responses of nonnative invaders can differ from native taxa. Indeed, we identified that the environmental drivers of biodiversity can differ among native and nonnative taxa, and stark contrasts were observed in future predictions. On average, overall increases were predicted across all metrics of nonnative biodiversity in the US, under all examined climate change scenarios. On the other hand, native species richness and evolutionary diversity (PD and MPD) were predicted to decrease on average, with counties becoming more phylogenetically clustered (i.e. decreased MPD_S). At the regional scale, nonnative species introductions have far outweighed native extinctions, especially in well-surveyed temperate zones such as the US (Vellend *et al.*, 2017). This suggests that overall changes in plant diversity in the US could be disproportionately driven by increases in nonnative taxa, possibly at the expense of native taxa (Knapp *et al.*, 2017). For instance, TWF were predicted to simultaneously experience large losses in MPD^N and large gains in MPD^I, suggesting the environment will shift to favor nonnative taxa. As our results suggest, this may increase the homogenization of biodiversity throughout the country, where not only do differences in cumulative biodiversity become smaller across regions, but the phylogenetic diversity within regions is reduced as well. As assemblages of species in ecological communities reflect interactions among, as well as between organisms and their abiotic environments (Walther *et al.*, 2002), such changes in the composition of plant communities can alter ecosystem properties in ways that feed back into other components of global change (Dukes & Mooney, 1999). Similarly, previous studies have suggested that climate change is likely to favor nonnative species (Dukes & Mooney, 1999; Prentice *et al.*, 2007; Thuiller *et al.*, 2008; but see Bezeng *et al.*, 2017) and native endemic taxa may be especially vulnerable as many have evolved long-term under relatively stable climatic conditions (Jansson, 2003; Linder, 2008). Along these lines, current patterns of native biodiversity show strong negative associations with seasonality in temperature and precipitation.

Our predictive modeling results generate insights into the capacity of an area to support a certain amount of plant diversity given specific environmental conditions. These models assume that most communities are currently at capacity, which is not likely to be the case. Many communities have accommodated the establishment of exotics over the last century without substantial losses of native species, resulting in a net increase in diversity (Stohlgren *et al.*, 2003). Thus, our predictions provide a baseline, conservative estimate of future native and nonnative plant diversity, but may underestimate true regional carrying capacity.

Drivers of native and nonnative diversity

Models built using machine learning often employ complex and opaque algorithms that render the internal components of the models something of a 'black box'. To address this knowledge gap, we analyzed current patterns of diversity using more traditional approaches focusing on key climatic and geophysical attributes. Broadly, patterns of taxonomic and phylogenetic diversity were associated more strongly with seasonal variation in climate than 'favorableness' indices (e.g. mean annual temperature or total annual precipitation). Seasonal variation in precipitation and temperature accounted for 7.8–16.3% of the variation in plant diversity between counties, with less variable counties being more diverse. This trend likely reflects, in part, the seasonal impacts of winter in northern areas. However, the impact of seasonality cannot be dismissed entirely as a byproduct of winter extremes *per se*, as mean annual temperature, temperature range, and latitude were all controlled for in our model. Rather, seasonality itself is known to act as a filter on plant diversity in North America (Swenson *et al.*, 2012), because it requires unique physiological adaptations that are not present in all plant lineages (Kreft & Jetz, 2007).

Both taxonomic and phylogenetic diversity tended to decrease with greater elevation. However, the association with elevation was relatively weak when compared to other geophysical factors, including soil clay content, soil pH, and variation in elevation. Of particular note was the positive association of taxonomic diversity with more acidic and clay-rich soils – two factors that are not generally considered favorable to plant growth. These associations are likely the result of averaging soil characteristics at the county level, which would mask edaphic heterogeneity. Environmental heterogeneity can drive increased alpha diversity, driven by turnover between microenvironment (beta diversity; Ricklefs, 1977), even at small geographic scales (Willis *et al.*, 2010). We observe this process with the positive relationship between standard deviation in elevation per county and diversity. Unfortunately, we were not able to test this same hypothesis with soil heterogeneity because we did not have similar data on within-county variation.

