Author/s:
Valavi, R;Guillera-Arroita, G;Lahoz-Monfort, JJ;Elith, J

Title:
Predictive performance of presence-only species distribution models: a benchmark study with reproducible code

Date:
2022-02-01

Citation:
Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J. & Elith, J. (2022). Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. Ecological Monographs, 92 (1), https://doi.org/10.1002/ecm.1486.

Persistent Link:
https://hdl.handle.net/11343/308040

Ecological Monographs, 92(1), 2022, e01486

Predictive performance of presence-only species distribution models: a benchmark study with reproducible code

Roozbeh Valavi [1,3], Gurutzeta Guillera-Arroita [2], José J. Lahoz-Monfort [2], and Jane Elith [2]

[1]*School of Biosciences, University of Melbourne, Parkville, Victoria 3010 Australia*
[2]*School of Ecosystem and Forest Sciences, University of Melbourne, Parkville, Victoria 3010 Australia*

Citation: Valavi, R., G. Guillera-Arroita, J. J. Lahoz-Monfort, and J. Elith. 2022. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. Ecological Monographs 92(1):e01486. 10.1002/ecm.1486

Abstract. Species distribution modeling (SDM) is widely used in ecology and conservation. Currently, the most available data for SDM are species presence-only records (available through digital databases). There have been many studies comparing the performance of alternative algorithms for modeling presence-only data. Among these, a 2006 paper from Elith and colleagues has been particularly influential in the field, partly because they used several novel methods (at the time) on a global data set that included independent presence–absence records for model evaluation. Since its publication, some of the algorithms have been further developed and new ones have emerged. In this paper, we explore patterns in predictive performance across methods, by reanalyzing the same data set (225 species from six different regions) using updated modeling knowledge and practices. We apply well-established methods such as generalized additive models and MaxEnt, alongside others that have received attention more recently, including regularized regressions, point-process weighted regressions, random forests, XGBoost, support vector machines, and the ensemble modeling framework biomod. All the methods we use include background samples (a sample of environments in the landscape) for model fitting. We explore impacts of using weights on the presence and background points in model fitting. We introduce new ways of evaluating models fitted to these data, using the area under the precision-recall gain curve, and focusing on the rank of results. We find that the way models are fitted matters. The top method was an ensemble of tuned individual models. In contrast, ensembles built using the biomod framework with default parameters performed no better than single moderate performing models. Similarly, the second top performing method was a random forest parameterized to deal with many background samples (contrasted to relatively few presence records), which substantially outperformed other random forest implementations. We find that, in general, nonparametric techniques with the capability of controlling for model complexity outperformed traditional regression methods, with MaxEnt and boosted regression trees still among the top performing models. All the data and code with working examples are provided to make this study fully reproducible.

Key words: boosted regression trees; down sampling; ecological niche model; ensemble modeling; imbalanced data; independent test data; machine learning; maxent; model evaluation; point process weighting; presence-background; random forest.

Introduction

Receiving much attention in the past few decades, correlative species distribution models (SDMs) are well known to many researchers in ecology, evolution, biogeography, and conservation. SDMs are used in a wide range of theoretical and practical applications to understand the relationship between species and the environment (Guisan and Thuiller 2005), and map their geographic distribution (Franklin 2010). They often underpin real-world management decisions such as conservation prioritization and planning (Guisan et al. 2013, Whitehead et al. 2017). Frequently modelers have to collate existing data rather than gathering expensive survey-based samples (Powney and Isaac 2015, Johnston et al. 2020), so it is of interest to know how models perform on the types of data widely available. In addition, SDM predictions can vary depending on the fitted model (Hallgren et al. 2019), which could result in substantial change in decisions made based on their prediction (Muscatello et al. 2021). Consequently, there is ongoing interest in assessing the predictive performance of different modeling methods to understand whether some tend to perform generally better than others. Several studies have tested a variety of models in the past (Elith et al. 2006, Bahn and McGill 2012, Shabani et al. 2016). However, many new modeling methods have

Manuscript received 16 October 2020; revised 7 May 2021; accepted 4 June 2021. Corresponding Editor: David Nogués-Bravo.
[3] E-mail: valavi.r@gmail.com

emerged recently, emphasizing the necessity of an up-to-date assessment of their general performance for ecologists. Here we explore the performance of several state-of-the-art modeling methods likely to be of interest for species distribution modelers on a global data set of 225 species (Elith et al. 2020) from different taxa. Our analysis provides a detailed report on different aspects of the predictive performance of these models, tested on a wide range of species, that facilitates the choice of modeling methods for ecologists.

There are many aspects to be considered when building an SDM, including appropriateness of underlying model assumptions, choice of modeling algorithms, tuning of model parameters and complexity, selection of background data, and availability of species data and environmental predictors (Araújo et al. 2019). These considerations can substantially impact predictions. Among these aspects, the choice of modeling algorithm is often prominent because there are many techniques available. Methods vary in predictive success (Pearson et al. 2006, Thuiller et al. 2009), so there is ongoing interest in identifying general trends in predictive performance across methods.

Of previous studies exploring patterns in the predictive performance of modeling algorithms (Segurado and Araujo 2004, Prasad et al. 2006, Meynard and Quinn 2007, Shabani et al. 2016), the study by Elith et al. (2006) has been highly influential in the field and is often quoted to justify the choice of modeling technique. We hereafter refer to that study as the 2006 NCEAS models, to emphasize the contribution of many modelers and to acknowledge the support of the National Centre for Ecological Analysis and Synthesis (NCEAS). Based on modeling methods of the early 2000s, it provided a detailed comparison of traditional and newer modeling methods (16 modeling methods) on several data sets of different taxa (226 species of plants, mammals, reptiles, and birds) from six different geographic regions. The performance of models fitted to presence-only species records was evaluated using independently collected presence–absence data sets.

Modeling methods have developed considerably since the 2006 NCEAS analysis, and there is now also much more attention to reproducibility of studies (Fidler et al. 2017). Furthermore, a range of different software packages were used in 2006 (some with manual steps and now outdated software) and several modelers were involved in fitting models. Therefore, there is substantial scope to build on the 2006 NCEAS study, using updated best practices for existing techniques, new modeling approaches and reproducible methods, and that is the aim of the present paper. Since the NCEAS data are now publicly released (Elith et al. 2020), our study will act as a new, reproducible, benchmark that expands our knowledge of SDM performance.

In choosing new modeling methods to include, we focused on methods that might aid model selection and model fitting. For instance, regularization techniques (Friedman et al. 2010) can improve the predictive performance of models by penalizing and shrinking regression coefficients, leading to a substantial reduction or complete removal of unimportant variables (James et al. 2013). Regularization methods provide a form of model selection that is known to solve common issues in other traditional model selection approaches like stepwise or best subset selection based on information criteria (Marra and Wood 2011). Some methods use weights across the response data, and point process weighting has been suggested by Fithian and Hastie (2013) as the proper way of fitting regression models for presence-only data. It has been reported to improve model performance (El-Gabbas and Dormann 2017). Model averaging or ensemble modeling is often considered to have higher predictive power and to be more reliable than single models (Araújo and New 2007, Marmion et al. 2009), and is popular among species-distribution modelers (Hao et al. 2019). Other methods not tested by the 2006 NCEAS modelers and that we assess here include Random Forests (RF), now a widely used method for modeling species distributions (Zhang et al. 2019), support vector machines (SVM), and extreme gradient boosting (XGBoost).

We will show briefly how the implemented models work, highlight their main differences, and provide example code showing how to fit them on species presence-only data. We also introduce new ways of comparing many modeling methods across several species data set via their performance rank rather than absolute mean, and assess statistical significance of the results. All the modeling code and species data are provided to serve as a baseline for future studies. As a comprehensive analysis of the performance of predictive models, our results are relevant to many researchers in ecology, evolution, and biogeography.

Materials and Methods

Data for modeling and evaluation

The species data we used for model fitting and evaluation is the data set assembled and used by the 2006 NCEAS modelers, excluding one species only represented by two records in the training data set. It thus represents 225 species from six regions of the world: birds and plants of the Australian Wet Tropics (AWT); birds of Ontario, Canada (CAN); plants, birds, mammals and reptiles of northeast New South Wales, Australia (NSW); plants of New Zealand (NZ); plants from five countries of South America (SA); and plants of Switzerland (SWI). The species data are detailed in Elith et al. (2020). They are provided in two independent data sets: (1) a set of presence-only data, generally from opportunistic records, ranging from 5 to 5,822 presence sites per species, and hereafter called "training data" and (2) a set of presence–absence data, gathered in designed surveys in each region, ranging from 102 (AWT) to

19,120 (NZ) survey sites, and hereafter called "testing data."

The environmental predictors were collated by the 2006 NCEAS working group, aiming to represent ecologically relevant factors for the species of each region, with 11–13 candidate variables per region. These predictor data are a set of continuous (e.g., temperature) and categorical (e.g., soil classes) raster files provided in spatial resolutions varying from 100 to 1,000 m depending on the region (see Appendix S1: Table S1). The 2006 NCEAS modelers variously used all candidate variables or subsets of them, depending on their preferred approach to modeling as influenced by their experience and restrictions imposed by the modeling method (e.g., some could not use categorical variables). Their choices are documented in their paper. Our study excluded variables with a pairwise Pearson correlation higher than 0.8, from the candidate data sets. Files containing both species and environmental data, exactly as used by the 2006 NCEAS group, are now available as CSV files on Open Science Framework (OSF) and also in an R package; environmental raster data are available on OSF. See all details of availability and metadata in Elith et al. (2020).

Selecting background data

Despite recent trends toward improvement of the quality of species data (Araújo et al. 2019), the vast majority of available data are still presence-only data available on online digital databases (Anderson 2012, Johnston et al. 2020). Presence-only data consist of the locations of observed species presence and lack information about locations where a species does not occur, i.e., absence (Renner et al. 2015). Several strategies have been used to allow models to be fitted to such data. A common technique is to sample a relatively large number of random samples from the landscape, termed *background* or sometimes *pseudo-absence* samples (Franklin 2010, Hefley and Hooten 2016, Araújo et al. 2019). Following the methods of the 2006 NCEAS study, and consistent with recommendations of recent statistical papers (Renner et al. 2015), we sampled background points irrespective of the location of species records, allowing that a presence and a background sample may occur at the same site. We sampled the background randomly despite the fact that some of our data sets may have spatial bias (see Phillips et al. [2009] for more information), aiming to reproduce the general approach of the 2006 NCEAS study.

The NCEAS group used a sample of 10,000 background points in each region, but recent research has explained why more may be necessary (Warton and Shepherd 2010). The number of background points should be large enough to comprehensively sample (and hence represent) all environments in the region of interest. Having a very large number of background samples, however, increases computational burden. Here, we used an incremental approach to choose the number of background points for our modeling (Fig. 1). We ran preliminary analyses on a few species using MaxEnt (v3.4.1; Phillips et al. 2006) to assess the influence of the number of background samples on predictive performance, using

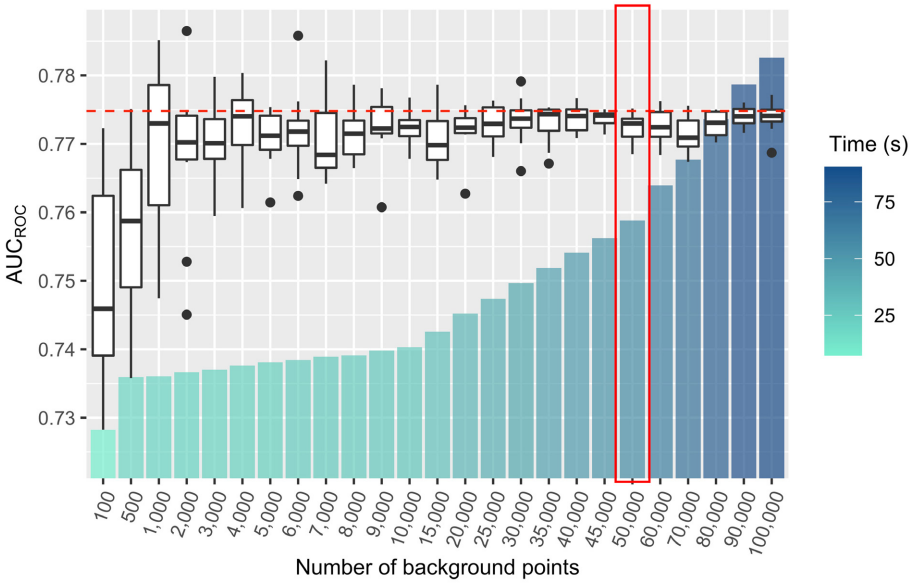

FIG. 1. Number of background points and the area under the receiver operating characteristic curve (AUC_{ROC}) of the MaxEnt model for a widespread species in AWT. Bars show the average computation time in seconds. The red horizontal dashed line indicates the AUC_{ROC} (0.775) obtained with a model fitted using all available cells as background samples. Box plot components are the median (mid line), the first and third quartiles (box hinges), and extend from the hinges by 1.5 * inter-quartile range (whiskers) of the AUC_{ROC} values.

area under the receiver operating characteristic curve (AUC_{ROC}) calculated on the testing data set. We assumed that when the number of background points becomes sufficiently large, and it properly represents the variation in the environmental covariates, model predictive performance would converge to a stable value. MaxEnt is a good candidate for this analysis as it is developed to model presence-background data, it is not a stochastic model (i.e., results do not vary each time it is run), and it is fast computationally. Two species from each region (one widespread and one with limited range) were selected for this purpose. We fitted models to 25 different sample sizes of background points (100–100,000, Fig. 1), with each sampling-then-modeling repeated 10 times to account for the variability in the selected background samples.

We present the outcome of these tests here, since it is central to the methods. Fig. 1 shows the change in evaluation metric (AUC_{ROC}) and computation time for different numbers of background points for a species in AWT region, chosen for its representative AUC_{ROC} values. The variation of AUC_{ROC} with background sample size is typical of results across species and regions. Based on these, we chose to use 50,000 background points in our modeling, because the variation in evaluation gets close to the AUC_{ROC} produced by using all cells in the landscape as background (the "gold standard," represented by a horizontal red dashed line). We acknowledge that the actual choice of 50,000 rather than say 35,000 or 70,000 is somewhat arbitrary, but it addresses the general issue of using enough background points. The final background sample for each region is provided in Data S1.

Modeling methods

We conducted all analyses in the R programming language (R Core Team 2020) and provide code and examples useful for new researchers. The free and open-source R programming language is commonly used among ecologists and SDM modelers, and provides a platform to keep together the whole modeling workflow and analysis of SDM outputs. We fitted models using several common modeling methods implemented in R, or able to be run through R. Where possible, for those methods used by the 2006 NCEAS modelers we use the same R packages as used by them (Table 1), or otherwise choose a current, popular alternative. We deliberately focused on methods that model species one at a time, not using any other species or community data to fit the models (in contrast see Norberg et al. 2019). Our chosen models include both traditional parametric and semi-parametric regression models and newer machine-learning methods (e.g., tree-based models and support vector machine). Many of these have been extensively used in SDM. In general, these models differ in the way they determine the fitted function (on a spectrum from largely user-defined to largely data-driven), whether they include interactions, and the way they handle model complexity and overfitting (i.e., how they address the bias-variance trade-off) (Hastie et al. 2009, Merow et al. 2014). All models use background samples in model fitting (i.e., none use the presence data only, as for instance in methods like BIOCLIM; Booth et al. 2014). For ease of discussion, we divide the models into three groups: *regression*, *tree-based*, and *other* methods.

Table 1. Modeling methods and their implementation in R packages.

