A comparison of predictive performance of joint species distribution models for presence-absence data

David P. Wilkinson[1*], Nick Golding[2,3,4], Gurutzeta Guillera-Arroita[1,5], Reid Tingley[6,7], Michael A. McCarthy[1]

1. School of Agriculture, Food and Ecosystem Sciences, University of Melbourne, Parkville, 3010, Victoria, Australia

2. Telethon Kids Institute, Perth Children's Hospital, 15 Hospital Ave, Nedlands, 6009, Western Australia, Australia

3. Curtin University, Kent St, Bentley, 6102, Western Australia, Australia

4. Melbourne School of Population and Global Health, University of Melbourne, Parkville, 3010, Victoria, Australia

5. Pyrenean Institute of Ecology, Spanish National Research Council, Jaca, 22700, Huesca, Spain

6. EnviroDNA Pty Ltd, 95 Albert Street, Brunswick, Victoria 3056 Australia

7. School of Biological Sciences, Monash University, Clayton, 3800, Victoria, Australia

*Corresponding author: dwilkinson@unimelb.edu.au

ORCiDs:

D.P.W. 0000-0002-9560-6499

N.G. 0000-0001-8916-5570

G.G.A. 0000-0002-8387-5739

R.T. 0000-0002-7630-7434

M.A.M. 0000-0003-1039-7980

Acknowledgements: We thank Peter Vesk, Brendan Wintle and Jian Yen for insightful discussions. N.G. was supported by an Australian Research Council (ARC) Discovery Early Career Researcher Award (DE180100635). G.G.A. is currently supported by a ´Ramón y Cajal´ grant (RYC2020-028826-I) funded by the Spanish Ministry of Science and Innovation, the Agencia Estatal de Investigación (10.13039/501100011033) and "ESF Investing in your future".

Data availability statement: The code and data for this analysis can be found on GitHub and is archived by Zenodo (Doi90/JSDM_Prediction; Wilkinson, 2019).

Conflicts of interest: No conflict of interest is declared.

Author contribution statement: All the authors conceived the ideas and methodology; D.P.W. implemented the analysis; D.P.W. led writing the manuscript but all the authors contributed significantly throughout and gave final approval before submission.

ABSTRACT

1. While there has been substantial literature on the evaluation of predictions from single species distribution models, the topic of prediction has only recently begun to be addressed for joint species distribution models (JSDMs). These studies have covered only limited aspects of prediction: limited selection of models being compared, limited number of evaluation metrics, and/or not comparing the different prediction types available to JSDMs.
2. In this study, we perform a large-scale comparison of the predictive performance of eight model types: two stacked species distribution models (SSDMs) and six JSDMs. We fit these models to 22 real and simulated datasets, make four types of JSDM predictions, and evaluate up to 32 metrics from five different classes that quantify different aspects of performance of predictions about species distributions and the community assemblage process.
3. We found that likelihood-based metrics indicated the JSDMs were better fit to the data than the standard SSDM, but most other metric classes showed the SSDM outperforming the JSDMs by generally small amounts. The spatial and non-spatial implementations of the hierarchical multivariate probit regression model with latent factors typically performed better than the other JSDMs, but overall still performed worse than the SSDM. The SSDM predictions constrained with the spatially-explicit species assemblage modelling framework (SESAM) consistently outperformed both the standard SSDM and all JSDMs for both species- and community-level metrics.
4. Our results indicate that despite the additional inference they provide about the community assemblage process by accounting for the residual association between species, JSDMs generally yield worse predictions than stacked single species models when evaluated at either the species or community level. The performance of the

SESAM framework suggests that exploring similar approaches to constrain JSDM predictions is an interesting future avenue of research.

Keywords: biotic interactions, community assemblage, evaluation metrics, joint species distribution models, prediction, species richness

1. INTRODUCTION

Species distributions are influenced by the response of a species to both the abiotic and biotic conditions it encounters, but the development of species distribution models (SDMs) has historically not accounted for the (biotic) effect of species interactions, largely focusing on single-species approaches. Over the past decade, joint species distribution models (JSDMs) have seen rapid development for modelling multiple species simultaneously while accounting for both environmental responses and residual species associations (Kissling et al., 2012; Warton et al., 2015; Wilkinson et al., 2019; Wisz et al., 2013). Research on JSDMs initially focused on the development of different statistical modelling approaches (Clark et al., 2017; Golding & Purse, 2016; Harris, 2015; Hui, 2016; Ovaskainen, Roy, et al., 2016; Pollock et al., 2014), then extended the framework to account for additional factors such as the effect of spatial scale or environmental gradients on species associations (Ovaskainen, Abrego, et al., 2016; Thorson et al., 2016; Tikhonov et al., 2017). Only recently has the question of prediction using JSDMs begun to be explored in any detail (Norberg et al., 2019; Wilkinson et al., 2021; Zhang et al., 2018; Zurell et al., 2019).

A model's ability to accurately predict species distributions or communities needs to be evaluated to be confident that the model performs well in practice. While there is substantial literature on the evaluation of single species SDMs (Fielding & Bell, 1997; Lawson et al., 2014; Liu et al., 2009), there is little research into the evaluation of the multi-species predictions of JSDMs. Whilst single species model evaluation metrics can also be applied to the predictions of multi-species models, the multivariate nature of JSDMs also invites the evaluation of predictions using the community dissimilarity indices widely used in community ecology (Legendre & De Cáceres, 2013).

The community assemblage process can be viewed as the result of two processes: (1) an abiotic filter where species-environment relationships influence which species can occur in a given environment, and (2) a biotic filter where between-species relationships influence which species are more or less likely to co-occur (Cornell & Harrison, 2014; Götzenberger et al., 2012). Some accounting of species interactions inside the single-species framework is possible, by using additional species as predictor variables (Araújo & Luoto, 2007; Meier et al., 2010; Zhang et al., 2020) or constraining predicted distributions to that of another species it depends on (Schweiger et al., 2012), but this only reflects unidirectional interactions where the direction of the relationship is already known (Kissling et al., 2012; Pollock et al., 2014; Wisz et al., 2013).

Stacked species distribution models (SSDMs) represent community assemblages by stacking the predictions of multiple *independent* single-species models (Guisan & Rahbek, 2011; Thuiller et al., 2015). In contrast, joint species distribution models can capture both the biotic and the abiotic factors impacting species and thus link distribution modelling and community ecology. Wilkinson *et al.* (2021) detailed the different ways that JSDMs can make predictions: environment-only marginal predictions at the species-level, joint predictions at the community-level predictions that leverage the additional information on residual species occurrence, and conditioning both marginal or joint on the known occurrence state of some species at a site.

JSDMs have been presented as a modelling approach with the potential to make better predictions, particularly at the community level, than methods that do not account for the residual correlations between species. However, the use of JSDMs for prediction has only begun to be addressed in the literature (Gelfand & Shirota, 2021; Norberg et al., 2019; Ovaskainen, Roy, et al., 2016; Poggiato et al., 2021; Wilkinson et al., 2021; Zhang et al., 2018; Zurell et al., 2019). These studies have been limited in what aspects of JSDM predictive performance they have considered. Only Norberg *et al*. (2019) considered prediction types that

account for the residual correlations between species in practice, while conditional prediction types have only been theoretically discussed (Gelfand & Shirota, 2021; Poggiato et al., 2021). Only Norberg *et al*. (2019) has compared multiple JSDM implementations, and they all have compared predictive performance for only a select few evaluation metrics. In this study we compare the predictive performance of six JSDMs and two SSDMs when fit to two real and 20 simulated datasets. We use 32 metrics over five metric classes to evaluate SSDM predictions against the four different prediction types available to JSDMs.

