Ecological Informatics 79 (2024) 102406

Contents lists available at ScienceDirect

Ecological Informatics

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

Practical guide for retaining correlated climate variables and unthinned samples in species distribution modeling, using random forests

Brice B. Hanberry

USDA Forest Service, Rocky Mountain Research Station, Rapid City, SD 57702, USA

ARTICLE INFO

Keywords:
Collinearity
Correlation
Feature selection
Misconceptions
Thinning
Random forests

ABSTRACT

Species distribution models contain bias, or inaccurate predictions, due to predictors and species occurrences. Collinearity of predictor variables is a concern limited to standard error of estimates for data models, although issues remain about identification of important variables and model transferability, whereas thinning of species occurrences is not likely to improve samples without information about species characteristics. Here, I present a case study of selection of correlated climate variables and thinning of species samples for distributions of 80 tree species in North America, modeled with random forests. Robust important variable identification and temporal transferability with correlated climate variables by random forests was displayed by 1) greatest accuracy with all 14 input predictor variables, at 0.99 sensitivity and 0.97 specificity, whether from current or future climate, compared to variable selection approaches to reduce correlation, 2) greater accuracy for two variable models from foremost ranked important variables than least important variables, 3) models with more than two fixed input variables agreed on the four most important variables, 4) models with all 14 input variables from current and future climates identified the same five most important variables, and 5) mapped models of current species observations were similar with use of current or future climates. Unconstrained distributions were apparent in mapped models with only one preselected temperature and precipitation variable. Thinning samples produced no clear benefits to unthinned samples. Following standard guidelines to remove correlated variables before modeling will result in information loss and may lead to incorrect omission of relevant variables, at least with random forests. Predictive modeling is a practical approach for informing variable selection, including retention of correlated climate variables, which create a reliable basis for accurate predictions.

1. Introduction

1.1. Uncertainty in species distribution models

Modeling of species distributions based on whether a species is present or absent is relatively simple and accessible, but a variety of issues produce uncertainty and errors (Santini et al., 2021; Zurell et al., 2020). Modeling depends on numerous choices about species datasets, predictor variable datasets, and methods, resulting in a range of recommendations from knowledgeable and skilled modelers about best practices (Araújo et al., 2019; Santini et al., 2021; Sofaer et al., 2019; Zurell et al., 2020). To be clear, modeling assumptions of unbiased sampling that comprehensively covers the environmental space of species and species in equilibrium with the environment, that is, occupying the full environmental space, are violated immediately by most datasets of species occurrences (Faurby and Araújo, 2018). Species only occupy a portion of the environmental space where they can grow, survive, and reproduce, due to constraints of land use, disturbance, biotic interactions, dispersal limitations, and other barriers. Consequently, native species rarely fill their modeled potential climate distributions, even if ranges have expanded since historical times. Indeed, range expansions likely result in models of increased potential distributions. Therefore, almost all models will be flawed, regardless of following modeling guidelines.

1.2. Species samples

Many species distributions are clustered, naturally, and non-random patterns and autocorrelation are characteristics of distributions (Hawkins, 2012; Kühn and Dormann, 2012). Some locations are better than others for supporting species; greater concentrations of samples produce the conceptual basis for species distribution models (Hawkins, 2012; Kühn and Dormann, 2012). Spatial autocorrelation in the residuals may be problem for data models (i.e., conventional statistical models; Kühn

E-mail address: brice.hanberry@usda.gov.

https://doi.org/10.1016/j.ecoinf.2023.102406
Received 24 August 2023; Received in revised form 29 November 2023; Accepted 1 December 2023
Available online 2 December 2023

and Dormann, 2012). Nonetheless, residual plots may not indicate whether the model is a good fit to the data (Breiman, 2001). For predictive modeling, the most practical guideline for model validation is to model the data and measure predictive accuracy (with machine learning algorithmic models, which are built for accurate predictions; Breiman, 2001).

For landscape studies, sampling of species typically is not uniform, arising from sampling bias. Even when sampling schemes are in place to reduce variation, landscapes will contain different amounts of land cover, which produces unequal sampling intensity. For example, sampling effort calculated relative to total land area is different than effort calculated for total area of suitable ecosystems (e.g., forests; Hanberry, 2023a). Manipulating samples by subsampling presence data is one approach to reduce unknown sampling bias (Aiello-Lammens et al., 2015; Varela et al., 2014; Vollering et al., 2019). Varela et al. (2014) found that spatial thinning by geographic filters consistently downgraded models, albeit slightly, but that spatial thinning by environmental filters improved model accuracy, again slightly. Without guidelines to differentiate sampling bias from species distributions, removing information contained in samples may not produce clear benefits.

Within sampling constraints, samples become critical. Small samples may indicate that sample errors or unusual outliers (the 'noise') are incorporated into the model. However, the major problem is if samples do not cover the species range of predictive space, due to limitations in modeled extent (e.g., administrative boundaries) or sampling records or due to species range contractions from larger historical distributions. That is, the number of samples is not as important as whether the samples capture the full range of environmental variables of the species.

However, species ranges have changed over time in response to human activities. No species is likely to have the same distribution currently as before the Industrial Revolution and human use of engine-powered machines, which resulted in intensive and extensive land use, disturbance change, and overexploitation (Faurby and Araújo, 2018; IPBES, 2019; WWF, 2018). Globally, monitored populations of birds, mammals, fish, reptiles, and amphibians have declined in abundance by 60% on average between 1970 and 2014 (WWF, 2018). Thousands of samples from a contracted range are not likely to produce good models, except for endemic species with concentrated locations or if extirpation removed subspecies genetically suited to the climate of their location. Species that have been extirpated from historical wide ranges cannot be modeled correctly in terms of assessing physiological tolerances or environmental associations by any number of samples from current contracted distributions. Conversely, under human activities of transplanting, forestation of historically non-forested ecosystems, exclusion of historical disturbances, and differential filters and competitive dynamics, some native species have expanded in range, similar to non-native species when released from constraints (Hanberry, 2021). Models may supply useful information about expanding, novel distributions, but expanding distributions are not stable, which may require multiple models to account for incomplete associations.

1.3. Predictor variables

Equally, predictor variables for species distribution models may contain some of the same sampling errors as species data and important predictors for species boundaries often are climate variables. Climate data are generated by differing models from measurements of varying sampling intensities (Karger et al., 2017; Lopez-Cantu et al., 2020). Due to the spatial structure of climate, climate data consistently will produce the best species distribution models and be the most important types of variables, with effectiveness evidenced by the number of studies that use climate data compared to other types of variables (Fourcade et al., 2018). Climate variables were used in 88% of 190 species distribution studies during 2012–2016 (Fourcade et al., 2018) and may be the only predictor variables in up to 85% of studies (Titeux et al., 2016). Ecologically, climate has generally outlined species boundaries over time and latitudinal gradients align with physiological tolerances of dominant plant species (Hanberry, 2023b).

Nevertheless, this does not mean that climate is most influential in realized species distributions (Fourcade et al., 2018), resulting in the need for isolation and integration of predictor variables beyond climate, and which can be projected to the future for modeling future species distributions (Arenas-Castro and Sillero, 2021; Leitão and Santos, 2019; Regos et al., 2022). Beyond variables related to climate, emerging products from remote sensing eventually may prove as effective as climate at modeling current species distributions, or at least, species abundances within distributions (Arenas-Castro and Sillero, 2021; Leitão and Santos, 2019; Regos et al., 2022). Easily accessible current land use and land cover classes do not particularly match well with current species distributions (Hanberry, 2023a). Consequently, land use and disturbance variables rarely are applied or effective in species distribution modeling relative to climate models (Fourcade et al., 2018; Hanberry, 2023a). Seemingly stable associations, such as with soil and topography, have changed since historical times (Hanberry et al., 2013, 2014). Soil and topography tend to modulate effects of climate and disturbance; for example, soil and topography can support a range of shifting species over time (Hanberry et al., 2013). Novel species distributions, after range contractions or expansions due to land use and disturbance change, may result in novel associations with environmental variables that did not occur in the past and may not persist in the future, which may require multiple models to account for spurious associations (Goring and Williams, 2017; Hanberry et al., 2013, 2014). Predictors for explaining current novel ranges may differ from predictors of historical ranges or future ranges (Hanberry et al., 2013).

Numerous variable selection approaches exist, which can result in generation of vastly different models. Modeling packages may incorporate algorithms to reduce the number of variables, such as recursive feature elimination, which ranks variables through resampling and backward selection (Kuhn, 2008). Still, standard guidelines are to select ecologically relevant variables, but then remove one variable from each pair of (ecologically relevant) correlated variables before modeling, albeit other options include combining variables (Araújo et al., 2019; Fourcade et al., 2018; Santini et al., 2021).

