

# ECOGRAPHY

## Research article

### Modeling the rarest of the rare: a comparison between multi-species distribution models, ensembles of small models, and single-species models at extremely low sample sizes

Kelley D. Erickson<sup>ID</sup>✉ and Adam B. Smith<sup>ID</sup>

Center for Conservation and Sustainable Development, Missouri Botanical Garden, Saint Louis, MO, USA

Correspondence: Kelley D. Erickson ([kerickson22@gmail.com](mailto:kerickson22@gmail.com))

Ecography

2023: e06500

doi: [10.1111/ecog.06500](https://doi.org/10.1111/ecog.06500)

Subject Editor: Daniel Stouffer  
Editor-in-Chief:  
Jens-Christian Svenning  
Accepted 14 February 2023



Species distribution models are useful for estimating the distribution and environmental preferences of rare species, but these same species are challenging to model on account of sparse data. We contrast a traditional single-species approach (generalized linear models, GLMs) with two promising frameworks for modeling rare species: ensembles of small models (ESMs), which average across simple models; and multi-species distribution models (MSDMs), which allow rarer species to benefit from statistical ‘borrowing of strength’ from more common species. Using a virtual species within a community of real species, we evaluated how model accuracy was influenced by the number of occurrences of the rare species ( $N=2-64$ ), niche breadth, and similarity to more numerous species’ niches. For discriminating between presence and absence, ESMs with just linear terms (ESM-L) performed best for  $N \leq 4$ , whereas for GLMs and ESMs with polynomial terms (ESM-P) were best for  $N \geq 8$ . For calibrating the species’ response to influential variables, the MSDM hierarchical modeling of species communities (HMSC) and ESM-P were best for species with niches similar to those of other species. For species with dissimilar niches, ESM-P did best for  $N \geq 8$ , but no model was well calibrated for smaller sample sizes. For identifying uninfluential variables, ESM-L and species archetype models (SAMs), a type of MSDM, did well for  $\leq 4$ , and ESM-L for  $N \geq 8$ . Models of species with narrow niches dissimilar to others had the highest discrimination capacity compared to models for generalist species and/or species with niches similar to other species’ niches. ‘Borrowing of strength’ in MSDMs can assist with some inference tasks, but does not necessarily improve predictions for rare species; simpler, single-species models may be better at a given task. The best algorithm depends on modeling goal (discrimination versus calibration), sample size, and niche breadth and similarity.

Keywords: borrowing of strength, calibration, data-deficient species, discrimination, presence–absence, rare species



[www.ecography.org](http://www.ecography.org)

---

© 2023 The Authors. Ecography published by John Wiley & Sons Ltd on behalf of Nordic Society Oikos

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Page 1 of 12

## Introduction

In the face of unprecedented threats to biodiversity, understanding species' niche preferences and distributions is as critical for conservation as ever (Guisan et al. 2013). Species distribution models (SDMs), which correlate species' presences with environmental conditions, are widely used to determine the current and past distribution of species as well as to predict how they will respond to global change (Elith and Leathwick 2009). Of special concern are rare species. Although there are a variety of situations that give rise to 'rarity' (Rabinowitz 1981), as the term is used here, 'rare' species have few known occurrences and are especially challenging to model (Jeliazkov et al. 2022). Indeed, many critically endangered species have very few occurrence records (Lomba et al. 2010, Zizka et al. 2018) and low sample sizes have been identified as one of the factors most often found to reduce model accuracy (Stockwell and Peterson 2002, Hernandez et al. 2006, Wisz et al. 2008, Thibaud et al. 2014, Santini et al. 2021, Butler and Sanderson 2022). The suggested minimum numbers of occurrences range from approximately 5 to > 200 for 'traditional' single-species SDMs (Wisz et al. 2008, van Proosdij et al. 2016, Santini et al. 2021). Sample size is often only somewhat under a modeler's control, either due to the inherent scarcity of a species, limited resources, or difficulty in sampling. Given the abundance of species with very little data and the urgency of understanding their niche preferences and distributions, how should modelers proceed with the rarest of the rare (Yoccoz 2022)?