2. MATERIALS AND METHODS

2.1. MODELS

The two SSDM models are generalised linear models with a probit link, fit individually to each species but differ in their approach to stacking predictions. The first is a standard stacked approach (SSDM) where individual species predictions are summed together to obtain species richness predictions. The second approach is spatially explicit species assemblage modelling (SESAM; Guisan & Rahbek, 2011) that selects species as present at a site up to a calculated maximum limit (e.g. from a macroecological model).

All six JSDMs are based on the multivariate probit regression model of Chib and Greenberg (1998). The first is the standard multivariate probit regression model (MPR) implemented in the R package *BayesComm* (Golding et al., 2015). Second, the hierarchical multivariate probit regression model (HPR) of Pollock *et al* (2014). Third, the multivariate probit regression with latent factors (LPR) implemented in the R package *boral* (Hui, 2016). Fourth, the multivariate generalised regression model (DPR) implemented in the R package *gjam* (Clark et al., 2017). The fifth and sixth models are the spatial (HLR-S) and non-spatial (HLR-NS) implementations

of the hierarchical multivariate probit regression model with latent factors implemented in the R package *HMSC* (Ovaskainen, Roy, et al., 2016). All of these models are implemented using a Bayesian framework and fit using Markov chain Monte Carlo (MCMC) sampling, using their default or suggested settings as defined in their source articles or the documentation of the software implementing them. Model equations, default priors, and MCMC regimes are defined in greater detail in Wilkinson *et al* (2019) and in Appendix S1.

2.2. Prediction Methods

Single-species models generate environment-only predictions that ignore species associations. For SSDMs, a prediction is obtained independently for each species using the estimated regression coefficients and the corresponding measured variables. This provides a predicted probability of presence for each species at each site. Binary predictions can be generated by taking draws from a Bernoulli distribution, using the predicted probabilities. The SESAM only provides binary predictions and constrains them to a site-specific species richness upper limit using an estimated species richness, either from a macroecological model or alternatively, as we have done here, using the probability rank rule (selecting species in decreasing order of probability, considering as present as many as the total sum of probabilities rounded down).

JSDMs, in contrast, provide additional information on residual correlations between species occurrence in their model outputs which can inform predictions. Marginal JSDM predictions are species-level, environment-only predictions where the information about species co-occurrence is ignored. Predictions are again probabilistic (a probability for each species at each site), and binary predictions can be generated by drawing Bernoulli samples, as for the standard SSDM. To quantify prediction uncertainty in binary predictions, Bernoulli samples can be drawn for each species, site, and sample of predicted probability from the model posterior in a Bayesian model fitting framework.

Joint predictions are obtained from the estimated joint probability distribution (a multivariate normal distribution) over plausible community assemblages. This probability distribution defines the probability of occurrence of an observed or hypothesised assemblage, or can be used to simulate binary community assemblage predictions. Averaging the presence-absence of individual species across simulated community assemblages approximates the marginal distribution.

Conditional predictions include the additional information of *known* occurrence states for one or more species in the community. The known occurrence state of a species truncates the multivariate normal distribution on one axis. As the total probability must sum to one, this affects the probability of the remaining possible community assemblages.

Conditional marginal predictions are similar to the conditional predictions in that they truncate the multivariate normal distribution based on the occurrence state of one or more species. But rather than representing the joint distribution of the remaining species, the distribution over the remaining species is marginalised, to yield a single probability of presence for each species. Thus, these are predictions conditional on one or more species but marginal to the remainder.

An in-depth explanation of these prediction methods can be found in Wilkinson *et al.* (Wilkinson et al., 2021) or Appendix S2.

2.3. DATASETS

For this comparison we have used two real datasets, on frogs and eucalypts, and twenty simulated presence-absence datasets. The frog dataset comprises 9 species, 104 sites, and 3 covariates from Melbourne, Australia (Parris, 2006). The eucalypt dataset comprises 12 species, 458 sites, and 7 covariates from Grampians National Park, Australia (Pollock et al., 2014). The simulated datasets all have 10 species, 100 sites, and 5 covariates (3 continuous, 2

binary), with the species correlation matrix generated using three latent factors. Species presence absence data was generated using the *HMSC::communitySimul* function (Blanchet et al., 2019).

For all datasets, the continuous variables were standardised. For evaluation, we implemented two different cross-validation approaches depending on the dataset. For the eucalypts and half of the simulated datasets, we used five-fold spatial block cross-validation using the *blockCV::spatialBlock* function (Valavi et al., 2019). For the frog and half of the simulated datasets, we implemented five-fold random cross-validation, using the *caret::createFolds* function (Kuhn et al., 2019), as the spatial scale of the frog dataset was too small for practical implementation of spatial block cross-validation.

2.4. Evaluation Metrics

The metrics available for evaluating JSDM predictions can be broadly classified into five groups based on the aspects of performance they consider. Threshold-independent metrics evaluate continuous predicted probabilities against observed presence-absence data. A common example in SDMs is the Area Under the Receiver Operating Characteristic Curve (AUC), but also includes root mean square error (RMSE), and the coefficient of determination (R^2).

Threshold-dependent metrics compare binary predictions against observed presence-absence data. Predicted probabilities are converted to presences if they exceed a set threshold value or absences if they do not. A confusion matrix that contrasts observed and predicted species occurrence states can then be built. Example metrics here include precision, sensitivity, and true/false positive/negative rates. However, thresholding predictions is a contentious topic in the SDM literature (Freeman & Moisen, 2008; Guillera-Arroita et al., 2015; Lawson et al.,

2014; Liu et al., 2005). In addition to debates about the use of thresholds in general, there are also debates about how to determine the threshold value. Thresholds are commonly set to an arbitrary value of 0.5 (Freeman & Moisen, 2008), but an alternative is to make the threshold equivalent to the observed prevalence of the species(Hanberry & He, 2013). A logical extension of this debate for JSDMs is species-specific or community-wide thresholds. However, Lawson *et al* (2014) showed that we can evaluate threshold-dependent metrics without thresholding predictions by calculating a probabilistic confusion matrix. We have used this approach here to avoid any influence of threshold choice impacting our analysis.

As JSDMs are multi-species in nature we can use additional evaluation metrics from community ecology in the form of community dissimilarity indices. These metrics compare how dissimilar our observed and predicted species assemblages are. Common examples are Bray-Curtis dissimilarity and Jaccard distance. These metrics are restricted to evaluating binary predictions. To evaluate these metrics on probabilistic predictions, we simulated binary community assemblages from the appropriate probability distribution. For JSDMs, a community assemblage was drawn per posterior sample; for SSDMs, the same number of community assemblages were simulated.

Species richness metrics consider a model's ability to predict a single aspect of community composition: the number of species present at a site. We consider species richness difference - the predicted richness minus the observed richness.

Likelihood-based metrics assess model fit by computing the probability of observing a given community assemblage, assuming a particular model structure, and given the set of model parameter estimates representing the prediction. It is common to work with the log of the likelihood for numerical stability reasons. The independent log-likelihood represents the typical log-likelihood metric used in SSDMs. This metric independently assesses each species

across all sites– computing the probability of observing a species' presence/absence observations– and combines them into a single metric, assuming the species' distributions to be independent. The joint log-likelihood simultaneously assesses all species, as an assemblage, at each site and accounts for the correlation structure encoded in the core JSDM formulation – the multivariate probit model.

More detail on the metrics including how they are calculated, which prediction types they are appropriate for, and how to interpret them can be found in Appendix S3.