Method	Description	R package
GAM	generalized additive model	*mgcv*
GLM	generalized linear model	*stats::glm* and *gam::step.Gam*
Lasso	regularized regression (L1 regularization)	*glmnet*
Ridge regression	regularized regression (L2 regularization)	*glmnet*
MARS	multivariate adaptive regression spline	*earth*
MaxEnt	maximum entropy	*dismo::maxent* (needs maxent.jar)
MaxNet	maximum entropy new implementation	*maxnet*
BRT/GBM	boosted regression trees	*dismo::gbm.step* (relies on the *gbm* package)
cforest	unbiased conditional inference forest	*party::cforest*
RF	random forest	*randomForest*
XGBoost	extreme gradient boosting	*xgboost*
biomod	ensemble framework with up to 10 different models	*biomod2*
SVM	support vector machine	*e1071*

Our aims were (1) to repeat some of the original methods used by the 2006 NCEAS modelers to test the general reproducibility of their results, (2) to add other methods likely to be of interest to distribution modelers and suited to these amounts of presence-only data, and (3) to provide code as a baseline for future studies. Overall, 13 modeling methods were used (Table 1), but some were implemented with more than one variant of model fitting, resulting in 21 approaches (Table 2). These models are described briefly in this section, followed by details of our choices of settings for fitting models. The details of models and how to fit them in R can be found in Appendix S2 and Data S1. We have made our modeling approach explicit so that others wishing to test additional methods can build on this basis.

Regression-based models.—Among the regression approaches, Generalized Linear Models (GLMs) and Generalized Additive Models (GAMs) are commonly used in species distribution modeling. GLMs use parametric functions such as linear or higher-degree

Table 2. A summary of model implementation settings.

Method and weights	Parameters	Values	Description
GAM			
DW	method	REML	smoothing parameter estimation method
GLM unweighted			
None	direction	both	step-selection direction (forward and backward) based on AIC
GLM			
DW	direction	both	step-selection direction (forward and backward) based on AIC
IWLR-GLM			
IWLR	same as GLM	same as GLM	same as GLM
IWLR-GAM			
IWLR	same as GAM	same as GLM	same as GLM
Lasso†			
DW	alpha	1	the lasso penalty; linear and quadratic terms allowed
Ridge regression†			
DW	alpha	0	the ridge penalty; linear and quadratic terms allowed
MARS unweighted			
None	nprune	2–20	number of terms
	degree	1	degree of interaction (1 means no interaction allowed)
MaxEnt			
NA	args	no threshold	auto select feature and exclude threshold feature
MaxEnt tuned			
NA	betamultiplier	0.5, 1, 2, 3, 4	regularization multiplier
	feature types	L, LQ, H, LQH, LQHP	transformations of input covariates: L, linear; Q, quadratic; H, hinge; P, product
BRT			
DW	tree.complexity	1 or 5	the complexity of individual trees
	learning.rate	0.001	shrinkage or the weight applied to individual trees
	bag.fraction	0.75	proportion of observations sampled to train each tree
	n.folds	5	number of cross-validation folds
cforest			
None	mtry	sqrt(n. covars)	number of variables randomly selected at each split
cforest weighted			
DW	mtry	sqrt(n. covars)	number of variables randomly selected at each split
RF			
None	mtry	sqrt(n. covars)	number of variables randomly selected at each split
RF down-sampled			
None	mtry	sqrt(n. covars)	number of variables randomly selected at each split
	sampsize	n. presences	number of bootstrap samples taken from each class
	ntrees	1,000	number of trees
	replace	TRUE	samples are taken by replacement
XGBoost			
None	nrounds	from 500 to 15,000 by 500	number of iterations (trees)
	eta	0.001	learning rate or shrinkage parameter
	max_depth	5	maximum number of terminal nodes allowed
	subsample	0.75	proportion of observations sampled to train each tree
	gamma	0	minimum loss reduction required to make a further partition

Table 2. Continued.

Method and weights	Parameters	Values	Description
	colsample_bytree	0.8	proportion of variables randomly selected at each split
	min_child_weight	1	minimum number of samples at each terminal node
biomod			
None	50,000 background samples		ANN, GLM, GAM, MARS, FDA, CTA, BRT, RF and MaxEnt models
Ensemble			rescale and average of individual modes implemented here: GAM, Lasso, MaxEnt, BRT and RF down-sampled
SVM			
None	kernel	radial	the radial basis kernel
SVM weighted			
DW	kernel	radial	the radial basis kernel

Notes: More details are provided in the Appendix S2 with example code (and complete code to reproduce our analysis in Data S1). The "Method and weights" column shows the method in normal font followed by weights in italic font. The possible weights are down-weighting (DW), Poisson process weighting (IWLR), or no weighting. NA indicates the model does not accept weights. The "parameters" column shows model arguments that are selected in the modeling process. The "values" column shows the value or ranges of values selected for model fitting and tuning. AIC, Akaike information criterion.

† Lasso and ridge regression are here used with GLMs.

polynomials to model the relationship between the response and predictors. GAMs use nonparametric smooth functions to allow nonlinearity in the fitted functions.

There are several options for fitting GLMs, with modern regularization methods often performing well (Reineking 2006). Lasso and ridge regression (L1 and L2 regularization, respectively) penalize the coefficients and shrink them toward zero (Friedman et al. 2010). This shrinkage reduces the variance of the regression model (i.e., its stability over different data samples), hence the fitted model may generalize better. Unlike ridge regression, lasso can reduce the coefficient of variables to exactly zero, de facto excluding those variables and resulting in sparser models (Hastie et al. 2009). A recent comparison showed penalized regression to perform as well as MaxEnt (Gastón and García-Viñas 2011). Since MaxEnt is often regarded as having strong predictive performance, this suggests these regularization approaches may be useful for predicting species distributions.

Multivariate Adaptive Regression Splines (MARS) is a flexible nonparametric regression similar to GAMs but using piecewise linear basis functions instead of smooth functions (Elith and Leathwick 2007). The complexity of the model varies with how many of these piecewise linear functions are fitted across each predictor variable, and that is determined with fast, inbuilt internal cross-validation methods. MARS fits interactions if that option is allowed (Leathwick et al. 2006); here we did not test that option.

MaxEnt is a popular modeling method for predicting species distributions, specifically developed for modeling presence-only species data (Phillips et al. 2006). We include it in the "regression-based" section due to its known links to regression methods, and particularly point process approaches (Renner and Warton 2013, Renner et al. 2015). MaxEnt has the flexibility to fit more or less complex models depending on the number of species records and user-defined settings. Complexity is controlled by use of transformed features of the predictor covariates (including *linear*, *quadratic*, *product*, *hinge*, and *threshold*) and also by choice of regularization settings (Elith et al. 2011). MaxNet is a new, alternative implementation of MaxEnt (Phillips et al. 2017), motivated by new understandings of the link between MaxEnt and Poisson point process models (Renner and Warton 2013). It uses infinitely-weighted logistic regression (Fithian and Hastie 2013) to fit the MaxEnt model and it is developed as an R package with no need for external software. Both MaxEnt and MaxNet use L1 regularization (Elith et al. 2011), similar to Lasso, but with potentially more flexible fitted functions (via transformed features).

Tree-based models.—Classification and regression trees are nonlinear (and nonparametric) models that recursively partition ("split") the predictor space into sections with similar values of the response variable (Elith 2019). This is a conceptually simple method that has several advantages such as reliably selecting influential covariates and allowing automatic fitting of interactions between covariates (Strobl et al. 2009). Single trees are high-variance methods, changing with each training data set. They are also poorly suited to estimating smooth functions. This limits their predictive performance, but they are commonly used as the base learner in ensembles of trees, often highly effective for prediction (Hastie et al. 2009). Hence, here we test ensembles rather than single trees. Tree-based models are often categorized as machine-learning models.

In Boosted Regression Trees (BRT) hundreds to thousands of regression trees are selected into an ensemble in a forward stagewise fashion. At each step of model fitting, the algorithm focuses on the weakest parts of the model built so far (the observations that so far are not predicted accurately) by fitting each new tree to the residuals of the previously fitted trees (Elith et al. 2008). Here, we use an implementation of BRT that has been widely used in SDMs (Table 1) and that constructs the models using stochastic gradient boosting (Friedman 2002).

The XGBoost algorithm (Chen and Guestrin 2016) is a new and slightly different form of gradient boosting with several features intended to improve its scalability and control over-fitting (Prasad 2018). Despite the successful usage of XGBoost in other disciplines (Chen and Guestrin 2016), the application of XGBoost in species distribution modeling is rare (examples in ecology include Doren and Horton [2018], Huang et al. [2018]. and Herdter [2019]). This might be due to the fact that XGBoost is relatively new and that, as it is highly flexible, it includes many hyperparameters that need careful tuning (Muñoz-Mas et al. 2019).

Random Forests (RF) have become popular for SDMs. This modeling method approaches ensemble creation differently to BRT, not using a stagewise approach but instead using bagging (bootstrap aggregation) to combine many trees. Bagging involves taking many bootstrap samples from the training data, fitting a tree to each sample and making an average prediction over all fitted trees (Strobl et al. 2009). Unique to this method, RF uses only a random subset of the predictor variables (parameter mtry) on each split while growing each tree. This creates decorrelated trees and reduces the variance of the final model, with consequent gains for predictive performance (Hastie et al. 2009). An advantage of RF compared to similar methods such as BRT and XGBoost, is that it is not very sensitive to tuning the model parameters (Strobl et al. 2009, Freeman et al. 2016). These characteristics have made RF a relatively common SDM approach (Mi et al. 2017, Harris et al. 2018).

Conditional Inference Forest (cforest) is a variant of RF that uses a different form of decision trees called ctree (Hothorn et al. 2006). This method was originally developed to deal with known problems in common splitting methods used in recursive binary partitioning, particularly, that there is a selection bias toward predictor variables with many possible splits or with missing values. In contrast to RF, cforest does not grow trees to maximum size and instead applies a stopping criterion. The cforest model also uses subsampling without replacement instead of bootstrap sampling to calculate variable importance in an unbiased way (Hothorn et al. 2006, Strobl et al. 2008). The process of fitting ctrees is costly and, as a result, creating ensembles of many ctrees in cforest is computationally expensive.

Other models.—Support Vector Machine (SVM) is a nonparametric machine-learning technique for regression and classification problems that has been used for modeling species distributions (Guo et al. 2005, Drake et al. 2006, Ashraf et al. 2017). SVMs work by defining linear hyperplanes that best separate different classes in the data. Similar to linear regression models, SVM can use nonlinear forms of the predictor variables for increased flexibility. This is done through a kernel function (e.g., polynomial or radial basis; Hastie et al. 2009: chapter 12).

Model averaging is a popular technique for reducing the uncertainty of model predictions (Dormann et al. 2018). For SDMs, it has become popular to average across predictions from different methods, based on the idea that prediction uncertainty due to the choice of method is decreased (Araújo and New 2007). The package *biomod* is specifically written for modeling species distributions (Thuiller et al. 2009), building ensembles across several modeling methods. It includes 10 algorithms, some of them used in our study (including GAM, GLM, BRT, RF, and MaxEnt), and combines the prediction of these models (e.g., by weighted averaging of the predictions). The biomod model has become a popular modeling approach since the 2006 NCEAS study, with widespread use but narrower dedicated exploration of its predictive performance (Hao et al. 2019).

In addition to using biomod, we averaged several models to build our own "ensemble" model. We selected the component models before models were fitted and evaluated, and used no knowledge of the testing data set to select them. Given evidence in the tree ensemble literature that ensembles work best when the component models are not highly correlated (Elith 2019), we chose for our self-selected ensemble a set of models with a breadth of fitted functions and model fitting approaches. We targeted methods we expected (based on our experience) to do well. The chosen models were Lasso, GAM, MaxEnt, BRT, and one of the RF variants (down-sampled; explained in the next section). Their predictions were all rescaled between 0 and 1 and their (unweighted) average used to build the ensemble model.

Model fitting and tuning, including weights

Model performance will reflect decisions made for model fitting and tuning, so we specify and justify our choices here (and see Table 2). This is an aspect of species distribution modeling for which different users hold different attitudes: some prefer using software defaults, whereas others emphasize careful model tuning. In the 2006 NCEAS modeling, the authors aimed to optimize the performance of methods by having researchers experienced with each approach choosing model settings and other aspects of modeling the data (e.g., choice of variables). Predictions from alternative feature class settings were produced by the author of MaxEnt, which was

relatively new at the time and at that stage had no hinge features (Elith et al. 2006, Phillips and Dudík 2008). Also, MARS was applied with and without interactions allowed. The 2006 NCEAS group did not allow access to the test species data at the model fitting stage, so all modelers chose settings "blind" to performance on the test set.

Regarding the shape of fitted relationships, parametric statistical models are often supplied with user-specified functions (e.g., linear or quadratic), and, among these, those that best fit the data are used to represent the unknown true structure of the data. Typically in ecology, the realized relationships would not be expected to follow strict functional shapes and assumptions (e.g., additive, linear, or stepwise; Austin and Meyers 1996). On the other hand, nonparametric models allow fitting very complex functions, which can easily over-fit to the training data and, as a result, perform poorly on independent data sets (Merow et al. 2014). There is ongoing interest in avoiding overfitting by considering the bias-variance trade-off in model fitting (Hastie et al. 2009). Here we tuned several models to control over-fitting, with tuning based on the presence-background training data. This was achieved through cross-validation for some methods (BRT, MaxEnt tuned, MARS unweighted, XGBoost, ridge regression, and Lasso; Table 2), and AIC for GLM (and GLM unweighted). We allowed both linear and quadratic terms in GLMs (GLM, GLM unweighted, IWLR-GLM, Lasso, and ridge regression). We explain them in more detail below. The biomod models were fitted using the package defaults, since that is the most common way that biomod is used (Hao et al. 2019).

Since the authors of this study are not equally experienced in all methods, we used both training and testing data for one species (species nsw09, a diurnal bird species from NSW) to explore the nuances of parameter settings of algorithms. This was simply done to understand the methods. Once we had a grasp on that all species were modeled using only species presence-background data, with specific tuning and model fitting per species estimated on the presence-background data, as described above. Covariates were normalized to have a mean of zero and standard deviation of one for all models (either manually, or internally in the functions that implement methods, e.g., for Lasso, ridge regression, MaxEnt, and MaxEnt tuned). The model fitting and tuning parameters of the models are summarized in Table 2. More detail on each model's settings and the versions of all software is presented in the supplementary materials (Appendix S1), along with coding examples on one species in R (Appendix S2; with complete code provided in Data S1).

Modeling presence data requires using a large number of background points, much larger than the number of presences. So, the training sample usually has a very small ratio of presence to background points. For a variety of reasons (e.g., producing very small predicted values) some users suggest weighting the records to balance their contribution (King and Zeng 2001, Guisan et al. 2017). The NCEAS 2006 study applied a weighting approach by which the background points were down-weighted to have a total (summed) weight equal to the total weight of the presences. This weighting strategy is viewed as statistically naïve (Renner et al. 2015), thus other well motivated weighting schemes are also implemented here. In addition to down-weighting, we applied the weighting scheme proposed by Fithian and Hastie (2013) to approximate an inhomogeneous Poisson process (IPP) by a logistic GLM. This is done by the so called "infinitely weighted logistic regression" (IWLR) method that gives a very large weight to the background samples. The statistical link of IWLR to IPPs provides a sound background for implementing weighting, as the IPP is recently identified as the proper way to model presence-only data (Warton and Shepherd 2010; Fithian and Hastie 2013). We implemented this weighting on GLM and GAM models and called them IWLR-GLM and IWLR-GAM, respectively.

For this study, our initial intention was to reproduce the NCEAS 2006 down-weighting approach, and apply it to all the regression models. We applied it to GLM, Lasso, ridge regression, GAM, BRT, and SVM weighted. (Table 2). We include variations for GLM and MARS. We realized that weighting impacts the GLM discrimination result, so we used both weighted and unweighted implementations. For MARS, we could not apply weights because they caused an error during model fitting for many species. This could be an issue with the current versions of the R package we used, *earth* (v4.7.0). We use no weights in *biomod*, since that is the package default (v3.3-7.1).