Removing correlated variables may be more of a cost than a benefit. Concern about collinearity in predictor variables is more properly limited to linear data models, for which correlated predictor variables may cause (appropriately) large errors of coefficient estimates (Goldberger and Goldberger, 1991; Lindner et al., 2020). Regarding prediction, coefficient estimates are not pertinent but rather accurate predictions are material because leaving out relevant predictor variables will reduce accuracy (Goldberger and Goldberger, 1991; Lindner et al., 2020; Vatcheva et al., 2016; Williams et al., 2013). Indeed, greater collinearity of variables, acting together to explain variation in a response variable, results in greater consequences of omitting relevant predictors (Lindner et al., 2020). Nonetheless, concerns linger about identification of the important variables when variables are correlated. If variables have different spatiotemporal structures of collinearity, collinearity likely will be a problem when a model is trained on data from one region or time and predicted to another (Dormann et al., 2013).

Specifically, climate variables are derivatives of temperature and precipitation that are correlated currently, in the past, and in the future. Moreover, past and future climates models include correction methods to preserve the same trends in magnitude and direction between historical and future periods (Hui et al., 2020; Nahar et al., 2017). Climate variables reflect order in the environment and generally should increase or decrease together, e.g., under climate warming, temperature variables will increase. If climate inherently was random with unfounded correlations among variables, weather conditions would no longer represent a defined climate and species distribution modeling may not be possible due to lack of species to model.

1.4. Modeling approaches

Modeling also includes selection of the classifier (or probability of occurrence estimator), which encompasses hundreds of different options (Fernández-Delgado et al., 2014), although general linear models, MaxEnt, and random forests are among the most commonly applied classifiers (Santini et al., 2021). For the selected classifier, modeling package options enable automated tuning for parameters through resampling and standard default values to reach the most accurate final model configuration. Multiple methods may be applied to detect overlapping outcomes, allowing some strength of evidence within a study rather than only comparing support from other studies, but multiple methods may be too burdensome when hundreds or thousands of models are generated due to numerous species, predictor variable combinations, and time intervals.

For other modeling choices, many different recommendations are available (Santini et al., 2021) and simpler may be better at least until an overwhelming consensus is reached. True absences generally are unknown; therefore, background samples that cover the entire study extent to provide a full range of environmental values can act as pseudoabsences, just as species occurrences ideally cover the full range of the species. Generally, when class membership probabilities are unknown, an equal number of known presences to pseudoabsences helps to standardize accuracy of predictions across classifiers, providing a balance between predicting presence and absences or among classes. Model evaluation metrics typically occur by predicting to samples withheld from modeling, through split samples of training and testing data. Some sort of accuracy measure is needed, and single score evaluation statistics have come in and out of favor, from the kappa statistic, to the area under the curve, and the true skill statistic (Fourcade et al., 2018). The phi coefficient, which was published in 1912 (also known as the mean square contingency coefficient or Matthew's correlation coefficient; Yule, 1912) currently may be considered the best single score metric of a binary classifier prediction in a confusion matrix context, due to accounting for imbalanced data, whereas scoring rules may be best for probabilities. While 'the search for an unerring test yields a series of tests that achieve partial success' (Youden, 1950), reporting accuracy through two scores of sensitivity and specificity is uncomplicated and straightforward to interpret.

Models are imperfect but may be informative in terms of providing a baseline and also magnitude of change. The accuracy measure is not the conclusion of model performance. Models predicted to maps generate another evaluation tool, at least as a check for obvious large departures from realized ranges. Results are discussed in an ecological context, with modeling caveats and comparison to other studies.

1.5. Objectives and overview of feature selection comparisons

Although disagreement occurs for any modeling approach, removal of correlated climate variables and species samples generally may be unwarranted at best. Despite lack of species-specific information to guide altering samples, a step of thinning samples is sometimes applied in an attempt to even out species distributions and potential sampling bias. Spurious associations with species and removal of relevant predictor variables through feature selection rather than multicollinearity are problems for model predictions (Goldberger and Goldberger, 1991; Lindner et al., 2020; Shmueli, 2010; Vatcheva et al., 2016; Williams et al., 2013). Decisions arise regarding which temperature or precipitation variables to retain and at which threshold of correlation, while concerns linger about identification of the important variables when variables are correlated and effects of collinearity when a model is trained on data from one region or time and predicted to another. However, future climate data are available and can be modeled with current species occurrences to determine if the same important variables and mapped models hold true, or temporal transferability. Testing for spatial transferability of continental tree species distributions is not an option because if the species distributions are subdivided, then the full ecological space for tree species will be lost from models, creating sample issues and violating assumptions of species distribution modeling. Here, my objectives were to examine the issues of variable and sample selection through modeling, with the random forests classifier, to determine the costs and benefits of reducing predictor variables and species occurrence samples, for 80 tree species that occurred but were not limited to Mexico in a study extent of North America. Predictive modeling can provide practical recommendations for variable and sample selection in species distribution modeling.

As an expansion to a correlation assessment of hindcasting tree species (Hanberry, 2023b), I applied seven different variable selection approaches for seven temperature and seven precipitation variables, modeled with the random forests classifier, to examine the effects of variable selection on model accuracy, correlation, and identification of important variables, for which I used near current climate and future climate. One approach was to use all 14 input variables, one approach was modeling to reduce to a number of variables that varied by species, two approaches reduced models to two fixed input variables preselected before modeling, two approaches reduced models to two fixed input variables after modeling, while one approach was modeling to reduce models to two input variables that varied by species. Reduction to two input variables, one of temperature and one of precipitation, was necessary to reduce high correlation (i.e., achieving $r \leq 0.61$) among temperature variables and among precipitation variables but will not represent the full complexity of species distributions. For equivalence to isolate comparisons among two variable models, holding all else equal, I also provided mapped model examples from two highly correlated temperature variables ($r = 0.97$), although this option is unrealistic. Comparisons used current or future climate predictors and were based on accuracy, stability of important variables, and mapped models relative to realized distributions of species occurrences, as kernels that contain 95% of observations.

2. Methods

2.1. Species samples and thinning

Tree species encompassed 80 tree species, which at least occurred in Mexico but were not endemic, and abundantly surveyed with 1000 to 5000 observations in the study extent of North America. The Global Biodiversity Information Facility (GBIF; Global Biodiversity Information Facility, 2023) is a centralized data repository containing 65,000 datasets of 2 billion species observations. I filtered tree species that were observed since 1990 in North America with coordinate uncertainty <1000 m (GBIF occurrence download https://doi.org/10.15468/dl.22w5sf on 17 March 2023). For pseudoabsences, I used an equal number of random points from the study extent, or background. I extracted climate variables for the occurrences and background absence points. Additionally, I also thinned the samples to a maximum of one sample per grid cell within the climate raster layers, with an equal number of pseudoabsences from the study extent of North America (i.e., gridSample function of the R dismo package; Hijmans et al., 2011; Varela et al., 2014). An outcome of reducing sample sizes is that minimum sample size thresholds need to be lowered to retain species or else species need to be removed, as they no longer meet the threshold of 1000 to 5000 observations. Almost all species (78) met a sample size threshold of 400 minimum observations, which I modeled with all variables and compared to models of all samples and all variables.

2.2. Climate variables

For climate, I downloaded climatologies during 1900–1990 provided by CHELSA-TraCE21k v1.0 (resolution 30 arc *sec*; Karger et al., 2020, 2021) for the study extent of the North American continent. The years 1900 to 1990 account for the long lifespan of trees and observations

recorded since 1990 typically are older trees, rather than seedlings. I obtained 14 bioclimatic variables that quantify temperature and precipitation at annual, seasonal, and monthly intervals: mean annual temperature, maximum temperature of the hottest month, minimum temperature of the coldest month, mean temperature of the coldest and hottest three months, mean temperature of the driest and wettest three months, annual precipitation, precipitation during the driest month and driest three months, precipitation during the wettest month and wettest three months, and precipitation during the coldest and warmest three months. Additionally, I downloaded the same variables during 2071–2100 (resolution 30 arc sec) from projections for GFDL-ESM4 (Geophysical Fluid Dynamics Laboratory, USA), IPSL-CM6A-LR (Institut Pierre Simon Laplace, France), UKESM1–0-LL (Met Office Hadley Centre, UK) and the moderate SSP3–7.0 scenario from the latest phase of climate simulations, Coupled Model Intercomparison Project Phase 6 (Karger et al., 2017, 2018). Temperature variables of the near current climate and three end-of-century climates had 253 out of 255 pair-wise correlations that exceeded 0.7, and a mean *r* value of 0.90, excluding the variable of mean temperature of the wettest quarter for all four climates (ESRI ArcMap 10.8.2, Redlands, California). Precipitation variables of the near current climate and three end-of-century climates had 254 out of 261 pair-wise correlations that exceeded 0.7, and a mean *r* value of 0.78, excluding the variable of precipitation of the warmest quarter during 1900–1990 (ESRI ArcMap 10.8.2, Redlands, California).