Traditional, single-species modeling methods typically perform poorly for species with few occurrences (Wisz et al. 2008, Santini et al. 2021). A general recommendation is to include ~ 10 presences per coefficient, which limits the number of terms allowable in a model when sample size is low (Brun et al. 2020). For species with few occurrences, this can severely reduce model complexity and realism. Low sample size also increases model variance (uncertainty in coefficient estimates and sensitivity of coefficient estimates to small changes in the data) and bias (mis-estimation of coefficients; James et al. 2013, Winship and Western 2016). These issues are further exacerbated when models are mis-specified (e.g. inclusion of unimportant predictors).

Several techniques hold promise for modeling species with few occurrences. One newer method employs ensembles of small models (ESMs), in which multiple, simple models with just a few terms are fit and then weighted by model performance to create an ensemble prediction (Lomba et al. 2010, Breiner et al. 2015). By fitting simple models one-by-one, ESMs obviate problems related to increased variance arising from fitting multiple coefficients with sparse data, but may not be able to address biases inherent in how small samples represent environmental space. ESMs have been shown to be more accurate than traditional species distribution models at sample sizes, although systematic assessments to date have only explored sample sizes ranging down to 10 (Breiner et al. 2015).

Another general promising approach for modeling rare species is through multi-species distribution models (MSDMs),

which model multiple species at once. MSDMs employ statistical 'borrowing of strength' in which information from more numerous species helps estimate coefficients for less numerous ones (Zipkin et al. 2009, Hui et al. 2013, Zhang et al. 2020). Borrowing of strength can be implemented in a variety of ways within frequentist or Bayesian frameworks, using model terms shared across species, including random effect variances, joint priors or hyperpriors, and latent factors. For example, species archetype models (SAMs) are a frequentist approach that uses a mixture distribution to assign each species to a group of species which share the same general form of response to environmental gradients (Hui et al. 2013). Similarly, hierarchical modeling of species communities (HMSC) is a Bayesian approach that employs joint priors on latent factors to estimate species' responses to the environment (Ovaskainen et al. 2010, Ovaskainen and Soininen 2011, D'Amen et al. 2017, Ovaskainen and Abrego 2020). On the assumption that borrowing of strength enables reliable estimation of coefficients of rare species, MSDMs have been used to model species with very few – sometimes even zero – observed occurrences (Ovaskainen and Soininen 2011).

Species' niche breadth and position can also influence model performance, and this can interact with modeling algorithm. Species that prefer a restricted set of environments and/or environments that are marginal to the sampled environmental space have niches that tend to be easier for models to differentiate from available environmental space (Santika 2011, Smith et al. 2013, Soultan and Safi 2017, Connor et al. 2018). In contrast, within the MSDM framework, borrowing of strength presumably most benefits rare species when they have responses that are close to the community-wide mean (Zipkin et al. 2009, Clark et al. 2014). Hence, MSDMs should be able to model species with 'average' niches more accurately than species with niches that are more marginal to the community.

Determining the best approach to modeling very rare species depends on whether the goal is to predict the geographical distribution of the species (species distribution modeling) or to estimate responses to environmental gradients (ecological niche modeling; Peterson et al. 2011, Guillera-Arroita et al. 2015). For example, models can be used to locate previously unknown populations (Guisan et al. 2006, Le Lay et al. 2010) or to estimate species' current distributions for identifying areas of conservation concern (Carroll et al. 2010). In both situations, the focus is on determining the geographic distribution of the species, so discrimination accuracy (ability to differentiate between presence and absence) is important (Norberg et al. 2019). Alternatively, modelers may be more interested in estimating species' environmental tolerances, for example to identify suitable habitat for establishing new populations (Albrecht and Long 2019), or predicting how a species may respond to a changing climate (Thuiller et al. 2006). In these cases, calibration accuracy (ability to identify suitable environments, regardless of whether a species occurs there) is more important (Norberg et al. 2019).

Here, we use virtual species (Meynard and Kaplan 2013), simulated on a real landscape within a community of real

species, to determine how the number of occurrences and its interaction with niche breadth and similarity to other species (niche position) influences the discrimination and calibration accuracy of single-species and multi-species SDMs. Specifically, we compared performance of ESMs, HMSC, SAMs, and a ‘traditional’ single-species distribution model (a generalized linear model, or GLM). We hypothesized that borrowing of strength would allow MSDMs to out-perform ESMs and GLMs when the focal species was similar in its niche position to the other species in the community. However, as sample size increased and as the species’ niche became more distinct from the other species, we expected GLMs and especially ESMs to perform better.