Finally, we implemented more than one approach for some models. For instance, we used cross-validation on presence-background training data to tune the regularization multiplier and feature types in MaxEnt and called this version MaxEnt-tuned (Table 2). RF is known to be sensitive to low ratios of presence and background, as used here (we discuss this in *New insights to model fitting*). We used down-sampling (Chen et al. 2004) to account for this issue. In RF down-sampled (Table 2), we fitted each tree with a bootstrap sample of presences and the same number of background points (Valavi et al. 2021). The equivalent of this approach can be applied to cforest by weighting the presence and background points (i.e., samples for fitting each tree are taken in proportion to the weights of the observations; as implemented in the version of the party package v1.3-1 at the time of this analysis), we call this other implementation cforest weighted. We also applied an equivalent approach to SVM, by weighting the presence and background points (SVM weighted; also discussed in *New insights to model fitting*).

Model evaluation

The NCEAS dataset includes independently collected presence–absence data available for evaluation, as

detailed in Elith et al. (2020). We used three threshold-independent measures of predictive performance: (1) area under the receiver operating characteristic curve (AUC_{ROC}); (2) area under the precision-recall gain curve (AUC_{PRG}); and (3) Pearson correlation between the predicted likelihood of presence and the presence–absence testing data (COR). We used these three evaluation metrics to cover different aspects of the modeling performance. For different ecological applications, different aspects of performance are more or less relevant, and therefore having a range of metrics is more informative about the wider applicability of the methods. Even though it is common in SDM evaluations to threshold predictions and test those with a metric relevant to binary predictions, we chose not to do that here for the main analysis. We had two reasons. First, we have presence–absence data available for predicting to, and evaluating against, supporting metrics such as the ones we have chosen. Second, there is growing evidence that thresholding is usually not required, and has negative impacts on the information content of the predictions (Calabrese et al. 2014, Lawson et al. 2014, Guillera-Arroita et al. 2015). However, since providing such metrics, e.g., the True Skill Statistic (TSS; Allouche et al. 2006) makes the results more directly comparable with other published studies such as Barbet-Massin et al. 2012, we also estimated TSS and present those methods and results in Appendix S1. We avoided using metrics that only use presence records for model evaluation (e.g., Boyce index; Hirzel et al. 2006) as they are specifically designed for evaluating performance when absence data are unavailable. Absence records bring important information, so our chosen metrics are relevant to those.

AUC_{ROC} is a widely used statistic in species distribution modeling. It assesses the ability of models to discriminate presence from absence sites. AUC_{ROC} is calculated considering "1 – specificity" (the proportion of wrongly predicted absences or false positive rate) with respect to "sensitivity" (the proportion of correctly predicted presences or true positive rate, also known as "recall"), across many thresholds that can be used to classify the output probability into 0 and 1 (Pearce and Ferrier 2000). AUC_{ROC} ranges from 0 to 1, with 1 showing perfect discrimination and 0.5 indicating no better discrimination than a random classification. Values <0.5 are generally considered worse than random classification, though it is worth noting that 0.5 is only the average estimated AUC_{ROC} of an uninformative model, and errors around this can be large, particularly if samples are small (Raes and ter Steege 2007).

The ROC curve includes the number of true absences in its calculation. In ecology, there are cases where modelers might be more interested in focusing on accurate prediction of the presences, e.g., when the costs of distinct error types are different (Franklin 2010), such as in the application of SDMs in conservation prioritization (Elith and Leathwick 2009). Area under the precision-recall curve (AUC_{PR}) is another discrimination metric, commonly applied in machine-learning when the emphasis is not on the true negative rate (Hughes-Oliver 2018). The focus is on the predicted presences, whether they capture the true presences and do not include false positives. This is a common technique to rank the predictive performance of modeling methods in machine-learning literature. Like AUC_{ROC}, AUC_{PR} provides a single measure of performance across all possible threshold values. AUC_{PR} is calculated considering "precision" (or positive predictive value; the proportion of presence predictions that are true species presences) with respect to "recall" (sensitivity; Sofaer et al. 2019). AUC_{PR} is often preferred to AUC_{ROC} where the number of negatives (absences) is much larger than positives (presences; Flach and Kull 2015, Hughes-Oliver 2018). Recent studies have recommended using AUC_{PR} for evaluating SDMs for rare species, i.e., species with low prevalence (Johnson et al. 2012 and Sofaer et al. 2019).

Unlike the ROC curve that has a fixed baseline of 0.5, the baseline in the precision-recall curve depends on the prevalence in the testing data (Saito and Rehmsmeier 2015). This makes it difficult to compare directly the AUC_{PR} between species. Several corrections have been suggested in the machine-learning literature to deal with this characteristic (Boyd et al. 2012, Flach and Kull 2015). Flach and Kull (2015) propose to plot PR curves in a new coordinate system, and call the new plot precision-recall gain (PRG) curves. Positive AUC_{PRG} values indicate discrimination better than random, with AUC_{PRG} of 1 indicating perfect discrimination. Negative values suggest predictions worse than random. In this study we estimated AUC_{PRG}.

In our evaluation, we avoided most measures of model calibration. Models fitted on presence-only data (with background samples) have no information about prevalence of the species and thus cannot estimate probability of occurrence, except under strong parametric assumptions about the structure of the true probability of presence (Ward et al. 2009, Hastie and Fithian 2013, Phillips and Elith 2013). These assumptions are likely violated with real-world data (Yackulic et al. 2013, Guillera-Arroita et al. 2015). Calibration, the agreement between predicted probabilities of occurrence and observation of presence and absence, is usually a concept only applied to models fitted to presence–absence data (Pearce and Ferrier 2000). However, in some instances it is relevant to ask whether a presence-background model is as well calibrated as it could be, on the proviso that it is understood that without knowledge of prevalence it cannot be calibrated in any absolute sense (Phillips and Elith 2010). In this study, we estimated COR, which gives information beyond that in purely rank-based discrimination measures, as it assesses the difference between the values of the prediction and the observations (0s and 1s; Elith et al. 2006), thus providing insight into how well predictions are calibrated in relative terms. As it takes into account the actual prediction values, COR reports some aspects of calibration (Phillips and Elith 2010). A

model that performs well on AUC_{ROC} but poorly on COR is likely to be poorly calibrated, even in relative terms. While we recognize that we could further test the calibration of these models with methods such as those based on logistic regression (Cox 1958, Pearce and Ferrier 2000), we have not pursued that aspect of evaluation since predictions based on presence-only species data are not expected to be well calibrated. We provide predictions from all our models so, if readers are interested, other evaluation measures can be estimated (see Section 11 in Appendix S1).

Statistical comparison of models

To check if the differences in model performance across methods are not due to chance, we applied nonparametric statistical tests known to be suitable for statistical comparison of predictive models over multiple data sets (Demšar 2006, García and Herrera 2008). The nonparametric tests require fewer assumptions than their parametric counterparts and they are safer options for statistical comparison of model performance (Demšar 2006, García et al. 2010).

We used Friedman's Aligned Rank test (García et al. 2010) to assess the statistical significance of differences in model performance, based on the evaluation metrics, i.e., AUC_{ROC}, AUC_{PRG}, and COR. This test determines, for each metric, whether there is any statistical difference in performance among all of the models, but does not provide any information about the pairwise differences (García and Herrera 2008, García et al. 2010). After that, we applied the Friedman's Aligned Rank post hoc test, to conduct pairwise comparisons of the same performance metrics as above. The *P* values obtained from this post hoc test were adjusted using the Shaffer correction (Shaffer 1986) to take into account the effect of multiple comparisons (see García and Herrera 2008). The adjusted *P* values provide information on whether the statistical hypothesis of "equal performance" of pairs of models is significant or not, and it also shows how significant the result is: the lower the *P* value, the stronger evidence against the null hypothesis of equal performance (García et al. 2010). We used the R package *scmamp* v0.2.55 (Calvo and Santafé 2016) to calculate these statistics.

Similar to the analysis by NCEAS 2006, to capture the variation in evaluation metrics for graphical presentation, we used a Generalized Linear Mixed Model (GLMM; Bolker et al. 2009) with the metrics (AUC_{ROC} and COR) as the response variable and the modeling method as the fixed effect. The species identity and the interaction between the methods and regions were fitted as random effects, the interaction term allowing for differing performance of methods across regions. Analyses were performed in the Bayesian framework of inference, using *JAGS* v4.3.0 (Plummer 2003) called from R.

Results and Discussion

We assessed the predictive performance of 21 modeling approaches (Table 2) fitted to presence-only species records with 50,000 randomly selected background points, testing them on independently collected presence–absence data. The NCEAS 2006 study analyzed their results from several viewpoints including assessing patterns of model performance per species and the impacts of prevalence, environmental and geographic distances between sites on predictive performance. However, in this study we concentrate on the overall predictive performance of modeling methods by comparing some of the previous models with newly emerged ones, and expanding the evaluation with the new precision-recall statistics. With the training and testing data sets now freely available (Elith et al. 2020), we have also provided sufficient code that our modeling can be repeated by others, for future benchmarks in other explorations (Appendix S2 and Data S1).

Overview of performance

The overall distribution of AUC_{ROC} values across methods and species follows closely that of the 2006 study (see Appendix S1: Fig. S6, and Elith et al. [2006: Fig. 2] for comparison). This is in line with the idea that, for a certain data set (having the same sets of species, observations, and covariates), there is an achievable bound on predictive accuracy (García et al. 2010). This could be due to species characteristics (such as taxa and trophic mode), study extent, number of presence records, lack of using proximal environmental predictors or a combination of these factors (Soininen and Luoto 2014). The range of AUC_{ROC} was from 0.139 to 0.996 (0.07 to 0.97 for the 2006 NCEAS study) with a mean 0.709 and median 0.715 for all methods and species. About 42% of the models and species had AUC_{ROC} of 0.75 or higher (compared to 40% for the 2006 NCEAS study) and 54% of them were 0.7 or higher. Nine percent of the models (spread over 58 species in all regions except SWI) had AUC_{ROC} below 0.5, indicating predictions worse than random. Judging by AUC_{PRG}, 86% of the models and species had a predictive performance better than random (positive AUC_{PRG}), see later.

Fig. 2 gives the first overview of patterns of predictive performance of the models across all 225 species. The concentric rings indicate methods, and the outer histogram bars indicate numbers of training presence records. Some species are inherently harder to predict, and the AUC_{ROC} for all modeling methods for those species is low (orange and yellow colors). To the contrary, some other species are predicted very well with all the methods (blue colors). Except in SWI and NZ, most high AUC_{ROC} species are species with less than 100 presence records in the training set. We will present details

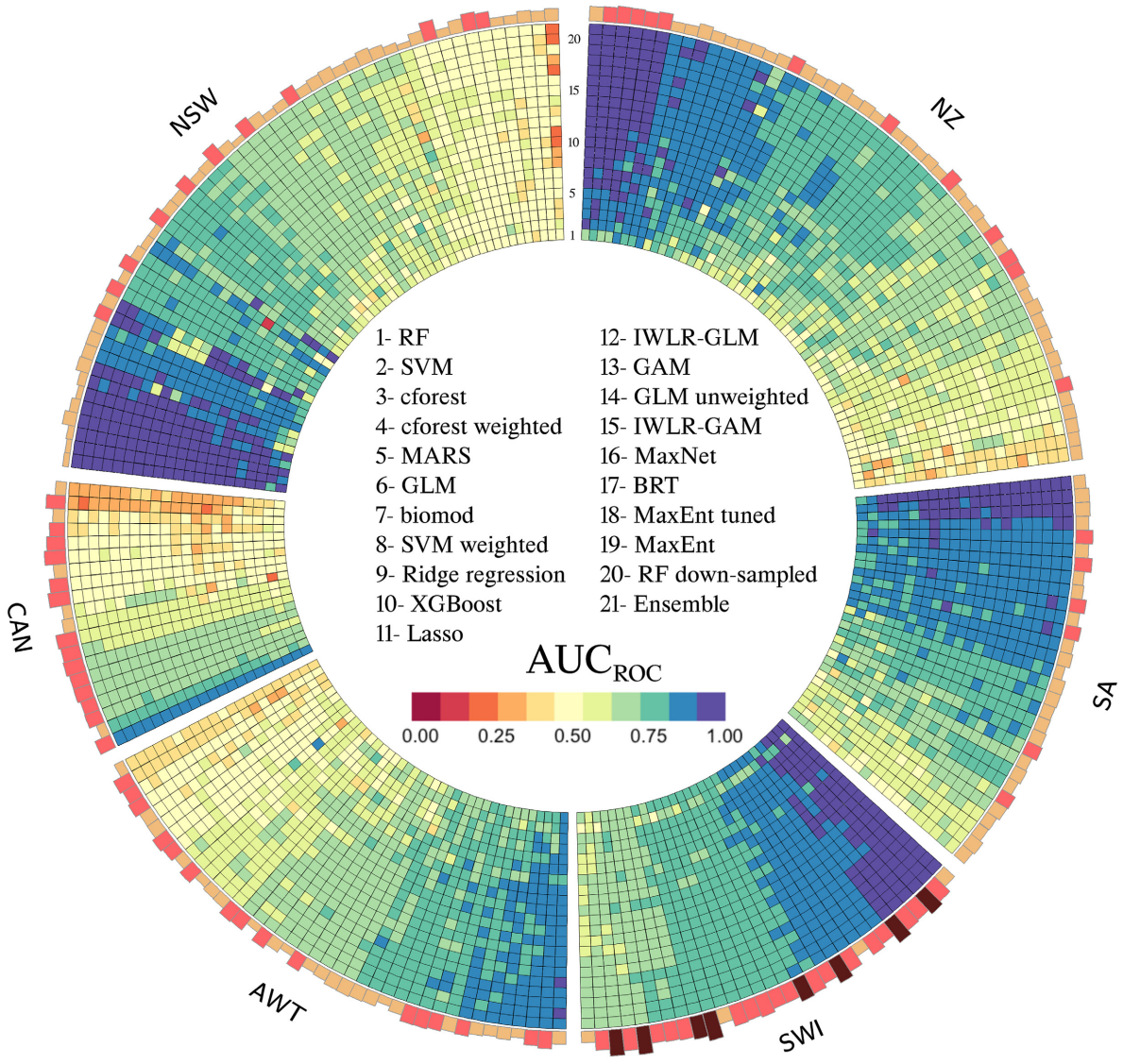

FIG. 2. AUC_{ROC} values for all modeling methods across all species. Each circular track is the AUC_{ROC} of a model for all species. Models and species are ordered by average AUC_{ROC}. The outer text labels indicate the regions. The numbers between NSW and NZ regions label the modeling method as mentioned in middle of the figure. The height of the outer histogram shows the $\log_{10}$(number of species presence-only records) in the training data set. Histogram colors follow three categories: light orange. <100; pink, 100–1,000; and dark brown, >1,000 presences. This figure was created in Circos software (Krzywinski et al. 2009).

and discuss the main themes, of differing performance across methods and regions, in the following sections.

Results averaged across all regions and species

Our results (Figs. 2, 3) show a gradient in performance across methods. For ease of discussion, we categorized the modeling methods into three groups based on both AUC_{ROC} and COR, using hierarchical clustering: (1) models with lower predictive performance, i.e., lower discrimination and lower correlation (left side of Fig. 3); (2) models with moderate performance; and (3) high-performance models with higher values for both discrimination and correlation (upper right side of Fig. 3). The Friedman's Aligned Rank test rejected the null hypothesis of "no difference between the performance of the models" by a high level of significance, i.e., very small *P* values for all three evaluation metrics (section 3 in Appendix S1). Therefore, we proceed with the post hoc test to analyze the pairwise differences, and present its adjusted *P* values in the next section.