Patterns of native and nonnative diversity were associated with different combinations of climate and geophysical variables,

© 2020 The Authors
New Phytologist © 2020 New Phytologist Trust

suggesting that climate change will likely impact the regional capacity of native vs nonnative diversity differently. In particular, geophysical factors including soil pH and clay content tended to have a greater effect on native diversity. Though climate is changing rapidly, geophysical factors are relatively fixed and not likely to change significantly over the timescales that we examine here. This may result in fewer areas with both suitable climates and the geophysical conditions that native species have evolved to require or prefer. The differences in the drivers of native vs nonnative diversity may also reflect the possibility that nonnative species may not have yet reached equilibria with their new environmental conditions as they have only existed on the landscape for a relatively short amount of time. Along these lines, our models do not implicitly take into account the long-term evolutionary processes that have influenced current patterns of biodiversity. Biodiversity in a region is the result of both shorter-term ecological processes such as environmental filtering and longer-term evolutionary processes that have generated the diversity of species on which filtering acts. Evolutionary and biogeographic history, including past diversification processes and environmental change, may have influenced the distributions of lineages and populations, creating deep-time legacy effects that influence the patterns of diversity we observe today (Cavender-Bares *et al.*, 2018). This is especially likely to be true for native diversity that has evolved *in situ*. Thus, serious consideration of a historical perspective is needed, and should improve our understanding of evolutionary and geographic mechanisms that link patterns of biodiversity across spatial scales (Qian & Ricklefs, 2004).

Machine learning biodiversity

When we projected the explanatory models we generated onto current environmental conditions, we identified that these more traditional models performed relatively well, but were not as effective as the best machine learning models (Fig. 1). The predictive improvements gained by using more complex, machine learning models are two-fold. First, these models are likely using macroecological factors not included in our explanatory model that might have a small but significant effect on predictive accuracy. Second, and more importantly, the improved accuracy of our predictive models most likely reflects the fact that they apply different combinations of macroecological factors when predicting diversity in different regions of the US. Though climatic factors will shift across the landscape as a result of climate change, geophysical factors will remain largely consistent. New combinations of macro-environments are likely to be created, resulting in novel assemblages of native and nonnative species. The advantage of the using a top-down, machine learning approach to predict biodiversity as we have done here is that these combinations need not be determined *a priori*. How changing patterns of diversity will be reflected in terms of the actual composition of the flora, however, will require a more targeted approach.

Our models do not consider the ecology of individual taxa, and by extension, do not directly consider possible range expansions and contractions via dispersal and local extinctions (Currie, 1991, 2001; Sommer *et al.*, 2010). As our results alone do not provide information on taxon identity or the functional roles and endangerment status of individual taxa, conservation strategies must also take complementary taxon-specific data into account to be as effective as possible. For instance, our results do not discern between terrestrial plant biodiversity and that of exceptionally vulnerable aquatic and wetland species. Also, while our models account for particular aspects of climate change, they do not address the complexity of biotic interactions, potential additional environmental constraints, and changes in included and additional nonclimatic environmental variables, all of which can influence changes in patterns of biodiversity (Hutchinson, 1959; MacArthur, 1965; Macarthur & Levins, 1967; Brown, 1981; Wright, 1983). For instance, individuals assumed to shift their distributions following the climatic conditions they are adapted to may not encounter adequate photoperiods or necessary mutualists, rendering our predictions overestimations (Visser, 2008). Alternatively, associations with certain mutualists can expand the environmental tolerance of plant species, potentially mitigating the effects of climate change, in which case our predictions could be underestimating biodiversity (Peay, 2016; Gerz *et al.*, 2018). Similarly, our predictions do not account for potential climatic refugia at small spatial scales, where species may be able to persist even as conditions become unsuitable in the overall area.

Additionally, though their relative importance is debated, stochastic factors related to demographic fluctuations and genetic drift, or environmental variability (e.g. extremes) and disturbances can influence patterns of biodiversity and community assembly (Watt, 1947; Wiens, 1977; Den Boer, 1981; Strong *et al.*, 1984; Hubbell, 2001; Tilman, 2004; McPeek & Gomulkiewicz, 2005; Guisan & Rahbek, 2011; Rosindell *et al.*, 2012). Species with larger effective population sizes may be able to adapt to changing climates in situ while those with smaller sizes may face local extirpation if they are not able to disperse to more favorable conditions. Along these lines, though our data represent a comprehensive inventory of the US flora, it is difficult to gauge whether local plant diversity is at capacity, especially in terms of nonnative species (Stohlgren *et al.*, 2003). Lastly, it is uncertain whether or how the relationships between plant biodiversity and the abiotic factors examined here may change over time and to what degree future novel environmental conditions could influence these patterns (Williams *et al.*, 2007).

In the near future, it may very well become possible to incorporate information regarding biotic interactions, genetic diversity, ecological traits, biogeographical history, and variable relationships between facets of biodiversity and climate into machine learning approaches as more data become available. Nonetheless, climatic variables are assumed to be the strongest influence on the distribution of biodiversity (Wright, 1983; Currie & Paquin, 1987; Adams & Woodward, 1989; Ricklefs, 1990; O'Brien, 1993; Araújo & Rahbek, 2006), and our results represent one of the most comprehensive uses of climatic variables in addition to edaphic and geographic factors to predict regional patterns of plant biodiversity to date. We thus demonstrate the potential of machine learning approaches for predicting complex biodiversity patterns and show that the consequences of climate change can vary markedly across different facets of biodiversity.