## Material and methods

### Study system

The data for our study are derived from [Norberg et al. \(2019\)](#), who found that MSDMs, and especially HMSC, generally out-performed single-species models for real communities. We used the tree community data from the United States Forest Service’s Forest Inventory and Analysis (as prepared by [Norberg et al. 2019; https://zenodo.org/record/2637812](#)), which comprises 1200 plots, with presence/absence data for 63 species. The original dataset had 38 environmental covariates, which [Norberg et al. \(2019\)](#) reduced to three principal components (PCs). To more realistically approximate sampling effort for a rare species, as well as to decrease computational time, we subset our data to only include the 474 plots located in the southeastern USA (Oklahoma, Arkansas, Tennessee, North Carolina, South Carolina, Texas, Louisiana, Mississippi, Alabama, Georgia, Florida). [Norberg et al. \(2019\)](#) discarded species that occurred at fewer than ten sites, as well as species that were not represented by at least one presence in each of their cross-validation training data sets. Our subsampling resulted in eight species with fewer than ten occurrences but more than zero (Supporting information), but we opted to retain them in the set of ‘real’ species (except for one species that was only present at one site) since inclusion of rare species can improve results from multi-species modeling ([Ovaskainen and Soininen 2011](#)) and because our focus was on very rare species. Our real community thus comprised 62 species.

### Simulated species

We considered two axes along which species could differ from each other: niche breadth (defined by the standard deviation across environmental covariates determining a species’ fundamental niche) and niche position within the environmental space available to the community (distance from the community-average niche). We simulated four different species types: a ‘central generalist’ with a broad fundamental niche centered on the community average, a ‘central specialist’ with a narrow niche breadth centered on the community average,

a ‘marginal generalist’ with a broad niche removed from the community average, and a ‘marginal specialist’ with a narrow niche removed from the community average.

In preliminary modeling of the real community (without any virtual species), we found that many of the species only responded to one or two of the environmental PCs. We therefore defined our simulated species as responding to PC1 and PC3 while being agnostic toward PC2, but provided PC2 to the SDMs to recreate the common case where a modeler is only partially knowledgeable about the factors driving species’ distributions. We defined the functional responses of our four species as the product of independent normal distributions for PC1 and PC3, normalized so the maximum probability of presence across sites was one.

For our generalist species with a broad niche, we set fundamental niche breadth (the standard deviation of the normal curves used to generate the niche) equal to the 90th percentile of realized niche breadths across the real species in each of the two niche axes ( $\sigma_{PC1} = 2.03$ ,  $\sigma_{PC3} = 2.43$ ; [Fig. 1](#)). Central species had their optimum located at the average of the mean environment occupied across species ( $\mu_{PC1} = -4.72$ ,  $\mu_{PC3} = 0.37$ ). To choose the marginal and specialist niche breadth values, we chose a combination where the suitability function still yielded a relatively high probability of occurrence across the most suitable 128 sites to ensure that there was enough variation between replicates ( $\mu_{PC1} = -3.5$ ,  $\mu_{PC3} = 2$ ,  $\sigma_{PC1} = 0.94$ ,  $\sigma_{PC3} = 1.58$ ; [Fig. 1](#)). Visual inspections of occupied habitats across replicates ensured that presences were likely to occur at sites with high suitability, while still allowing for variation among replicates ([Meynard and Kaplan 2013](#), [Meynard et al. 2019](#)).

To generate occurrence data using a given niche, we used the habitat suitability function as the sampling probability to randomly select  $N + 64$  sites, where  $N = \{2, 4, 8, 16, 32, \text{ or } 64\}$ . The remaining sites were set as absences. We randomly withheld 64 presences and 64 absences as test data to be used in cross-validation for each of the replicates, and used the remainder of the presences ( $N$ ) and absences ( $474 - 2 \times 64 - N$ ) for model training. For each of the four niche types, we generated 100 replicates for each of the six levels of  $N$ , for a total of 2400 species.

### Generalized linear models (GLMs)

We used a single-species generalized linear model (GLM) with binomial errors. Models included both linear and quadratic terms, and were implemented using the `glm` function in the ‘stats’ R package ([www.r-project.org](#)). We used GLMs only for cases where  $N \geq 8$ .