2.5. Model Prediction Comparisons

For the standard SSDM, we obtained binary and probabilistic predictions of community assemblages, and for SESAM only binary predictions. For the JSDMs, we evaluated nine prediction types: binary and probabilistic marginal predictions, binary joint predictions, binary conditional predictions for low-, middle-, and high-prevalence known species scenarios, and probabilistic conditional marginal predictions for low-, middle-, and high-prevalence known species scenarios. For real datasets, the low-prevalence species were randomly selected from those within the bottom 20% of prevalence, medium-prevalence species the middle 30%, and high-prevalence the top 20%. For simulated datasets, this was the species with lowest, median, and highest prevalence. We evaluated a suite of 32 evaluation metrics in total, but some applied only to binary or probabilistic prediction types.

2.6. Analysis of Results

We used linear mixed effects models (MEMs) to analyse the predictive performance of the eight models for the nine prediction methods. We fit an MEM to each combination of evaluation metric and prediction type, to assess the relationship between the response variable (the evaluation statistic) and three explanatory variables (model, dataset, and cross-validation

fold), with a random effect on the intercept of either species or site (depending on the test statistic's calculation method). The performance results for any given model are relative to the standard SSDM approach, which is set as the reference class. A partial interaction between model and dataset was included for the HPR and DPR models to account for observed patterns in the residuals. The MEMs explicitly considered different residual variances for the eight model types to account for evident inhomogeneity of variance. We assessed whether the model residuals met the model assumptions of being normally-distributed with homogenous variance (after accounting for inhomogeneity between model types) with a Kolmogorov-Smirnov test (Massey Jr, 1951), hereafter referred to as a KS-test. The KS-test determines if the distribution of the residuals is significantly different from a normal distribution with the same mean and standard deviation. As we performed a large number of comparisons, we used a Bonferroni-corrected p-value, $\frac{0.05}{405} = 1.2 * 10^{-4}$ to consider whether normality assumptions were violated; to reduce the sensitivity of these tests (Dunn, 1961).

2.7. SOFTWARE

All models were fit using R v3.5.2 (R Core Team, 2018). R packages for model fitting include *BayesComm* v0.1-2 (Golding & Harris, 2015), *boral* v1.7 (Hui, 2018), *gjam* v2.2.5 (Clark & Taylor-Rodríguez, 2018), and *HMSC* v2.2-0 (Blanchet et al., 2019). R packages required for prediction and prediction evaluation are *mvtnorm* v1.0-10 (Genz et al., 2019), *tmvtnorm* v1.4-10 (Wilhelm & B G, 2015), *TruncatedNormal* v1.0 (Zdravko, 2015), *Metrics* v0.1.4 (Hamner & Frasco, 2018), *caret* 6.0-84 (Kuhn et al., 2019), *vegan* v2.5-5 (Oksanan et al., 2018), and *psych* v1.8.12 (Revelle, 2018). Analyses were run on The University of Melbourne's Spartan HPC infrastructure (Meade et al., 2017).

3. RESULTS

No model outperformed the others across all prediction types, but there were consistent trends within each class of validation metrics. The relative performance of the models compared to the standard SSDM ($(metric_{model} - metric_{SSDM}) / |metric_{SSDM}| * 100$) are summarised in Figure 1 for marginal predictions and Figure 3 for joint predictions. Likelihood-based metrics showed that the JSDMs were better fit to the data than the SSDM. Both threshold-dependent and threshold-independent metrics indicated better performance by the SSDM, although the difference was small for the majority of metrics. The SSDM generally outperformed JSDMs for community dissimilarity metrics, with a greater difference for joint prediction types than marginal ones. The JSDMs almost always overpredicted species richness compared to the SSDM for marginal predictions, but HLR-S and HLR-NS had more accurate estimates than the SSDM for most joint prediction types. The SESAM model outperformed both the SSDM and JSDMs for threshold-dependent and community dissimilarity metrics for binary prediction types (Figure 3).

Only a selection of figures is presented in the main article. Forest plots for the absolute value of model performance for each evaluation metric are presented in Appendix S4. Individual heatmaps for the different prediction types are presented in Appendix 5.

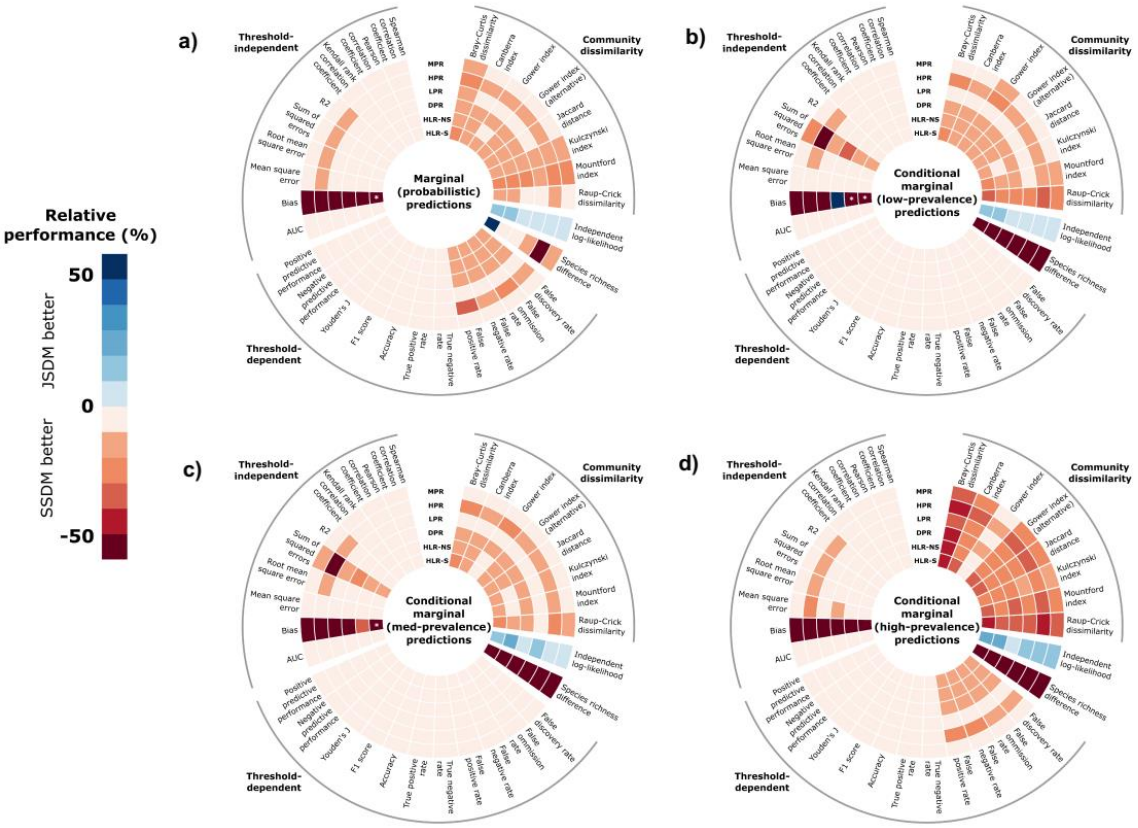
As the JSDM underperformance results are unexpected we performed checks to ensure the MEMs were not returning erroneous results. We found no evidence to suggest our results are an artefact of the model fitting process. We present the result of these checks in Appendix 6.

