Received: 1 November 2021 | Accepted: 9 May 2022

DOI: 10.1111/2041-210X.13915

Methods in Ecology and Evolution

RESEARCH ARTICLE

Realising the Promise of Large Data and Complex Models

Improving the predictability and interpretability of co-occurrence modelling through feature-based joint species distribution ensembles

Francisca Powell-Romero[1] | Nicholas M. Fountain-Jones[2] | Anna Norberg[3] | Nicholas J. Clark[1]

[1]School of Veterinary Science, The University of Queensland, Gatton, Qld, Australia

[2]School of Natural Sciences, University of Tasmania, Hobart, TAS, Australia

[3]Centre for Biodiversity Dynamics, Department of Biology, Norwegian University of Science and Technology, Trondheim, Norway

Correspondence
Francisca Powell-Romero
Email: francisca.powell@outlook.com; francisca.powell@uq.net.au

Funding information
ARC Discovery Early Career Researcher Award, Grant/Award Number: DE210101439; Australian Research Council Discovery Project Grant, Grant/Award Number: DP190102020

Handling Editor: Laura Graham

Abstract

1. Species Distribution Models (SDMs) are vital tools for predicting species occurrences and are used in many practical tasks including conservation and biodiversity management. However, the expanding minefield of SDM methodologies makes it difficult to select the most reliable method for large co-occurrence datasets, particularly when time constraints make designing a bespoke model challenging. To facilitate model selection for practical out-of-sample prediction, we consider three major challenges: (a) the difficulty of incorporating multiple functional forms for species associations; (b) the limited knowledge on how characteristics of co-occurrence data impact model performance; and (c) whether individual model predictions could be combined to obtain optimised community predictions without the need for bespoke models.
2. To address these gaps, we propose an ensemble method that uses descriptive features of binary co-occurrence datasets to predict model weightings for a set of candidate SDMs. We demonstrate how this method may be applied through a simple case study that uses five independent Joint Species Distribution Models (JSDMs) and Stacked Species Distribution Models (SSDMs) to predict out-of-sample observations for a diversity of co-occurrence datasets. Moreover, we introduce a novel SSDM that offers the potential to include multiple functional forms for each species while delivering robust community predictions.
3. Our case study highlights two major findings. First, the ability for the feature-based ensemble to offer more robust species co-occurrence predictions compared to other candidate SDMs while providing insights into the data features that impact model performance. Second, the competitiveness of the novel SSDM method for forecasting species co-occurrences, even when using a simple univariate generalised linear model (GLM) as the base model prior to stacking.

4. We conclude that feature-based ensembles can provide ecologists with a useful tool for generating species distribution predictions in a way that is reliable and informative. Moreover, the flexibility of the ensemble and the novel SSDM method both offer exciting prospects for incorporating a diversity of functional forms while prioritising out-of-sample prediction.

KEYWORDS

community modelling, conservation, co-occurrence, ensemble, feature-based, joint species distribution model, predictive modelling, stacked species distribution model.

1 | INTRODUCTION

Ecologists are increasingly faced with the task of fitting hundreds or thousands of species distribution models (SDMs) to large species community datasets for applied purposes such as biodiversity conservation and management (Palacio et al., 2021; Velásquez-Tibatá et al., 2019). Such tasks require methods that offer reliable predictions for unsampled areas under time constraints that inhibit the design of bespoke models, that is, models customised to a particular dataset. A major advance to aid in this task has been the development of multivariate approaches, which more realistically capture the co-occurrence of species and their possible interspecific biotic associations (Araújo & Luoto, 2007; Heikkinen et al., 2007; Leathwick et al., 2006; Ovaskainen et al., 2017). Multivariate models that estimate the occurrences of species jointly, that is, simultaneously for all species in a dataset, are referred to as joint species distribution models (JSDMs). JSDMs include nonparametric methods that learn from patterns in the data and utilise classification algorithms to predict species co-occurrence (Ingram et al., 2020), and parametric methods that model species' responses to environmental variables and account for co-occurrence patterns in the residuals (Norberg et al., 2019; Pollock et al., 2014; Wilkinson et al., 2019) or estimate them as joint responses to latent factors (Hui & Poisot, 2016; Ovaskainen et al., 2017; Warton et al., 2015). Alternatively, parametric and nonparametric methods that predict the occurrence of each species individually and aggregate the outcomes to enable multispecies predictions are described as stacked species distribution models (SSDMs; Algar et al., 2009; Calabrese et al., 2014; Distler et al., 2015; Harris et al., 2018; Zurell et al., 2020). Selecting the most appropriate method for predicting species distributions is no simple task, requiring the user to navigate an expanding field of alternative approaches whose advantages and disadvantages are not immediately clear. We consider three major gaps in the estimation and application of species distribution models for out-of-sample prediction. First, it is challenging to incorporate different functional forms for each species while producing coherent community predictions. Second, there is little consensus on which aspects of observed data impact model performance. Finally, few studies have described how to combine model predictions into ensemble forecasts, a practice that is widely known to reduce prediction bias in other fields.

Although JSDMs can offer reliable predictions of species co-occurrences in some ecological contexts (Franklin, 1998; Norberg et al., 2019; Thuiller et al., 2003), parametric methods often make assumptions about the functional form of species (Vayssières et al., 2000). This can be problematic when estimating the occurrence of multiple species simultaneously, as this is not necessarily a characteristic that is uniform across all species. Nonparametric methods offer an alternative modelling approach by utilising classification algorithms that require no assumptions about the distributions of the data or model residuals, and thus better cater for species with various functional forms (Vayssières et al., 2000). However, a common pitfall of standard SSDMs is overpredicting outcomes by not accounting for shared responses between species (Dubuis et al., 2011; Guisan & Rahbek, 2011; Calabrese et al., 2014; D'Amen, Dubuis, et al., 2015; Zurell et al., 2020). A recent development fills this gap by allowing each binary vector of species occurrences to be modelled independently, using whichever base univariate model is appropriate, after which the predictions are aggregated ('stacked'). The stacking is done by learning possible nonlinear multivariate associations (Xing et al., 2020). The authors propose that a weak learner, that is, a method that performs better than random, is appropriate for modelling the associations between the fitted values of other vectors in the dataset and the residuals of the focal vector. Adjusted predictions are then stacked to obtain multivariate predictions (Xing et al., 2020).

There are several reasons why the method proposed by Xing et al. could be advantageous for species distribution modelling. First, it allows for different functional models for each species, meaning that users can freely incorporate relevant domain expertise without being restricted by a single set of assumptions. Incorporating different functional forms for each species simultaneously in JSDMs is challenging, yet this could easily be accommodated by the stacking approach by fitting univariate nonlinear models to each species prior to aggregating the outcomes. Second, the stacking algorithm can potentially estimate complex, nonlinear species' associations without the need for large variance covariance matrices or latent factors, both of which typically assume linearity and can be computationally demanding when modelling many species.

Another challenge in model selection is the limited understanding on what aspects of observed data impact model performance. Studies that have compared different types of models to aid in the model selection process have found inconsistent results considering the predictive performance of SDMs, JSDMs and SSDMs (Baselga & Araújo, 2010; D'Amen, Pradervand, & Guisan, 2015; Harris et al., 2018; Leathwick et al., 2006; Maguire et al., 2016; Moisen & Frescino, 2002; Norberg et al., 2019; Zhang et al., 2018). In particular, a comparison between all three model types highlighted the need for researchers to undertake the computationally and time-demanding task of fitting subsets of data to various complimentary models before undertaking analysis (Norberg et al., 2019). For tasks when time and data constraints are not an issue, bespoke modelling is a highly suitable approach. However, the applied ecologist working with continuously updated datasets that increase in size and complexity may require more feasible alternatives. In such cases, modelling large datasets without fitting customised models requires a deeper understanding of how variation in underlying data structures impacts model performance. While this has been done to understand the structural properties of the models themselves (Elith et al., 2006; Norberg et al., 2019; Wisz et al., 2008), few studies have delved into the structure of the observed data. Data structures can be quantified through features that measure species and community-level characteristics. These may include species characteristics such as growth rate, elevational distribution range and maximum elevation (Guisan, Graham, et al., 2007; Guisan, Zimmermann, et al., 2007), or extrinsic parameters, such as location error and sample size (Guisan, Graham, et al., 2007; Guisan, Zimmermann, et al., 2007; Norberg et al., 2019; Wisz et al., 2008), features of time-series data (Montero-Manso et al., 2020), or network-based features (Azhagesan et al., 2018). These metrics can provide deeper insights into correlations between data structure and model performance. For example, a dataset with a more sparsely connected co-occurrence network may be more effectively modelled by univariate than multivariate methods. To our knowledge, extensive exploration on how data structures impact SDM predictive performance has not yet been undertaken.