Six of the 21 models are in the high-performance group (BRT, RF down-sampled, MaxEnt, MaxEnt-tuned, MaxNet, and Ensemble). Most models in the moderate group have a mean AUC_{ROC} higher than 0.7, indicating an average acceptable level of discrimination

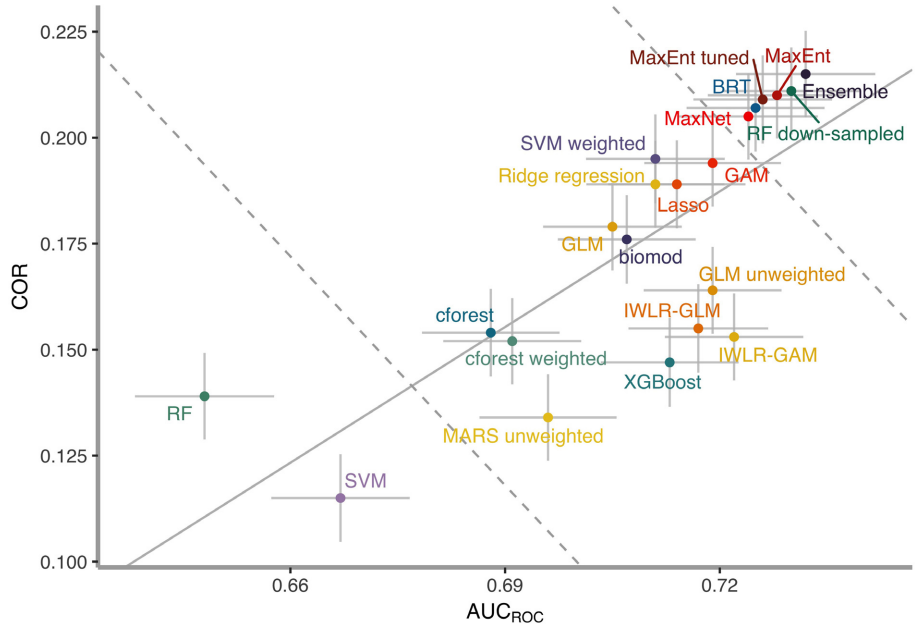

FIG. 3. Mean AUC_{ROC} vs. mean COR for the models, summarized across all species. The gray bars (around each model) are standard errors estimated in a Bayesian mixed model, reflecting variation for an average species in an average region. The solid gray line is the least-squares line fitted on the models. The dashed gray lines delimit regions of low-, moderate-, and high-performance models, following the hierarchical clustering (represented in the top of the plot in Fig. 4). As a reminder, the ensemble model is the average of Lasso, GAM, MaxEnt, BRT, and RF down-sampled.

on presence–absence data (Guisan et al. 2017). Variants of the statistical models with smooth surfaces (GLMs, GAMs) are all in the upper right section of this moderate group, whereas a couple of tree-based models (cforest and XGBoost) tended toward lower performance. RF (with default settings) and SVM were the only two methods classified as of low performance. Below we first interrogate these new results, focusing on new methods and interesting comparisons. We later compare results for methods repeated from the NCEAS 2006 study.

New stand-out methods

Ensemble modeling (such as biomod) and RF are two modeling methods prominent in recent SDM literature for their reputation of achieving good predictive performance (Marmion et al. 2009, Thuiller et al. 2009, Liu et al. 2013, Zhang et al. 2019). While these techniques are indeed the best performing modeling approaches in our study, they only achieved this status under unusual implementations. Often models are fitted following default procedures. For instance, ensemble modeling is typically done through packages such as biomod (Thuiller et al. 2009, Hao et al. 2019). In our comparison, biomod (with default parameters) performed not better than average models such as GLM, whereas our selected Ensemble model was the top-performing approach.

RF only performed well when using down-sampling (RF down-sampled). This result is noteworthy, because RFs are generally considered to be robust to the settings used (Freeman et al. 2016, Probst et al. 2019) and often shown to predict well (Liu et al. 2013, Beaumont et al. 2016). However, this is not always the case and clearly must depend on what other models are in the comparison, and on evaluation data. Shabani et al. (2016) compared the performance of several presence-background modeling methods on an independent data set and reported a poor predictive performance of their RF model. While our standard RF model performed poorly, the use of down-sampling improved its performance dramatically, from the lowest performance in AUC_{ROC} (0.648) to a place among the top performing models in both AUC_{ROC} and COR metrics (0.730 and 0.216, respectively; Fig. 3). The result of these implementations of RF and ensemble have interesting nuances that we will highlight in the "new insights to model fitting" section.

Tree-based methods

Ensembles of trees are among the best performers when well-tuned (BRT, RF down-sampled) but, despite efforts to tune others well (XGBoost and cforest), performance was not strong. BRT was one the best models in the NCEAS 2006 study and is still among the top-ranking models here, and not significantly different from other models in this group (Fig. 4). XGBoost has been successful in other disciplines (Chen and Guestrin

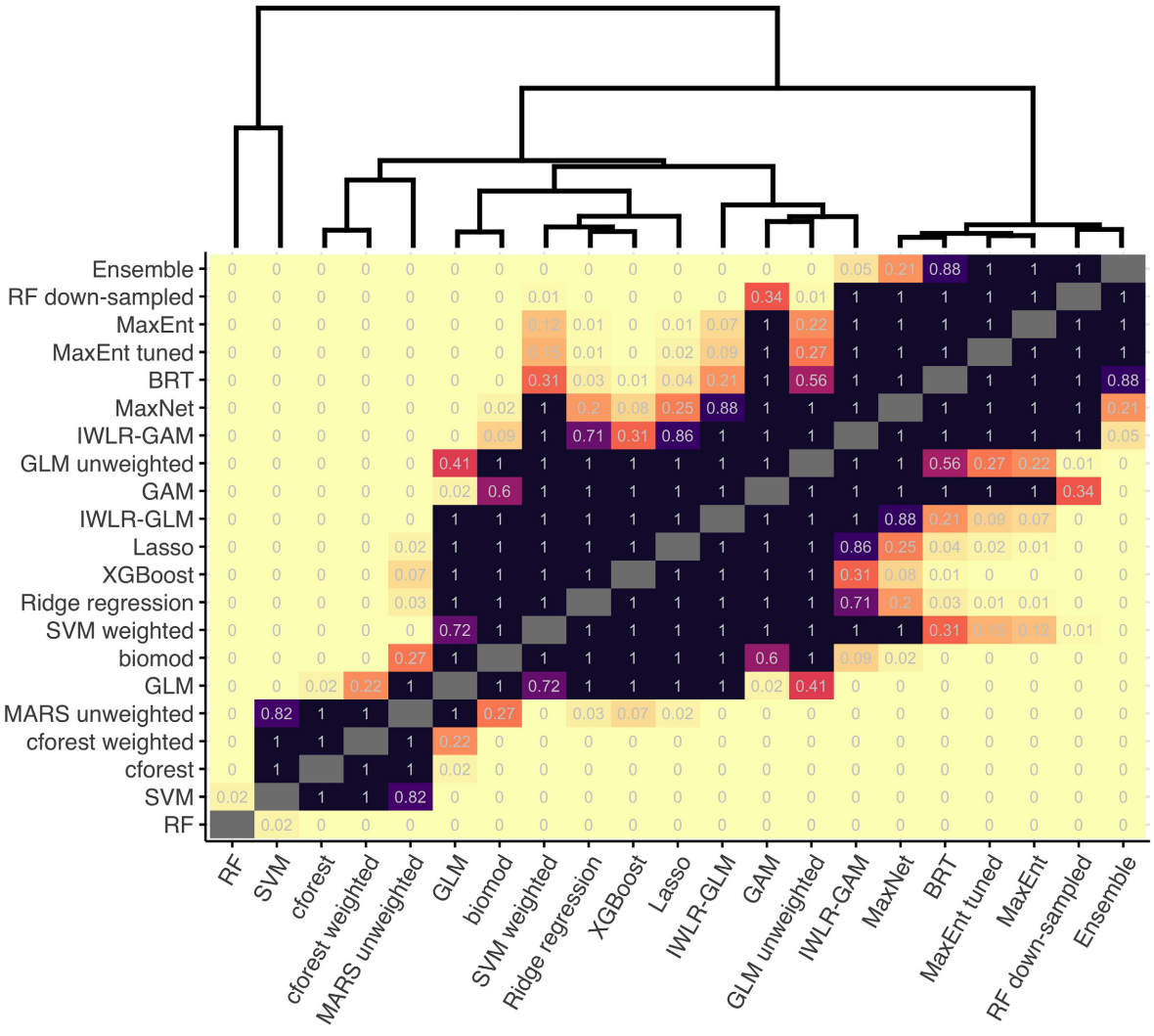

FIG. 4. *P* values of the post-hoc of the Friedman Aligned Rank test with Shaffer adjustments for AUC_{ROC} for a null hypothesis of no difference between pairs of models (values are rounded). The colors and the text inside each cell show different *P* values. Darker colors indicate higher *P* values (no significant difference in performance between models). The dendrogram on the top is based on both AUC_{ROC} and COR that clusters models in three main groups. The height of each bar shows the relative difference between each model/group.

2016), but here was only a mid-level performer. In a recent ecological example, Herdter (2019) modeled relationships between juvenile fish recruitment and environment for one species, and found XGBoost outperformed GLM with stepwise selection. In contrast, most of our GLMs (GLM-unweighted, IWLR-GLM, and Lasso) outperformed XGBoost, although XGBoost was still within one standard error of the GLM-unweighted and IWLR-GLM models. XGBoost is a very flexible algorithm, but needs intensive model tuning (Muñoz-Mas et al. 2019). While we attempted to tune our XGBoost models by using a grid search over the model parameters (Kuhn and Johnson 2013), we could not afford to explore the full potential parameter space when modeling 225 species. XGBoost predictions could possibly improve if all parameters were carefully tuned; this would make an interesting further study. Furthermore, the performance of XGBoost might be different when modeling presence–absence data. For instance, Muñoz-Mas et al. (2019) showed that a carefully tuned XGBoost fitted to presence–absence data of invasive fish species performed better than or equal to other tree-based models such as BRT and RF. cforest and cforest-weighted performed similarly, at the lower range of moderate performers (Fig. 4). Both cforest models have much better performance than RF (Figs. 3, 4), despite using similar settings (see Table 2). This could be due to the fundamental differences in the tree types used in these algorithms, where unlike RF, cforest does not grow very deep trees.

Regression-based methods

Some studies show that MaxEnt models perform better if the *regularization multiplier* and *feature types* are tuned (Muscarella et al. 2014, Radosavljevic and Anderson 2014). Here we tuned MaxEnt on the training data using cross-validation (MaxEnt-tuned). MaxEnt-tuned showed no statistically significant difference in performance to the default MaxEnt model (Figs. 3, 4). Note that the default MaxEnt settings were selected over 10 years ago based on performance on the presence–absence data in this same data set (Phillips and Dudík 2008), and this may be the reason that the tuned version of MaxEnt did not perform better.

Since MaxEnt, MaxNet, and IWLR-GLM are so closely related (Fithian and Hastie 2013, Renner and Warton 2013, Phillips et al. 2017), one might expect indistinguishable differences in performance. MaxEnt and MaxNet models showed comparable performance (Fig. 4, no evidence of statistical difference in AUC_{ROC}), both notably stronger than the infinitely weighted logistic regression (IWLR-GLM). This is likely because the former models use a wide array of transformed features instead of original variables, allowing more flexible nonlinear relationships. Our GLMs only allowed relatively smooth models (linear and quadratic terms). This is evident when comparing a more flexible infinitely weighted regression (IWLR-GAM) with MaxEnt and MaxNet (no evidence of statistical difference in AUC_{ROC}; adjusted *P* values in Fig. 4). The other pairs of methods that might be expected to be close in performance are Lasso and MaxNet, since MaxNet is using regularized regression to fit its model (even using same underlying R package as we used for Lasso here). Again, the performance of MaxNet was higher, probably again because it has more flexibility in fitted functions. Note that the results discussed here are based on AUC_{ROC}; the pairs of methods (MaxEnt/IWLR-GAM and MaxNet/Lasso) are more separated by COR, with the first-named performing better in each pair (Fig 3).

Comparing results for weighted versions of GLMs, our results demonstrate a trend of slightly better performance for regularization methods (Lasso, ridge regression) over stepwise GLM model selection (GLM, IWLR-GLM) when assessed with AUC_{ROC} and COR (Fig. 3), although these differences are not statistically significant (Fig. 4). The infinite weighting (IWLR-GLM) did not significantly improve the accuracy of GLM-unweighted. They are within one standard error in both AUC_{ROC} and COR (Fig. 3). However, IWLR-GLM had slightly higher average AUC_{ROC} than GLM, but lower COR. These differences were also not significant (*P* values: 1 for AUC_{ROC}, 0.06 for COR; Fig. 4 and Appendix S1: Fig. S3). The unweighted GLM achieved higher AUC_{ROC} than GLM (down-weighted) but lower in COR (Fig. 3) and TSS (Appendix S1: Fig. S9). This is in line with the result by Barbet-Massin et al. (2012), where they find that GLM with equal weight (named down-weighted here) gets the best result when evaluated with true skill statistics (Appendix S1: Fig. S9).

New insights to model fitting

The way one fits a model matters. We cannot cover the details of model tuning and spatial prediction of all models, rather we emphasize the importance of understanding the way models work and considering the nature of the data. Here we focus on three of the most extreme examples from this current study that highlight how simple modifications, grounded in theory, can make a big difference to model performance. RF and SVM models with default parameters performed poorly on our presence-background data (Fig. 5), putting them as the worst performing models. Background samples need to be large to sample all environments (Renner et al. 2015), but this necessarily leads to a large disparity in the number of presence records compared with the number of background records. This phenomenon (large difference in number of records between classes) is often referred to as class imbalance. In SVM and RF, techniques such as weighting are viewed as a way of addressing it. In our study, the performance of SVM improved dramatically with weighting, as it did using down-sampling for RF.

The sensitivity of RF to imbalanced data sets is often attributed to the unequal representation of the classes (here presence and background classes; He and Garcia 2009, Khalilia et al. 2011, Liu et al. 2013). However other issues are also relevant, and one that is clearly at play with presence-background data is that of class overlap, i.e., where the two classes sample similar environmental conditions (Prati et al. 2004, Ali et al. 2015). We explain and explore these issues in detail elsewhere (Valavi et al. 2021), it is sufficient here to note that presence-background data present unique challenges to classification methods like RF, requiring adjustments to how the models are fitted. Other ecologists have noticed the sensitivity of RF to many background samples and addressed it by using a very low number of background points (Barbet-Massin et al. 2012, Liu et al. 2013). However, that approach leads to small samples of the background, which is far from ideal (Renner et al. 2015). For imbalanced presence–absence data, previous studies have demonstrated the improvement of RF predictions using down-sampling techniques (Evans and Cushman 2009, Robinson et al. 2018, Shaeri Karimi et al. 2019). Here, we used RF with and without down-sampling, and the results clearly show the benefit of down-sampling (Fig. 5).

SVM models have been shown to have poor performance when the number of negative cases (background samples here) heavily outnumber positives (presence records here; Akbani et al. 2004). We believe that this is also largely due to class overlap rather than different representation of classes (see also Japkowicz and Stephen 2002). SVM works by constructing a series of

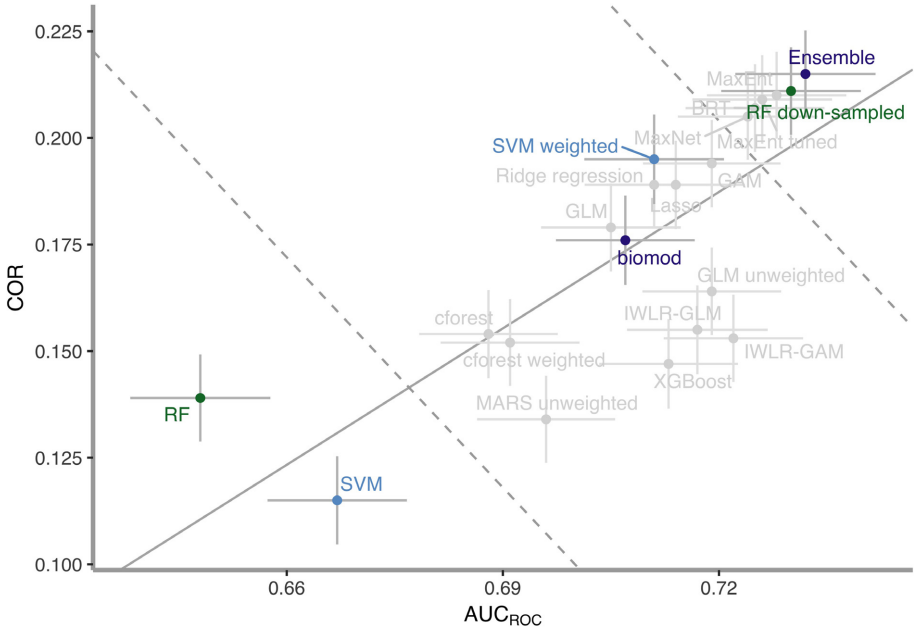

FIG. 5. The difference between the performance of the default model with modified models in RF, SVM, and ensemble models. Other methods are presented in gray for reference. Bars and lines as in Fig. 3.

hyperplanes to separate data points based on their class. The observations that lie very close to the decision boundaries (hyperplanes) are called support vectors, and affect their position. To maintain a good ability to generalize, SVM allows some of the support vectors to end up on the wrong side of the boundary, i.e., be misclassified (James et al. 2013). In a default setting with equal misclassification costs, the plethora of overlapping background points pushes the decision boundary toward the presence points (i.e., more misclassification of presences) as the overall cost of misclassification of the small number of presences is much less than many background samples (Akbani et al. 2004). By applying weights, we greatly increase the cost of misclassification of the presence points so that the decision boundary is defined in way that has a balanced misclassification over both presence and background classes. SVM improved substantially by applying weights (Fig. 5).