3.1. Probabilistic Predictions

3.1.1. Single Species Metrics

The SSDM outperformed all JSDMs in marginal predictions for almost every threshold-independent metric (Figure 1). The only exception to this was the bias metric for DPR in low-

prevalence conditional marginal predictions. The relative performance of the JSDMs for the bias metric suggests very poor performance of the JSDMs, but on an absolute scale these differences are actually quite small. As the optimum bias value is 0, good performing models may show high relative differences in this metric when they actually have little absolute difference (Figure 2). Across all probabilistic predictions, the SSDM had a mean bias of $-2.7*10^{-4}$, while the JSDMs average $-1.6*10^{-3}$, i.e. all models were largely unbiased. For all other metrics, the average relative JSDM performance across all marginal prediction types was -6.3% [-20.7, -0.3]. HPR was the worst performer for the error-based metrics with a mean relative difference of -18.3% [-61.8, -2.2]. For low- and medium-prevalence scenarios of conditional marginal predictions, there was a large relative difference between SSDM and the JSDMs in sum of squared errors (-26.9% [-62.0, -13.0]). The mean difference over all marginal prediction types for AUC between the SSDM and the worst performing JSDM is only 0.01.



311

312

313 Figure 1: Heat maps showing the relative performance of the different JSDMs compared to the SSDM
314 for all metrics applicable to *probabilistic* marginal predictions (see Appendix S4 for binary marginal
315 predictions). Each ring represents a particular model, while each ray represents a different evaluation
316 metric, which are clustered by class. Values in blue indicate better performance relative to the SSDM,
317 while red values indicate worse performance relative to the SSDM. The four prediction types shown
318 are a) probabilistic marginal, b) low-prevalence conditional marginal, c) medium-prevalence
319 conditional marginal, and d) high-prevalence conditional marginal prediction. These heat maps were
320 generated using Circos (Krzywinski et al., 2009).

321

322 JSDMs performed worse than the SSDM for threshold-dependent metrics but differences were
323 small. Across all probabilistic predictions, the relative JSDM performance was -3.9% [-18.2, -
324 0.2]. The only notable relative differences were in HPR, DPR, HLR-S, and HLR-NS for the

false rate metrics for probabilistic marginal predictions (-17.8% [-28.7, -13.6]) and high-prevalence conditional marginal predictions (-16.5% [-25.8, -13.5]).

3.1.2. Community Dissimilarity Metrics

The SSDM outperformed the JSDMs for all community dissimilarity metrics on marginal prediction types (Figure 1). The mean relative JSDM performance was -14.5% [-25.5, -4.9] for probabilistic marginal predictions, -14.7% [-34.4, -5.7] for low-prevalence, -12.4% [-22.4, -6.1] for medium-prevalence, and -27.8% [-44.1, -6.1] for high prevalence conditional marginal predictions. For high-prevalence predictions the largest differences were for Bray-Curtis dissimilarity (-40.4% [-46.9, -34.6]) and Raup-Crick dissimilarity (-36.2% [-42.5, -30.9]), while the smallest was for Gower index (-9.3% [-17.8, -5.15]). The HPR model was the worst performer overall with a mean relative difference of -23.1% [-43.7, -14.7] across all marginal prediction types. LPR and MPR were the best JSDMs with a mean relative performance for all marginal prediction types of -12.7% [-32.5, -4.8], compared to -19.7% [-41.2, -7.8] for the other JSDMs.

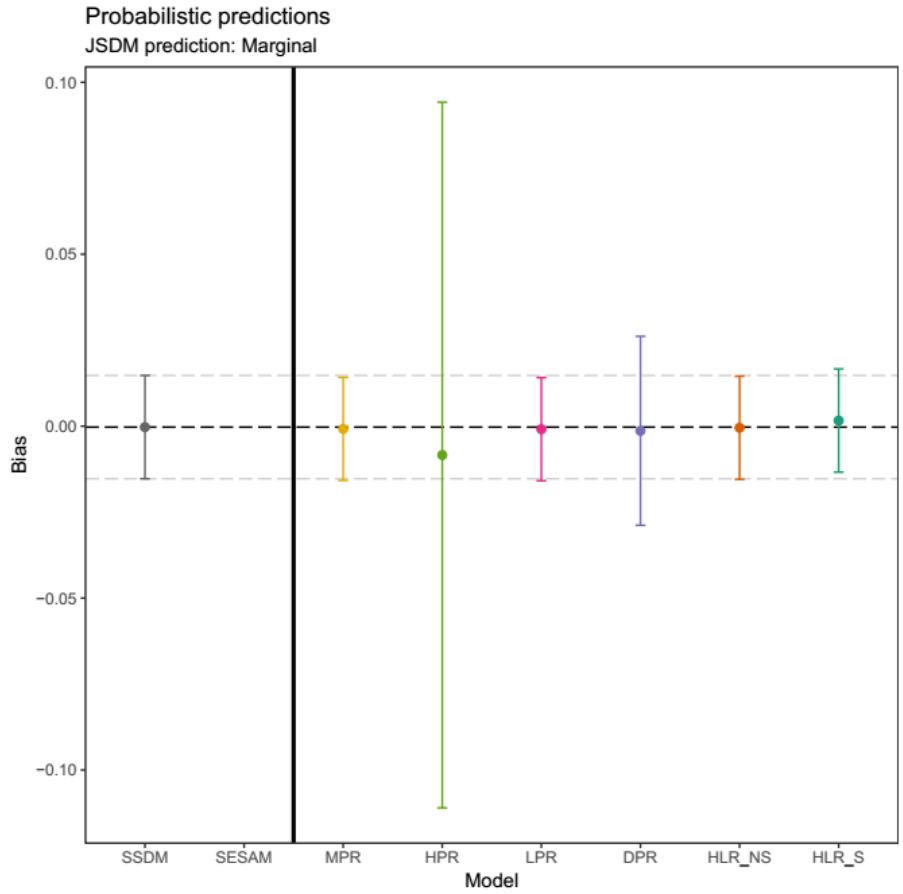


Figure 2: Model performance for the bias (mean error) evaluation metric for probabilistic marginal predictions. Performance estimates are shown as the mean and 95% confidence intervals, over species in each dataset, after accounting for dataset and fold using an MEM. The dashed dark grey line corresponds to the SSDM mean bias, and the dashed light grey line corresponds to 95% confidence intervals from SSDM predictions. SSDMs are shown to the left of the black vertical line, and JSDMs to the right. This figure is illustrative of the absolute metric plots provided for all metrics and prediction types as supplementary information in Appendix S4. SESAM is not plotted here as it does not make predictions for individual species.