Despite evidence that combining multiple models can improve predictions in diverse scenarios (Atiya, 2020; Gneiting & Raftery, 2005; Murray, 2018; Wang & Srinivasan, 2017), few ecological studies have applied ensemble methods (Araújo & New, 2007). Ensembles offer great advantages for ecological forecasting, as they allow the properties of multiple models to be combined into a single weighted prediction that often reduces prediction error relative to any of its constituent models (Araújo & New, 2007). Moreover, combining models into an ensemble algorithm often allows for a fast approach to yield optimised predictions (Lemke & Gabrys, 2010). No SDM correctly captures the true data generating process, suggesting it can be useful to hedge bets against model misspecification by combining predictions. This is especially true when using a diverse set of candidate models, as combinations from models with different degrees of flexibility should, on average, outperform predictions from individual models across heterogeneous environments. Determining appropriate model weights for ensemble models can be challenging. However, the calculation of features that describe structural differences among observed data offers a direct way to estimate model weights (Kang et al., 2017). Recently, a promising novel time-series ensemble approach was proposed, which uses a suite of descriptive features for each response variable in a multivariate dataset as predictors when training a machine learning algorithm to predict the relative weights of simple forecast models (Montero-Manso et al., 2020). The method won second place in M4, a highly competitive global forecasting competition (Makridakis et al., 2020) and has been applied to both aid in the selection of individual models and to build weighted ensemble models (Kück et al., 2016; Lemke & Gabrys, 2010; Talagala et al., 2018). However, while this method has been applied for economic forecasting purposes, to our knowledge, a similar approach has not been applied in ecological modelling. We propose that using binary community features to both understand why some models outperform others and to combine model predictions into a weighted ensemble is a useful avenue of research for building better predictions for communities of species.

Our study explores the effects of the underlying data structure of binary datasets, and how this can be used to predict model weights within an ensemble model to optimise predictions of species distributions. We do this by evaluating the predictive performance of five candidate models and use the deviance residuals from each respective model to generate optimised model weights. Using the optimised weights as the response variables, we build an ensemble algorithm that learns from features describing the composition of the species communities to predict ensemble weights for generating out-of-sample forecasts, a novel approach not yet applied in the field of ecology. We suggest that our framework can be useful for applied modellers seeking to predict the distributions of large sets of species for practical tasks where it is not feasible to undergo the lengthy process of fitting bespoke models.

2 | MATERIALS AND METHODS

2.1 | Data collection and preparation

We used a total of 30 binary presence–absence co-occurrence datasets across pathogen, vegetation and animal communities. Datasets originally containing abundance or count measures for species occurrence were converted to binary data, where any species with a value equal or >1 was considered to be present, that is, assigned a value of 1. Descriptions, number of species and observations, median prevalence and prevalence range are summarised for each dataset in Table 1 (See Supplementary File 1 for more detailed descriptions). To reduce the risk of overfitting, all covariates for each dataset were standardised using principal component analysis (PCA), and the first five principal components (PCs) were selected as predictors, unless fewer PCs were required to explain at least 80% of the variation in the covariate space, as per Norberg et al. (2019), or if fewer PCs were available (See Supplementary File 1 for description on number of PCs included for each dataset and the cumulative variation explained by these PCs). All individual models were fitted using the same PCs as covariates for each species. For each model apart from the GBM stacking models and the MVRF, covariates were included as additive linear effects. The MVRF can learn nonlinear effects, while the GBM models did not use covariates (the fitted values

TABLE 1 Description of datasets used for analyses. Values reported for number of species (*N* species), number of observations (*N* obs.), number of covariates (*N* covariates) and prevalence (median and range) are the values after data cleaning for analysis. For original values, see Supplementary File 1

No.	Dataset	Species	Test/train	*N* species	*N* obs.	*N* covariates	Prevalence (median and range)	Reference
1	bird_parasites	Malaria parasites in birds	Test	4	449	1	0.156 (0.098–0.265)	Clark et al. (2016)
2	helminths	Soil-Transmitted Helminths in School Children	Train	4	8786	19	0.139 (0.021–0.375)	Ruberanziza et al. (2019)
3	fennoscandia_birds	Birds	Train	141	1,800	21	0.122 (0.010–0.944)	Norberg et al. (2019)
4	uk_butterflies	Butterflies	Test	47	1,800	34	0.452 (0.023–0.948)	Norberg et al. (2019)
5	victoria_plants	Plants	Train	162	1,800	19	0.025 (0.005–0.148)	Norberg et al. (2019)
6	usa_trees	Trees	Train	63	1,800	38	0.043 (0.012–0.339)	Norberg et al. (2019)
7	norway_vegetation	Vegetation	Test	242	1,800	6	0.058 (0.007–0.750)	Norberg et al. (2019)
8	eelgrass	Species found in eelgrass communities	Train	32	96	15	0.276 (0.042–0.885)	Stark et al. (2020)
9	shrews	European Shrews	Train	7	2,921	8	0.163 (0.117–0.687)	(Neves et al., 2022)
10	mussel_parasites	Parasites in mussels	Train	13	720	6	0.200 (0.014–0.731)	Brian and Aldridge (2021)
11	lion_infections	Various infectious pathogens in lions	Train	5	105	11	0.533 (0.333–0.562)	Fountain-Jones et al. (2019)
12	eucalyptus	Eucalyptus	Train	20	327	33	0.090 (0.003–0.284)	Pollock et al. (2015)
13	grassland_birds	Birds	Test	30	560	4	0.040 (0.002–0.421)	Han et al. (2020)
14	mulu_birds	Birds	Test	84	166	3	0.136 (0.036–0.500)	Burner et al. (2019)
15	usa_birds	Birds	Train	101	1,284	28	0.031 (0.001–0.450)	Steen et al. (2020)
16	swiss_birds	Birds	Test	56	1,774	53	0.240 (0.029–0.726)	Zurell et al. (2020)
17	swiss_forest	Trees	Train	63	4,816	45	0.055 (0.012–0.792)	Zurell et al. (2020)
18	fish_parasites	Parasites in fish	Train	42	3,966	8	0.028 (0.001–0.364)	Bolnick et al. (2020)
19	brazil_fish	Fish	Train	66	52	12	0.077 (0.019–0.481)	Vieira et al. (2020)
20	reptiles	Reptiles	Train	104	455	11	0.015 (0.002–0.411)	Escoriza (2020)
21	canopy_ants	Ants	Train	99	153	5	0.039 (0.007–0.693)	Adams et al. (2017)
22	swissalps_plants	Plants	Train	175	912	7	0.080 (0.024–0.476)	D'Amen et al. (2018)
23	earthworms	Earthworms	Test	97	1,352	4	0.004 (0.001–0.708)	Mathieu and Jonathan Davies (2014)
24	vines	Vines	Test	42	50	16	0.070 (0.020–0.780)	Delgado and Restrepo (2019)
25	buffalo_infections	Various infectious pathogens in buffalo	Train	6	343	10	0.106 (0.088–0.185)	Glidden et al. (2021)
26	andean_birds	Birds	Test	159	358	2	0.022 (0.003–0.411)	Montaño-Centellas (2020)
27	finland_beetles	Beetles	Train	239	152	16	0.118 (0.026–0.941)	Burner et al. (2021)

(Continues)

TABLE 1 (Continued)

No.	Dataset	Species	Test/train	*N* species	*N* obs.	*N* covariates	Prevalence (median and range)	Reference
28	germany_beetles	Beetles	Train	75	386	11	0.031 (0.013–0.277)	Burner et al. (2021)
29	norway_beetles	Beetles	Test	125	1111	14	0.023 (0.005–0.369)	Burner et al. (2021)
30	nz_forest	Trees	Train	205	964	2	0.004 (0.001–0.500)	Popovic et al. (2019)

and residuals from the univariate GLM were already conditioned on covariates prior to their inclusion in the GBM models).

Training our ensemble required measures of model predictive performance across a large number of datasets with a diversity of binary feature profiles. Training datasets were selected by stratifying the number of species in each community, number of PCs and median prevalence into three groups (low, medium and high values). These values were used to select 10 of the 30 datasets to be withheld for testing. One dataset from each combination of three stratified variables was withheld. In cases where only one combination was present, the dataset was withheld as a testing dataset to enable extrapolation. Datasets retained for training and withheld for testing are described in (Table 1). A total of 20 datasets, containing 1,622 binary vectors (64.67%) were used as training datasets, and 10 datasets, containing 886 binary vectors (35.33%) were withheld as testing datasets for the final ensemble model. The median prevalence for the training and testing data was 5.23% (Q1 = 1.81%; Q3 = 13.34%) and 5.17% (Q1 = 1.61%; Q3 = 16.03%) respectively (See also Supplementary File 2 for a visualisation of feature diversity in training and testing datasets). Although we acknowledge that not every vector is necessarily a different species since some species may be present in multiple datasets, the features will vary at the species level when measured in different communities, and therefore for clarity, binary vectors will be referred to as 'species'.

2.2 | Fitting multivariate models and obtaining predictive performance metrics

We fitted a total of five individual models to the 20 training datasets, to replicate what modellers may be faced with if modelling hundreds of species with limited resources. Three of the models will likely be familiar to quantitative ecologists. They included (a) a generalised linear model (Bernoulli outcomes with a logit link function) to be used as the univariate baseline predictions for comparison (GLM-BASE), which was fitted by applying iteratively reweighted least squares; (b) a Multivariate Random Forest model (MVRF) fitted using the Fast Unified Random Forests for Survival, Regression and Classification function (Ishwaran et al., 2008), using a node size of 8 to define the average number of observations in a terminal node; and (c) a Hierarchical Modelling of Species Communities model (HMSC; Tikhonov et al., 2021), which was fitted using two MCMC chains with a burn-in of 2,000 and 1,000 iterations, and with default priors for all model parameters (see Table 2 and Supplementary File 3 for further descriptions on these methods and R packages used).