© 2020 The Authors
New Phytologist © 2020 New Phytologist Trust

Such approaches can especially be useful for conservation efforts when species-specific data are unavailable and where the goal is to identify regions that will gain and lose the capacity to support biodiversity.

Acknowledgements

We express gratitude to the many collectors and curators of biodiversity data which have made this research possible, as well as the anonymous reviewers who provided invaluable feedback. We also thank Sharon Qi and the USGS for providing glaciation data. This research was supported by the Harvard University Herbaria and NSF-DEB 1754584. The authors declare no competing interests.

Author contributions

DSP and SW designed the study. JTK contributed plant inventories and status data. DSP, ZX and CCD conducted phylogenetic analyses. SW implemented models and created figures and tables. DSP, CGW and SW analyzed model outputs. DSP, CGW and SW wrote the initial draft of the manuscript, and all authors contributed substantially to revisions.

ORCID

Charles C. Davis https://orcid.org/0000-0001-8747-1101
Daniel S. Park https://orcid.org/0000-0003-2783-530X
Steven Worthington https://orcid.org/0000-0001-9550-5797
Charles G. Willis https://orcid.org/0000-0003-2543-246X
Zhenxiang Xi https://orcid.org/0000-0002-2851-5474

References

Adams JM, Woodward FI. 1989. Patterns in tree species richness as a test of the glacial extinction hypothesis. *Nature* **339**: 699.

Algar AC, Kharouba HM, Young ER, Kerr JT. 2009. Predicting the future of species diversity: macroecological theory, climate change, and direct tests of alternative forecasting methods. *Ecography* **32**: 22–33.

Araújo MB, Rahbek C. 2006. How does climate change affect biodiversity? *Science* **313**: 1396–1397.

Ayres MP, Lombardero MJ. 2000. Assessing the consequences of global change for forest disturbance from herbivores and pathogens. *Science of the Total Environment* **262**: 263–286.

Bakkenes M, Alkemade JRM, Ihle F, Leemans R, Latour JB. 2002. Assessing effects of forecasted climate change on the diversity and distribution of European higher plants for 2050. *Global Change Biology* **8**: 390–407.

Bates D, Mächler M, Bolker B, Walker S. 2014. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67**: 1–48.

Bell CD, Soltis DE, Soltis PS. 2010. The age and diversification of the angiosperms re-revisited. *American Journal of Botany* **97**: 1296–1303.

Bezeng BS, Morales-Castilla I, van der Bank M, Yessoufou K, Daru BH, Davies TJ. 2017. Climate change may reduce the spread of non-native species. *Ecosphere* **8**: e01694.

den Boer PJ. 1981. On the survival of populations in a heterogeneous and variable environment. *Oecologia* **50**: 39–53.

Boucher-Lalonde V, Kerr JT, Currie DJ. 2013. Does climate limit species richness by limiting individual species' ranges? *Proceedings of the Royal Society B: Biological Sciences* **281**: 20132695–20132695.

Bradley BA, Blumenthal DM, Wilcove DS, Ziska LH. 2010. Predicting plant invasions in an era of global change. *Trends in Ecology & Evolution* **25**: 310–318.

Brown JH. 1981. Two decades of homage to Santa Rosalia: toward a general theory of diversity. *American Zoologist* **21**: 877–888.

Cadotte MW, Cardinale BJ, Oakley TH. 2008. Evolutionary history and the effect of biodiversity on plant productivity. *Proceedings of the National Academy of Sciences, USA* **105**: 17012–17017.

Cadotte MW, Dinnage R, Tilman D. 2012. Phylogenetic diversity promotes ecosystem stability. *Ecology* **93**: S223–S233.

Cavender-Bares J, Kothari S, Meireles JE, Kaproth MA, Manos PS, Hipp AL. 2018. The role of diversification in community assembly of the oaks (*Quercus* L.) across the continental US. *American Journal of Botany* **105**: 565–586.

Collins WJ, Bellouin N, Doutriaux-Boucher M, Gedney N, Halloran P, Hinton T, Hughes J, Jones CD, Joshi M, Liddicoat S *et al.* 2011. Development and evaluation of an Earth-System model – HadGEM2. *Geoscientific Model Development* **4**: 1051–1075.

Cornell HV. 1985. Species assemblages of cynipid gall wasps are not saturated. *The American Naturalist* **126**: 565–569.

Cornell HV, Karlson RH. 1996. Species richness of reef-building corals determined by local and regional processes. *Journal of Animal Ecology* **65**: 233–241.