2.3. Modeling framework

For modeling species distributions, I applied the random forests classifier with 10-fold cross-validation, repeated three times, for one run of each model that varied by input variables, and all modeling used the same random number generator (Kuhn, 2008; R Core Team, 2023; the caret package is well-documented, e.g., https://cran.r-project.org/web/packages/caret/vignettes/caret.html). The random forests classifier, which is widely used and highly accurate in predictions (i.e., > 90% in 102 out of 121 of the data sets; Fernández-Delgado et al., 2014), is a nonlinear ensemble method that aggregates results of many decision trees or rule-based models to output the most optimal result, helping to minimize the influence of error. The random forests classifier runs models in parallel (i.e., bagging) and averages results to reduce variance (i.e., overfitting). To help distinguish the influence of redundant, highly correlated predictors, only a subset of variables is considered for each tree and the rule at each branch is constructed with a single variable threshold. In the case of a reduced set of two input variables, random forests still 1) bootstraps samples from training data, 2) selects a feature at each split in the tree, 3) determines a split point from many possible values, and 4) averages trees. In random forests, the standard is to make use of bagging and tree averaging, rather than pruning trees or limiting tree size, to prevent over-fitting (Díaz-Uriarte and Alvarez de Andrés, 2006; Goldstein et al., 2011). Breiman (2000:4) wrote that 'the largest trees possible result in the best accuracy.' Indeed, typically fine-tuning parameters is unnecessary to achieve accurate performance with random forests (Díaz-Uriarte and Alvarez de Andrés, 2006).

Accuracy metrics occurred on separate testing data, or withheld known classes, to determine how well the classifier assigned classes using predictor variables. The testing data were 25% of samples for this modeling, a randomly sampled split that is standard modeling practice (performed through the createDataPartition function in the caret package). The true positive rate or sensitivity is a measure of the number of predicted observations that were accurately assigned to a class, while the true negative rate or specificity is a measure of the number of predicted observations where the response variable of interest did not occur.

2.4. Variable reduction for correlation

Seven variable reduction approaches were applied to the 14 input climate variables, covering a full range of options from all (fully correlated) variables, variable reduction, and preselected sets of two variables, which have reduced correlation. Input variables for four models were derived from modeling whereas input variables were predetermined before modeling for three selections: two general variables, two most important variables from prior modeling in the U.S., and all 14 variables (Table 1). In addition, five selections reduced models to two input variables. Variable input approaches for the 14 input climate variables were 1) removing correlated variables before modeling to select only the general variables of mean temperature and annual precipitation ('general 2'), 2) removing correlated variables before modeling to select only mean temperature of the warmest quarter and precipitation of the driest month based on prior knowledge from modeling tree species in the U.S. ('prior 2'; Hanberry, 2023a), 3) modeling with all variables ('all 14'), 4) performing variable reduction in two modeling steps ('recursive feature elimination'), 5) performing variable reduction to keep only the foremost important temperature and precipitation variables for each species ('reduced 2'), 6) modeling with all variables to select the overall most important one temperature variable and one precipitation variable as the inputs for modeling ('foremost 2'), and 7) for comparison, modeling with all variables to select the overall least important one temperature variable and one general precipitation variable as the inputs for modeling ('least 2').

Table 1
Input variables for seven different models of 80 tree species in North America. Four models had fixed one temperature and one precipitation variables, of which two input variables were preselected based on the most general variables or prior modeling for tree species in the United States and two variables were selected after modeling all variables. In contrast, three models used or started with all input variables, and the count is the number of tree species for which the variable is most important.

Two variables	All 14 variables		Recursive feature elimination		Reduced 2	
	Variable	Count	Variable	Count	Variable	Count
General (preselected)	Annual Mean Temperature	8	Annual Mean Temperature	7	Annual Mean Temperature	7
Annual Mean Temperature	Max Temperature Warmest Month	6	Max Temperature Warmest Month	6	Max Temperature Warmest Month	3
Annual Precipitation	Min Temperature Coldest Month	10	Min Temperature Coldest Month	10	Min Temperature Coldest Month	11
Prior (preselected)	Temperature Wettest Quarter	3	Temperature Wettest Quarter	3	Temperature Wettest Quarter	4
Temperature Warmest Quarter	Temperature Driest Quarter	5	Temperature Driest Quarter	3	Temperature Driest Quarter	5
Precipitation Driest Month	Temperature Warmest Quarter	6	Temperature Warmest Quarter	7	Temperature Warmest Quarter	6
Foremost	Temperature Coldest Quarter	34	Temperature Coldest Quarter	36	Temperature Coldest Quarter	33
Temperature Coldest Quarter	Annual Precipitation	1	Annual Precipitation	1	Annual Precipitation	1
Precipitation Warmest Quarter	Precipitation Driest Month	1	Precipitation Driest Month	1	Precipitation Driest Month	2
Least	Precipitation Driest Quarter	2	Precipitation Driest Month	2	Precipitation Driest Month	2
Temperature Wettest Quarter	Precipitation Warmest Quarter	3	Precipitation Warmest Quarter	3	Precipitation Warmest Quarter	5
Precipitation Wettest Month	Precipitation Coldest Quarter	1	Precipitation Coldest Quarter	1	Precipitation Coldest Quarter	1
	Precipitation Wettest Month	0	Precipitation Wettest Month	0	Precipitation Wettest Month	0
	Precipitation Wettest Quarter	0	Precipitation Wettest Quarter	0	Precipitation Wettest Quarter	0

Regarding 'recursive feature elimination', I performed variable reduction in two modeling steps. Firstly, recursive feature elimination, which builds a model based on all variables, rank-orders variable importance, and removes variables with the least importance based on model evaluation metrics (i.e., accuracy) for resampled data, reduced the number of variables to at most half of the initial number of 14 variables. Then, the modeling process is repeated again, retaining only variables that carried an importance value ≥40 (out of 100 scale; note that other classifiers will assign weights differently). For 'reduced 2', recursive feature elimination modeling was applied to determine the foremost important temperature and precipitation variables for each species. For 'foremost 2' and 'least 2' (as a baseline), modeling occurred with all variables to select the overall most and least important one temperature variable and one precipitation variable.

2.5. Comparisons

To assess effects of correlation, I compared sensitivity and specificity for accuracy, importance of variables for stability and temporal transferability, and mapped models relative to realized distributions, which were kernels that contained 95% of observations, for accuracy and transferability. For species samples in models with two input variables, I calculated correlation and number of variable pairs that exceeded $r = 0.4$ to $r = 0.8$, in six bins. For models with more than two input variables, I assessed stability of the identified foremost important variables for all 80 tree species. As a check of variable stability over time, I modeled current tree samples with three future climates and ranked the overall importance of variables by summing the variable importance scores for all tree species. To isolate the effects of two variable models with high correlation relative to the low correlation approaches, I also provided mapped examples of a two variable model that contained only highly correlated variables, of annual temperature and temperature of the coldest quarter, but did not present this unrealistic option as one of the variable reduction approaches.

3. Results

3.1. Greatest accuracy and similar mapped models occurred with all 14 input variables, using current or future climate

Model accuracy, for 80 tree species, was greatest when informed by all input variables (Fig. 1). Models with all 14 variables had the greatest mean accuracy, at 0.985 sensitivity and 0.968 specificity, while models from variable reduction of two steps including recursive feature elimination to a mean of 3.38 variables had the next greatest mean accuracy, at 0.980 sensitivity and 0.960 specificity. Models of current observations with all 14 variables from each of the three future climates also matched greatest accuracies of the models with all 14 current climate variables, at 0.986 to 0.987 sensitivity and 0.965 to 0.968 specificity. Additionally, mapped models of current species observations were similar with use of current or future climates (Fig. 2).

3.2. Greater accuracy for models with the two foremost important variables than the two least important variables

For two input variables, mean accuracy also was high, but remained closer to the greatest mean accuracy of the all variable models if informed by most important variables of the all variable models. The two variable models that made use of information about the most important two variables from either the model with all 14 variables (i.e., 'foremost 2' or two variables of temperature of the coldest quarter and precipitation of the warmest quarter) or through recursive feature elimination to two variables ('reduced 2' variables varied by species) had equivalent accuracies, at 0.97 sensitivity and 0.95 specificity. Model inputs of general temperature and precipitation also had 0.97 sensitivity but 0.94 specificity. Accuracy for model inputs from prior modeling information was most similar to the model inputs from the least important variables, which were temperature of the wettest quarter and precipitation of the wettest month identified from modeling all 14 variables, at 0.94 sensitivity and 0.92 specificity.