To our knowledge, the two remaining models have not been previously used in ecological applications, hence we describe them in more detail here. These models take the original in-sample predictions from a univariate model (in our case, a generalised linear regression model) and learns from the errors in a stacking algorithm to adjust the out-of-sample predictions. In our approach the errors (i.e. residuals from a focal species' GLM) are modelled as a function of the fitted values from other species' univariate GLMs. This allows

TABLE 2 Description of models used to predict species occurrence

Model	Abbreviation	Type	Multi-outcome method	Parametric/nonparametric	R packages	Source
Generalised Linear Model (Baseline)	GLM-BASE	Univariate	Stacked Species Distribution Model	Parametric	STATS	R Core Team (2021)
Gradient Boosted Model – Pearson Residuals	GBM-PR	Univariate	Stacked Species Distribution Model	Nonparametric	GBM	Greenwell et al. (2020)
Gradient Boosted Model – Deviance Residuals	GBM-DR	Univariate	Stacked Species Distribution Model	Nonparametric	GBM	Greenwell et al. (2020)
Multivariate Random Forest	MVRF	Multivariate	Joint Species Distribution Model	Nonparametric	RANDOMFORESTSRC	Ishwaran et al. (2008)
Hierarchical Modelling of Species Communities	HMSC	Multivariate	Joint Species Distribution Model	Parametric	HMSC	Tikhonov et al. (2021)

the model to uncover potentially nonlinear species associations, avoids the need to parameterise a covariance matrix or set of latent factors, and ensures that out-of-sample predictions can be made for the entire community. We included two versions of the stacking model: one that uses the Pearson Residuals (PR) and another that uses Deviance Residuals (DR) as per Xing et al. (2020) from the individual species as the outcome, with the fitted values from the other species included as features in the stacking algorithm. These residuals are defined as:

Pearson Residual:

$$r_{ik} = \frac{\left(Y_{ik} - \hat{P}_{ik}\right)}{\sqrt{\hat{P}_{ik}\left(1 - \hat{P}_{ik}\right)}}.$$

Deviance Residual:

$$r_{ik} = \begin{cases} \sqrt{-2\log\left(\hat{P}_{ik}\right)} \text{ if } Y_{ik} = 1 \\ -\sqrt{-2\log\left(\hat{P}_{ik}\right)} \text{ if } Y_{ik} = 0 \end{cases},$$

where r_{ik} denotes the residual for the k-th outcome for the i-th sample in the univariate GLM model, $\hat{P}_{ik}$ denotes the predicted probability from the GLM model and Y_{ik} denotes the binary outcome. Following the fitting of the GBM stacking model, the adjustment of the GLM original univariate predictions was made following Xing et al. (2020). Both stacking models were fitted using the GBM package in R (Greenwell et al., 2020), and parameters were tuned using 50 trees, a maximum depth for each tree of 2, and the default shrinkage parameter of 0.1. We specifically used a weak learner to reduce overfitting and prioritise species associations that may be important for out-of-sample predictions.

Each of the training datasets were split into training and testing folds, whereby 70% of the data was randomly selected for fitting the models and the remaining 30% was used for evaluating model predictions. This process was repeated three times for each dataset to capture heterogeneity in model performance among testing folds. To measure predictive accuracy of the five individual models, we binarised predicted probabilities of occurrence into presence/absence using a standard threshold of 0.5 for simplicity, since our datasets contained a range of median prevalence values and species, and as models were not optimised to improve predictions for a particular community but rather compare model performance. Using the binarised predictions, we calculated out-of-sample recall (the ratio of correctly predicted species present to all observations where the species is actually present), precision (the ratio of correctly predicted species present to the total species predicted to be present) and the F1 statistic (the weighted average of precision and recall). We used the F1 score instead of accuracy (total number of correctly predicted observations over the total number of observations) or area under the receiver operating characteristic curve (AUROC; uses the area under the ROC curve that plots sensitivity and specificity to quantify the performance of a model) due to the likely unequal distribution of false positives and false negatives resulting from the large proportion of rare species. This metric is calculated as:

$$\text{F1score} = \frac{2(\text{Recall} \times \text{Precision})}{(\text{Recall} + \text{Precision})}.$$

2.3 | Ensemble model

Our goal was to find a weighted ensemble of model predictions (on the probability scale) that could minimise an appropriate binary loss function. In practise, for each species in each evaluation set (i.e. containing the with-held 30% of observations), we optimised weights that minimised the mean squared deviance residual. We accounted for class imbalance by weighting residuals for positive and negative observations by their respective frequencies in the test set when calculating the final mean residual. Optimisations of the unknown model weights were performed using the L-BFGS-B algorithm (Byrd et al., 1995) in the R function OPTIM of the STATS package (R Core Team, 2021). For all species we used five separate optimisations with different random starting weights to ensure the parameter space was adequately explored. Final model weights for each species were calculated by taking the mean from the three sets used for training.

Our ensemble model was a multivariate random forest that was trained to predict optimal model weights for a set of binary observations based on features that described the structures and community contexts of those observations. We calculated 23 features to describe the characteristics of species individually and within their community, as well as features to describe the overall nature of community structure (Table 3). These features included three measures of prevalence, the numbers of observations and species, network analysis metrics, measures of species 'uniqueness', measures describing characteristics of the Markov Random Field (MRF) Networks, and features that describe the predictors and covariates for each of the datasets (See Supplementary File 4 for histograms showing the distribution of features across all, training, and testing datasets). Note that this set of features is not exhaustive, and it would be fruitful and ecologically interesting to consider other features to describe variation among species' observation vectors.

2.4 | Ensemble model performance

We used the 10 datasets excluded from the model training to test the predictive accuracy of our ensemble model relative to the individual models. We again used a 70–30 split for validation. For the training dataset containing 70% of the data, we fit the candidate models as described above. We then calculated the 23 features to use as new data in the ensemble algorithm ('ENS') to predict weights for each species to generate weighted ensemble predictions. We also generated a null ensemble model ('NULL-ENS') for comparison that assigned equal weightings for each candidate model. We then calculated performance metrics as above for the five individual models as well as the two ensemble models.

As our case study aimed to describe a proof-of-concept, all models used in our study were fitted using default configurations. However, it is important to note that an ensemble could just as easily be fitted to bespoke models to capture domain knowledge and tune model parameters, which would likely increase prediction performance. All models were implemented in the R environment, version 4.0.2 (R Core Team, 2021).

3 | RESULTS

3.1 | Variability among individual model performance

Models were compared based on their predictive performance using classification metrics (recall, precision and F1) for a total of 1,622 binary vectors (referred to as 'species' here), which we grouped into four prevalence groups for initial exploration: rare, with prevalence <10% ($n = 1{,}110$), uncommon (prevalence 10 to 30%; $n = 339$), common (prevalence 30 to 75%; $n = 160$) and very common (prevalence >75%; $n = 13$). For rare species, out-of-sample F1 performance was comparable between the GBM-DR and HMSC methods, which both performed substantially better than the GLM-BASE by 52.34% and 48.11% respectively (Figure 1). Similarly, for uncommon species HMSC (70.50% average net improvement) and GBM-DR (59.88% improvement), along with the GBM-PR (44.54% improvement), performed better than the base, while MVRF performed slightly worse (by 1.77%). The relative performances of HMSC and GBM-DR were highest for uncommon species and decreased as prevalence increased, with GBM-DR performance falling below the GLM-BASE model performance for common species (by 6.25%) and for both GBM-DR and HMSC for very common species (by 84.62% and 100.00% respectively). HMSC and GBM-DR both showed higher recall values compared to the GLM-BASE model across all prevalence categories except for 'Very Common', where they both performed significantly worse than the GLM-BASE in terms of recall (both by 100.00%). HMSC and GBM-DR also showed improvements over the GLM-BASE in terms of precision for 'Rare' species (by 24.23% and 35.67% respectively). See Supplementary File 5 for all comparisons for precision and recall, as well as values used to calculate percentages of net improvement for F1 by prevalence category.