3.1.3. Species Richness

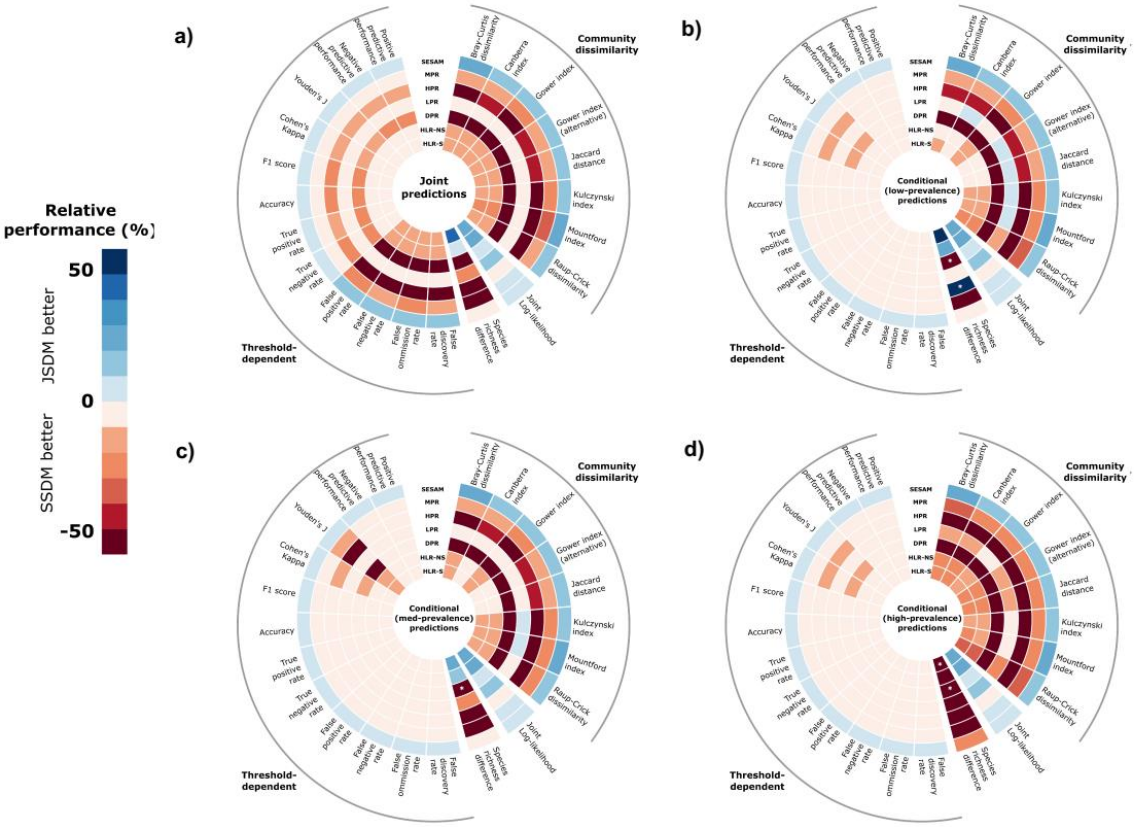
The SSDM generally outperformed JSDMs in species richness difference estimates for marginal predictions. However, the difference in performance is not as dramatic as suggested by the relative performance results in Figure 1 as relative differences can be exaggerated where the metric's optimum value is 0, so we discuss the absolute differences here. The SSDM had a mean species richness difference of 0.06 [0.02, 0.08] across the MEMs fitted for marginal

prediction results. For probabilistic marginal predictions, the JSDMs had a similar mean species richness difference of 0.09 [0.02, 0.15], with HLR-S outperforming the SSDM with a value of just 0.007. For low-, medium-, and high-prevalence conditional marginal predictions, the JSDMs had a difference of 0.86 [0.79, 0.93], 0.63 [0.57, 0.70], and 0.34 [0.28, 0.39] species respectively. HLR-S exhibited the smallest differences of all of the JSDMs for each marginal prediction type.

3.2. Binary Predictions

3.2.1. Single Species Metrics

For threshold-dependent metrics evaluated on joint predictions, the JSDMs performed worse than the SSDM (Figure 3). For binary marginal predictions (Appendix S4), the JSDMs had a mean relative difference of -3.4% [-6.9, -0.7] from the SSDM except for the false rate metrics for HPR, DPR, HLR-S, and HLR-NS with a mean of -17.9% [-29.0, -13.5]. The mean relative difference between JSDMs and SSDMs for conditional predictions was -4.1% [-15.0, -0.2] across all metrics, with a larger relative difference for HPR and DPR on the Cohen's Kappa and Youden's J metrics of -21.8 [-66.7, -11.9]. For all JSDM joint predictions, there was a large relative difference from SSDMs of -36.3 [-97.7, -3.5] for false rate metrics. This was strongest for HPR and DPR with means of -70.4% and -88.7% respectively. LPR had a comparatively small relative difference of -4.0%. Across all threshold-dependent metrics for joint prediction, HPR and DPR had a large relative difference of -35.7% [-82.2, -12.0] and -45.2% [-100.9, -15.0] respectively.



374

375 Figure 3: Heat maps showing the relative performance of the different JSDMs and SESAM compared
376 to the SSDM for all metrics applicable to joint (binary) predictions. Each ring represents a particular
377 model, while each ray represents a different evaluation metric, which are clustered by class. Values in
378 blue indicate better performance relative to the SSDM, while red values indicate worse performance
379 relative to the SSDM. The four prediction types shown are a) joint, b) low-prevalence conditional, c)
380 medium-prevalence conditional, and d) high-prevalence conditional prediction.

381 ### 3.2.2. Community Dissimilarity Metrics

382 For community dissimilarity metrics on binary marginal predictions, the SSDM outperformed
383 the JSDMs. The relative performance difference was largest for HPR, DPR, HLR-S, and HLR-
384 NS (-14.2% [-24.1, -10.0]), and smallest for MPR and LPR (-2.8% [-4.1, -2.0]). For joint
385 prediction types, HPR and DPR performed worst with a mean relative performance difference
386 of -75.3% [-132.7, -44.7]. LPR performed best with a mean difference of -4.3% [-20.5, 6.3],

and performed slightly better than the SSDM for several metrics in low- or medium-prevalence scenarios.

3.2.3. Species Richness

The results for species richness were mixed. Across all joint prediction types, the mean difference between predicted and observed species richness was 0.06 [0.01, 0.09] for the SSDM. HLR-S had a small mean species richness difference of 0.02 [-0.01, 0.04] while HLR-NS performed similarly to the SSDM with a mean difference of 0.06 [0.03, 0.09]. LPR, HPR, and MPR exhibited slightly larger mean species richness differences than the SSDM with values of 0.07 [0.02, 0.1], 0.09 [-0.03, 0.17], and 0.14 [0.1, 0.17] respectively. DPR was the only JSDM to underpredict species richness with a mean difference value of -0.12 [-0.38, 0.14].

3.3. Likelihood-based Metrics

Both likelihood metrics (Figures 1 and 3) indicate that the JSDMs were better fit to the out-of-sample data than the SSDMs, under the assumptions of the univariate and multivariate probit models. For the probabilistic marginal predictions, the JSDMs on average performed 12.2% [7.9, 18.4] better than the SSDM, and for the low-, medium-, and high-prevalence conditional marginal predictions the JSDMs on average performed 12.5% [8.0, 19.0], 13.6% [8.5, 20.4], and 15.2% [10.4, 22.1] better than the SSDM. HLR-S and HLR-NS outperformed the other JSDMs across all probabilistic prediction types with a mean relative independent log likelihood of 19.3% [17.0, 22.0] compared to 9.4% [7.9, 12.0] for the remaining models. Results were similar for the joint log-likelihood except that HPR performed worse relative to the SSDM by 7.8%. Relative to the SSDM, HLR-S and HLR-NS performed better than the other JSDMs by an additional ~10%.

3.4. SESAM Performance

SESAM outperformed the SSDM and all JSDM joint predictions for the threshold-dependent and community dissimilarity metrics. SESAM had a mean relative performance difference of 3.5% [0.7, 15.6] across threshold-dependent metrics for all joint prediction types. This performance is strongest in the false rate metric for binary marginal and joint predictions with a mean relative difference of 13.2% [10.0, 16.7]. For the community dissimilarity metrics, SESAM had a mean relative performance across all joint prediction types of 17.8% [14.4, 22.5]. SESAM and SSDM richness predictions were functionally equivalent, as expected, as SESAM uses SSDM predictions to set it's species richness limit.