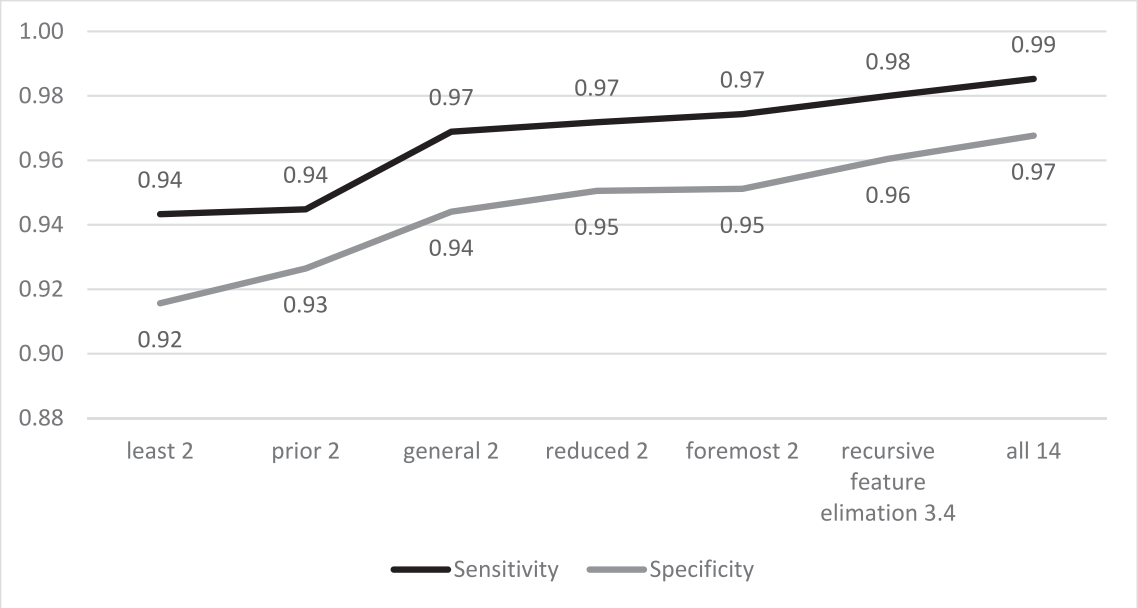

Fig. 1. Accuracy of different input variables for 80 tree species. Number in the x-axis labels indicates the number of input variables. General 2 = preselected general fixed variables of mean temperature and annual precipitation, prior 2 = preselected fixed mean temperature of the warmest quarter and precipitation of the driest month based on prior modeling in the United States, all 14 variables for both current climate and future climate predictors, recursive feature elimination 3.4 = variable reduction in two modeling steps that resulted in a mean number of 3.4 variables, reduced 2 = variable reduction to retain the foremost important temperature and precipitation variables for each species, foremost 2 = fixed overall most important one temperature variable and one precipitation variable based on modeling of all variables, least 2 = fixed overall least important one temperature variable and one precipitation variable based on modeling of all variables.

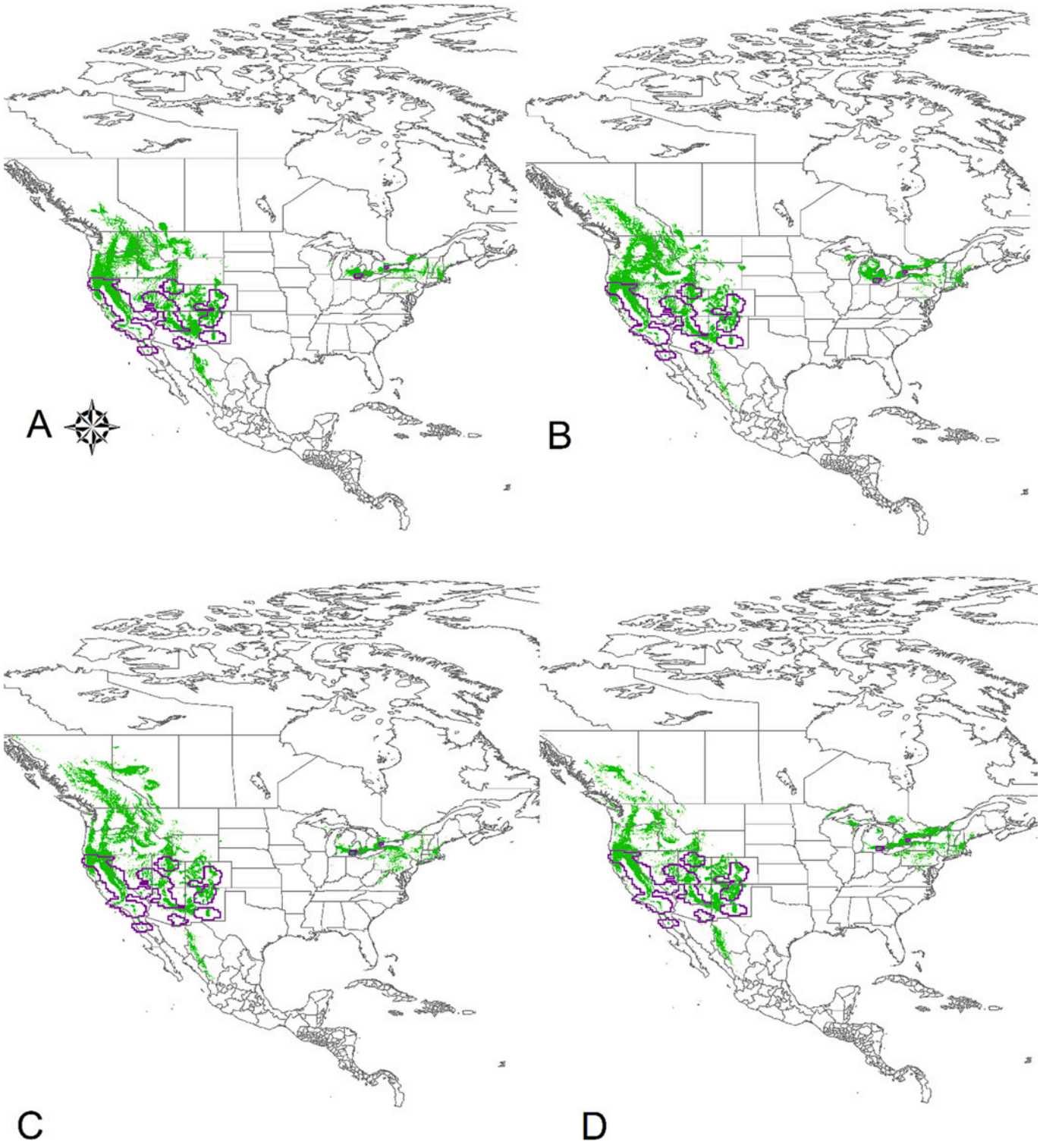

Fig. 2. Similar mapped models of *Abies concolor* (light green) that approximate the patchy realized distribution (purple outline is the kernel that contains 95% of observations) for all 14 current climate predictor variables (A) and all 14 future climate predictor variables from general circulation models of GFDL-ESM4 (B), IPSL-CM6A-LR (C), and UKESM1–0-LL (D). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.3. Evident mapped error for two input variable models

Despite high accuracy, inaccurate predictions in two variable models may be apparent when mapped in comparison to kernels of observed locations (Figs. 3 and 4). In alphabetical order, the distributions of *Abies concolor* and *Arbutus xalapensis* were not constrained by temperature of the warmest quarter in the 'prior' input variables. Conversely, the all variable model was able to limit predictions to approximate the patchy realized distributions of the two tree species, although climate alone typically does not limit distributions to exact ranges.