Cramer MD, Verboom GA. 2017. Measures of biologically relevant environmental heterogeneity improve prediction of regional plant species richness. *Journal of Biogeography* **44**: 579–591.

Currie DJ. 1991. Energy and large-scale patterns of animal- and plant-species richness. *The American Naturalist* **137**: 27–49.

Currie DJ. 2001. Projected effects of climate change on patterns of vertebrate and tree species richness in the conterminous United States. *Ecosystems* **4**: 216–225.

Currie DJ, Paquin V. 1987. Large-scale biogeographical patterns of species richness of trees. *Nature* **329**: 326.

Danielson JJ, Gesch DB. 2011. *Global multi-resolution terrain elevation data 2010 (GMTED2010).* Reston, VA, USA: US Geological Survey.

Daru BH, Park DS, Primack RB, Willis CG, Barrington DS, Whitfeld TJS, Seidler TG, Sweeney PW, Foster DR, Ellison AM *et al.* 2018. Widespread sampling biases in herbaria revealed from large-scale digitization. *New Phytologist* **217**: 939–955.

Daru BH, le Roux PC, Gopalraj J, Park DS, Holt BG, Greve M. 2019. Spatial overlaps between the global protected areas network and terrestrial hotspots of evolutionary diversity. *Global Ecology and Biogeography* **28**: 757–766.

Dinnage R, Cadotte MW, Haddad NM, Crutsinger GM, Tilman D. 2012. Diversity of plant evolutionary lineages promotes arthropod diversity. *Ecology Letters* **15**: 1308–1317.

Dukes JS, Mooney HA. 1999. Does global change increase the success of biological invaders? *Trends in Ecology & Evolution* **14**: 135–139.

Elith J, Leathwick JR. 2009. Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics* **40**: 677–697.

Ferrier S, Guisan A. 2006. Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology* **43**: 393–404.

Fischer G, Nachtergaele FO, Prieler S, van Velthuizen HT, Verelst L, Wiberg D. 2008. *Global agro-ecological zones assessment for agriculture (GAEZ 2008).* Laxenburg, Austria: IIASA, and Rome, Italy: FAO.

Flynn DFB, Mirotchnick N, Jain M, Palmer MI, Naeem S. 2011. Functional and phylogenetic diversity as predictors of biodiversity–ecosystem-function relationships. *Ecology* **92**: 1573–1581.

Francis AP, Currie DJ. 2003. A globally consistent richness-climate relationship for angiosperms. *The American Naturalist* **161**: 523–536.

Gerz M, Guillermo Bueno C, Ozinga WA, Zobel M, Moora M. 2018. Niche differentiation and expansion of plant species are associated with mycorrhizal symbiosis. *Journal of Ecology* **106**: 254–264.

González-Orozco CE, Pollock LJ, Thornhill AH, Mishler BD, Knerr N, Laffan SW, Miller JT, Rosauer DF, Faith DP, Nipperess DA. 2016. Phylogenetic approaches reveal biodiversity threats under climate change. *Nature Climate Change* **6**: 1110.

Goodwin BJ, McAllister AJ, Fahrig L. 1999. Predicting invasiveness of plant species based on biological information. *Conservation Biology* **13**: 422–426.

© 2020 The Authors
New Phytologist © 2020 New Phytologist Trust

Grodowitz MJ, Stewart RM, Cofrancesco AF. 1991. Population dynamics of waterhyacinth and the biological control agent *Neochetina eichhorniae* (Coleoptera: Curculionidae) at a southeast Texas location. *Environmental Entomology* **20**: 652–660.

Grömping U. 2009. Variable importance assessment in regression: linear regression versus random forest. *The American Statistician* **63**: 308–319.

Guisan A, Rahbek C. 2011. SESAM–a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *Journal of Biogeography* **38**: 1433–1444.

H-Acevedo D, Currie DJ. 2003. Does climate determine broad-scale patterns of species richness? A test of the causal link by natural experiment. *Global Ecology and Biogeography* **12**: 461–473.

Haj AE, Soller DR, Buchwald CA, Kauffman LJ, Heisig PM, Reddy JE. 2018. Databases used to develop a hydrogeologic framework for Quaternary sediments in the glaciated conterminous United States. *US Geological Survey data release*, doi: 10.5066/F71R6PQG.

Hawkins BA, Field R, Cornell HV, Currie DJ, Guégan J-F, Kaufman DM, Kerr JT, Mittelbach GG, Oberdorff T, O'Brien EM. 2003. Energy, water, and broad-scale geographic patterns of species richness. *Ecology* **84**: 3105–3117.

Hawkins BA, Rodríguez MÁ, Weller SG. 2011. Global angiosperm family richness revisited: linking ecology and evolution to climate. *Journal of Biogeography* **38**: 1253–1266.