Finally, the contrast in performance between the two types of ensemble-across-methods is striking. Ensemble modeling using software such as biomod is popular for modeling species distributions (Hao et al. 2019). We used biomod with default parameters since this is a common choice among modelers (Hao et al. 2019). In our results, biomod with default parameters performed worse than a number of single models, and ended up roughly mid-field in performance (Figs. 5, 3). In contrast, ensembling a set of well-tuned models performed relatively strongly (Ensemble model in Fig. 5). Many modelers take default settings for granted and automatically use them for modeling. Our results demonstrate that ensembles per se are not effective (many single models performed better than biomod) but ensembles of well-tuned models can perform stronger than any of the components.

Comparison with the previous studies

Since part of the intent of this paper is to provide a reproducible benchmark for future comparisons, we compare the current results with those in previous publications (Elith et al. 2006, Phillips et al. 2017). Fig. 6 shows all common models used across these studies, noting that we could not always reproduce the way the model fitting was done, giving previous use of software no longer available.

Average performance for MaxEnt in Phillips et al. (2017) was slightly better than for the present study (Phillips et al. [2017] used MaxEnt v3.4.1, but we used v3.4.4). This small difference could be attributed to the different selection and number of random background data, and/or a different approach to choosing the candidate sets of predictors (we used a subset, Appendix S1: Table S1, whereas Phillips et al. [2017] used all). It is unlikely that more background points will detract from predictive performance (as evidenced below for GLMs and GAMs; also in Fig. 1), so it is likely due to the choice of predictor variables. This is an interesting result and suggests that wider testing of predictor selection could be worthwhile. MaxEnt results for the NCEAS 2006 study were slightly poorer again (but still strong). In 2006 MaxEnt was relatively new, lacking hinge features and the extensive tuning it now has (Phillips and Dudík 2008). In addition, the default

output is now slightly different (*cloglog* compared with *logistic*; Phillips et al. 2017). BRT had a very similar result in NCEAS 2006 and the present study. Similarly, MARS (with no interaction) achieved a similar AUC_{ROC}, however COR was very different. This is likely because no weights were used during the MARS model fitting in the present study, and a different package was used to implement MARS (*earth* in this paper vs. *mda* in 2006).

Noticeably, the improvement in GAM for the current study is substantial, putting it very close to the MaxEnt result in NCEAS 2006 (one of the best models in that study; Fig. 6). Likewise, GLM performance has improved compared to the NCEAS 2006. To assess whether this improvement (in GLM and GAM) is due to the different number/set of background points or the newer implementation of these models, we refitted both GLM and GAM with the same 10,000 background points from NCEAS 2006 and then compared the result with the one reported on the NCEAS 2006 and our current study (fitted also on 50,000 background samples).

The GLM with 10,000 background points in both studies (present modeling and 2006) obtained identical AUC_{ROC} with a slight change in the correlation. Increasing the background points to 50,000, improved accuracy only slightly (Table 3; Fig. 6). This shows that, averaged over all species and regions, there is no measurable benefit, but also no penalty to predictive performance, by using larger background samples in this case. Although the software used for modeling GLM in 2006 NCEAS study was different from this study (the GRASP package in S-PLUS vs. *stats::glm* in R), the implementation in terms of model selection was quite similar and similar covariates were used. On the other hand, focusing on the results for 10,000 background points, GAM notably improved in both correlation and discrimination since 2006. There is no measurable improvement in the performance of the GAM model by increasing the number of background points. The difference between years is likely due to new software packages used of this model (*mgcv* in R vs. the GRASP package in S-PLUS; see Elith et al. [2006] for full details).

TABLE 3. Comparison of GLM (Generalized Linear Models) and GAM (Generalized Additive Models) models with 10,000 and 50,000 background points to the results presented in the NCEAS 2006 study.

Models	Study/Modeling	AUC_{ROC}	COR
GLM	present fitting, 50,000 background	0.705	0.179
GLM	present fitting, 10,000 background	0.695	0.174
GLM	NCEAS 2006, 10,000 background	0.695	0.177
GAM	present fitting, 50,000 background	0.719	0.194
GAM	present fitting, 10,000 background	0.719	0.195
GAM	NCEAS 2006, 10,000 background	0.700	0.176

Note: All models have the same down-weighting scheme.

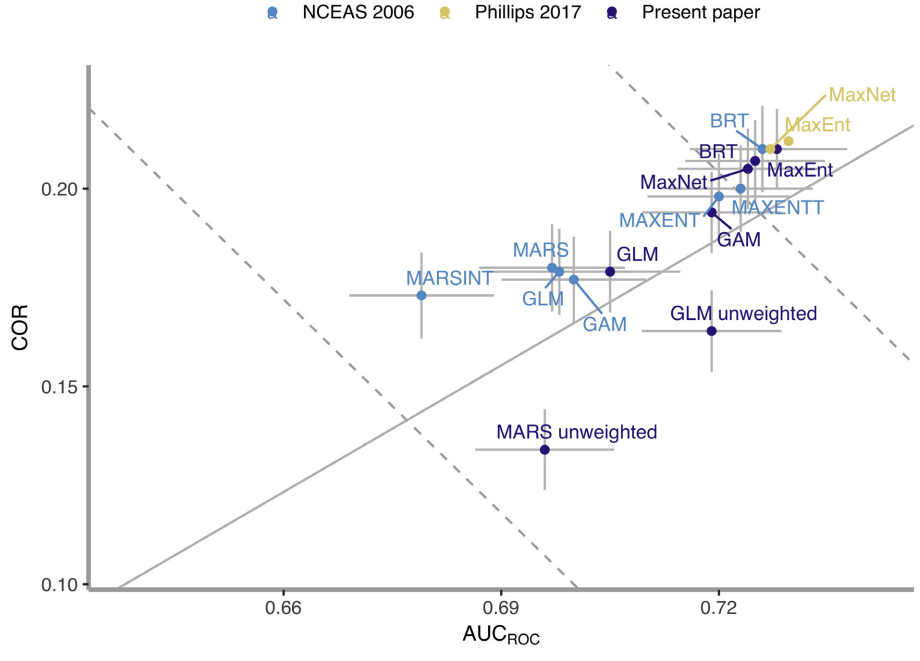

FIG. 6. Performance of models in NCEAS 2006, Phillips et al. (2017), and this study. For ease of comparison, this plot uses the same axes ranges and division lines as Fig. 3. The dashed and solid gray lines separate the low-, moderate-, and high-performance models, as in Fig. 3. The Phillips et al. (2017) models have no standard errors attached, as only their average performance was available to us. The MAXENTT from NCEAS 2006 study is a MaxEnt model with "threshold" feature (this feature is not included in the other studies).

Other approaches to evaluation: model ranks

So far, we assessed and compared the predictive performance of the models by averaging their accuracy measures (AUC and COR) across species. When using averages, we can measure the size of the overall magnitude of difference in performance. However, we cannot distinguish whether a modeling method might have a big advantage for just a few species or a small advantage over many species. Also, averaging is sensitive to outliers, so failure in a few species may skew results. Since model performance comparisons are done on just a sample of species, with the hope that results generalize to other species, one may prefer to choose a method that tends to perform well on most species, rather than a method that may be a lot better on only a few species, but worse in most other cases. Accordingly, we can use average ranks of models to summarize their performance across species (Fig. 7). This approach is widespread in machine learning (Demšar 2006, García et al. 2010, Kull et al. 2019), but rarely used in ecology. To implement this idea, we calculated the rank of each model by sorting the accuracy of all models for each species, using the average rank in case of ties (e.g., two models with the same AUC_{ROC}).

The overall average ranks of AUC_{ROC} vs. COR (Fig. 7A) shows a similar arrangement of the models to the averages of original values (Fig. 3) with the top models showing a bigger advantage over the middle performing models (Fig. 7 vs. Fig. 3). For ease of comparison with the previous graphs, we reversed the axis of the average rank plots to have the best performing model on the top right of the plot and the worst performing on the lower left. The average ranking better reflects the *P* values in the post hoc Friedman's Aligned Rank test as they are both based on rank rather than absolute values. One noticeable thing here is that, other than the high performing models identified by clustering and shown in Fig. 3, only GAM and SVM weighted have a better rank than the mean ranks of both AUC_{ROC} and COR (rank 11, the upper right of the dotted lines).

We also show the average ranks of AUC_{ROC} vs. AUC_{PRG} in Fig. 7B. Across both panels, this figure focuses on the discrimination power of the models as both AUCs are a measure of discrimination but AUC_{PRG} (Fig. 7B) focuses more on presences. Ensemble and RF down-sampled still show the best results in all evaluation metrics (Fig. 7). BRT is the third best model based on ranking AUCs, with a slightly better performance than all variants of MaxEnt (Fig. 7B), although this difference is not large and there is no evidence that this difference is statistically significant (Fig. 4 and Appendix S1: Fig. S2). IWLR-GAM, SVM weighted and XGBoost are the only models other than the high-ranking models (ensemble, RF down-sampled, BRT, MaxEnts, and MaxNet), that have an average rank better than the mean rank of all models (upper-right corner, Fig. 7B). The infinitely weighted GAM (IWLR-GAM) is ranked very close to MaxEnt and MaxNet, showing its comparable discrimination power.

Similar to the previous results, RF and SVM with default parameters are ranked the worst. Other than these two models, MARS, GLM, cforests, biomod, and regularized regressions are ranked lower than the mean rank (Fig. 7B). Noticeably, GLM has a worse rank in both AUCs compared to its point process weighted counterpart, i.e., IWLR-GLM. Similarly, El-Gabbas and Dormann (2017) found higher performance of GLMs when applied with point process weighting (in AUC_{ROC} and a threshold dependent metric). IWLR-GAM also has a moderately better rank than (down-weighted) GAM in AUC_{PRG}, but less difference in AUC_{ROC}. This result requires further exploration to understand the links between scaling of the outputs and the errors across presences and absences.

Best performing models per species

Here we continue with the rankings, but rather than focusing on the average rank we focus on how frequently a method is ranked top, in the top 2 or in the top 3 models for any given species. We order methods (left to right) on overall performance (average AUC_{ROC}), to show how these results compare with those discussed so far. Instead of the average of ranks or values, we show the percentage of species in the top 1 to 3 ranks. All methods performed best (top 1) for at least a few species. While the ensembles of our chosen five models (the Ensemble model) achieved the highest overall performance, they were the best models (top one, Fig. 8) only for 3–5% of the species in both AUCs and less than 2% for COR (Fig. 8). However, among the top 3 models, they were second or third for about one-quarter of species (25% in AUC_{ROC}, 21% in AUC_{PRG} and 20% for COR; Fig. 8) and the highest average rank among all (Fig. 7). This indicates that model averaging might not be the best performer for all species, but overall it performs well over many species.

Note that when measured by whether they are among the top 1–3 models, ensembles of models fitted with default settings (biomod, in this case) did not perform particularly strongly, not better than a standard GLM. Similar evidence is provided by Hao et al. (2020) when comparing the performance of biomod ensembles to that of individual models on 14 eucalypt tree species in New South Wales, Australia. They applied the untuned biomod ensemble and individual models in addition to tuned BRT and found no evidence of superiority of the biomod ensemble over the top-performing individual models.

RF with down-sampling achieved the best discrimination and correlation results in terms of being among the top-ranked models, yielding the highest AUC_{ROC}, AUC_{PRG}, and COR for all three groups (top 1, 2, and 3 in Fig. 8). Similarly, BRT and SVM weighted showed consistent strong performance across all groups and evaluation metrics. A particularly noticeable outcome is that the order of the methods in Fig. 8 does not necessarily follow

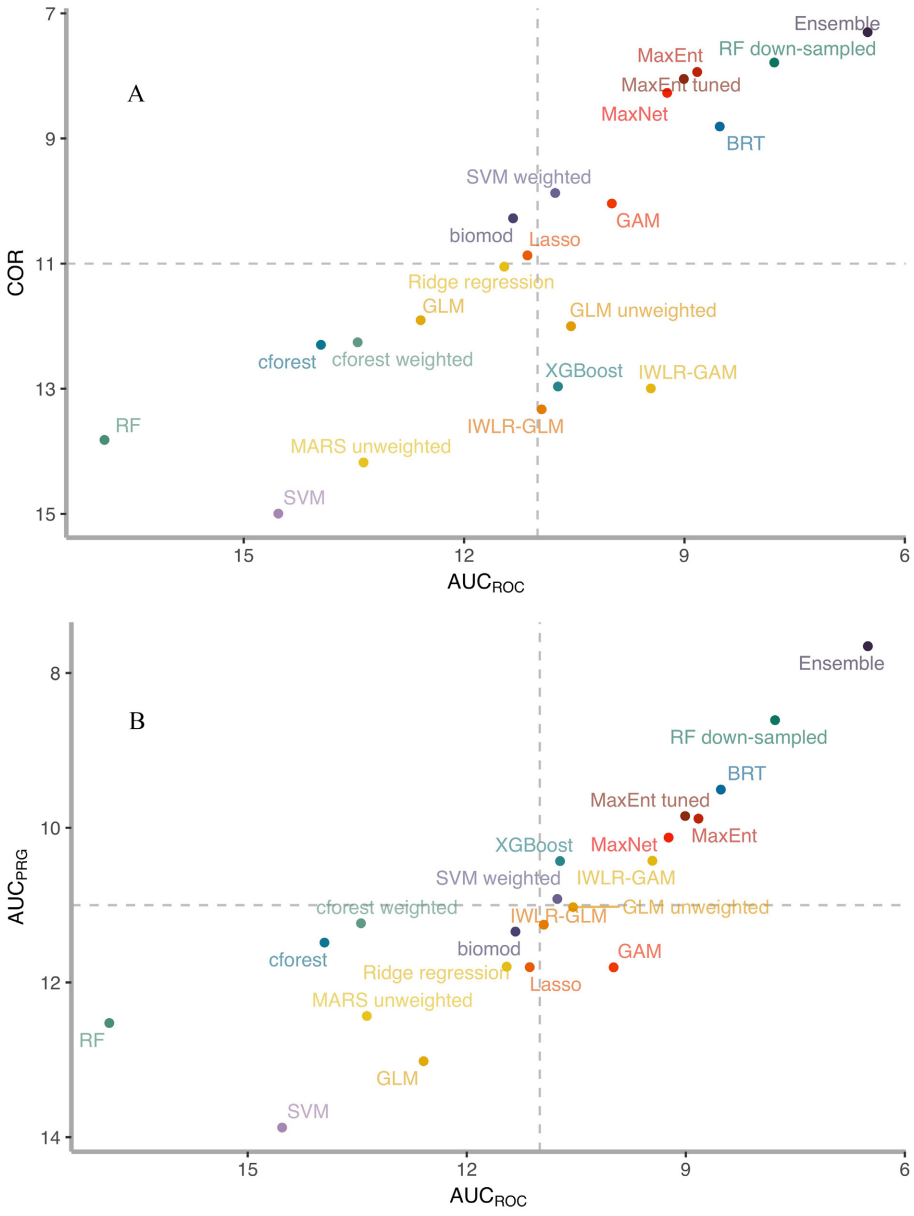

FIG. 7. The average rank of AUC_{ROC} vs. (A) Pearson correlation between the predicted likelihood of presence and the presence–absence testing data (COR) and (B) area under the precision-recall gain curve (AUC_{PRG}) for all the models. The gray dotted lines are the mean ranks. The values of the *x* and *y*-axis are reversed to have the lowest rank (better models) on the top-right corner and vice versa.

the order of the models by their average correlation and discrimination, raw or ranked (in Figs. 3, 7). Specifically, SVM weighted, with medium average AUCs (e.g., rank 11 in AUC_{ROC}) and correlation over all species, is the second or fourth model in percentage of times it achieves top 1, 2, or 3 rankings; its AUC_{ROC} puts it in the top 3 models for 40 species (20%). These results suggest that SVM weighted does very well for some species but presumably quite poorly for others (since its mean performance is not strong; examples in row 8 of Fig. 2). Methods differ in their abilities to model patterns in data (Merow et al. 2014). Depending on the underlying patterns in the data, a very easy modeling task for one technique might be hard for another. For instance, perfectly separable response classes, which are problematic for GLMs (as the estimated coefficient becomes infinite), are very easy for classification trees (Strobl et al. 2009, James et al. 2013).