Indeed, an equivalent contrast of a two variable model that contained only highly correlated variables, of annual temperature and temperature of the coldest quarter, was better able to delineate realized boundaries of kernels than the relatively uncorrelated temperature and

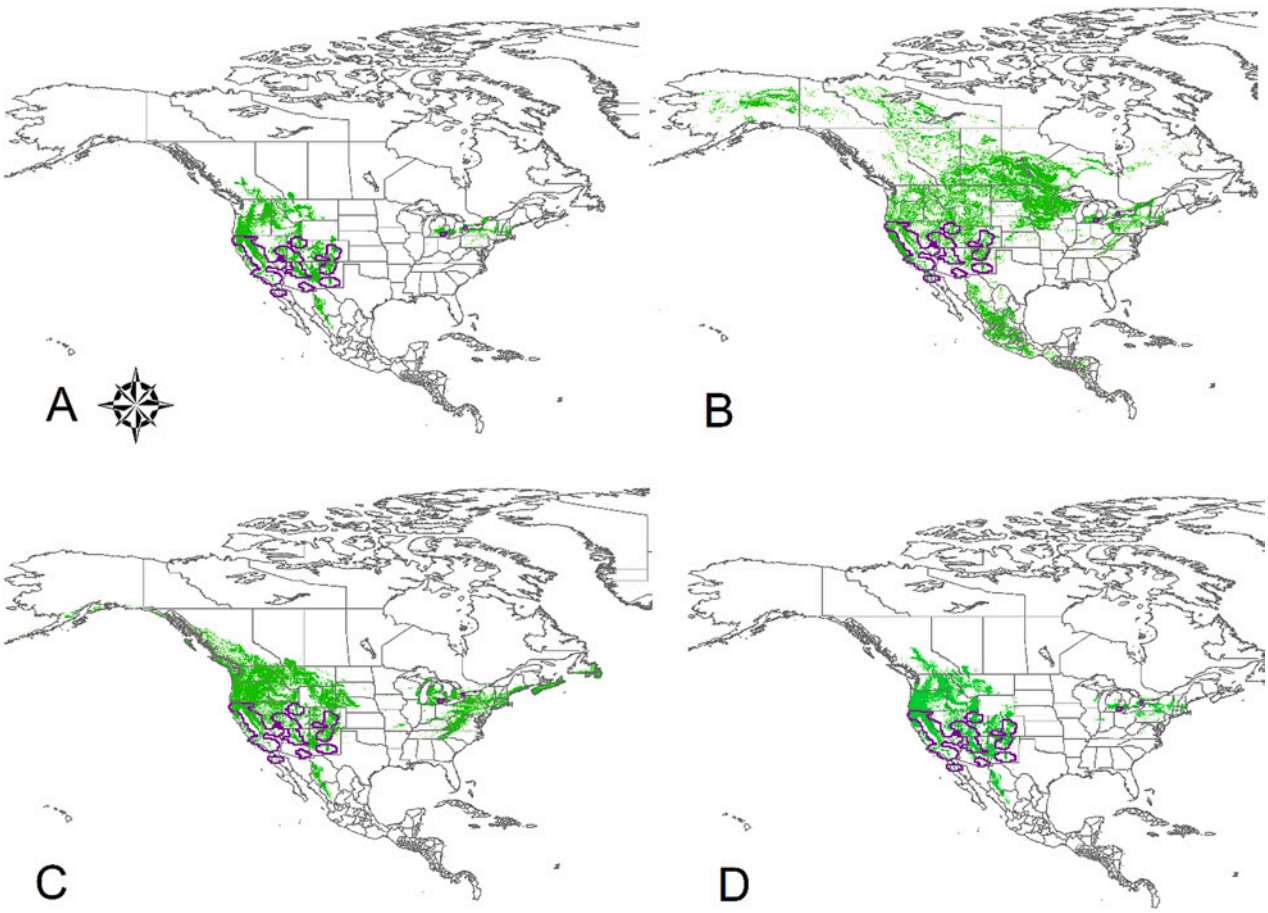

Fig. 3. Mapped models for *Abies concolor* (light green) compared to the realized distribution (purple outline is the kernel that contains 95% of observations) from the all variable model (A), the prior two input variables ($r = 0.19$; B), two highly correlated temperature variables ($r = 0.97$; C), and thinned samples with all input variables (D). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

precipitation variables of the prior input model (Figs. 3 and 4). Variable selection to two highly correlated temperature variables was not included as one of the seven variable selection approaches because reduction to only two highly correlated temperature variables, excluding precipitation variables, is not likely to be recommended. However, this comparison of feature selection, while holding all else equal, demonstrated that two highly correlated variables can be more effective than two uncorrelated variables for predictive models.

3.4. Robust identification of important correlated variables by random forests

The most important variables, or foremost variables for the 80 species, identified by all variable models remained very similar to most important variables identified by the two models through reduction ('recursive feature elimination' and 'reduced 2'; Table 1). Most critically, the first four variables had the same rank. These three models that started with more than two input variables all agreed that temperature of the coldest quarter was by far the most important variable, for 33 to 36 of the 80 total species. The three models concurred that minimum temperature of the coldest month was the next most important model, for 10 to 11 species. Annual mean temperature and temperature of the warmest quarter were the thirdmost and fourthmost important variables, respectively.

Likewise, modeling current occurrences of the 80 species with three future climate projections demonstrated stability in isolation of the most important variables, from ranked sums of importance values (Table 2). The first five most important variables had the same rank among the four climates, albeit the second and third most important variables were reversed between near current climate and the three future climates. The first five most important variables were all temperature variables, which are highly correlated.

3.5. High correlation can occur in species samples between two uncorrelated input variables

Regarding correlation in the 80 species samples in the five models with one temperature and one precipitation variable, when all the correlated pairs were combined (i.e., 400 pairs), mean (absolute) correlation was 0.4 and the maximum correlation was 0.86. Ten pairs, or 2.5% of samples, had an $r \geq 0.8$, whereas 36 pairs, or 9% of samples, had an $r \geq 0.7$, while 53% of pairs had an $r \geq 0.4$. The 'prior 2' input variable model was the only model that did not have any variable pairs that exceeded $r \geq 0.7$. For entire layers of variables, 'general' variables of mean temperature and annual precipitation had a correlation of 0.61, 'prior' variables of temperature of the warmest quarter and precipitation of the driest month had a correlation of 0.49, 'foremost' variables of temperature of the coldest quarter and precipitation of the warmest quarter had a correlation of 0.54, and 'least' variables of temperature of the wettest quarter and precipitation of the wettest month had a correlation of 0.41 (ESRI ArcMap 10.8.2, Redlands, California).

3.6. Models with thinned samples similar to unthinned samples

For thinned samples, accuracy and variable importance, along with mapped models (Figs. 3 and 4), remained similar to unthinned samples.