Hector A, Schmid B, Beierkuhnlein C, Caldeira MC, Diemer M, Dimitrakopoulos PG, Finn JA, Freitas H, Giller PS, Good J. 1999. Plant diversity and productivity experiments in European grasslands. *Science* **286**: 1123–1127.

Hedrick B, Heberling M, Meineke E, Turner K, Grassa C, Park D, Kennedy J, Clarke J, Cook J, Blackburn D. 2020. Digitization and the future of natural history collections. *BioScience* **70**: 243–251.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. 2005. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology* **25**: 1965–1978.

Holdridge LR. 1947. Determination of world plant formations from simple climatic data. *Science* **105**: 367–368.

Hubbell SP. 2001. *The unified neutral theory of biodiversity and biogeography*. Princeton, NJ, USA: Princeton University Press.

Hutchinson GE. 1959. Homage to Santa Rosalia or why are there so many kinds of animals? *The American Naturalist* **93**: 145–159.

IPCC. 2014. *Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the intergovernmental panel on climate change* (Core writing team, Pachauri RK, Meyer LA, eds.). Geneva, Switzerland: IPCC.

Iverson LR, Prasad AM. 1998. Predicting abundance of 80 tree species following climate change in the eastern United States. *Ecological Monographs* **68**: 465–485.

Iverson LR, Prasad AM. 2001. Potential changes in tree species richness and forest community types following climate change. *Ecosystems* **4**: 186–199.

Jansson R. 2003. Global patterns in endemism explained by past climatic change. *Proceedings of the Royal Society of London B: Biological Sciences* **270**: 583–590.

Jiao Y, Wickett NJ, Ayyampalayam S, Chanderbali AS, Landherr L, Ralph PE, Tomsho LP, Hu Y, Liang H, Soltis PS. 2011. Ancestral polyploidy in seed plants and angiosperms. *Nature* **473**: 97.

Kamilar JM, Beaudrot L, Reed KE. 2015. Climate and species richness predict the phylogenetic structure of African mammal communities. *PLoS ONE* **10**: 1–16.

Kartesz JT. 2015. *The biota of north america program (BONAP). North american plant atlas*. Chapel Hill, NC, USA. (Maps generated from Kartesz, JT 2015. Floristic Synthesis of North America, v.1.0. Biota of North America Program (BONAP)).

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**: 772–780.

Kelling S, Hochachka WM, Fink D, Riedewald M, Caruana R, Ballard G, Hooker G. 2009. Data-intensive science: a new paradigm for biodiversity studies. *BioScience* **59**: 613–620.

Kerkhoff AJ, Moriarty PE, Weiser MD. 2014. The latitudinal species richness gradient in New World woody angiosperms is consistent with the tropical conservatism hypothesis. *Proceedings of the National Academy of Sciences, USA* **111**: 8125–8130.

Khun M. 2018. *Caret: classification and regression training. Caret. R package v.6.0-80*. [WWW document] URL https://cran.r-project.org/web/packages/caret/index.html.

Knapp S, Winter M, Klotz S. 2017. Increasing species richness but decreasing phylogenetic richness and divergence over a 320-year period of urbanization. *Journal of Applied Ecology* **54**: 1152–1160.

Kozlov AM, Aberer AJ, Stamatakis A. 2015. ExaML version 3: a tool for phylogenomic analyses on supercomputers. *Bioinformatics* **31**: 2577–2579.

Kreft H, Jetz W. 2007. Global patterns and determinants of vascular plant diversity. *Proceedings of the National Academy of Sciences, USA* **104**: 5925–5930.

Lima-Ribeiro MS, Moreno AKM, Terribile LC, Caten CT, Loyola R, Rangel TF, Diniz-Filho JAF. 2017. Fossil record improves biodiversity risk assessment under future climate change scenarios. *Diversity and Distributions* **23**: 922–933.

Linder HP. 2008. Plant species radiations: where, when, why? *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **363**: 3097–3105.

MacArthur RH. 1965. Patterns of species diversity. *Biological Reviews* **40**: 510–533.

Macarthur R, Levins R. 1967. The limiting similarity, convergence, and divergence of coexisting species. *The American Naturalist* **101**: 377–385.

McKenney DW, Pedlar JH, Lawrence K, Campbell K, Hutchinson MF. 2007. Potential impacts of climate change on the distribution of North American trees. *AIBS Bulletin* **57**: 939–948.

McPeek MA, Gomulkiewicz R. 2005. Assembling and depleting species richness in metacommunities: insights from ecology, population genetics, and macroevolution. In: Holyoak M, Leibold MA, Holt RD, eds. *Metacommunities: spatial dynamics and ecological communities*. Chicago, IL, USA: The University of Chicago Press, 355–373.