3.2 | Predicted model performance based on data features

Across the datasets used to train the ensemble, the mean weighting as a percentage for each model in the ensemble were: 8.80% for GLM-BASE, 23.70% for GBM-DR, 7.95% for GBM-PR, 70.39% for HMSC and 10.52% for MVRF. Predicted response functions from the ensemble can be used to interrogate how model performance is related to particular features of a community dataset, providing useful insights for improving both domain knowledge and model

TABLE 3 Description of features used to define community structures for inclusion in ensemble as predictors of model weights. Value range shows the min and max values for each feature across both training and testing datasets

No.	Feature	Description	Level	Value range
1	Prevalence	Describes how rare or common a species is	Species	0.001, 0.948
2	Prevalence Rank	Describes how rare or common a species is relative to the other species within a community	Species	0.004, 1
3	Prevalence Standard Deviation	Describes how much variation in prevalence there is within a community	Community	0.026, 0.326
4	Number of observations	Describes how many sampling units are present in the dataset	Community	50, 8786
5	Number of Species	Describes how many species are present within a community	Community	4, 242
6	Degree Centrality	Describes the number of species with which one species co-occurs	Species	0, 1
7	Eigenvector Centrality	Describes how influential one species is within the community	Species	<0.001, 1
8	Betweenness Centrality	Describes how influential one species is within a community	Species	0, 1.415
9	Modularity (Newman's Q)	Describes the structure of the species network in terms of clustering	Community	−1.459, 0.515
10	Mean Jaccard Distance	Describes how unique individual species are relative to others	Species	0.659, 1
11	Mean Jaccard Distance Standard Deviation	Describes the variation in how unique species in a community are	Community	0.004, 0.119
12	Mean Sørensen–Dice Distance	Describes how unique individual species are relative to others	Species	0.539, 1
13	Mean Sørensen–Dice Distance Standard Deviation	Describes the variation in how unique species in a community are	Community	0.010, 0.138
14	Mean Sørensen Index	Describes the similarity between two samples of binary observations	Species	0.355, 0.962
15	Mean Sørensen Index Standard Deviation	Describes the variation of the Sørensen Index within the community	Community	0.093, 0.345
16	MRF Intercept	Describes the probability of occurrence (on the logit scale) when all other species are equal to 0	Community	−49.943, 4.066
17	MRF Network Information	Describes how connected the MRF graph is overall. This metric is normalised by the number of species in the data	Community	0.641, 85.825
18	MRF Network Information Standard Deviation	Describes the variation in the MRF Network Information within a community	Community	0.134, 2.076
19	MRF Trace	Describes the total amount of dispersion of the variables in the MRF network	Community	−2.734, 4.387
20	Log Determinant	Describes the correlations among pairs of variables in the MRF network	Community	−0.943, 0.177
21	Number of Covariates	The number of raw predictors in the dataset used to run the PCA to prepare covariates for analysis	Community	1, 53
22	Number of PCs	The number of PCs included as covariates in the analysis	Community	1, 5
23	Cumulative Variation Explained by PCs	The cumulative variation explained by the PCs included in the analysis	Community	0.407, 1

performance. In our case study, prevalence, eigenvector centrality and degree centrality were the top three most important predictors of variation in performance across all five models, while betweenness centrality was the least informative (Figure 2). Across all metrics, HMSC was consistently attributed the highest weights, however showed greatest variability across prevalence values. For rare species, HMSC was the clearly prioritised method (Figure 3). For common species and, in particular, species with mid-range prevalence values, the differences in weights between HMSC and MVRF were much less pronounced (See Supplementary File 6 for the response functions for the remaining 20 features included in our case study). With the exception of prevalence, which ranks as the most important predictor of model weighting for GLM-BASE. GBM-PR, HMSC and MRF, the most influential features on model weights were co-occurrence network features, with eigenvector centrality surpassing prevalence as the most important predictor for GBM-DR. GBM-DR and HMSC were most influenced by the 23 features overall, with higher relative importance values across multiple features compared to the other models. In particular, the contrast between the two models in terms of feature importance highlights that individual features will influence performance differently for each model (Figure 2).

3.3 | Ensemble model performance comparable with best performing models

We tested the predictive performance of our ensemble (ENS) and an equally weighted ensemble (NULL-ENS). Overall, the GBM-DR performed the best based on the F1 statistic, followed by the ENS and HMSC (Figure 4). Out of the 886 species included in the final validation set, the ENS had the greatest net improvement (51.13%), followed by HMSC (48.87%) and GBM-DR (46.50%; Table 4). Of all six models tested, GBM-PR and the ENS provided the most robust predictions by yielding the lowest number of F1 metrics that were worse than GLM-BASE (3.95% and 5.87% respectively), followed by the MVRF (7.79 %) and the GBM-DR (8.92%). The GBM-DR method showed the highest improvement in precision (34.20%), followed by ENS (33.30%) and HMSC (24.60%). In contrast, HMSC showed the highest improvement in recall (68.85%), followed by the ENS (59.26%) and the GBM-DR (48.98%; see Supplementary File 7 for the tabulated results for precision and recall values and boxplots, as well as results for accuracy and deviance residual performance metrics).

4 | DISCUSSION

Given the overwhelming volume of SDMs available and their high variability in performance for predicting species distributions, selecting an appropriate model for analysis is not a straight-forward task and often requires the lengthy process of fitting several models with complementary performance. This is not always feasible for ecologists seeking to model hundreds or thousands of species under time constraints. We proposed an ensemble approach that could be used to determine a weighted value for the performance of each

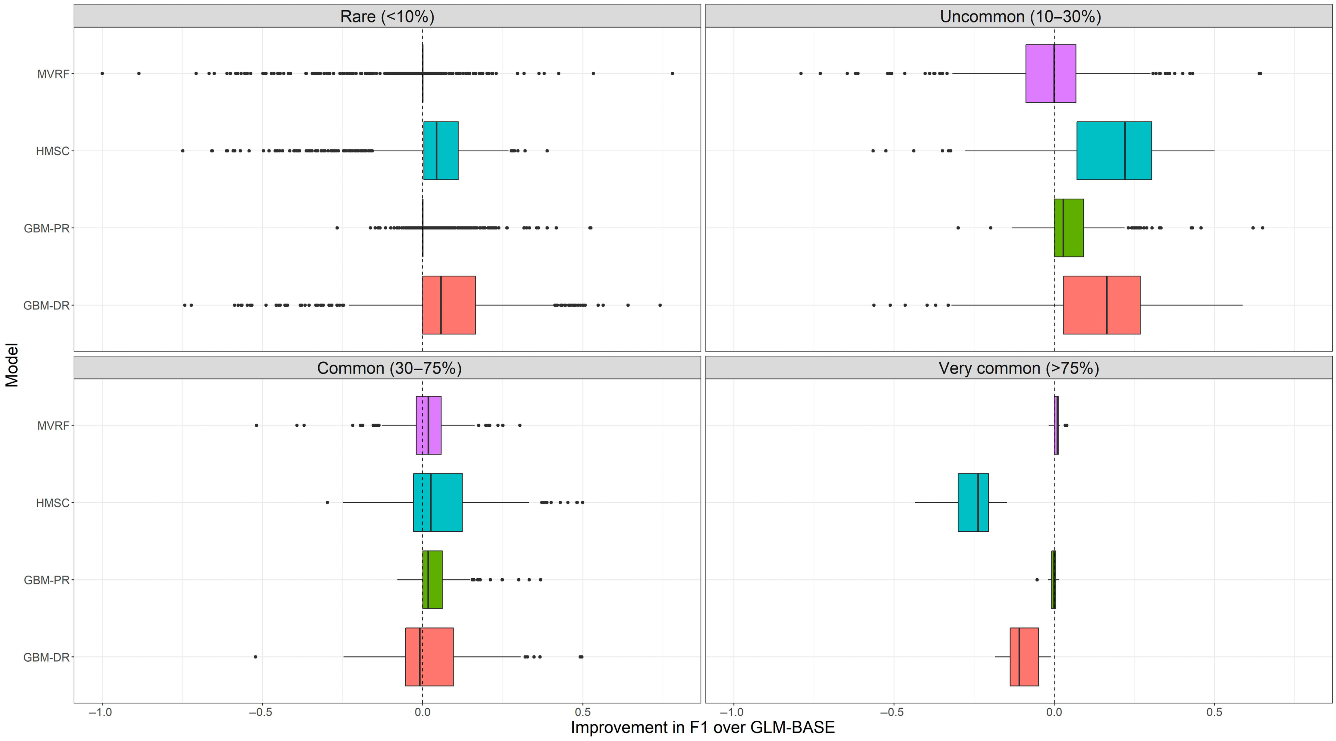

FIGURE 1 Relative performance of the compared models MVRF, HMSC, GBM-PR and GBM-DR (see Table 2 for details) measured using F1 metric, describing the weighted average of precision and recall, compared to the baseline GLM-BASE model by species prevalence for 1,622 species. Species prevalence is classified into four categories: Prevalence <10% classified as 'rare' (1110 species), between 10% and 30% as 'uncommon' (339 species), between 30% and 75% as 'common' (160 species), and >75% as 'very common' (13 species). Performance for HMSC and GBM-DR is highest for 'rare' species, highest for HMSC, GBM-DR and GBM-PR for 'uncommon' species, similar performance across all models for common species, and inferior performance of the HMSC and GBM-DR relative to GLM-BASE model for very common species.