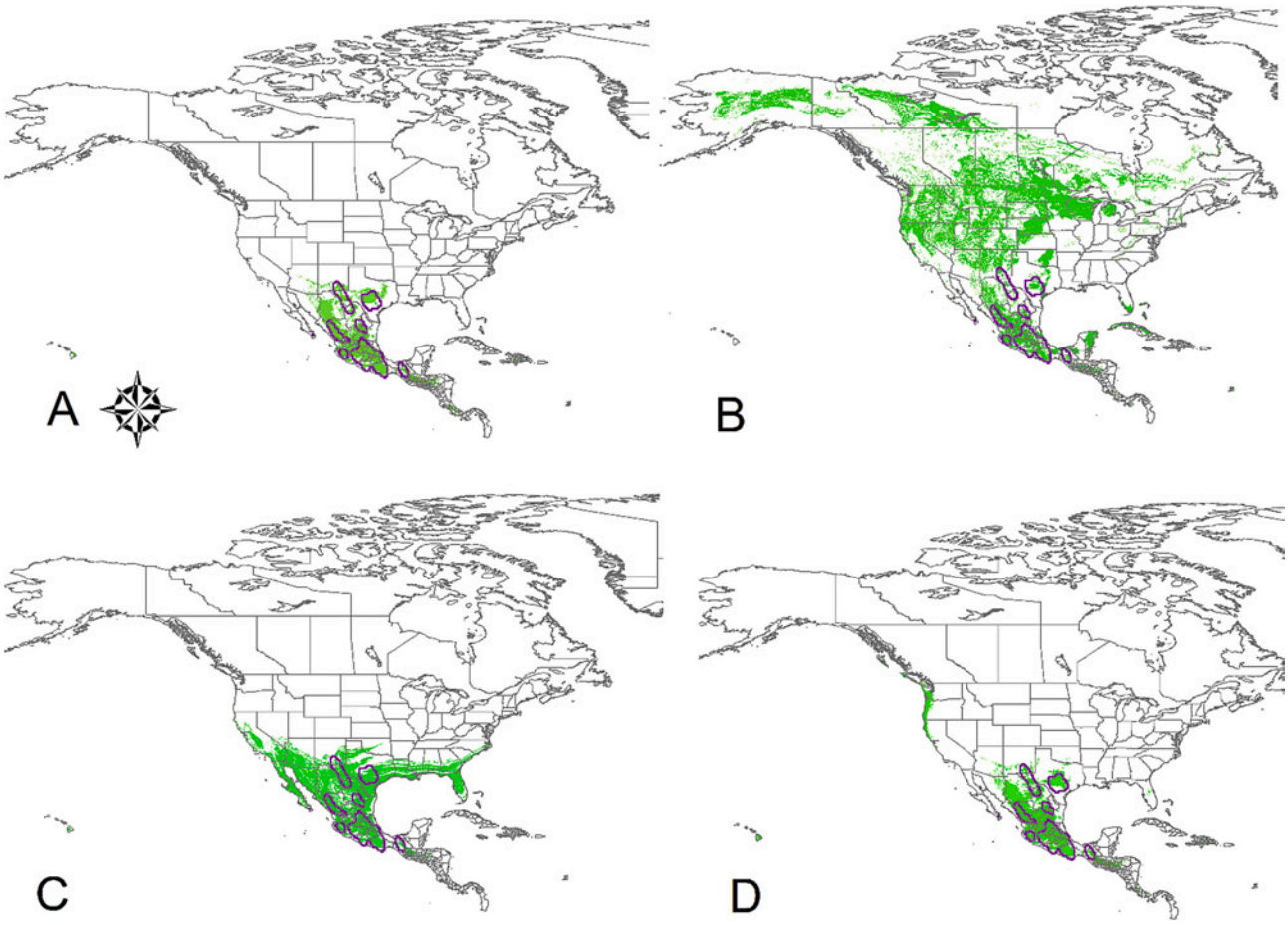

Fig. 4. Mapped models for *Arbutus xalapensis* (light green) compared to the realized distribution (purple outline) from the all variable model (A), the prior two input variables (r = 0.34; B), two highly correlated temperature variables (r = 0.97; C), and thinned samples with all input variables (D). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Ranked sums of importance values for 14 climate variables of models under current climate (1900 to 1990) and three future climates (2071–2100), for 80 tree species in North America. GFDL = GFDL-ESM4; Geophysical Fluid Dynamics Laboratory, USA; IPSL = IPSL-CM6A-LR; Institut Pierre Simon Laplace, France; UKESM1 = UKESM1–0-LL (Met Office Hadley Centre, UK.

	Current	GFDL	IPSL	UKESM1
Temperature Coldest Quarter	1	1	1	1
Annual Mean Temperature	2	3	3	3
Min Temperature Coldest Month	3	2	2	2
Temperature Driest Quarter	4	4	4	4
Temperature Warmest Quarter	5	5	5	5
Max Temperature Warmest Month	6	6	7	9
Temperature Wettest Quarter	7	7	6	6
Precipitation Driest Month	8	8	8	7
Annual Precipitation	9	11	11	11
Precipitation Driest Quarter	10	9	9	8
Precipitation Wettest Month	11	14	14	13
Precipitation Wettest Quarter	12	13	13	14
Precipitation Warmest Quarter	13	10	10	10
Precipitation Coldest Quarter	14	12	12	12

For models of 78 tree species that met the new minimum threshold of 400 observations with all climate predictor variables, compared to models of 80 tree species that met the minimum threshold of 1000 observations with all climate predictor variables, mean sensitivity remained the same and mean specificity was slightly reduced to 0.96 from 0.97. If the two species that did not meet the new minimum threshold of 400 observations were removed from comparison, then overall accuracy was nearly identical at 0.967 for thinned samples compared to 0.968 for all samples. Equally, ranking of summed variable importance was almost identical, with a reversal between the eighth and ninth important variables and a different ordering of the tenth, twelfth, and thirteenth important variables, when 78 models of thinned samples were compared with either 78 or 80 models with all samples.

4. Discussion

4.1. Key findings

Species distribution models contain bias from predictors and species occurrences that supply the models. Collinearity has received elaborate attention, stemming from large standard errors of estimates of the coefficients for linear regression models. This is not an issue for predictive modeling, but secondary concerns are that correlated predictor variables may interfere with measuring the importance of variables and that relationships among correlated predictors may differ through time and space. To test if relationships hold true over time, current and future climate data, which contain highly correlated variables, are available to be modeled with current species occurrences. In this study with a North American extent, temperature variables of the near current climate and three end-of-century climates had a mean r value of 0.9, excluding the variable of mean temperature of the wettest quarter for all four climates. Nonetheless, for models of current observations from 80 tree species by the random forests classifier, correlated predictor variables did not interfere with measuring the importance of variables and relationships among predictors remained similar under current and future climates.

Robustness in important variable identification and temporal transferability with correlated climate variables was revealed by 1) greatest accuracy in predictions to withheld training data for models with all 14 input variables, whether with use of current climate or future climate predictors, 2) greater accuracy for models with the two foremost important variables than the two least important variables, as informed by variable importance rankings in all variable models, 3) models with more than two input variables agreed on the four most important variables, 4) the same five most important variables of highly correlated temperature derivatives were identified under current climate and future climates for models with all 14 input variables, and 5) mapped models of current species observations were similar with current climate predictors or future climate predictors. In contrast after feature selection, omitted variable bias was revealed by unconstrained mapped distributions from modeling with only one temperature variable, temperature of the warmest quarter, which was preselected. Although still accurate in predicting to reserved testing samples, models that used two variables consisting of one temperature and one precipitation predictor to limit correlation (i.e., $r \leq 0.61$) were less accurate and accuracy decreased with preselected variables, that is, if modeling did not inform variable selection. Collinearity remained between uncorrelated one temperature and one precipitation variables in species samples, as 9% of samples had an $r \geq 0.7$. These results of accurate predictions and identification of important variables with correlated predictor variables agree with other publications, including statistical textbooks (Chatterjee et al., 2000; Genuer et al., 2010; Goldberger and Goldberger, 1991; Kiers and Smilde, 2007; Shmueli, 2010).

4.2. Correlated climate variables compared to spurious variables

With multiple correlated input variables, accuracy was greatest, mapped models approximated realized species ranges, and stable identification of the same important variables occurred. Models from the two foremost important variables, which were identified by the all variable model, had greater accuracy in predictions than the two least important variables. Moreover, the same four most important variables were identified by models that started with more than two variables: all variable models, recursive feature elimination to a mean of 3.38 input variables, and recursive feature elimination to one temperature and one precipitation input variable. Genuer et al. (2010) also found that equivalent orders of important variables appeared in models from highly correlated data.

Correlated future climate predictor variables modeled current species distributions similarly to current climate predictors, even though species have not been exposed to future climate. Accuracy in predictions was the same for all variable models under current climate as under the three future climates. The same five most important variables were identified for models under near current climate and the three projected climates, albeit the order of the secondmost and thirdmost variable was reversed. Mapped species distributions were similar using any of the four climates as well. These results aligned with results from comparison of current and past climates, using different sources of species observations and climate data (Hanberry, 2023b). Climate data retain strong correlated structure over time; therefore, collinearity in climate predictor variables is not likely to result in poor predictions when trained on data from one time and predicted to another, all else being held equal (e.g., same variables). Constant structure is perhaps reinforced by bias correction methods for modeled climate data (Hui et al., 2020; Nahar et al., 2017).

The presence of multicollinearity in climate variables did not affect relationships with species occurrences when extrapolated to future climate because climate predictor variables follow similar spatial patterns under future climate as in near current climate. Therefore, models can take advantage of the spatial structure to understand species distributions, which have long-standing relationships with climate, e.g., spruce associations with boreal climate (Williams et al., 2004). The correlated structure among predictor variables may indicate enduring non-random spatial patterns, which, rather than randomness, maintains species associations over time and makes climate more reliable than other predictors for forecasting or hindcasting species distribution models. As described by Hawkins (2012:2), 'the existence of spatial autocorrelation is not bias or artefact, it is what biogeographers and geographical ecologists want to understand. Given the spatially structured distributions of species arising from biological processes operating in a spatially patterned environment, any set of samples or representations of nature must also contain this structure if it is accurate.'

Rather than focusing on correlated variables, variables that have spurious associations with species are problematic in predicting to space and time. Chatterjee et al. (2000:235) state: 'To forecast we must be confident that the character and strength of the overall relationship will hold into future periods. This matter of confidence is a problem in all forecasting models whether or not multicollinearity is present.' If the association is transient, then the variable cannot be predictive. For example, substrates, topography, elevation, and biotic interactions are used to model species. While these variables may appear to have strong associations with current species distributions, it is important to realize that before human land use and disturbance, a completely different suite of species and vegetation structures likely occurred under the same substrates and topography, resulting in different species associations (Hanberry et al., 2013). Near complete turnover in the historical ecosystems of the eastern U.S. has occurred due to surface fire exclusion and chronic overstory tree removal (Bragg et al., 2020). Generally, most soils and topography are appropriate to support tree species, although tree species may not be competitive under certain conditions, whereas disturbances and human land uses directly cause tree mortality. Additionally, non-native species in the process of expansion may have chance associations that change as their distributions develop. Depending on length of associations, two models (e.g., 'here and now' predictor variables and climate-only predictor variables from complete native and non-native ranges) may be required rather than one model of prediction from one interval to another or else unreliable constraints will be incorporated in predictions, causing inaccuracies.