Conversely, MaxEnt, which was among the best performing models for averaged accuracies (Fig. 3), did not appear often in the top models, especially for the AUC statistics. In other words, MaxEnt predicts strongly overall but is not necessarily the best model for many species.

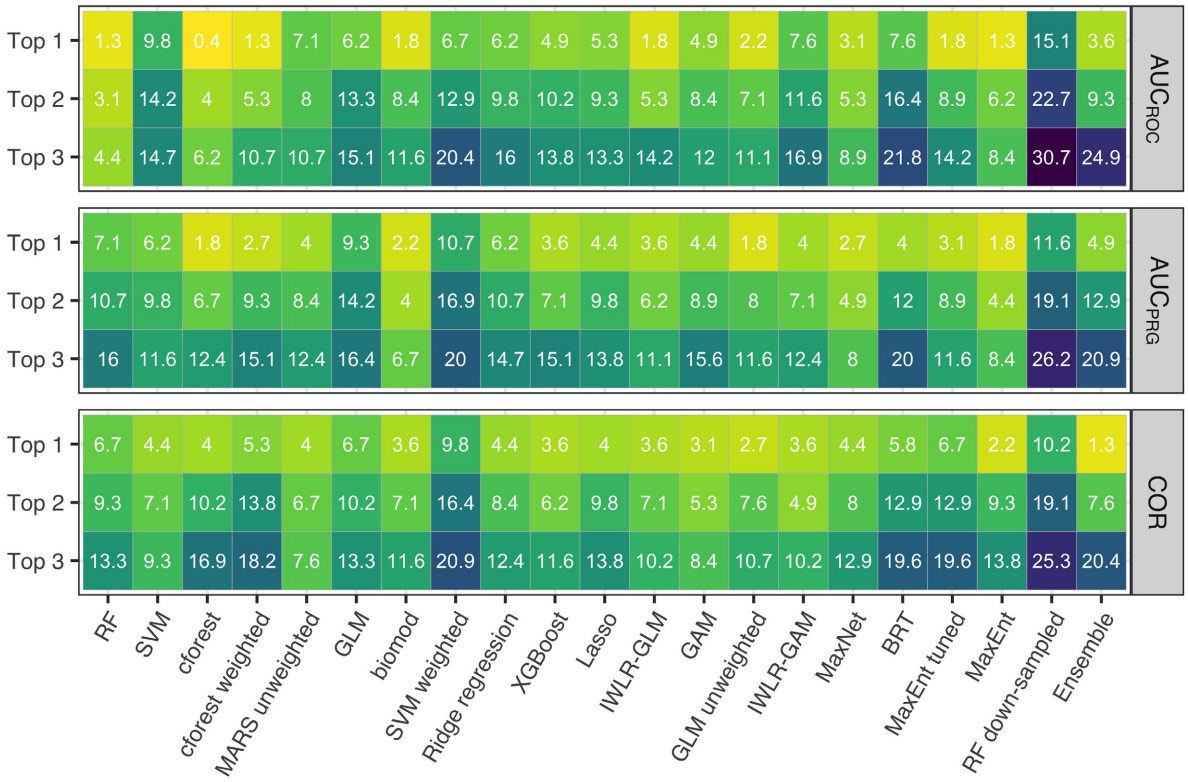

FIG. 8. The percentage of species for which a model was the top model or among the top two or three models. The models are the sorted by overall average AUC_{ROC} from the left (lowest) to the right (highest). The gradient of colors reflects the percentages, from yellow (low) to dark green (high).

This confirms that no one method is superior in all situations (Elith et al. 2006, Pearson et al. 2006, Miller 2010) and, depending on the species and the area of study, some models might obtain a better result, despite not being the best model overall (i.e., mean over all species/locations). It also emphasizes the idea of consistency: perhaps a more reliable method is one that performs consistently well across all regions and species, rather than one that happens to do well on some species (see Fig. 7). Consistency is particularly important because one does not know a priori whether the particular species and region being modeled at any moment in time is likely to be well modeled or poorly modeled by a method. Hence the methods to the far right of Fig. 8 are a safer bet than those like SVM that perform well for a few but poorly on average. Another way of looking at this it to explore whether methods that show promise for some (e.g., SVM) could be made more reliable with different implementations. It is often a good strategy to evaluate the performance of several models to assess which one is performing the best for a specific species.

Performance with very low occurrences

The AUC_{ROC} performance of the models for the species with a low (<30) and moderate-high (≥30) number of presences in the training set is presented in Fig. 9 (with 61 and 164 species in each group, respectively). The threshold of 30 is chosen arbitrarily; we show another threshold in Appendix S1: Fig. S1. This is a very low number of occurrences for the models to capture species distribution properly.

The difference between the low (<30) and high (≥30) species presences is relatively larger for most of the more complex models, e.g., cforest, XGBoost, and SVM, though RF down-sampled is an exception. This difference is smaller for the regression-based models. This is expected as tree-based models are completely data driven, so they need more data to accurately predict the distribution. Several examples can be seen in Fig. 2. For instance, species 15 (from left) in the NZ data set, with low number of presences, performed poorly in all tree-based methods, but not in regression methods. Regression models with parametric or semi-parametric functions predict relatively better than complex nonparametric models with lower number of presences (James et al. 2013). Similarly, MaxEnt performs relatively well, as MaxEnt by default controls the complexity of the fitted functions according to the number of presences, namely, linear is always used, quadratic with at least 10 samples, hinge with at least 15, and product with at least 80 (Elith et al. 2011).

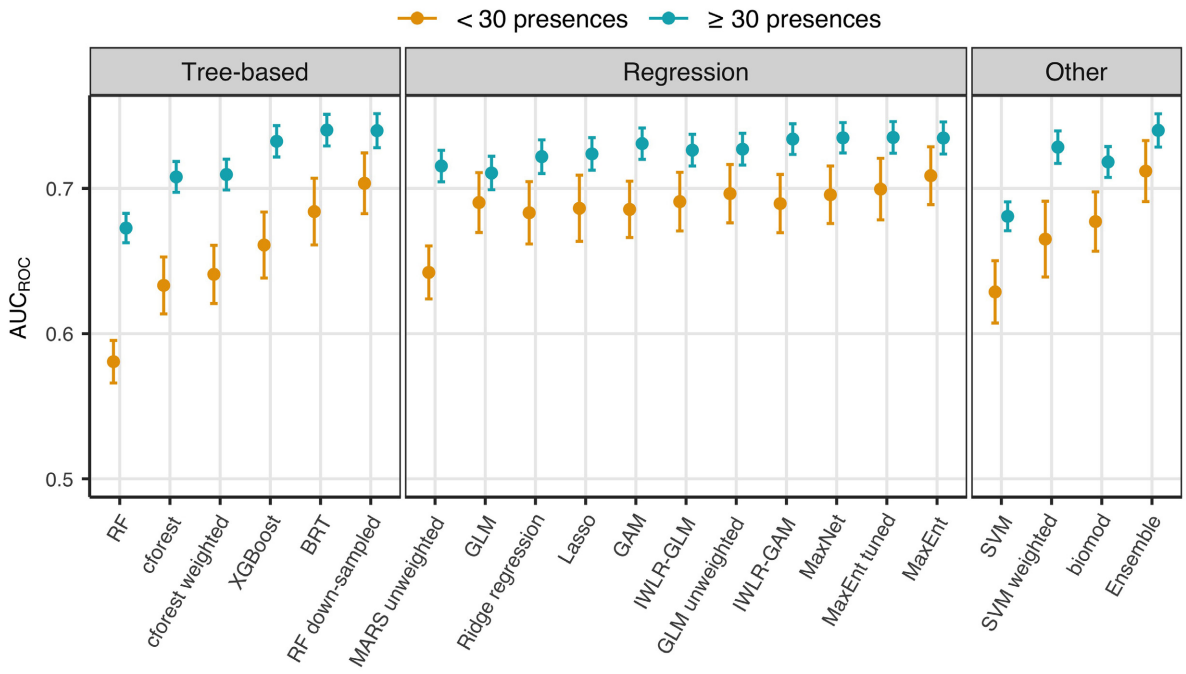

FIG. 9. Mean and standard error of the AUC_{ROC} for the species with <30 and ≥30 presence points (with 61 and 164 species, respectively). Models are arranged based on median AUC_{ROC} in their corresponding group: tree-based, regression, and others.

Regional performance

Fig. 10 shows AUC_{ROC} performance at the regional level (see AUC_{PRG} and COR plots in Appendix S1: Fig. S8). Models for species in SWI and SA achieved the highest AUC_{ROC}, followed by NZ, NSW, AWT, and finally CAN. This is consistent with the result from the 2006 NCEAS study. The ranking is quite similar for AUC_{PRG} but different for COR, which is reasonable as they measure different things. The low performance in CAN is known to be due to the strong sampling bias in this data set (see Fig. 4 in Philips et al. [2009]). With no adjustment for a strong sampling bias, presence-background models will model a combination of environmental suitability and sampling intensity, as the two cannot be untangled (Phillips et al. 2009, Fithian et al. 2015). The presence–absence test data are valuable because they show what has been sampled, allowing a well-informed evaluation of predictions.

The best overall-performing models also have good performance in each region but not necessarily in the same order (Fig. 10). For instance, XGBoost performed almost as well as BRT (a top model) in SWI, RF down-sampled had the highest AUC_{ROC} in SWI and SA. In AWT, CAN, NSW, and NZ the highest AUC_{ROC} is achieved by SVM-weighted, IWLR-GAM, ensemble, GAM, and BRT, respectively. Both RF down-sampled and ensemble performed consistently well in all three evaluation metrics at this regional level (Fig. 10). It is not surprising that the standard RF performs relatively well on the SWI data since this region has the highest number of presence points in our data sets (for presence-only data: mean 1,170, range from 36 to 5,822; Elith et al. 2020).

Spatial prediction

The statistics used for accuracy assessment (AUC_{ROC}, AUC_{PRG}, and COR) report the accuracy of the models on the location of the evaluation points. Visual assessments of the prediction maps are useful for checking whether any of the models predict unlikely patterns. For instance, when models are highly overfitted to the training data, the predicted map is extremely conservative and results in predicting high likelihood/probabilities only around the occurrence data. On the other hand, a map with smooth variation in prediction could be a sign of a simpler model (Merow et al. 2014). For so many species, this visual assessment is not practicable, so instead here we simply illustrate the mapped predictions of several methods (Fig. 11) for one of the modeled species from NZ with 101 occurrences spread over both north and south islands (there was no preference on choosing any species over others). In these maps, predictions reflect the spatial pattern of training presence points, with higher predictions generally aligning with presence locations. The AUC_{ROC} of these models on this species are 0.782, 0.793, 0.797, 0798, 0.805, 0.819, and 0.821 for SVM weighted, Lasso, GAM, RF down-sampled, MaxEnt, BRT, and Ensemble, respectively. AUCs simply report discrimination, testing whether the models tend to predict higher at presence sites rather

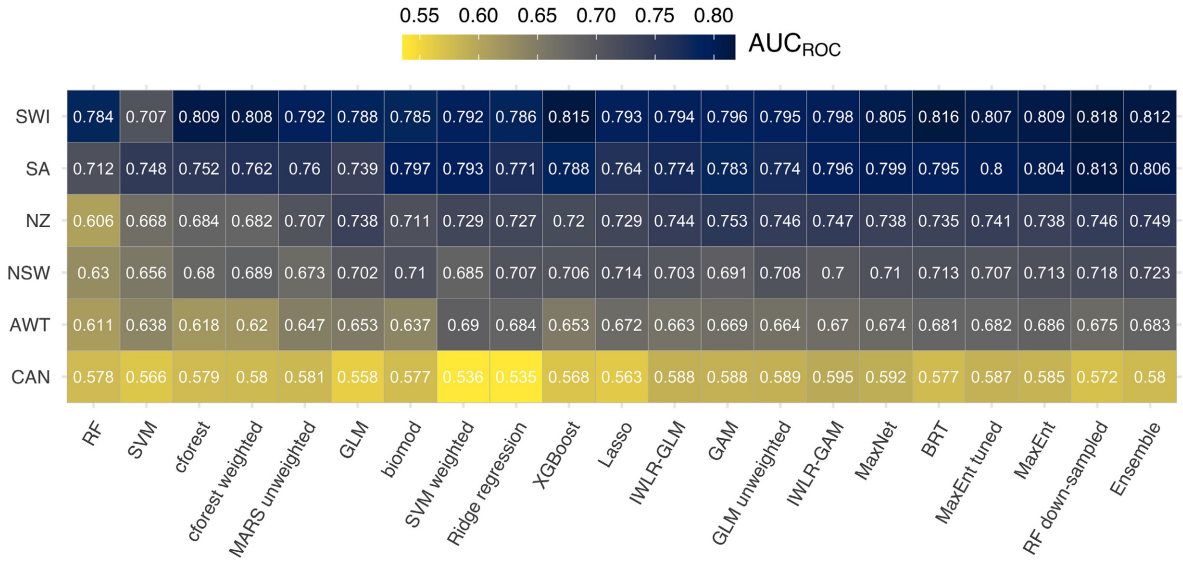

FIG. 10. Mean AUC_{ROC} of each model per region. Models are sorted by overall AUC_{ROC} from lowest (left) to highest (right). Regions are sorted by mean AUC_{ROC} across region, from highest (top) to lowest (bottom).