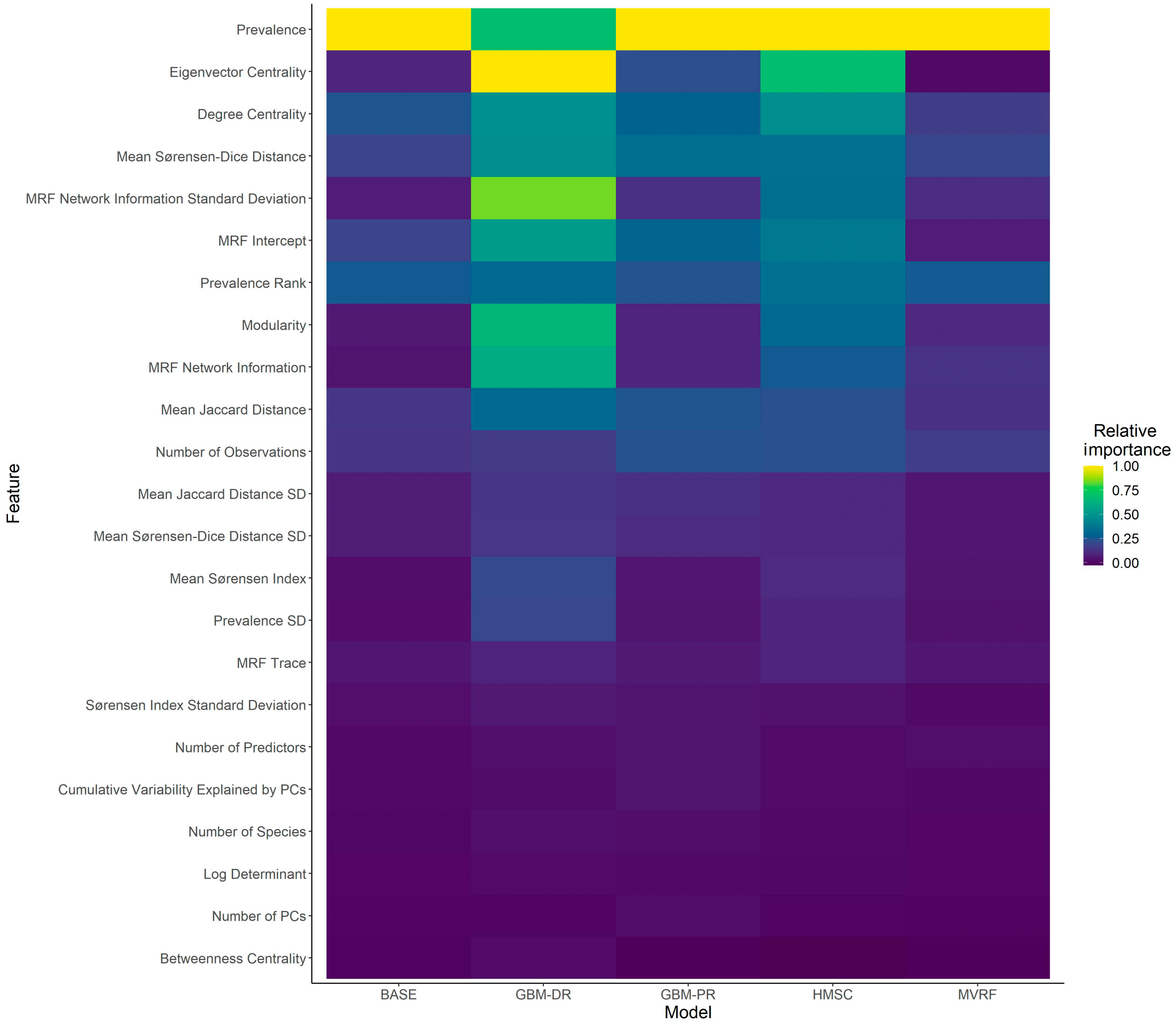

FIGURE 2 Heatmap showing relative importance of each of the 23 features (see Table 3 for details) by the compared models GLM-BASE, GBM-DR, GBM-PR, MVRF and HMSC (see Table 2 for details) in predicting model weights by the ensemble model. Features are ranked from highest to smallest relative importance across all five models.

desired model based on features of the data. While initial training of the proposed ensemble also requires fitting individual models, and as such will be equally as time-consuming, a continuously trained ensemble model could significantly reduce computational times for practitioners. Ultimately, this model could bypass the need for all constituent models to be fitted to new datasets, which may then be used as a tool to select a single model best suited to the dataset. Alternatively, over time this model could also be used to select a subset of models to be fitted as an ensemble and their respective weights, as a platform for providing more robust predictions than individual JSDMs or SSDMs, as demonstrated by our case study.

In practical settings, SDMs for hundreds or thousands of species are widely applied for management and conservation purposes (Palacio et al., 2021; Velásquez-Tibatá et al., 2019). In this case study, we illustrate a basic example of how a feature-based ensemble may be applied to a small subset of SDMs to improve species occurrence predictions. Our findings demonstrated a net improvement over GLM-BASE as measured by the F1 statistic of 51.13% for ENS model, 2.26% higher than the second-best performing model, the HMSC, and 4.63% higher than the GBM-DR net performance (Table 4). These findings support the idea that combining predictions of multiple models within an ensemble algorithm helps to reduce the biases from individual constituent models, offering predictions that are both robust and reliable (Araújo & New, 2007). The competitiveness of the ENS against the other models was also reflected across other performance metrics estimated from the binary predictions (precision and recall) as the second-best performing model, highlighting the ability for the ENS to detect true presence values. Similarly, the competitiveness of the ENS model was also highlighted by the performance metrics estimated from probability predictions (deviance

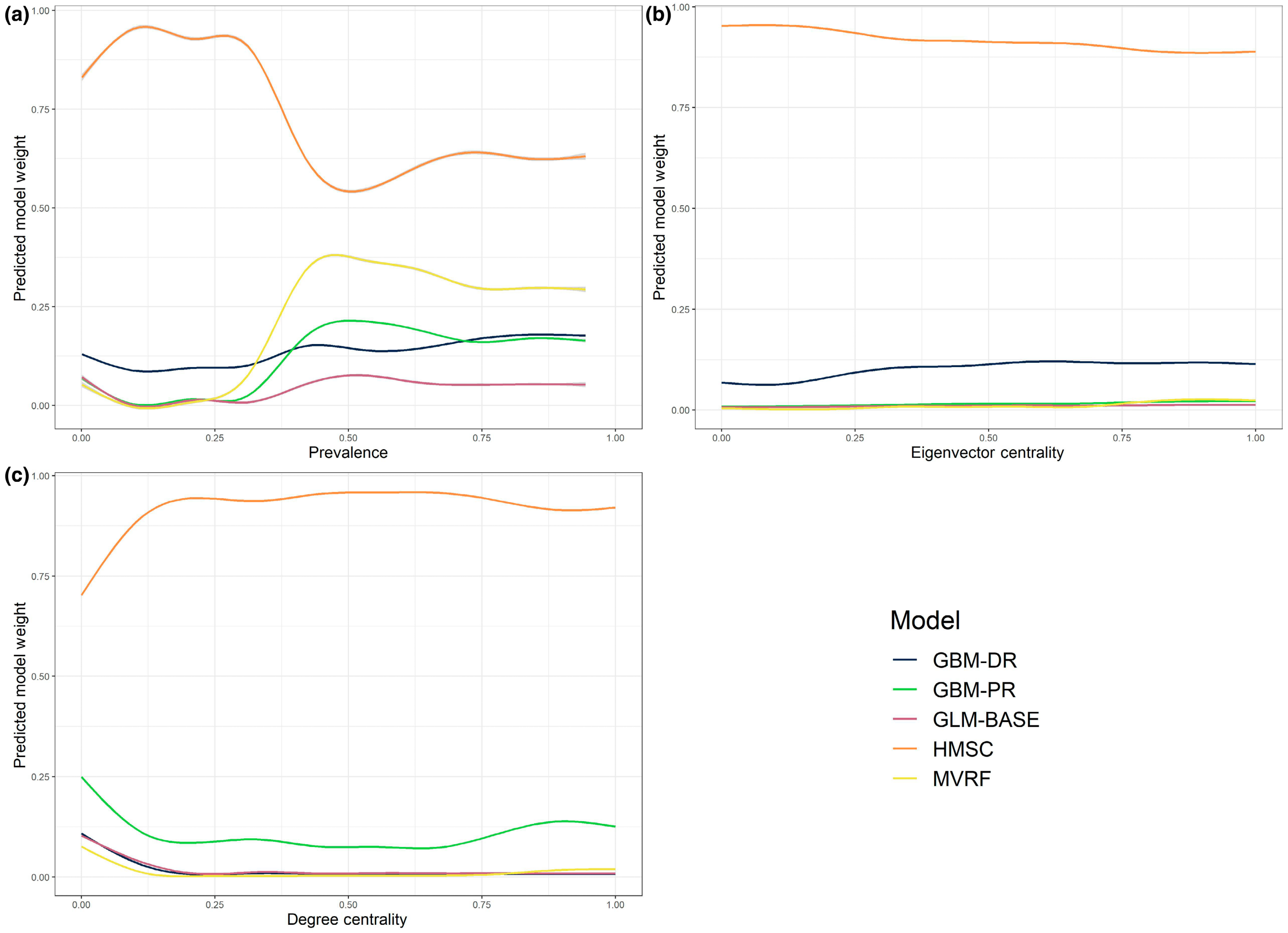

FIGURE 3 Model weight response functions for the three features with highest relative importance (prevalence, eigenvector centrality and degree centrality) for the GLM-BASE (pink), GBM-DR (blue), GBM-PR (green), HMSC (orange) and MVRF (yellow). Functions were estimated by holding all other feature predictors at their mean value and predicting from the ensemble random forest. (a) Trend shows that HMSC receives the highest model weighting by the ensemble model. MVRF is attributed the second highest weighting based on prevalence, and peaks at mid-range prevalence values. At a prevalence of 0.5, HMSC and MVRF are assigned mid-range weighting values (b) response function shows that HMSC is attributed highest weightings across all eigenvector centrality values peaking at lower values. GBM-DR is the second highest best performing model. (c) Response function shows that HMSC and GBM-PR are attributed the highest and second highest weightings across all degree centrality values, respectively, with complimentary performance for number of observations.

residuals), however, performed relatively poorly in terms of accuracy, suggesting that the ENS may be unable to predict absences as accurately as other models (given that the median prevalence value for the testing datasets is 5.17%; See Supplementary File 7 for tabulated results for the various performance metrics). These findings suggest that consideration of the most appropriate performance metric for the data is important when selecting a model for use.