4.3. Variable selection through random forests modeling

The random forests classifier, which is designed to detect boundaries (Breiman, 2000), specifically can isolate important correlated predictors, as evidenced through constancy of important variable selection. Similar identification and ranking of important variables ensued, whether for 14 input variables or reduced to two input variables through feature selection by modeling and whether for current or future climate variables. Accuracy decreased if modeling did not inform variable selection, because starting from preselected variables reduced the explanatory power of modeling. These results support that this classifier is robust in isolating influence of redundant, highly correlated predictors (Genuer et al., 2010; Kuhn and Johnson, 2019).

Interpretation of important variables may require consideration of all variables correlated with the important variables (Dormann et al., 2013). Nevertheless, application of many variables will result in numerous rules for fitting the model, which may be too specific (i.e., overfitting through high variance and little bias) for the species and reduce interpretability. Overfitting by random forests did not occur with 14 climate predictor variables based on generalizing in time, accuracy in fitting the test set, and mapping potential climate space beyond realized ranges.

4.4. Standard guidelines and guidelines based on predictive modeling

Standard recommendations are to select ecologically relevant variables that are not highly correlated (Araújo et al., 2019; Fourcade et al., 2018; Santini et al., 2021). However, reduction of correlated climate variables was costly, with increasing costs when variable selection was

not informed by modeling, which was apparent upon visualization of models that used only temperature of the warmest quarter compared to realized species distributions from occurrence samples. Modeling with one or two uncorrelated climate variables after variable selection may result in inaccurate models and loss of information extracted from modeling, which is a less desirable outcome than accurate models with retention of correlated variables.

If the issue of correlation remains worth removing predictor variables, then decisions are required to remove highly correlated variables (Araújo et al., 2019; Dormann et al., 2013). Nonetheless, correlation is continuous, and multicollinearity is a matter of magnitude, not a matter of presence or absence. Complete elimination of multicollinearity is not a simple matter of removing variables before modeling. For this study, despite that the two input variables for two variable models did not exceed 0.6 in correlation, samples for some species contained collinearity. Setting a threshold of 0.8 is practical in being likely to avoid correlation in most samples, when adhering to standard guidelines regarding removal of correlated variables.

Regarding ecological relevance, maximum temperature of the warmest month and mean temperature of the warmest quarter may appear to be the most critical variables to retain under warming climate change and in warmer regions, such as Mexico. Maximum temperatures that exceed an optimal physiological range can be detrimental to regeneration, growth, and mortality of species. Under future climate change, heat wave temperatures may exceed a month in length (Hanberry, 2022), making three-month measurements relevant. Based on prior knowledge from modeling tree species in the U.S., mean temperature of the warmest quarter and precipitation of the driest month were important variables ('prior 2'; Hanberry, 2023a). Moreover, previous work and concepts support the idea that warm temperatures are more influential in warm regions, whereas cold temperatures are more influential in cold regions (Deutsch et al., 2008). Boreal species have traits for frost tolerance whereas species in lower latitudes have traits for growth and competition (Loehle, 1998). Temperate and boreal species may be in climates that are currently cooler than their physiological optima, while tropical species may be close to thermal thresholds (Deutsch et al., 2008).

However, selecting ecologically relevant variables for modeling presumes that all modeling results can be anticipated, but modeling space is different than ecological space. Instead of warm temperatures, modeling identified that temperature of the coldest quarter was most important overall for this set of 80 tree species present in Mexico. For models, cold temperatures were more effective at describing species distributions located in warmer regions and warm temperatures were more effective at defining species distributions located in cooler regions. Furthermore, for modeling, warm temperatures alone created predictions that did not limit species to locations based on year-round temperature and instead predicted potential climate in seasonally appropriate regions, resulting in evident omitted variable bias in mapped models. Temperature variables particularly are correlated, but multiple variables such as maximum and minimum temperatures may be ecologically relevant and necessary information for accurate models.

In contrast to standard guidelines, practical guidelines based on results of predictive modeling are to retain important variables that achieve greatest modeling accuracy, which may be highly correlated variables (Breiman, 2001). A guide to retain correlated variables and avoid omitted variable bias may be to first remove the least important third of variables, and then convey species-specific models by selecting based on importance values by species, which will result in a different number of variables, to isolate variables most relevant for a species. Nonetheless, modeling with less than a full suite of climate variables may result in a unique combination of climate for ranges of some species that may generate unusually limited or expansive extents in the future. While this may indicate future ranges, climate limitations on species distributions remain unknown and the model simply is trying to best approximate species samples. For atypical predictions, a full suite of input variables may be desirable to represent an alternative prediction. True physiological tolerances of species are unspecified along with an unknown ability of species to compete against other species under future climate and land use and disturbance change, which is information not captured by current species samples. Although modeling is available to answer questions about modeling decisions, species predictions may be useful as an indicator of change over time but not likely to be correct.

4.5. Sample thinning

Sample thinning had very little effect on accuracy or variable importance, or mapped models, indicating no clear benefits for tampering with samples that may provide information about species characteristics and responses to the environment. It is not necessary to add another modeling step that involves many choices, although thinning may be beneficial if thousands of samples were taken from one study site. An outcome of reducing samples is that minimum sample size thresholds need to be lowered to maintain the same species pool. Sample thinning does provide a confirmation that sample size is not as critical as capturing ecological amplitude for species distribution.

5. Conclusions

Species distribution models are widely generated, creating predictions of potential current and future distributions that may be useful. Generally, model inaccuracies arise from the species samples and variables, including variable selection, that supply the models rather than relatively minor modeling concerns, such as unthinned species samples or correlated climate variables. Future climate projections and models of past climate are highly correlated with current climate; high correlation particularly among temperature variables is expected in any climate dataset. To alleviate concerns about effects of collinearity on identification of important variables and temporal transferability, models with current or future climate and the random forests classifier demonstrated that retention of correlated climate variables was beneficial due to robust identification of important variables, consistent identification of important variables in models, and approximated realized distributions. Conversely, mapped models exhibited the consequences of avoiding highly correlated variables, based on evident bias in mapped models from two preselected variables. Collinearity was not an issue for climate variables, which have persistent associations in time and space, and without clear support for altering samples, thinning of samples does not provide benefits. Following a standard guideline to remove correlated variables before modeling may result in incorrect omission of relevant variables for species distribution models. Rather than preset guidelines, predictive modeling is a practical approach for making modeling decisions, including variable selection.

CRediT authorship contribution statement

Brice B. Hanberry: Conceptualization, Formal analysis, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The author declares no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All data are publicly available.

Acknowledgements

This research was supported by the USDA Forest Service, Rocky Mountain Research Station. The findings and conclusions in this publication are those of the authors and should not be construed to represent any official USDA or U.S. Government determination or policy.

References

Aiello-Lammens, M.E., Boria, R.A., Radosavljevic, A., Vilela, B., Anderson, R.P., 2015. spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. Ecography 38, 541–545.

Araújo, M.B., Anderson, R.P., Márcia Barbosa, A., Beale, C.M., Dormann, C.F., Early, R., Garcia, R.A., Guisan, A., Maiorano, L., Naimi, B., O'Hara, R.B., 2019. Standards for distribution models in biodiversity assessments. Science. Advances 5 (1), eaat4858.

Arenas-Castro, S., Sillero, N., 2021. Cross-scale monitoring of habitat suitability changes using satellite time series and ecological niche models. Sci. Total Environ. 784, 147172.

Bragg, D.C., Hanberry, B.B., Hutchinson, T.F., Jack, S.B., Kabrick, J.M., 2020. Silvicultural options for open forest management in eastern North America. For. Ecol. Manag. 474, 118383.

Breiman, L., 2000. Some infinity theory for tree ensembles. Available at https://www.stat.berkeley.edu/users/breiman/some_theory2000.pdf Accessed 27 October 2023.

Breiman, L., 2001. Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat. Sci. 16 (3), 199–231.

Chatterjee, S., Hadi, A.S., Price, B., 2000. Regression Analysis by Example, Third ed. John Wiley and Sons https://archive.org/details/regressionanalys0000chat_q4i3/page/234/mode/2up?q=multicollinearity.

Deutsch, C.A., Tewksbury, J.J., Huey, R.B., Sheldon, K.S., Ghalambor, C.K., Haak, D.C., Martin, P.R., 2008. Impacts of climate warming on terrestrial ectotherms across latitude. Proc. Natl. Acad. Sci. 105, 6668–6672.