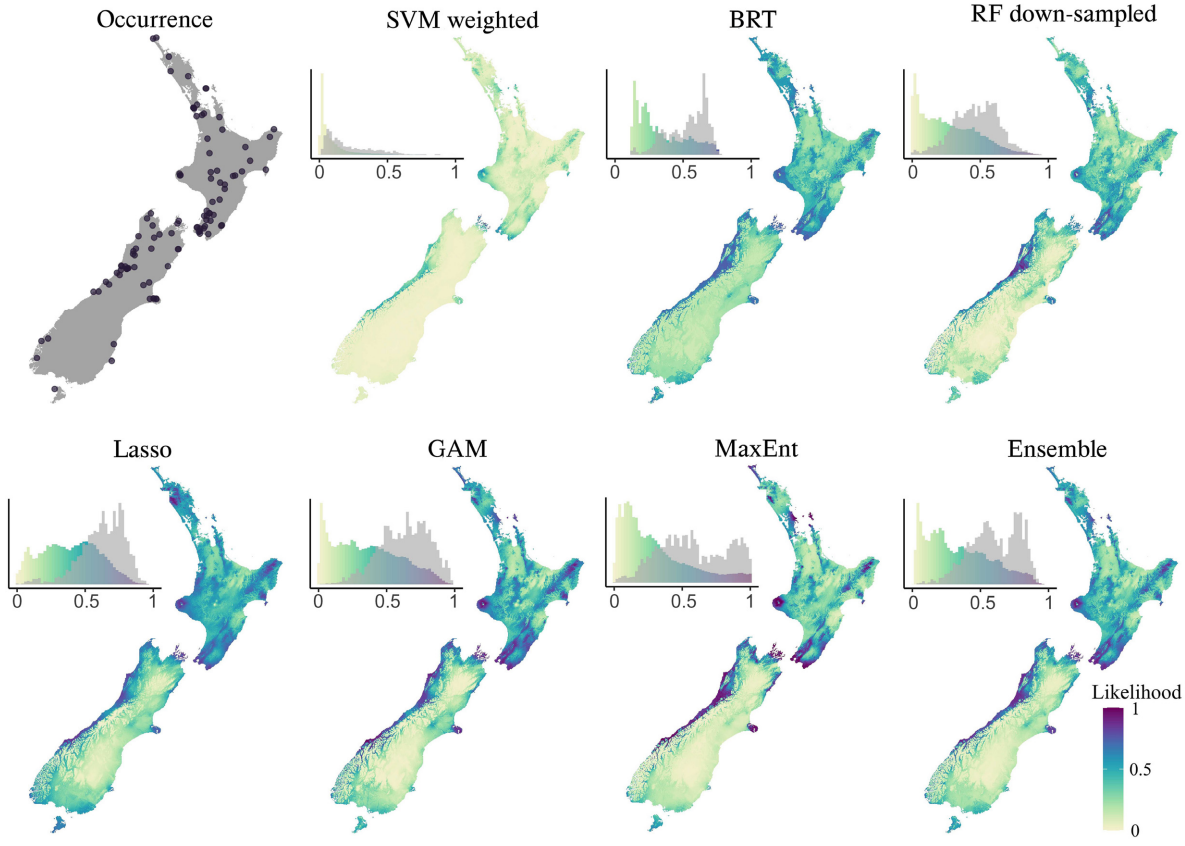

FIG. 11. Map of predicted relative likelihood of occurrence for the species nz30 in the NZ data set. The top-left figure shows the occurrence data used for fitting the models. The colored histogram to the left of each map shows the distribution of predicted values on all grid cells and the transparent gray histogram on top of it, demonstrates the distribution of predicted values on presence points (from testing data). The SVM weighted predicted values were very low, so its map is linearly rescaled between 0 and 1.

than absence sites. The maps and histograms give more insights into the spread of predictions both across the range 0–1 (of all grid cells and testing presence sites), and across the landscape.

Model-fitting computation time

Computation time is another important aspect of modeling methods, often not reported in distribution modeling studies (but see, e.g., Breiner et al. 2017 and Ingram et al. 2020). Due to the high number of species and models, we fitted the models on different machines. Therefore, to obtain a fair comparison of the runtime of the models, we fitted all models on five randomly chosen species from each region (30 species in total; Fig. 12) in a single platform: an online platform with 16 GB of allocated memory and eight CPU cores (similar to a desktop computer). Some methods are very similar regarding their predictive performance, but their computational costs are much different. For instance, RF down-sampled, MaxEnt, and MaxNet achieved a remarkable predictive performance in much shorter time than similar methods, e.g., Ensemble or BRT (Fig. 7). XGBoost was the most computationally expensive model to fit in this study, with more than 100 minutes per species on average. This method with full parameter tuning takes much longer (several hours based on our experience). In contrast, RF down-sampled had an average of 6 s per model fitting, making it the fastest model to fit. The ensemble model, as our best model in all evaluation metrics, had an average of ~30 minutes per model fitting. The regression methods were among the fastest models with less than 2 minutes on average for model fitting (except GAMs and Lasso). As a side note, prediction time of all models are reasonably fast, except the cforest model with a very slow prediction (even slower than model-fitting time).

Conclusion and Perspectives

There are numerous methods available for modeling species geographic distributions using environmental covariates. No single model is superior in all situations, although some modeling approaches generally perform better than others. Natural systems often show complex and nonlinear relationships, autocorrelation and variable interaction across spatial scales. Nonparametric models often outperform traditional parametric models in these situations (Evans and Cushman 2009). Our results show that models capable of fitting complex functions and interactions between covariates tend to obtain higher overall performance when modeling species distributions. BRT, MaxEnt, and RF have strategies to avoid overfitting while they can have a fairly complex response. These capabilities put them among the best performing models in our comparison. It is possible that a GLM

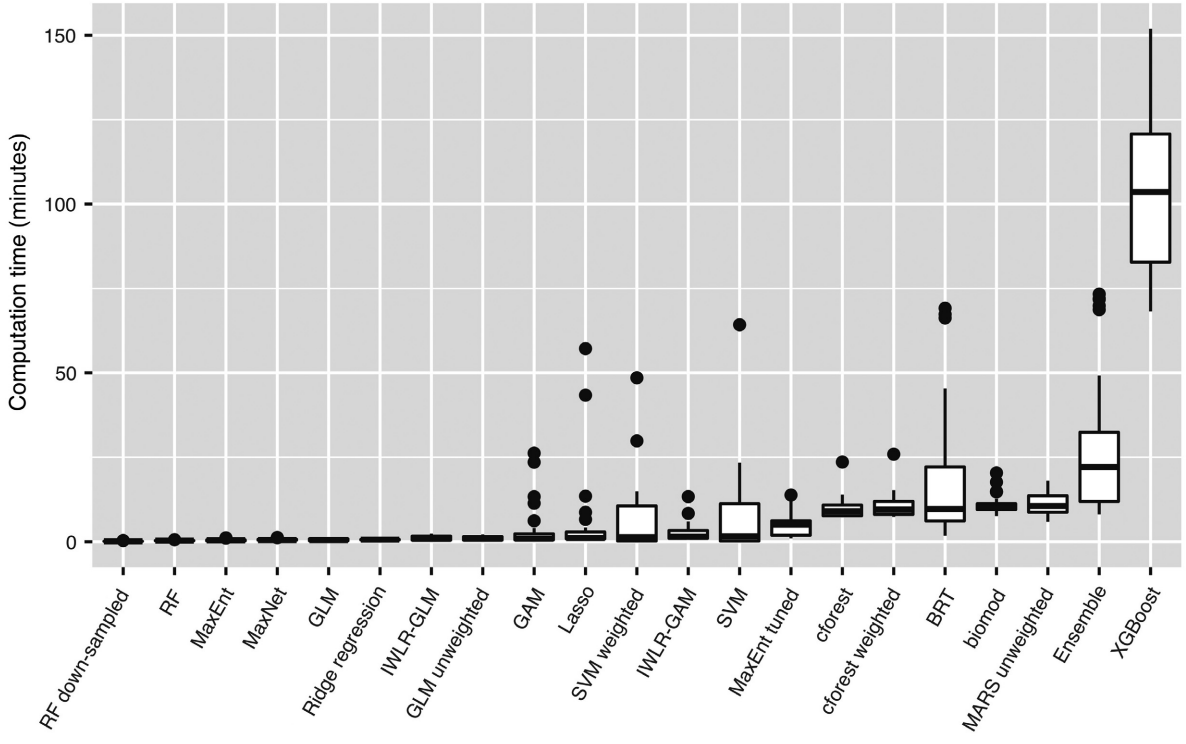

Fig. 12. Model-fitting computation time for each method. Methods are arranged along the *x*-axis based on the median of times (from short to long). The ensemble model runtime was calculated by summing the computation times for all its underlying models (see *Modeling methods*). XGBoost and MARS models were run with multiple CPU cores in parallel (eight cores).

with higher order polynomials and interactions than the one we used here could achieve comparable results. On the other hand, parametric and semi-parametric regression models (like GLM and GAM) can be a good choice especially when the number of occurrences is very low, when a complex and purely data-driven model might not provide a reliable fit. Moreover, GLM and GAM are useful when one wants to do multi-model inference (Burnham and Anderson 2003), ensemble of small models (Breiner et al. 2017), and community models or joint species-distribution models (Wilkinson et al. 2018, Norberg et al. 2019, Ingram et al. 2020).

Presence-background data have the peculiarity of being imbalanced, in that the proportion of background points is much larger than the proportion of presence records. This can be a difficult learning task for some machine-learning methods. Here we introduced weighting and down-sampling techniques for SVM and RF models that led to a substantial improvement in their prediction performance. Similarly, other weighting approaches for regression models i.e., infinitely weighted scheme, achieve better discrimination than the down-weighting approach.

The modeling framework biomod is a popular platform for ensemble modeling and most users take the default settings for granted (Hao et al. 2019). We showed that such settings can lead to suboptimal models, producing only average performance. On the other hand, ensembles of a selection of well-tuned models were the best performing models overall.

Our analysis used independently collected data for evaluating predictive performance of models. However, this does not necessarily guarantee that training and testing points are truly independent, and do not fall close to each other (Bahn and McGill 2012). The closer the points are, the more spatial dependence they have. The lack of independence may favor some methods more than others, particularly those methods that can fit more complex relationships tighter to the data (James et al. 2013). To compare models with truly independent data we would need spatially separated training and testing data. A potential avenue of research in this direction would be to test the predictive performance by accounting for the impact of spatial dependence, e.g., by using block cross-validation techniques (Roberts et al. 2017, Valavi et al. 2019).

Here we presented a comparison of a breadth of statistical and machine-learning models commonly used for species distribution modeling. In addition to specific findings about the performance of alternative techniques, our results emphasize the importance of thinking about the characteristics of presence-background data when choosing how to implement many methods. We fitted all models in the free R programming language and provide example code to facilitate their application to other data sets. As the data we use are now also public (Elith et al. 2020), our comparison is fully reproducible and can serve as basis for future extensions.

Acknowledgment

R. Valavi was supported by an Australian Government Research Training Program Scholarship and a Rowden White Scholarship; G. Guillera-Arroita by an Australian Research Council (ARC) Discovery Early Career Researcher Award (DE160100904), and J. J. Lahoz-Monfort and J. Elith by ARC Discovery Project 160101003. We thank Matthew Cantele for providing the code for Circos software. We also thank Meelis Kull, Nick Golding, Martin Ingram, and David Wilkinson for their helpful suggestions and advice, and two reviewers and our handling editor for insightful comments. This analysis uses the data collated for the working group "Testing alternative methodologies for modeling species' ecological niches and predicting geographic distributions" (project ID: 4980) funded and hosted by the National Centre for Ecological Analysis and Synthesis, Santa Barbara, California. We thank NCEAS for the funding, the project leaders and participants, and those who contributed the data.

Literature Cited

Akbani, R., S. Kwek, and N. Japkowicz. (2004) Applying support vector machines to imbalanced datasets. Pages 39–50 *in* J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, editors, Machine Learning: ECML 2004 Lecture Notes in Computer Science. Springer, Berlin, Germany.

Ali, A., S. M. Shamsuddin, and A. L. Ralescu. 2015. Classification with class imbalance problem: a review. International Journal of Advances in Soft Computing and Its Applications 7:176–204.

Allouche, O., A. Tsoar, and R. Kadmon. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). Journal of Applied Ecology 43:1223–1232.

Anderson, R. P. 2012. Harnessing the world's biodiversity data: promise and peril in ecological niche modeling of species distributions: Niche modeling to harness biodiversity data. Annals of the New York Academy of Sciences 1260:66–80.

Araújo, M. B., et al. 2019. Standards for distribution models in biodiversity assessments. Science Advances 5:eaat4858.

Araújo, M. B., and M. New. 2007. Ensemble forecasting of species distributions. Trends in Ecology & Evolution 22:42–47.

Ashraf, U., A. T. Peterson, M. N. Chaudhry, I. Ashraf, Z. Saqib, S. Rashid Ahmad, and H. Ali. 2017. Ecological niche model comparison under different climate scenarios: a case study of *Olea* spp. in Asia. Ecosphere 8:e01825.

Austin, M. P., and J. A. Meyers. 1996. Current approaches to modelling the environmental niche of eucalypts: implication for management of forest biodiversity. Forest Ecology and Management 85:95–106.

Bahn, V., and B. J. McGill. 2012. Testing the predictive performance of distribution models. Oikos 122:321–331.

Barbet-Massin, M., F. Jiguet, C. H. Albert, and W. Thuiller. 2012. Selecting pseudo-absences for species distribution models: how, where and how many?: How to use pseudo-absences in niche modelling? Methods in Ecology and Evolution 3:327–338.

Beaumont, L. J., et al. 2016. Which species distribution models are more (or less) likely to project broad-scale, climate-induced shifts in species ranges? Ecological Modelling 342:135–146.

Bolker, B. M., M. E. Brooks, C. J. Clark, S. W. Geange, J. R. Poulsen, M. H. H. Stevens, and J.-S.-S. White. 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology & Evolution 24:127–135.

Booth, T. H., H. A. Nix, J. R. Busby, and M. F. Hutchinson. 2014. bioclim: the first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. Diversity and Distributions 20:1–9.

Boyd, K., V. S. Costa, J. Davis, and C. D. Page. 2012. Unachievable region in precision-recall space and its effect on empirical evaluation. Technical Report TR1771. Department of Computer Sciences, University of Wisconsin, Madison, Wisconsin, USA.

Breiner, F. T., M. P. Nobis, A. Bergamini, and A. Guisan. 2017. Optimizing ensembles of small models for predicting the distribution of species with few occurrences. Methods in Ecology and Evolution 9:802–808.

Burnham, K. P., and D. R. Anderson 2003. Model selection and multimodel inference: a practical information-theoretic approach. Second edition. Springer Science & Business Media, Berlin, Germany.

Calabrese, J. M., G. Certain, C. Kraan, and C. F. Dormann. 2014. Stacking species distribution models and adjusting bias by linking them to macroecological models: Stacking species distribution models. Global Ecology and Biogeography 23:99–112.

Calvo, B., and G. Santafé. 2016. scmamp: Statistical comparison of multiple algorithms in multiple problems. R Journal 8:248.

Chen, C., A. Liaw, and L. Breiman. 2004. Using random forest to learn imbalanced data. Technical report. Department of Statistics, University of California, Berkeley, California, USA.

Chen, T., and C. Guestrin. 2016. Xgboost: A scalable tree boosting system. Pages 785–794 *in* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, USA.

Cox, D. R. 1958. Two further applications of a model for binary regression. Biometrika 45:562–565.

Demšar, J. 2006. Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research 7:30.

Doren, B. M. V., and K. G. Horton. 2018. A continental system for forecasting bird migration. Science 361:1115–1118.

Dormann, C. F., et al. 2018. Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference. Ecological Monographs 88:485–504.

Drake, J. M., C. Randin, and A. Guisan. 2006. Modelling ecological niches with support vector machines. Journal of Applied Ecology 43:424–432.

El-Gabbas, A., and C. F. Dormann. 2017. Improved species-occurrence predictions in data-poor regions: using large-scale data and bias correction with down-weighted Poisson regression and Maxent. Ecography 41(7):1161–1172.

Elith, J., et al. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29:129–151.

Elith, J. 2019. Machine learning, random forests, and boosted regression trees. Chapter 15 *in* L. A. Brennan, A. N. Tri, and B. G. Marcot, editors. Quantitative analyses in wildlife science. Johns Hopkins University Press, Baltimore, Maryland, USA.

Elith, J., et al. 2020. Presence-only and presence-absence data for comparing species distribution modeling methods. Biodiversity Informatics 15:69–80.

Elith, J., and J. Leathwick. 2007. Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. Diversity and Distributions 13:265–275.

Elith, J., and J. R. Leathwick 2009. The contribution of species distribution modelling to conservation prioritization. Pages 70–93 *in* A. Moilanen, K. A. Wilson, and H. Possingham, editors. Spatial conservation prioritization: quantitative methods. Volume 1. Oxford University Press, New York, USA; Oxford, UK.

Elith, J., J. R. Leathwick, and T. Hastie. 2008. A working guide to boosted regression trees. Journal of Animal Ecology 77:802–813.