Meineke EK, Davies JT, Daru BH, Davis CC. 2019. Biological collections for understanding biodiversity in the Anthropocene. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **374**: 20170386.

Meyer C, Weigelt P, Kreft H. 2016. Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters* **19**: 992–1006.

Mokany K, Ferrier S. 2011. Predicting impacts of climate change on biodiversity: a role for semi-mechanistic community-level modelling. *Diversity and Distributions* **17**: 374–380.

Mokany K, Richardson AJ, Poloczanska ES, Ferrier S, Elith J, Robinson LM, Reside A, Harwood TD, Dunstan PK, Williams KJ *et al.* 2010. Uniting marine and terrestrial modelling of biodiversity under climate change. *Trends in Ecology & Evolution* **25**: 550–551.

Morse LE, Kutner LS, Maddox GD, Honey LL, Thurman CM, Kartesz JT, Chaplin SJ. 1993. *The potential effects of climate change on the native vascular flora of North America. A preliminary climate envelopes analysis: final report*. Palo Alto, CA, USA: Electric Power Research Institute.

O'Brien EM. 1993. Climatic gradients in woody plant species richness: towards an explanation based on an analysis of southern Africa's woody flora. *Journal of Biogeography* **20**: 181–198.

Olden JD, Lawler JJ, Poff NL. 2008. Machine learning methods without tears: a primer for ecologists. *The Quarterly Review of Biology* **83**: 171–193.

Owens CS, Madsen JD. 1995. Low temperature limits of waterhyacinth. *Journal of Aquatic Plant Management* **33**: 63–68.

Owens CS, Smart RM, Stewart RM. 2004. Low temperature limits of giant salvinia. *Journal of Aquatic Plant Management* **42**: 91–94.

Park DS, Davis CC. 2017. Implications and alternatives of assigning climate data to geographical centroids. *Journal of Biogeography* **44**: 2188–2198.

Park DS, Potter D. 2015. Why close relatives make bad neighbours: phylogenetic conservatism in niche preferences and dispersal disproves Darwin's naturalization hypothesis in the thistle tribe. *Molecular Ecology* **24**: 3181–3193.

Park DS, Razafindratsima OH. 2018. Anthropogenic threats can have cascading homogenizing effects on the phylogenetic and functional diversity of tropical ecosystems. *Ecography* **42**: 148–161.

Park DS, Willis CG, Xi Z, Kartesz JT, Davis CC, Worthington S. 2020. Replication code and data for: "Machine learning predicts large scale declines in native plant phylogenetic diversity". *Zenodo*. doi: 10.5281/zenodo.3755913.