Díaz-Uriarte, R., Alvarez de Andrés, S., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinform. 7, 1–13.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36 (1), 27–46.

Faurby, S., Araújo, M.B., 2018. Anthropogenic range contractions bias species climate change forecasts. Nat. Clim. Chang. 8, 252–256.

Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn. Res. 15 (1), 3133–3181.

Fourcade, Y., Besnard, A.G., Secondi, J., 2018. Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. Glob. Ecol. Biogeogr. 27 (2), 245–256.

Genuer, R., Poggi, J.M., Tuleau-Malot, C., 2010. Variable selection using random forests. Pattern Recogn. Lett. 31 (14), 2225–2236.

Global Biodiversity Information Facility [GBIF], 2023. Free and Open Access to Biodiversity Data. Available at: www.gbif.org. Accessed 17 March 2023.

Goldberger, A.S., Goldberger, A.S.G., 1991. A COURSE in econometrics. Harvard University Press. Available at https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxlY29ub21ldHJpY3NpdGFtfGd4OjY5NmZiMDEwMzMwOTQyZTk accessed 3 March 2023.

Goldstein, B.A., Polley, E.C., Briggs, F.B., 2011. Random forests for genetic association studies. Stat. Appl. Genet. Mol. Biol. 10 (1), 32.

Goring, S.J., Williams, J.W., 2017. Effect of historical land-use and climate change on tree-climate relationships in the upper Midwestern United States. Ecol. Lett. 20 (4), 461–470.

Hanberry, B.B., 2021. Timing of tree density increases, influence of climate change, and a land use proxy for tree density increases in the eastern United States. Land 10, 1121.

Hanberry, B.B., 2022. Global population densities, climate change, and the maximum monthly temperature threshold as a potential tipping point for high urban densities. Ecol. Indic. 135, 108512.

Hanberry, B.B., 2023a. Effectiveness of land use and disturbance measures, compared to climate, soil, and topography, for modeling relative abundances of tree species in the eastern United States. Eco. Inform. 102110.

Hanberry, B.B., 2023b. Shifting potential tree species distributions from the last glacial maximum to the mid-Holocene in North America, with a correlation assessment. J. Quat. Sci. 38, 829–839.

Hanberry, B.B., Palik, B.J., He, H.S., 2013. Winning and losing tree species of reassembly in Minnesota's mixed and broadleaf forests. PLoS One 8, e61709.

Hanberry, B.B., Kabrick, J.M., He, H.S., 2014. Changing tree composition by life history strategy in a grassland-forest landscape. Ecosphere 5 (3), 1–16.

Hawkins, B.A., 2012. Eight (and a half) deadly sins of spatial analysis. J. Biogeogr. 39, 1–9.

Hijmans, R.J., Phillips, S., Leathwick, J., Elith, J., 2011. Package 'dismo'. Available at: http://cran.r-project.org/web/packages/dismo/index.html. Accessed 30 March 2023.

Hui, Y., Xu, Y., Chen, J., Xu, C.Y., Chen, H., 2020. Impacts of bias nonstationarity of climate model outputs on hydrological simulations. Hydrol. Res. 51 (5), 925–941.

Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services [IPBES], 2019. Global assessment report on biodiversity and ecosystem services of the intergovernmental science-policy platform on biodiversity and ecosystem services. IPBES Secretariat. https://doi.org/10.5281/zenodo.3831673. Available at.

Karger, D.N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W., Zimmermann, N.E., Linder, H.P., Kessler, M., 2017. Climatologies at high resolution for the earth land surface areas. Sci. Data 4, 170122.

Karger, D.N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W., Zimmermann, N.E., Linder, H.P., Kessler, M., 2018. Data from: climatologies at high resolution for the earth's land surface areas. EnviDat. https://doi.org/10.16904/envidat.228.v2.1.

Karger, D.N., Nobis, M.P., Normand, S., Graham, C.H., Zimmermann, N.E., 2020. CHELSA-TraCE21k: downscaled transient temperature and precipitation data since the last glacial maximum. EnviDat. https://doi.org/10.16904/envidat.211.

Karger, D.N., Nobis, M.P., Normand, S., Graham, C.H., Zimmermann, N.E., 2021. CHELSA-TraCE21k v1. 0. Downscaled transient temperature and precipitation data since the last glacial maximum. Clim. Past Discuss. 1–27.

Kiers, H.A., Smilde, A.K., 2007. A comparison of various methods for multivariate regression with highly collinear variables. JISS 16, 193–228.

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28, 1–26.

Kühn, I., Dormann, C.F., 2012. Less than eight (and a half) misconceptions of spatial analysis. J. Biogeogr. 39 (5), 995–998.

Kuhn, M., Johnson, K., 2019. Feature Engineering and Selection: A Practical Approach for Predictive Models. CRC Press. Available at https://bookdown.org/max/FES/feature-selection-simulation.html Accessed 6 July 2022.

Leitão, P.J., Santos, M.J., 2019. Improving models of species ecological niches: a remote sensing overview. Front. Ecol. Evol. 7, 9.

Lindner, T., Puck, J., Verbeke, A., 2020. Misconceptions about multicollinearity in international business research: identification, consequences, and remedies. J. Int. Bus. Stud. 51, 283–298.

Loehle, C., 1998. Height growth rate tradeoffs determine northern and southern range limits for trees. J. Biogeogr. 25 (4), 735–742.

Lopez-Cantu, T., Prein, A.F., Samaras, C., 2020. Uncertainties in future US extreme precipitation from downscaled climate projections. Geophys. Res. Lett. 47 (9), e2019GL086797.

Nahar, J., Johnson, F., Sharma, A., 2017. Assessing the extent of non-stationary biases in GCMs. J. Hydrol. 549, 148–162.

R Core Team, 2023. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Regos, A., Gonçalves, J., Arenas-Castro, S., Alcaraz-Segura, D., Guisan, A., Honrado, J.P., 2022. Mainstreaming remotely sensed ecosystem functioning in ecological niche models. Remote Sens. Ecol. Conserv. 8 (4), 431–447.

Santini, L., Benítez-López, A., Maiorano, L., Čengić, M., Huijbregts, M.A., 2021. Assessing the reliability of species distribution projections in climate change research. Divers. Distrib. 27 (6), 1035–1050.

Shmueli, G., 2010. To explain or to predict? Stat. Sci. 25, 289–310.

Sofaer, H.R., Jarnevich, C.S., Pearse, I.S., Smyth, R.L., Auer, S., Cook, G.L., Edwards Jr., T.C., Guala, G.F., Howard, T.G., Morisette, J.T., Hamilton, H., 2019. Development and delivery of species distribution models to inform decision-making. BioScience 69 (7), 544–557.

Titeux, N., Henle, K., Mihoub, J.B., Regos, A., Geijzendorffer, I.R., Cramer, W., Verburg, P.H., Brotons, L., 2016. Biodiversity scenarios neglect future land-use changes. Glob. Chang. Biol. 22, 2505–2515.

Varela, S., Anderson, R.P., García-Valdés, R., Fernández-González, F., 2014. Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. Ecography 37 (11), 1084–1091.

Vatcheva, K.P., Lee, M., McCormick, J.B., Rahbar, M.H., 2016. Multicollinearity in regression analyses conducted in epidemiologic studies. Epidemiology (Sunnyvale, Calif.) 6 (2).

Vollering, J., Halvorsen, R., Auestad, I., Rydgren, K., 2019. Bunching up the background betters bias in species distribution models. Ecography 42 (10), 1717–1727.

Williams, M.N., Grajales, C.A.G., Kurkiewicz, D., 2013. Assumptions of multiple regression: correcting two misconceptions. Pract. Assess. Res. Eval. 18 (1), 11.

Williams, J.W., Shuman, B.N., Webb III, T., Bartlein, P.J., Leduc, P.L., 2004. Late-Quaternary vegetation dynamics in North America: scaling from taxa to biomes. Ecol. 74 (2), 309–334.

WWF, 2018. In: Grooten, M., Almond, R.E.A. (Eds.), Living Planet Report - 2018: Aiming Higher. WWF, Gland, Switzerland.

Youden, W.J., 1950. Index for rating diagnostic tests. Cancer 3, 32–35.

Yule, G.U., 1912. On the methods of measuring association between two attributes. J. R. Stat. Soc. 75 (6), 579–652.

Zurell, D., Franklin, J., König, C., Bouchet, P.J., Dormann, C.F., Elith, J., Fandos, G., Feng, X., Guillera-Arroita, G., Guisan, A., Lahoz-Monfort, J.J., et al., 2020. A standard protocol for reporting species distribution models. Ecography 43, 1261–1277.