Elith, J., S. J. Phillips, T. Hastie, M. Dudík, Y. E. Chee, and C. J. Yates. 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions 17:43–57.

Evans, J. S., and S. A. Cushman. 2009. Gradient modeling of conifer species using random forests. Landscape Ecology 24:673–683.

Fidler, F., Y. E. Chee, B. C. Wintle, M. A. Burgman, M. A. McCarthy, and A. Gordon. 2017. Metaresearch for evaluating reproducibility in ecology and evolution. BioScience 67:282–289.

Fithian, W., J. Elith, T. Hastie, and D. A. Keith. 2015. Bias correction in species distribution models: pooling survey and collection data for multiple species. Methods in Ecology and Evolution 6:424–438.

Fithian, W., and T. Hastie. 2013. Finite-sample equivalence in statistical models for presence-only data. Annals of Applied Statistics 7:1917.

Flach, P., and M. Kull. 2015. Precision-recall-gain curves: PR analysis done right. Pages 838–846 *in* C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors. Advances in neural information processing systems. Volume 1. Massachusetts Institute of Technology (MIT) Press, Red Hook, New York, USA.

Franklin, J. 2010. Mapping species distributions: spatial inference and prediction. Cambridge University Press, Cambridge, UK.

Freeman, E. A., G. G. Moisen, J. W. Coulston, and B. T. Wilson. 2016. Random forests and stochastic gradient boosting for predicting tree canopy cover: comparing tuning processes and model performance. Canadian Journal of Forest Research 46:323–339.

Friedman, J. H. 2002. Stochastic gradient boosting. Computational Statistics & Data Analysis 38:367–378.

Friedman, J., T. Hastie, and R. Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33:1.

García, S., A. Fernández, J. Luengo, and F. Herrera. 2010. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. Information Sciences 180:2044–2064.

García, S., and F. Herrera. 2008. An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. Journal of Machine Learning Research 9:2677–2694.

Gastón, A., and J. I. García-Viñas. 2011. Modelling species distributions with penalised logistic regressions: A comparison with maximum entropy models. Ecological Modelling 222:2037–2041.

Guillera-Arroita, G., J. J. Lahoz-Monfort, J. Elith, A. Gordon, H. Kujala, P. E. Lentini, M. A. McCarthy, R. Tingley, and B. A. Wintle. 2015. Is my species distribution model fit for purpose? Matching data and models to applications: Matching distribution models to applications. Global Ecology and Biogeography 24:276–292.

Guisan, A., et al. 2013. Predicting species distributions for conservation decisions. Ecology Letters 16:1424–1435.

Guisan, A., and W. Thuiller. 2005. Predicting species distribution: offering more than simple habitat models. Ecology Letters 8:993–1009.

Guisan, A., W. Thuiller, and N. E. Zimmermann 2017. Habitat suitability and distribution models: with applications in R. Cambridge University Press, Cambridge, UK.

Guo, Q., M. Kelly, and C. H. Graham. 2005. Support vector machines for predicting distribution of Sudden Oak Death in California. Ecological Modelling 182:75–90.

Hallgren, W., F. Santana, S. Low-Choy, Y. Zhao, and B. Mackey. 2019. Species distribution models can be highly sensitive to algorithm configuration. Ecological Modelling 408:108719.

Hao, T., J. Elith, G. Guillera-Arroita, and J. J. Lahoz-Monfort. 2019. A review of evidence about use and performance of species distribution modelling ensembles like BIOMOD. Diversity and Distributions 25:839–852.

Hao, T., J. Elith, J. J. Lahoz-Monfort, and G. Guillera-Arroita. 2020. Testing whether ensemble modelling is advantageous for maximising predictive performance of species distribution models. Ecography 43:549–558.

Harris, D. J., S. D. Taylor, and E. P. White. 2018. Forecasting biodiversity in breeding birds using best practices. PeerJ 6: e4278.

Hastie, T., and W. Fithian. 2013. Inference from presence-only data; the ongoing controversy. Ecography 36:864–867.

Hastie, T., R. Tibshirani, and J. Friedman 2009. The elements of statistical learning: data mining, inference, and prediction. Second edition. Springer Series in Statistics. Springer, New York, New York, USA.

He, H., and E. A. Garcia. 2009. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 21:1263–1284.

Hefley, T. J., and M. B. Hooten. 2016. Hierarchical species distribution models. Current Landscape Ecology Reports 1:87–97.

Herdter, E. 2019. Using extreme gradient boosting (XGBoost) to evaluate the importance of a suite of environmental variables and to predict recruitment of young-of-the-year spotted seatrout in Florida. bioRxiv 543181.

Hirzel, A. H., G. Le Lay, V. Helfer, C. Randin, and A. Guisan. 2006. Evaluating the ability of habitat suitability models to predict species presences. Ecological Modelling 199:142–152.

Hothorn, T., K. Hornik, and A. Zeileis. 2006. Unbiased recursive partitioning: a conditional inference framework. Journal of Computational and Graphical Statistics 15:651–674.

Huang, H., F. Hong, J. Liu, C. Liu, Y. Feng, and Z. Guo. 2018. FVID: Fishing vessel type identification based on VMS trajectories. Journal of Ocean University of China 17:1–10.

Hughes-Oliver, J. M. 2018. Population and empirical PR curves for assessment of ranking algorithms. arXiv:1810.08635 [cs, stat].

Ingram, M., D. Vukcevic, and N. Golding. 2020. Multi-output Gaussian processes for species distribution modelling. Methods in Ecology and Evolution 11:1587–1598.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. An introduction to statistical learning, Springer, Berlin, Germany.

Japkowicz, N., and S. Stephen. 2002. The class imbalance problem: A systematic study. Intelligent Data Analysis 6:429–449.

Johnson, R. A., N. V. Chawla, and J. J. Hellmann 2012. Species distribution modeling and prediction: A class imbalance problem. Pages 9–16 *in* 2012 Conference on Intelligent Data Understanding. CIDU, Boulder, Colorado, USA.

Johnston, A., N. Moran, A. Musgrove, D. Fink, and S. R. Baillie. 2020. Estimating species distributions from spatially biased citizen science data. Ecological Modelling 422:108927.

Khalilia, M., S. Chakraborty, and M. Popescu. 2011. Predicting disease risks from highly imbalanced data using random forest. BMC Medical Informatics and Decision Making 11:51.

King, G., and L. Zeng. 2001. Logistic regression in rare events data. Political Analysis 9:137–163.

Krzywinski, M., J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. 2009. Circos: An information aesthetic for comparative genomics. Genome Research 19:1639–1645.

Kuhn, M., and K. Johnson 2013. Applied predictive modeling. Springer, Berlin, Germany.

Kull, M., M. P. Nieto, M. Kängsepp, T. Silva Filho, H. Song, and P. Flach. 2019. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with Dirichlet calibration. Pages 12295–12305 *in* Advances in Neural Information Processing Systems. NeurIPS, Vancouver, Canada.

Lawson, C. R., J. A. Hodgson, R. J. Wilson, and S. A. Richards. 2014. Prevalence, thresholds and the performance of presence–absence models. Methods in Ecology and Evolution 5:54–64.

Leathwick, J. R., J. Elith, and T. Hastie. 2006. Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions. Ecological Modelling 199:188–196.

Liu, C., M. White, G. Newell, and P. Griffioen. 2013. Species distribution modelling for conservation planning in Victoria, Australia. Ecological Modelling 249:68–74.

Marmion, M., M. Parviainen, M. Luoto, R. K. Heikkinen, and W. Thuiller. 2009. Evaluation of consensus methods in predictive species distribution modelling. Diversity and Distributions 15:59–69.

Marra, G., and S. N. Wood. 2011. Practical variable selection for generalized additive models. Computational Statistics & Data Analysis 55:2372–2387.

Merow, C., M. J. Smith, T. C. Edwards, A. Guisan, S. M. McMahon, S. Normand, W. Thuiller, R. O. Wüest, N. E. Zimmermann, and J. Elith. 2014. What do we gain from simplicity versus complexity in species distribution models? Ecography 37:1267–1281.

Meynard, C. N., and J. F. Quinn. 2007. Predicting species distributions: a critical comparison of the most common statistical models using artificial species: Comparison of species-distribution models. Journal of Biogeography 34: 1455–1469.

Mi, C., F. Huettmann, Y. Guo, X. Han, and L. Wen. 2017. Why choose Random Forest to predict rare species distribution with few samples in large undersampled areas? Three Asian crane species models provide supporting evidence. PeerJ 5: e2849.

Miller, J. 2010. Species distribution modeling. Geography Compass 4:490–509.

Muñoz-Mas, R., E. Gil-Martínez, F. J. Oliva-Paterna, E. J. Belda, and F. Martínez-Capel. 2019. Tree-based ensembles unveil the microhabitat suitability for the invasive bleak (*Alburnus alburnus* L.) and pumpkinseed (*Lepomis gibbosus* L.): Introducing XGBoost to eco-informatics. Ecological Informatics 53:100974.

Muscarella, R., P. J. Galante, M. Soley-Guardia, R. A. Boria, J. M. Kass, M. Uriarte, and R. P. Anderson. 2014. ENMeval: an R package for conducting spatially independent evaluations and estimating optimal model complexity for Maxent ecological niche models. Methods in Ecology and Evolution 5:1198–1205.

Muscatello, A., J. Elith, and H. Kujala. 2021. How decisions about fitting species distribution models affect conservation outcomes. Conservation Biology 35:1309–1320.

Norberg, A., et al. 2019. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. Ecological Monographs 89: e01370.

Pearce, J., and S. Ferrier. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecological Modelling 133:225–245.

Pearson, R. G., W. Thuiller, M. B. Araújo, E. Martinez-Meyer, L. Brotons, C. McClean, L. Miles, P. Segurado, T. P. Dawson, and D. C. Lees. 2006. Model-based uncertainty in species range prediction. Journal of Biogeography 33:1704–1711.

Phillips, S. J., R. P. Anderson, M. Dudík, R. E. Schapire, and M. E. Blair. 2017. Opening the black box: an open-source release of Maxent. Ecography 40:887–893.

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190:231–259.

Phillips, S. J., and M. Dudík. 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31:161–175.

Phillips, S. J., M. Dudík, J. Elith, C. H. Graham, A. Lehmann, J. Leathwick, and S. Ferrier. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications 19:181–197.

Phillips, S. J., and J. Elith. 2010. POC plots: calibrating species distribution models with presence-only data. Ecology 91:2476–2484.

Phillips, S. J., and J. Elith. 2013. On estimating probability of presence from use–availability or presence–background data. Ecology 94:1409–1419.

Plummer, M. 2003. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Pages 1–10 *in* Proceedings of the 3rd International Workshop on Distributed Statistical Computing, Vienna, Austria.

Powney, G. D., and N. J. B. Isaac. 2015. Beyond maps: a review of the applications of biological records: Applications of Biological Records. Biological Journal of the Linnean Society 115:532–542.

Prasad, A. M. 2018. Machine learning for macroscale ecological niche modeling-a multi-model, multi-response ensemble technique for tree species management under climate change. Pages 123–139 *in* Machine learning for ecology and sustainable natural resource management. Springer, Berlin, Germany.

Prasad, A. M., L. R. Iverson, and A. Liaw. 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9:181–199.

Prati, R. C., G. E. Batista, and M. C. Monard 2004. Class imbalances versus class overlapping: an analysis of a learning system behavior. Pages 312–321 *in* Mexican International Conference on Artificial Intelligence. Springer, Berlin, Germany.

Probst, P., A.-L. Boulesteix, and B. Bischl. 2019. Tunability: Importance of hyperparameters of machine learning algorithms. Journal of Machine Learning Research 20:1–32.

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

Radosavljevic, A., and R. P. Anderson. 2014. Making better Maxent models of species distributions: complexity, overfitting and evaluation. Journal of Biogeography 41:629–643.

Raes, N., and H. ter Steege. 2007. A null-model for significance testing of presence-only species distribution models. Ecography 30:727–736.

Reineking, B., and B. S. der. 2006. Constrain to perform: regularization of habitat models. Ecological Modelling 193:675–690.

Renner, I. W., J. Elith, A. Baddeley, W. Fithian, T. Hastie, S. J. Phillips, G. Popovic, and D. I. Warton. 2015. Point process models for presence-only analysis. Methods in Ecology and Evolution 6:366–379.

Renner, I. W., and D. I. Warton. 2013. Equivalence of MAXENT and Poisson point process models for species distribution modeling in ecology. Biometrics 69:274–281.

Roberts, D. R., et al. 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40:913–929.

Robinson, O. J., V. Ruiz-Gutierrez, and D. Fink. 2018. Correcting for bias in distribution modelling for rare species using citizen science data. Diversity and Distributions 24: 460–472.

Saito, T., and M. Rehmsmeier. 2015. The Precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS ONE 10: e0118432.

Segurado, P., and M. B. Araujo. 2004. An evaluation of methods for modelling species distributions. Journal of Biogeography 31:1555–1568.

Shabani, F., L. Kumar, and M. Ahmadi. 2016. A comparison of absolute performance of different correlative and mechanistic species distribution models in an independent area. Ecology and Evolution 6:5973–5986.

Shaeri Karimi, S., N. Saintilan, L. Wen, and R. Valavi. 2019. Application of machine learning to model wetland inundation patterns across a large semiarid floodplain. Water Resources Research 55:8765–8778.

Shaffer, J. P. 1986. Modified sequentially rejective multiple test procedures. Journal of the American Statistical Association 81:826–831.

Sofaer, H. R., J. A. Hoeting, and C. S. Jarnevich 2019. The area under the precision-recall curve as a performance metric for rare binary events. Methods in Ecology and Evolution 10:565–577.

Soininen, J., and M. Luoto. 2014. Predictability in species distributions: a global analysis across organisms and ecosystems: Predictability in species distributions. Global Ecology and Biogeography 23:1264–1274.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. 2008. Conditional variable importance for random forests. BMC Bioinformatics 9:307.

Strobl, C., J. Malley, and G. Tutz. 2009. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychological Methods 14:323–348.

Thuiller, W., B. Lafourcade, R. Engler, and M. B. Araújo. 2009. BIOMOD—a platform for ensemble forecasting of species distributions. Ecography 32:369–373.

Valavi, R., J. Elith, J. J. Lahoz-Monfort, and G. Guillera-Arroita. 2019. blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. Methods in Ecology and Evolution 10:225–232.

Valavi, R., J. Elith, J. J. Lahoz-Monfort, and G. Guillera-Arroita. 2021. Modelling species presence-only data with random forests. Ecography 44:1–12.

Ward, G., T. Hastie, S. Barry, J. Elith, and J. R. Leathwick. 2009. Presence-only data and the EM algorithm. Biometrics 65:554–563.

Warton, D. I., and L. C. Shepherd. 2010. Poisson point process models solve the "pseudo-absence problem" for presence-only data in ecology. Annals of Applied Statistics 4:1383–1402.

Whitehead, A. L., H. Kujala, and B. A. Wintle. 2017. Dealing with cumulative biodiversity impacts in strategic environmental assessment: a new frontier for conservation planning. Conservation Letters 10:195–204.

Wilkinson, D. P., N. Golding, G. Guillera-Arroita, R. Tingley, and M. A. McCarthy 2018. A comparison of joint species distribution models for presence–absence data. Methods in Ecology and Evolution 10:198–211.

Yackulic, C. B., R. Chandler, E. F. Zipkin, J. A. Royle, J. D. Nichols, E. H. Campbell Grant, and S. Veran. 2013. Presence-only modelling using MAXENT: when can we trust the inferences? Methods in Ecology and Evolution 4:236–243.

Zhang, L., F. Huettmann, S. Liu, P. Sun, Z. Yu, X. Zhang, and C. Mi. 2019. Classification and regression with random forests as a standard method for presence-only data SDMs: a future conservation example using China tree species. Ecological Informatics 52:46–56.

Supporting Information

Additional supporting information may be found online at: http://onlinelibrary.wiley.com/doi/10.1002/ecm.1486/full