4. DISCUSSION

4.1. SINGLE SPECIES METRICS

The SSDM routinely outperformed the JSDMs in marginal prediction for all threshold-independent metrics, but the magnitude of the difference was generally small on the absolute scale of the metric. This differs from Zurell *et al* (2019) which found the JSDM performed much worse than the stacked model. However, that study used a stacked ensemble SDM while we have just used a generalised linear model based SDM. Ensemble models have the potential to outperform single models (Dormann et al., 2018; Hao et al., 2019), which may have increased the observed difference in performance between the JSDM and SSDM. We also observed small differences on the absolute scale for both common error-based metrics (e.g. bias, MSE, RMSE) and the correlation metrics. The only exception to this was the SSE metric for low- and medium-prevalence conditional marginal metrics, but all models, including the SSDM, had significantly larger errors here compared to the other marginal prediction types. This suggests that, for species-level predictions, the JSDMs and SSDM are performing similarly in terms of threshold-independent metrics.

JSDMs performed slightly worse than, but generally similar to, the SSDM for all threshold-dependent metrics on both marginal and joint prediction types but, like for threshold-independent metrics, differences were minor. The relative performance of HPR, DPR, HLR-S, and HLR-NS suggests a large difference between them and the SSDM for the false rate metrics in probabilistic marginal and high-prevalence conditional marginal predictions, but on the absolute scale the performance difference is ≤4% for all metrics. The same trend can be seen in the joint prediction type exhibit larger relative differences for the false rate metrics. On the absolute scale these differences are ≤4% for most metrics, and usually ≤1% for MPR and LPR. The exception to this is that both HPR and DPR perform poorly in the joint predictions, with difference on the absolute scale of 15.5% and 19.6% respectively.

4.2. Community Dissimilarity Metrics

One of the purported benefits of JSDMs over single species models is that accounting for residual species co-occurrences during the model fitting process they will better predict community composition by accounting for species associations. These species associations, which could include species interactions or shared responses to unmeasured environmental variables, provide information on how likely species are to co-occur beyond their response to the measured variables. However, our community dissimilarity metrics results show the JSDMs predicting worse than the SSDM, although the difference was again minor on the absolute scale. Only LPR outperformed the SSDM for some conditional predictions, but just by a mean of 0.004, so the performance can in practice be considered identical.

4.3. Likelihood

In almost all cases JSDMs outperformed the SSDM for both of the likelihood-based metrics, indicating they were better fit to the in-sample data. This is consistent with Norberg *et al*

(2019), who also found better likelihood metrics for JSDMs when they contrasted model pairs that were identical in all aspects except accounting, or not, for residual correlations. The JSDMs exhibiting better likelihood performance is expected as they include additional useful model parameters (i.e. a residual correlation structure). The estimation of regression coefficients is fairly robust to the covariance matrix, so JSDMs and SSDMs would estimate similar species niches(Chib & Greenberg, 1998; Poggiato et al., 2021), therefore a possible explanation for JSDMs performing better on likelihood-based metrics and worse on the other classes is that the JSDMs have overfit the covariance matrix and are explaining some noise in the data (Poggiato et al., 2021). This could also explain why we see greater differences in performance between the JSDMs and the SSDM for the community-level joint predictions that leverage this information. Conditioning these predictions with known information still improves the predictions, which suggests the estimated correlations are too strong rather than sign-switched. A potential solution would be penalising the residual covariance matrix to prevent overfitting (Pichler & Hartig, 2021).

4.4. Species Richness

The species richness difference metric presented mixed results where the SSDM did not consistently perform better or worse than the JSDMs. All models generally over-predicted species richness, with the exception of DPR for conditional predictions, but the mean difference was minimal at $\sim \leq 0.1$ species per site. HLR-S and HLR-NS were the best performing JSDMs overall. For all prediction types, except the conditional marginal, HLR-S outperformed the SSDM and HLR-NS performed equivalently to it. As HLR-S was the only JSDM to regularly outperform the SSDM, it suggests that the effect of spatial scale and/or spatially-driven unmeasured variables could be a potential driver of this result. Species co-occurrence can be driven by several factors, including species interactions or shared responses to environmental

conditions, which operate at different scales. Two species can have the same broad environmental condition preferences but tend to rarely co-occur at a finer scale. If the estimated correlations between species are driven by shared responses to environmental variables at larger spatial scales this can lead to higher species richness estimates at the site level (Ovaskainen, Roy, et al., 2016). All JSDMs had much larger species richness difference estimates for the probabilistic conditional marginal predictions compared to the binary conditional marginal predictions, but it is unclear why. This result contrasts previous works that indicate binary predictions are more prone to overprediction than probabilistic ones (Calabrese et al., 2014; Thuiller et al., 2015; Zurell et al., 2019).

4.5. SESAM Performance

The SESAM predictions outperformed the SSDM and JSDMs for all evaluation metrics in the threshold-dependent and community dissimilarity metric classes. This suggests that constraining the number of species predicted at a site can improve predictive performance at both the species- and community-level, possibly acting as a carrying capacity proxy (Guisan & Rahbek, 2011). These results are consistent with Zurell *et al* (2019) who found that predictions constrained with SESAM's probability rank rule performed better than SSDMs for both species- and community-level metrics. Zurell *et al* (2019) found a small benefit to species richness metrics for SESAM predictions compared to an SSDM, but they used a macroecological model to set the species richness limit compared to using the summed SSDM predictions as we have in this study (thus being unsurprising that we identified no difference between the two in this aspect of predictive performance).

An interesting avenue of research not considered in our study is exploring whether constraining JSDM predictions can improve their performance. The superior performance of SESAM in our results is in concordance with that of Zurell *et al* (2019) who suggested that choosing how

504 species predictions are combined into community-level predictions is potentially more
505 important than choosing the underlying model used to generate them. JSDMs could be
506 predicting likely community assemblages in the absence of limiting factors like site carrying
507 capacities or dispersal limits, and thus could potentially benefit from the application of a
508 constraining framework. Research into how to incorporate suitable constraints into the
509 prediction process itself rather than applying them post-hoc is suggested.

510 5. CONCLUSION

511 While there were consistent trends within evaluation metric classes, we did not find evidence
512 to suggest that any one model outperformed all of the others across all prediction types. The
513 likelihood metrics indicated that the JSDMs were better fit to the data, but SSDMs generally
514 outperformed all of the JSDMs in the rest of evaluation metrics. On the absolute scale, the
515 difference in performance between models was generally small. HLR-S and HLR-NS were the
516 best performing JSDMs and were able to outperform the SSDM for most species richness
517 difference estimates in joint prediction types and generally had the smallest difference in
518 performance from the SSDM when underperforming. The SESAM model consistently
519 outperformed both the JSDMs and the SSDM for both binary species- and community-level
520 metrics which suggests that the application of frameworks to constrain JSDM prediction types
521 should be evaluated in the future.