To enable robust and optimised predictions, our methodological approach utilises simple descriptive features that describe species and their associated communities. As such, these features provide insights into why and when some models outperform others, improving the interpretability of model performance. In particular, our findings highlight the importance of features that relate to the co-occurrence network (Figure 2). This is particularly evident in the response functions for several network metrics, which show the variability in attributed weights as the association between species differs (see Supplementary File 6). For example, it can be seen that for the 'MRF Network Information' feature value increases, the attributed weighting to the GBM-DR model increases, while the weighting attributed to the HMSC model decreases within the ensemble (Supplementary Figure 6-17). This suggests that co-occurrence datasets with more or stronger associations between species, that is, the presence of species has a higher influence on the presence or absence of another species, tend to favour the GBM-DR method more, while favouring the HMSC method less. This provides important and useful evidence that multivariate structure in the observed data can be a key indicator of which models are likely to perform best. While previous studies have attempted to interpret why some models outperform others in particular situations, usually by using post-hoc descriptive statistics (e.g. Norberg et al., 2019), our study uniquely quantified these associations through features that describe characteristics of binary co-occurrence data. Thus, the

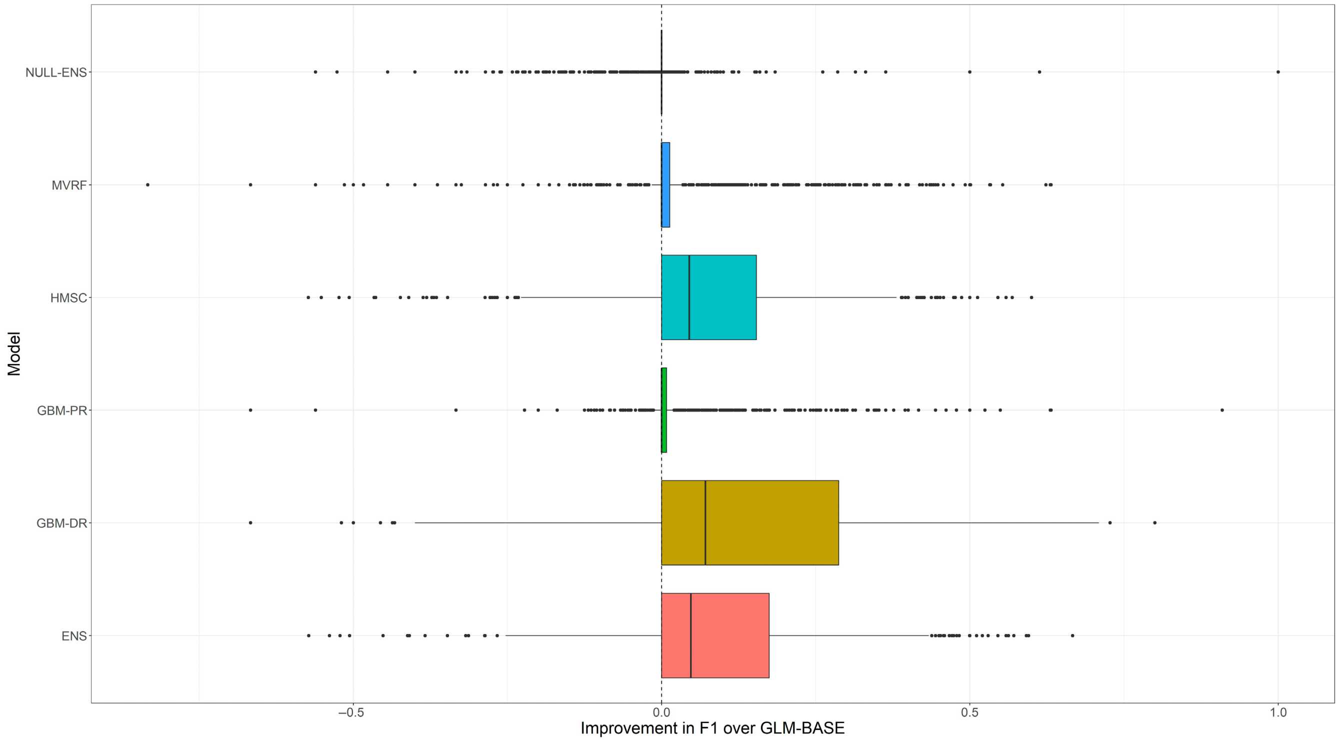

FIGURE 4 Relative performance of Null Ensemble (NULL-ENS), MVRF, HMSC, GBM-PR, GBM-DR and the Weighted Ensemble (ENS) relative to the base GLM model (GLM-BASE) as measured by the F1 statistic, describing the weighted average of precision and recall. GBM-DR, HMSC and ENS model perform significantly better than the GLM-BASE.

TABLE 4 Improvement of predictions over the GLM-BASE for each method based on the F1 statistic for 886 species in the test datasets

Method	Positive difference (adj. F1 > 0.02)	No difference (adj. F1 −0.02 −0.02)	Negative difference (adj. F1 < −0.02)	Net improvement (positive − negative)
ENS	505	329	52	453
NULL-ENS	52	695	139	−87
GBM-DR	491	316	79	412
GBM-PR	207	644	35	172
MVRF	209	608	69	140
HMSC	540	239	107	433

results together with the valuable insights into how models perform relative to features offer promise for the feature-ensemble method's broader applications.

While our example highlights the utility of ensemble modelling without necessarily having to fit a bespoke model, the flexibility of this approach means that users could incorporate more bespoke, knowledge-driven models. Bayesian models with context-specific prior information can readily be included (Clark et al., 2017; Ovaskainen & Soininen, 2011), as well as models that rely solely on expert opinion to estimate species occurrence (Velásquez-Tibatá et al., 2019). Beyond ecology, ensembles that combine a diversity of expert-driven predictions have demonstrated their superiority compared to individual models in many settings, such as forecasting weekly deaths from COVID-19 in the USA (https://viz.covid19forecasthub.org/). Evaluating the performance of the feature-based ensemble method using more specialised individual models offers exciting avenues for future investigations.

Our findings also highlight some of the strengths and limitations of the individual constituent models. Of particular note is the GBM-DR method, whose competitive performance offers some valuable insights into the importance of learning from other species to predict the occurrence of a focal species, adding to the growing body of evidence regarding the importance of accounting for biotic associations in species distribution modelling (Araújo & Luoto, 2007; Heikkinen et al., 2007; Leathwick et al., 2006; Ovaskainen et al., 2017). While our GBM-DR model only used GLMs as the base models for all species and a weak GBM learner as the stacker, in principle, a wide variety of models could be applied to each individual species prior to stacking. The flexibility of the approach means that users can potentially incorporate any model of any form, so long as they can generate fitted values and residuals, and there is opportunity to use other learners to optimise the stacking predictions (Xing et al., 2020).

Another advantage of the SSDM approach is the ability to estimate nonlinear species associations, rather than relying on additive-only associations described by loadings on latent factors, such as the HMSC approach, or estimated from the full covariance matrix (Clark et al., 2018; Ovaskainen et al., 2016), which can be slow and inefficient for large and complex datasets (Norberg et al., 2019; Pichler & Hartig, 2021). Inclusion of covariates within the stacking learner could also be done, which could in-principle capture how species associations change across environmental gradients. This ability to use recent advances from machine learning for the stacking model coincides with the rising need for interpretable machine learning processes to interrogate and understand these models. For example the recently developed Multi-response Interpretable Machine Learning (MrIML) framework offers a flexible approach that compares the performance of multivariate models and delivers interpretable outputs, which could be used to better understand the associations estimated in the stacking model (Fountain-Jones et al., 2021).

Beyond our case study, the feature-based ensemble framework could be manipulated to suit different end user requirements. For example, while we used deviance residuals to obtain the initial model weights to train the ensemble model, different loss functions including Pearson residuals or even classification metrics such as F1 scores could be used instead. Incorporating uncertainty could also be used by optimising on a penalised prediction interval rather than on a point metric such as the deviance residual, although this approach is more challenging when considering methods such as GBM-DR and GBM-PR as there is no convenient way to quantify prediction uncertainty. Alternatively, identifying more precise ways than using posterior means to calculate point predictions from Bayesian posterior distributions (as we did here) could allow for optimisation of the Bayesian methods where models do not allow for quantification of prediction uncertainty. For simplicity in our model, we optimised the binarisation threshold for species predictions to 0.5, but this arbitrary value could also be optimised to improve each model's predictive ability.