© 2020 The Authors
New Phytologist © 2020 New Phytologist Trust

Parker J. 1963. Cold resistance in woody plants. *Botanical Review* **29**: 123–201.
Parmesan C. 2006. Ecological and evolutionary responses to recent climate change. *Annual Review of Ecology Evolution and Systematics* **37**: 637–669.
Peay KG. 2016. The mutualistic niche: mycorrhizal symbiosis and community dynamics. *Annual Review of Ecology, Evolution, and Systematics* **47**: 143–164.
Peterson T, Ortega-Huerta M, Bartley J, Sánchez-Cordero V, Soberón J, Buddemeier RH, Stockwell DRB. 2002. Future projections for Mexican faunas under global climate change scenarios. *Nature* **416**: 626–629.
Pigott CD, Pigott S. 1993. Water as a determinant of the distribution of trees at the boundary of the mediterranean zone. *Journal of Ecology* **81**: 557–566.
Prentice IC, Bondeau A, Cramer W, Harrison SP, Hickler T, Lucht W, Sitch S, Smith B, Sykes MT. 2007. Dynamic global vegetation modeling: quantifying terrestrial ecosystem responses to large-scale environmental change. In: Canadell JG, Pataki DE, Pitelka LF, eds. *Terrestrial ecosystems in a changing world*. Heidelberg, Germany: Springer, 175–192.
Pyšek P, Prach K, Rejmánek M, Wade M. 1995. *Plant invasions: general aspects and special problems*. Amsterdam, the Netherlands: SPB Academic Publishing.
Qian H, Ricklefs RE. 2004. Taxon richness and climate in angiosperms: is there a globally consistent relationship that precludes region effects? *The American Naturalist* **163**: 773–779.
Qian H, Ricklefs RE. 2006. The role of exotic species in homogenizing the North American flora. *Ecology Letters* **9**: 1293–1298.
R Core Team. 2017. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
Rejmánek M, Richardson DM. 1996. What attributes make some plant species more invasive? *Ecology* 77: 1655–1661.
Ricklefs RE. 1977. Environmental heterogeneity and plant species diversity: a hypothesis. *The American Naturalist* **111**: 376–381.
Ricklefs RE. 1990. Seabird life histories and the marine environment: some speculations. *Colonial Waterbirds* **13**: 1–6.
Ricklefs RE. 2004. A comprehensive framework for global patterns in biodiversity. *Ecology Letters* 7: 1–15.
Rosindell J, Hubbell SP, He F, Harmon LJ, Etienne RS. 2012. The case for ecological neutral theory. *Trends in Ecology & Evolution* **27**: 203–208.
Sakai A, Weiser CJ. 1973. Freezing resistance of trees in North America with reference to tree regions. *Ecology* **54**: 118–126.
Smith SA, Beaulieu JM, Donoghue MJ. 2009. Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. *BMC Evolutionary Biology* **9**: 37.
Smith SA, Dunn CW. 2008. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. *Bioinformatics* **24**: 715–716.
Smith SA, O'Meara BC. 2012. treePL: divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* **28**: 2689–2690.
Sommer JH, Kreft H, Kier G, Jetz W, Mutke J, Barthlott W. 2010. Projected impacts of climate change on regional capacities for global plant species richness. *Proceedings of the Royal Society B: Biological Sciences* **277**: 2271–2280.
Srivastava DS, Cadotte MW, MacDonald AAM, Marushia RG, Mirotchnick N. 2012. Phylogenetic diversity and the functioning of ecosystems. *Ecology Letters* **15**: 637–648.
Staab M, Bruelheide H, Durka W, Michalski S, Purschke O, Zhu C-D, Klein A-M. 2016. Tree phylogenetic diversity promotes host–parasitoid interactions. *Proceedings of the Royal Society B: Biological Sciences* **283**: 20160275.
Stephenson NL. 1990. Climatic Control of vegetation distribution – the role of water-balance. *American Naturalist* **135**: 649–670.
Stohlgren TJ, Barnett DT, Kartesz JT. 2003. The rich get richer: patterns of plant invasions in the United States. *Frontiers in Ecology and the Environment* **1**: 11–14.
Strong DR, Lawton JH, Southwood SR. 1984. *Insects on plants. Community patterns and mechanisms*. Oxford, UK: Blackwell Scientific Publications.
Swenson NG, Enquist BJ, Pither J, Kerkhoff AJ, Boyle B, Weiser MD, Elser JJ, Fagan WF, Forero-Montaña J, Fyllas N *et al.* 2012. The biogeography and filtering of woody plant functional diversity in North and South America. *Global Ecology and Biogeography* **21**: 798–808.
Thuiller W, Lavorel S, Araujo MB, Sykes MT, Prentice IC. 2005. Climate change threats to plant diversity in Europe. *Proceedings of the National Academy of Sciences, USA* **102**: 8245–8250.
Thuiller W, Richardson DM, Midgley GF. 2008. Will climate change promote alien plant invasions? In: Nentwig W, ed. *Biological invasions. Ecological studies (analysis and synthesis), vol. 193*. Berlin/Heidelberg, Germany: Springer, 197–211.
Tilman D. 2004. Niche tradeoffs, neutrality, and community structure: a stochastic theory of resource competition, invasion, and community assembly. *Proceedings of the National Academy of Sciences, USA* **101**: 10854–10861.
Tilman D, Clark M, Williams DR, Kimmel K, Polasky S, Packer C. 2017. Future threats to biodiversity and pathways to their prevention. *Nature* **546**: 73–81.
Tsirogiannis C, Sandel B. 2016. PhyloMeasures: a package for computing phylogenetic biodiversity measures and their statistical moments. *Ecography* **39**: 709–714.
Vellend M, Baeten L, Becker-Scarpitta A, Boucher-Lalonde V, McCune JL, Messier J, Myers-Smith IH, Sax DF. 2017. Plant biodiversity change across scales during the Anthropocene. *Annual Review of Plant Biology* **68**: 563–586.
Venail P, Gross K, Oakley TH, Narwani A, Allan E, Flombaum P, Isbell F, Joshi J, Reich PB, Tilman D. 2015. Species richness, but not phylogenetic diversity, influences community biomass production and temporal stability in a re-examination of 16 grassland biodiversity studies. *Functional Ecology* **29**: 615–626.
Venevsky S, Veneskaia I. 2003. Large-scale energetic and landscape factors of vegetation diversity. *Ecology Letters* **6**: 1004–1016.
Visser ME. 2008. Keeping up with a warming world; assessing the rate of adaptation to climate change. *Proceedings of the Royal Society of London B: Biological Sciences* **275**: 649–659.
Walther GR, Post E, Convey P, Menzel A, Parmesan C, Beebee TJC, Fromentin JM, Hoegh-Guldberg O, Bairlein F. 2002. Ecological responses to recent climate change. *Nature* **416**: 389–395.
Watt AS. 1947. Pattern and process in the plant community. *Journal of ecology* **35**: 1–22.
Wiens JA. 1977. On competition and variable environments: populations may experience 'ecological crunches' in variable climates, nullifying the assumptions of competition theory and limiting the usefulness of short-term studies of population patterns. *American Scientist* **65**: 590–597.
Williams JW, Jackson ST, Kutzbach JE. 2007. Projected distributions of novel and disappearing climates by 2100 AD. *Proceedings of the National Academy of Sciences, USA* **104**: 5738–5742.
Willis CG, Halina M, Lehman C, Reich PB, Keen A, McCarthy S, Cavender-Bares J. 2010. Phylogenetic community structure in Minnesota oak savanna is influenced by spatial extent and environmental variation. *Ecography* **33**: 565–577.
Wilsey BJ, Teaschner TB, Daneshgar PP, Isbell FI, Polley HW. 2009. Biodiversity maintenance mechanisms differ between native and novel exotic-dominated communities. *Ecology Letters* **12**: 432–442.
Wright DH. 1983. Species-energy theory: an extension of species-area theory. *Oikos* **41**: 496–506.
Wuebbles DJ, Fahey DW, Hibbard KA, Dokken BC, Stewart BC, Maycock TK. 2017. *Climate Science Special Report: Fourth National Climate Assessment, vol* **I**. Washington, DC, USA: US Global Change Research Program.
Zhang J, Nielsen SE, Chen Y, Georges D, Qin Y, Wang S, Svenning J, Thuiller W. 2017. Extinction risk of North American seed plants elevated by climate and land-use change. *Journal of Applied Ecology* **54**: 303–312.
Zhang J, Nielsen SE, Stolar J, Chen Y, Thuiller W. 2015. Gains and losses of plant species and phylogenetic diversity for a northern high-latitude region. *Diversity and Distributions* **21**: 1441–1454.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Fig. S1 Standard errors of each biodiversity metric derived from the phylogenetic bootstrap replicates.