522

Reference List

Araújo, M. B., & Luoto, M. (2007). The importance of biotic interactions for modelling species distributions under climate change. *Global Ecology and Biogeography*, *16*(6), 743–753. https://doi.org/10.1111/j.1466-8238.2007.00359.x

Blanchet, F. G., Tikhonov, G., & Norberg, A. (2019). *HMSC: Hierarchical Modelling of Species Community* (2.2-0). github.com/guiblanchet/HMSC

Calabrese, J. M., Certain, G., Kraan, C., & Dormann, C. F. (2014). Stacking species distribution models and adjusting bias by linking them to macroecological models. *Global Ecology and Biogeography*, *23*(1), 99–112. https://doi.org/10.1111/geb.12102

Chib, S., & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, *85*(2), 347–361. https://doi.org/10.1093/biomet/85.2.347

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*(1), 34–56. https://doi.org/10.1002/ecm.1241

Clark, J. S., & Taylor-Rodríguez, D. (2018). *gjam: Generalised Joint Attribute Modelling* (2.2.5). https://cran.r-project.org/package=gjam

Cornell, H. V., & Harrison, S. P. (2014). What Are Species Pools and When Are They Important? *Annual Review of Ecology, Evolution, and Systematics*, *45*(1), 45–67. https://doi.org/10.1146/annurev-ecolsys-120213-091759

Dormann, C. F., Calabrese, J. M., Guillera-Arroita, G., Matechou, E., Bahn, V., Bartoń, K., Beale, C. M., Ciuti, S., Elith, J., Gerstner, K., Guelat, J., Keil, P., Lahoz-Monfort, J. J., Pollock, L. J., Reineking, B., Roberts, D. R., Schröder, B., Thuiller, W., Warton, D. I., … Hartig, F. (2018). Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference. *Ecological Monographs*, *88*(4), 485–504. https://doi.org/10.1002/ecm.1309

Dunn, O. J. (1961). Multiple Comparisons Among Means. *Journal of the American Statistical Association*, *56*(293), 52–64.

Fielding, A. H., & Bell, J. F. (1997). A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, *24*(1), 38–49. https://doi.org/10.1017/S0376892997000088

Freeman, E. A., & Moisen, G. G. (2008). A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling*, *217*(1–2), 48–58. https://doi.org/10.1016/j.ecolmodel.2008.05.015

Gelfand, A. E., & Shirota, S. (2021). The role of odds ratios in joint species distribution modeling. *Environmental and Ecological Statistics*. https://doi.org/10.1007/s10651-021-00486-4

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2019). *mvtnorm: Multivariate Normal and t Distributions* (1.0-10). http://cran.r-project.org/package=mvtnorm

562 Golding, N., & Harris, D. J. (2015). *BayesComm: Bayesian Community Ecology Analysis*
563 (0.1-2). https://cran.r-project.org/package=BayesComm

564 Golding, N., Nunn, M. A., & Purse, B. V. (2015). Identifying biotic interactions which drive
565 the spatial distribution of a mosquito community. *Parasites & Vectors*, *8*(1), 367.
566 https://doi.org/10.1186/s13071-015-0915-1

567 Golding, N., & Purse, B. V. (2016). Fast and flexible Bayesian species distribution modelling
568 using Gaussian processes. *Methods in Ecology and Evolution*, *7*(5), 598–608.
569 https://doi.org/10.1111/2041-210X.12523

570 Götzenberger, L., de Bello, F., Bråthen, K. A., Davison, J., Dubuis, A., Guisan, A., Lepš, J.,
571 Lindborg, R., Moora, M., Pärtel, M., Pellissier, L., Pottier, J., Vittoz, P., Zobel, K., &
572 Zobel, M. (2012). Ecological assembly rules in plant communities-approaches, patterns
573 and prospects. *Biological Reviews*, *87*(1), 111–127. https://doi.org/10.1111/j.1469-
574 185X.2011.00187.x

575 Guillera-Arroita, G., Lahoz-Monfort, J. J., Elith, J., Gordon, A., Kujala, H., Lentini, P. E.,
576 Mccarthy, M. A., Tingley, R., & Wintle, B. A. (2015). Is my species distribution model
577 fit for purpose? Matching data and models to applications. *Global Ecology and*
578 *Biogeography*, *24*(3), 276–292. https://doi.org/10.1111/geb.12268

579 Guisan, A., & Rahbek, C. (2011). SESAM - a new framework integrating macroecological
580 and species distribution models for predicting spatio-temporal patterns of species
581 assemblages. *Journal of Biogeography*, *38*(8), 1433–1444.
582 https://doi.org/10.1111/j.1365-2699.2011.02550.x

583 Hamner, B., & Frasco, M. (2018). *Metrics: Evaluation Metrics for Machine Learning* (0.1.4).
584 https://cran.r-project.org/package=Metrics

585 Hanberry, B. B., & He, H. S. (2013). Prevalence, statistical thresholds, and accuracy
586 assessment for species distribution models. *Web Ecol*, *13*, 13–19.
587 https://doi.org/10.5194/we-13-13-2013

588 Hao, T., Elith, J., Guillera-Arroita, G., & Lahoz-Monfort, J. J. (2019). A review of evidence
589 about use and performance of species distribution modelling ensembles like BIOMOD.
590 *Diversity and Distributions*, *25*(5), 839–852. https://doi.org/10.1111/ddi.12892

591 Harris, D. J. (2015). Generating realistic assemblages with a joint species distribution model.
592 *Methods in Ecology and Evolution*, *6*(4), 465–473. https://doi.org/10.1111/2041-
593 210X.12332

594 Hui, F. K. C. (2016). boral - Bayesian Ordination and Regression Analysis of Multivariate
595 Abundance Data in r. *Methods in Ecology and Evolution*, *7*(6), 744–750.
596 https://doi.org/10.1111/2041-210X.12514

597 Hui, F. K. C. (2018). *boral: Bayesian Ordination and Regression AnaLysis* (1.7).
598 https://cran.r-project.org/package=boral

599 Kissling, W. D., Dormann, C. F., Groeneveld, J., Hickler, T., Kühn, I., McInerny, G. J.,
600 Montoya, J. M., Römermann, C., Schiffers, K., Schurr, F. M., Singer, A., Svenning, J.-
601 C., Zimmermann, N. E., & O'Hara, R. B. (2012). Towards novel approaches to
602 modelling biotic interactions in multispecies assemblages at large spatial extents.

603 *Journal of Biogeography*, *39*(12), 2163–2178. https://doi.org/10.1111/j.1365-
604 2699.2011.02663.x

605 Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., &
606 Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics.
607 *Genome Research*, *19*(9), 1639–1645. https://doi.org/10.1101/gr.092759.109

608 Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T., Mayer,
609 Z., Kenkel, B., R Core Team, Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang,
610 L., Candan, C., & Hunt, T. (2019). *caret: Classification and Regression Training* (6.0-
611 84). https://cran.r-project.org/package=caret

612 Lawson, C. R., Hodgson, J. A., Wilson, R. J., & Richards, S. A. (2014). Prevalence,
613 thresholds and the performance of presence-absence models. *Methods in Ecology and*
614 *Evolution*, *5*(1), 54–64. https://doi.org/10.1111/2041-210X.12123

615 Legendre, P., & De Cáceres, M. (2013). Beta diversity as the variance of community data:
616 Dissimilarity coefficients and partitioning. *Ecology Letters*, *16*(8), 951–963.
617 https://doi.org/10.1111/ele.12141

618 Liu, C., Berry, P. M., Dawson, T. P., & Pearson, R. G. (2005). Selecting thresholds of
619 occurrence in the prediction of species distributions. *Ecography*, *28*(3), 385–393.
620 https://doi.org/10.1111/j.0906-7590.2005.03957.x

621 Liu, C., White, M., & Newell, G. (2009). Measuring the accuracy of species distribution
622 models: A review. *18th World IMACS Congress and MODSIM09 International*
623 *Congress on Modelling and Simulation: Interfacing Modelling and Simulation with*
624 *Mathematical and Computational Sciences, Proceedings*, *July*, 4241–4247.