5 | CONCLUSIONS

Improving the predictability and interpretability of species distribution model for practical applications requires more than comparisons between model performance across ecological contexts: it requires a deeper understanding of how co-occurrence data drives model performance and better ways for accounting for variations in species associations. In our study, we have demonstrated the utility of a flexible feature-based ensemble approach with the capacity to retrieve accurate and robust predictions rapidly over a range of ecological contexts, without necessarily needing to fit highly specialised models. Within our case study used to highlight the potential applications of our ensemble, we have also introduced a new SSDM approach with great potential for future applications in ecological modelling.

AUTHORS' CONTRIBUTIONS

F.P.-R. conducted the data analysis, prepared figures and lead the writing of the manuscript; N.J.C. led the project, conceived the ideas for the project and contributed critically to the syntax and design of the methodology; N.M.F.-J. provided valuable input for the methodological design and provided recommendations for improving results; A.N. provided valuable recommendations for refining models and results based on her expertise. All authors contributed critically to the drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

This project was funded by ARC Discovery Early Career Researcher Award (DE210101439) and by an Australian Research Council Discovery Project Grant (DP190102020). Open access publishing facilitated by The University of Queensland, as part of the Wiley - The University of Queensland agreement via the Council of Australian University Librarians.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13915.

DATA AVAILABILITY STATEMENT

All data were obtained from open-source databases. Original and cleaned versions of the datasets, code and guided workflow for the analysis of this study can be found on the Zenodo Repository https://doi.org/10.5281/zenodo.6565339 (Powell-Romero et al., 2022).

ORCID

Francisca Powell-Romero https://orcid.org/0000-0001-9800-3100
Nicholas M. Fountain-Jones https://orcid.org/0000-0001-9248-8493
Anna Norberg https://orcid.org/0000-0002-3520-1043
Nicholas J. Clark https://orcid.org/0000-0001-7131-3301

REFERENCES

Adams, B. J., Schnitzer, S. A., & Yanoviak, S. P. (2017). Trees as islands: Canopy ant species richness increases with the size of liana-free trees in a neotropical forest. *Ecography (Copenhagen)*, *40*, 1067–1075.

Algar, A. C., Kharouba, H. M., Young, E. R., & Kerr, J. T. (2009). Predicting the future of species diversity: Macroecological theory, climate change, and direct tests of alternative forecasting methods. *Ecography*, *32*, 22–33.

Araújo, M. B., & Luoto, M. (2007). The importance of biotic interactions for modelling species distributions under climate change. *Global Ecology and Biogeography*, *16*, 743–753.

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

Atiya, A. F. (2020). Why does forecast combination work so well? *International Journal of Forecasting*, *36*, 197–200.

Azhagesan, K., Ravindran, B., & Raman, K. (2018). Network-based features enable prediction of essential genes across diverse organisms. *PLoS ONE*, *13*, e0208722.

Baselga, A., & Araújo, M. B. (2010). Do community-level models describe community variation effectively? *Journal of Biogeography*, *37*, 1842–1850.

Bolnick, D., Ballare, K., Stuart, Y., Stutz, W., & Resetarits, E. (2020). Host patch traits have scale-dependent effects on diversity in a stickleback parasite metacommunity. *Ecography*, *43*, 1–13.

Brian, J. I., & Aldridge, D. C. (2021). Abundance data applied to a novel model invertebrate host shed new light on parasite community assembly in nature. *Journal of Animal Ecology*, *90*, 1096–1108.

Burner, R. C., Stephan, J. G., Drag, L., Birkemoe, T., Muller, J., Snäll, T., Ovaskainen, O., Potterf, M., Siitonen, J., Skarpaas, O., Doerfler, I., Gossner, M. M., Schall, P., Weisser, W. W., & Sverdrup-Thygeson, A. (2021). Traits mediate niches and co-occurrences of forest beetles in ways that differ among bioclimatic regions. *Journal of Biogeography*, *48*, 3145–3157.

Burner, R. C., Styring, A. R., Rahman, M. A., & Sheldon, F. H. (2019). Occupancy patterns and upper range limits of lowland Bornean birds along an elevational gradient. *Journal of Biogeography*, *46*, 2583–2596.

Byrd, R. H., Peihuang, L. U., Nocedal, J., & Ciyou, Z. H. U. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*, 1190–1208.

Calabrese, J. M., Certain, G., Kraan, C., & Dormann, C. F. (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.

Clark, J. S., Nemergut, D., Seyednasrollah, B., Turner, P. J., & Zhang, S. (2017). Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data. *Ecological Monographs*, *87*, 34–56.

Clark, N. J., Wells, K., Dimitrov, D., & Clegg, S. M. (2016). Co-infections and environmental conditions drive the distributions of blood parasites in wild birds. *Journal of Animal Ecology*, *85*, 1461–1470.

Clark, N. J., Wells, K., & Lindberg, O. (2018). Unravelling changing interspecific interactions across environmental gradients using Markov random fields. *Ecology*, *99*, 1277–1283.

D'Amen, M., Dubuis, A., Fernandes, R. F., Pottier, J., Pellissier, L., & Guisan, A. (2015). Using species richness and functional traits predictions to constrain assemblage predictions from stacked species distribution models. *Journal of Biogeography*, *42*, 1255–1266.

D'Amen, M., Mod, H. K., Gotelli, N. J., & Guisan, A. (2018). Disentangling biotic interactions, environmental filters, and dispersal limitation as drivers of species co-occurrence. *Ecography (Copenhagen)*, *41*, 1233–1244.

D'Amen, M., Pradervand, J.-N., & Guisan, A. (2015). Predicting richness and composition in mountain insect communities at high resolution: A new test of the SESAM framework. *Global Ecology and Biogeography*, *24*, 1443–1453.

Delgado, D. L., & Restrepo, C. (2019). Multi-driver and multi-scale assessment of vine community structure and composition across a complex tropical environmental matrix. *PLoS ONE*, *14*, e0215274.

Distler, T., Schuetz, J. G., Velásquez-Tibatá, J., & Langham, G. M. (2015). Stacked species distribution models and macroecological models provide congruent projections of avian species richness under climate change. *Journal of Biogeography*, *42*, 976–988.

Dubuis, A., Pottier, J., Rion, V., Pellissier, L., Theurillat, J.-P., & Guisan, A. (2011). Predicting spatial patterns of plant species richness: A comparison of direct macroecological and species stacking modelling approaches. *Diversity & distributions*, *17*, 1122–1131.

Elith, J., Graham, H. C., Anderson, P. R., Dudík, M., Ferrier, S., Guisan, A., Hijmans, J. R., Huettmann, F., Leathwick, R. J., Lehmann, A., Li, J. G., Lohmann, L. A., Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC, M., Overton, J., … Zimmermann, E. N. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*, 129–151.

Escoriza, D. (2020). Organization of Squamata (Reptilia) assemblages in Mediterranean archipelagos. *Ecology and Evolution*, *10*, 1592–1601.

Fountain-Jones, N. M., Kozakiewicz, C. P., Forester, B. R., Landguth, E. L., Carver, S., Charleston, M., Gagne, R. B., Greenwell, B., Kraberger, S., Trumbo, D. R., Mayer, M., Clark, N. J., & Machado, G. (2021). MrIML: Multi-response interpretable machine learning to model genomic landscapes. *Molecular Ecology Resources*, *21*, 2766–2781.

Fountain-Jones, N. M., Packer, C., Jacquot, M., Blanchet, F. G., Terio, K., Craft, M. E., & Ezenwa, V. (2019). Endemic infection can shape exposure to novel pathogens: Pathogen co-occurrence networks in the Serengeti lions. *Ecology Letters*, *22*, 904–913.

Franklin, J. (1998). Predicting the distribution of shrub species in southern California from climate and terrain-derived variables. *Journal of Vegetation Science*, *9*, 733–748.

Glidden, C. K., Coon, C. A. C., Beechler, B. R., McNulty, C., Ezenwa, V. O., & Jolles, A. E. (2021). Co-infection best predicts respiratory viral infection in a wild host. *Journal of Animal Ecology*, *90*, 602–614.

Gneiting, T., & Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science*, *310*, 248–249.

Greenwell, B., Boehmke, B., Cunningham, J., & GBM Developers (2020). *gbm: Generalized boosted regression models*. R package version 2.1.8. Retrieved from https://CRAN.R-project.org/package=gbm

Guisan, A., Graham, C. H., Elith, J., & Huettmann, F. (2007). Sensitivity of predictive species distribution models to change in grain size. *Diversity & distributions*, *13*, 332–340.

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

Guisan, A., Zimmermann, N. E., Elith, J., Graham, C. H., Phillips, S., & Peterson, A. T. (2007). What matters for predicting the occurrences of trees: Techniques, data, or Species' characteristics? *Ecological Monographs*, *77*, 615–630.

Han, Z., Zhang, L., Jiang, Y., Wang, H., & Jiguet, F. (2020). Unravelling species co-occurrence in a steppe bird community of Inner Mongolia: Insights for the conservation of the endangered Jankowski's Bunting. *Diversity & distributions*, *26*, 843–852.

Harris, D. J., Taylor, S. D., & White, E. P. (2018). Forecasting biodiversity in breeding birds using best practices. *PeerJ*, *2018*, e4278.