© 2020 The Authors
New Phytologist © 2020 New Phytologist Trust

Fig. S2 Illustration of the predictive modeling workflow.

Fig. S3 Predicted changes in total plant biodiversity and community structure in 2050 and 2070 relative to current values under RCP 2.6, 6.0 and 8.5.

Fig. S4 Predicted changes in native vs nonnative introduced plant biodiversity and community structure in 2050 and 2070 relative to current values under RCP 2.6, 6.0 and 8.5.

Fig. S5 Correlations between changes in different aspects of biodiversity under RCP 4.5.

Fig. S6 Predicted changes in native vs nonnative introduced plant biodiversity and community structure in 2050 and 2070 relative to current values under RCP 4.5.

Fig. S7 Predicted changes in total plant biodiversity and community structure in 2050 and 2070 relative to current values under RCP4.5.

Fig. S8 Predicted changes in native vs nonnative plant biodiversity and community structure in 2050 and 2070 for RCP4.5 by political divisions and ecoregions.

Fig. S9 Predicted changes in total plant biodiversity and community structure in 2050 and 2070 for RCP4.5 by political divisions and ecoregions.

Fig. S10 Current patterns of total plant diversity.

Fig. S11 Present day observed relationships between four climatic explanator variables (BIO 07 (mean), BIO 15 (mean), clay (mean), coast (binary)) and total taxonomic diversity.

Fig. S12 Present day observed relationships between four climatic explanator variables (BIO 07 (mean), BIO 15 (mean), clay (mean), coast (binary)) and native and nonnative diversity.

Fig. S13 Present day observed relationships between four climatic explanator variables (BIO 07 (mean), BIO 15 (mean), clay (mean), coast (binary)) and total phylogenetic diversity.

Table S1 Proportions of variance in current biodiversity explained by Cubist models.

Table S2 Pearson's correlation coefficients (R) between current amounts of biodiversity and predicted changes in the future (2050, 2070).

Table S3 Standardized effects of climatic and environmental variables on total taxonomic, phylogenetic diversity, and phylogenetic community structure.

Table S4 Standardized effects of climatic and environmental variables on nonnative taxonomic, phylogenetic diversity, and phylogenetic community structure.

Table S5 Standardized effects of climatic and environmental variables on native taxonomic, phylogenetic diversity, and phylogenetic community structure.

Table S6 The proportion of change in the standard deviation of biodiversity across the United States under various climate change scenarios in 2050 and 2070.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

© 2020 The Authors
New Phytologist © 2020 New Phytologist Trust