625 Massey Jr, F. J. (1951). The Kolmogorov-Smirnov Test for Goodness of Fit. *Journal of the*
626 *American Statistical Association*, *46*(253), 68–78.
627 https://doi.org/10.1080/01621459.1951.10500769

628 Meade, B., Lafayette, L., Sauter, G., & Tosello, D. (2017). *Spartan HPC-Cloud Hybrid:*
629 *Delivering Performance and Flexibility*. https://doi.org/10.4225/49/58ead90dceaaa

630 Meier, E. S., Kienast, F., Pearman, P. B., Svenning, J.-C., Thuiller, W., Araújo, M. B.,
631 Guisan, A., & Zimmermann, N. E. (2010). Biotic and abiotic variables show little
632 redundancy in explaining tree species distributions. *Ecography*, *33*(6), 1038–1048.
633 https://doi.org/10.1111/j.1600-0587.2010.06229.x

634 Norberg, A., Abrego, N., Blanchet, F. G., Adler, F. R., Anderson, B. J., Anttila, J., Araújo,
635 M. B., Dallas, T., Dunson, D., Elith, J., Foster, S. D., Fox, R., Franklin, J., Godsoe, W.,
636 Guisan, A., O'Hara, B., Hill, N. A., Holt, R. D., Hui, F. K. C., … Ovaskainen, O.
637 (2019). A comprehensive evaluation of predictive performance of 33 species distribution
638 models at species and community levels. *Ecological Monographs*, *89*(3), 1–24.
639 https://doi.org/10.1002/ecm.1370

640 Oksanan, J., Guillaume Blanchet, F., Friendly, M., Kindt, R., Legendre, P., McGlinn, D.,
641 Minchin, P., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E.,
642 & Wagner, H. (2018). *vegan: Community Ecology Package* (2.5-5). https://cran.r-
643 project.org/package=vegan

644 Ovaskainen, O., Abrego, N., Halme, P., & Dunson, D. (2016). Using latent variable models
645 to identify large networks of species-to-species associations at different spatial scales.
646 *Methods in Ecology and Evolution*, *7*(5), 549–555. https://doi.org/10.1111/2041-
647 210X.12501

648 Ovaskainen, O., Roy, D. B., Fox, R., & Anderson, B. J. (2016). Uncovering hidden spatial
649 structure in species communities with spatially explicit joint species distribution models.
650 *Methods in Ecology and Evolution*, *7*(4), 428–436. https://doi.org/10.1111/2041-
651 210X.12502

652 Parris, K. M. (2006). Urban amphibian assemblages as metacommunities. *Journal of Animal*
653 *Ecology*, *75*(3), 757–764. https://doi.org/10.1111/j.1365-2656.2006.01096.x

654 Pichler, M., & Hartig, F. (2021). A new joint species distribution model for faster and more
655 accurate inference of species associations from big community data. *Methods in Ecology*
656 *and Evolution*, *12*(11), 2159–2173. https://doi.org/10.1111/2041-210X.13687

657 Poggiato, G., Münkemüller, T., Bystrova, D., Arbel, J., Clark, J. S., & Thuiller, W. (2021).
658 On the Interpretations of Joint Modeling in Community Ecology. *Trends in Ecology and*
659 *Evolution*, *36*(5), 391–401. https://doi.org/10.1016/j.tree.2021.01.002

660 Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., Vesk, P.
661 A., & McCarthy, M. A. (2014). Understanding co-occurrence by modelling species
662 simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology*
663 *and Evolution*, *5*(5), 397–406. https://doi.org/10.1111/2041-210X.12180

664 R Core Team. (2018). *R: A Language and Environment for Statistical Computing* (3.5.2).

665 Revelle, W. (2018). *psych: Procedures for Personality and Psychological Research* (1.8-12).

666 Schweiger, O., Heikkinen, R. K., Harpke, A., Hickler, T., Klotz, S., Kudrna, O., Kühn, I.,
667 Pöyry, J., & Settele, J. (2012). Increasing range mismatching of interacting species
668 under global change is related to their ecological characteristics. *Global Ecology and*
669 *Biogeography*, *21*(1), 88–99. https://doi.org/10.1111/j.1466-8238.2010.00607.x

670 Thorson, J. T., Ianelli, J. N., Larsen, E. A., Ries, L., Scheuerell, M. D., Szuwalski, C., &
671 Zipkin, E. F. (2016). Joint dynamic species distribution models: a tool for community
672 ordination and spatio-temporal monitoring. *Global Ecology and Biogeography*, *25*(9),
673 1144–1158. https://doi.org/10.1111/geb.12464

674 Thuiller, W., Pollock, L. J., Gueguen, M., & Münkemüller, T. (2015). From species
675 distributions to meta-communities. *Ecology Letters*, *18*(12), 1321–1328.
676 https://doi.org/10.1111/ele.12526

677 Tikhonov, G., Abrego, N., Dunson, D., & Ovaskainen, O. (2017). Using joint species
678 distribution models for evaluating how species-to-species associations depend on the
679 environmental context. *Methods in Ecology and Evolution*, *8*(4), 443–452.
680 https://doi.org/10.1111/2041-210X.12723

681 Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2019). blockCV: An r
682 package for generating spatially or environmentally separated folds for k-fold cross-
683 validation of species distribution models. *Methods in Ecology and Evolution*, *10*(2),
684 225–232. https://doi.org/10.1111/2041-210X.13107

Warton, D. I., Foster, S. D., De, G., Stoklosa, J., & Dunstan, P. K. (2015). Model-based thinking for community ecology. *Plant Ecology*, 669–682. https://doi.org/10.1007/s11258-014-0366-3

Wilhelm, S., & B G, M. (2015). *tmvtnorm: Truncated Multivariate Normal and Student t Distribution* (1.4-10). http://cran.r-project.org/package=tmvtnorm

Wilkinson, D. P. (2019). *JSDM_Prediction v0.1.0.* Zenodo. https://doi.org/10.5281/zenodo.3514766

Wilkinson, D. P., Golding, N., Guillera-Arroita, G., Tingley, R., & McCarthy, M. A. (2019). A comparison of joint species distribution models for presence–absence data. *Methods in Ecology and Evolution*, *10*(2), 198–211. https://doi.org/10.1111/2041-210X.13106

Wilkinson, D. P., Golding, N., Guillera-Arroita, G., Tingley, R., & McCarthy, M. A. (2021). Defining and evaluating predictions of joint species distribution models. *Methods in Ecology and Evolution*, *12*(3), 394–404. https://doi.org/10.1111/2041-210X.13518

Wisz, M. S., Pottier, J., Kissling, W. D., Pellissier, L., Lenoir, J., Damgaard, C. F., Dormann, C. F., Forchhammer, M. C., Grytnes, J. A., Guisan, A., Heikkinen, R. K., Høye, T. T., Kühn, I., Luoto, M., Maiorano, L., Nilsson, M. C., Normand, S., Öckinger, E., Schmidt, N. M., … Svenning, J. C. (2013). The role of biotic interactions in shaping distributions and realised assemblages of species: Implications for species distribution modelling. *Biological Reviews*, *88*(1), 15–30. https://doi.org/10.1111/j.1469-185X.2012.00235.x

Zdravko, B. I. (2015). *TruncatedNormal: Truncated Multivariate Normal* (1.0). https://cran.r-project.org/package=TruncatedNormal

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*, *41*(11), 1876–1887. https://doi.org/10.1111/ecog.03571

Zhang, C., Chen, Y., Xu, B., Xue, Y., & Ren, Y. (2020). Improving prediction of rare species' distribution from community data. *Scientific Reports*, *10*(1), 1–9. https://doi.org/10.1038/s41598-020-69157-x

Zurell, D., Zimmermann, N. E., Gross, H., Baltensweiler, A., Sattler, T., & Wüest, R. O. (2019). Testing species assemblage predictions from stacked and joint species distribution models. *Journal of Biogeography*, *00*, 1–13. https://doi.org/10.1111/jbi.13608