Heikkinen, R. K., Luoto, M., Virkkala, R., Pearson, R. G., & Körber, J.-H. (2007). Biotic interactions improve prediction of boreal bird distributions at macro-scales. *Global Ecology and Biogeography*, *16*, 754–763.

Hui, F. K. C., & Poisot, T. (2016). Boral – Bayesian ordination and regression analysis of multivariate abundance data in R. *Methods in Ecology and Evolution*, *7*, 744–750.

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

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *Annals of Applied Statistics*, *2*, 841–860.

Kang, Y., Hyndman, R. J., & Smith-Miles, K. (2017). Visualising forecasting algorithm performance using time series instance spaces. *International Journal of Forecasting*, *33*, 345–358.

Kück, M., Crone, S. F., & Freitag, M. (2016). Meta-learning with neural networks and landmarking for forecasting model selection an empirical evaluation of different feature sets applied to industry data. In *2016 international joint conference on neural networks (IJCNN)* (pp. 1499–1506). Institute of Electrical and Electronics Engineers.

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

Lemke, C., & Gabrys, B. (2010). Meta-learning for time series forecasting and forecast combination. *Neurocomputing*, *73*, 2006–2016.

Maguire, K. C., Nieto-Lugilde, D., Blois, J. L., Fitzpatrick, M. C., Williams, J. W., Ferrier, S., & Lorenz, D. J. (2016). Controlled comparison of species- and community-level models across novel climates and communities. *Proceedings of the Royal Society B: Biological Sciences*, *283*, 20152817.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, *36*, 54–74.

Mathieu, J., & Jonathan Davies, T. (2014). Glaciation as an historical filter of below-ground biodiversity. *Journal of Biogeography*, *41*, 1204–1214.

Moisen, G. G., & Frescino, T. S. (2002). Comparing five modelling techniques for predicting forest characteristics. *Ecological Modelling*, *157*, 209–225.

Montaño-Centellas, F. A. (2020). Interaction networks of avian mixed-species flocks along elevation in the tropical Andes. *Ecography (Copenhagen)*, *43*, 930–942.

Montero-Manso, P., Athanasopoulos, G., Hyndman, R. J., & Talagala, T. S. (2020). FFORMA: Feature-based forecast model averaging. *International Journal of Forecasting*, *36*, 86–92.

Murray, S. A. (2018). The importance of ensemble techniques for operational space weather forecasting. *Space Weather*, *16*, 777–783.

Neves, T., Borda-De-água, L., Mathias, M. D. L., & Tapisso, J. T. (2022). The influence of the interaction between climate and competition on the distributional limits of European shrews. *Animals (Basel)*, *12*, 57.

Norberg, A., Abrego, N., Blanchet, F. G., Adler, F. R., Anderson, B. J., Anttila, J., Araújo, M. B., Dallas, T., Dunson, D., Elith, J., Foster, S. D., Fox, R., Franklin, J., Godsoe, W., Guisan, A., O'Hara, B., Hill, N. A., Holt, R. D., Hui, F. K. C., … Ovaskainen, O. (2019). A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*, *89*, e01370.

Ovaskainen, O., Abrego, N., Halme, P., Dunson, D., & Warton, D. (2016). Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods in Ecology and Evolution*, *7*, 549–555.

Ovaskainen, O., & Soininen, J. (2011). Making more out of sparse data: Hierarchical modeling of species communities. *Ecology*, *92*, 289–295.

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T., Abrego, N., & Chave, J. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters*, *20*, 561–576.

Palacio, R. D., Negret, P. J., Velásquez-Tibatá, J., & Jacobson, A. P. (2021). A data-driven geospatial workflow to map species distributions for conservation assessments. *Diversity and Distributions*, *27*, 2559–2570.

Pichler, M., & Hartig, F. (2021). A new joint species distribution model for faster and more accurate inference of species associations from big community data. *Methods in Ecology and Evolution*, *12*, 2159–2173.

Pollock, L. J., Bayly, M. J., & Vesk, P. A. (2015). The roles of ecological and evolutionary processes in Plant community assembly: The environment, hybridization, and introgression influence co-occurrence of eucalyptus. *The American Naturalist*, *185*, 784–796.

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., Vesk, P. A., McCarthy, M. A., & McPherson, J. (2014). Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, *5*, 397–406.

Popovic, G. C., Warton, D. I., Thomson, F. J., Hui, F. K. C., Moles, A. T., & Murrell, D. (2019). Untangling direct species associations from indirect mediator species effects with graphical models. *Methods in Ecology and Evolution*, *10*, 1571–1583.

Powell-Romero, F., Fountain-Jones, N., Norberg, A., & Clark, N. (2022). R code to replicate analyses in Powell-Romero et al. 2022. Improving the predictability and interpretability of co-occurrence modelling through feature-based joint species distribution ensembles. *Methods in Ecology and Evolution (v1.0.0). Zenodo*, https://doi.org/10.5281/zenodo.6565339

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ruberanziza, E., Owada, K., Clark, N. J., Umulisa, I., Ortu, G., Lancaster, W., Munyaneza, T., Mbituyumuremyi, A., Bayisenge, U., Fenwick, A., & Soares Magalhães, R. J. (2019). Mapping soil-transmitted helminth parasite infection in Rwanda: Estimating endemicity and identifying at-risk populations. *Tropical Medicine and Infectious Disease*, *4*, 93.

Stark, K. A., Thompson, P. L., Yakimishyn, J., Lee, L., Adamczyk, E. M., Hessing-Lewis, M., & O'Connor, M. I. (2020). Beyond a single patch: Local and regional processes explain diversity patterns in a seagrass epifaunal metacommunity. *Marine Ecology. Progress Series (Halstenbek)*, *655*, 91–106.

Steen, V., Tingley, M., Paton, P., & Elphick, C. (2020). Data from: Avian point-counts from Rhode Island and Connecticut used to test species distribution models. *Dryad Digital Repository*, https://doi.org/10.5061/dryad.8cz8w9gnp

Talagala, T., Hyndman, R., & Athanasopoulos, G. (2018). *Meta-learning how to forecast time series*. Monash University, Department of Econometrics and Business Statistics.

Thuiller, W., Araújo, M. B., & Lavorel, S. (2003). Generalized models vs. classification tree analysis: Predicting spatial distributions of plant species at different scales. *Journal of Vegetation Science*, *14*, 669–680.

Tikhonov, G., Ovaskainen, O., Oksanen, J., de Jonge, M., Opedal, O., & Dallas, T. (2021). *Hmsc: Hierarchical model of species communities*. R package version 3.0-11. Retrieved from https://CRAN.R-project.org/package=Hmsc

Vayssières, M. P., Plant, R. E., & Allen-Diaz, B. H. (2000). Classification trees: An alternative non-parametric approach for predicting species distributions. *Journal of Vegetation Science*, *11*, 679–694.

Velásquez-Tibatá, J., Olaya-Rodríguez, M. H., López-Lozano, D., Gutiérrez, C., González, I., & Londoño-Murcia, M. C. (2019). BioModelos: A collaborative online system to map species distributions. *PLoS ONE*, *14*, e0214522.

Vieira, T. B., Brasil, L. S., Silva, L. C. N., Tejerina-Garro, F. L., Aquino, P.d. P. U., Pompeu, P. S., & Marco, P. (2020). Elements of fish metacommunity structure in neotropical freshwater streams. *Ecology and Evolution*, *10*, 12024–12035.

Wang, Z., & Srinivasan, R. S. (2017). A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models. *Renewable and Sustainable Energy Reviews*, *75*, 796–808.

Warton, D. I., Blanchet, F. G., O'Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C., & Hui, F. K. C. (2015). So many variables: Joint modeling in community ecology. *Trends in Ecology & Evolution*, *30*, 766–779.

Wilkinson, D. P., Golding, N., Guillera-Arroita, G., Tingley, R., McCarthy, M. A., & Peres-Neto, P. (2019). A comparison of joint species distribution models for presence–absence data. *Methods in Ecology and Evolution*, *10*, 198–211.

Wisz, M. S., Hijmans, R. J., Li, J., Peterson, A. T., Graham, C. H., & Guisan, A. (2008). Effects of sample size on the performance of species distribution models. *Diversity & Distributions*, *14*, 763–773.

Xing, L., Lesperance, M., & Zhang, X. (2020). Simultaneous prediction of multiple outcomes using revised stacking algorithms. *Bioinformatics*, *36*, 65–72.

Zhang, C., Chen, Y., Xu, B., Xue, Y., & Ren, Y. (2018). Comparing the prediction of joint species distribution models with respect to characteristics of sampling data. *Ecography (Copenhagen)*, *41*, 1876–1887.

Zurell, D., Zimmermann, N. E., Gross, H., Baltensweiler, A., Sattler, T., & Wüest, R. O. (2020). Testing species assemblage predictions from stacked and joint species distribution models. *Journal of Biogeography*, *47*, 101–113.

SUPPORTING INFORMATION

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

How to cite this article: Powell-Romero, F., Fountain-Jones, N. M., Norberg, A., & Clark, N. J. (2023). Improving the predictability and interpretability of co-occurrence modelling through feature-based joint species distribution ensembles. *Methods in Ecology and Evolution*, *14*, 146–161. https://doi.org/10.1111/2041-210X.13915