### Ensembles of small models (ESMs)

As presented in [Lomba et al. \(2010\)](#) and [Breiner et al. \(2015\)](#), ESMs comprise an ensemble of small models each with just two linear terms. Using GLMs, we constructed two variants of ESMs, a version using up to six sub-models with just one or two predictors appearing in linear terms (ESM-L), and a version using up to 26 sub-models with linear and polynomial

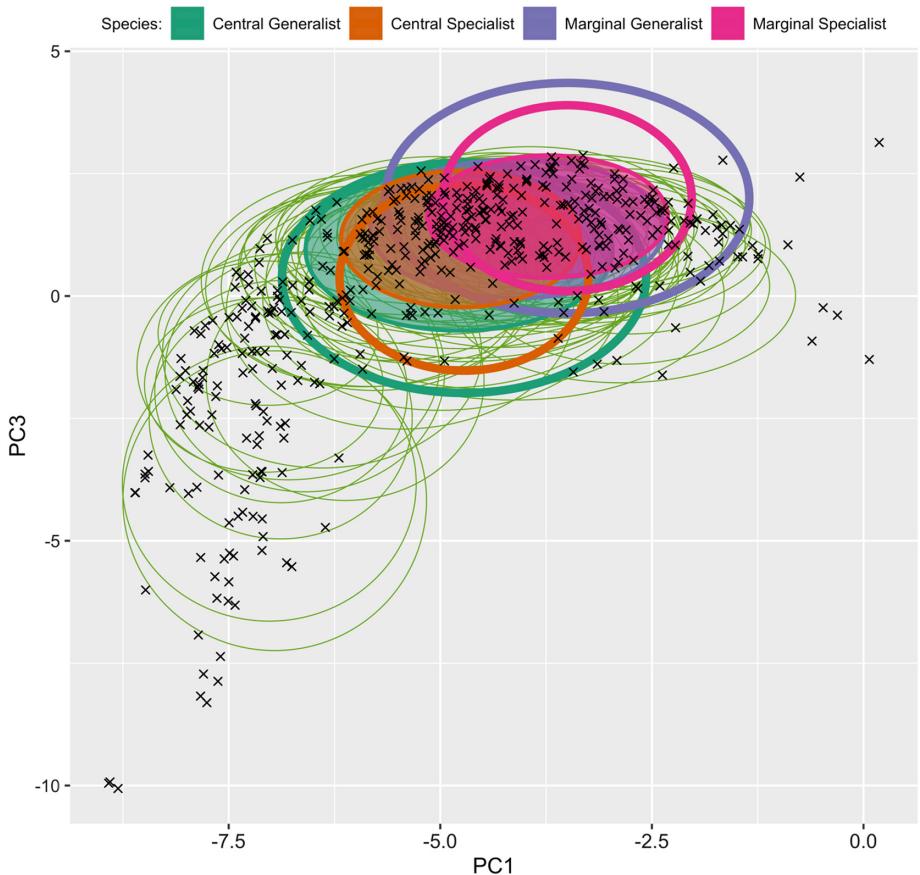


Figure 1. Real and simulated species in environmental space. Each cross represents a sampled site. Ellipses with thin borders represent the real species' realized niches and are centered on mean occupied environments, with axes equal to  $\pm 1$  SD across occupied sites. Ellipses with thick borders represent fundamental and realized niches for the four types of simulated species used in the study: a broad niche centered on the community average environment (central generalist), a narrow niche centered on the community average environment (central specialist), a broad niche centered on a point in environmental space more extreme than the community average (marginal generalist), and a narrow niche centered on the extreme point (marginal specialist). Each simulated species has two ellipses, one indicating fundamental niche (open ellipses), and one the average realized niche given the arrangement of sites (shaded ellipses). To generate this figure, realized niches for simulated species were calculated from the mean and standard deviation of 1000 iterations of choosing 64 presences for each species.

terms (ESM-P; see the Supporting information for model types). We trained ESM-L for all sample sizes, and ESM-P for cases with  $N \geq 8$ . To calculate the ensemble prediction, we weighted each sub-model's prediction by the Somer's D statistic ( $2 \times \text{AUC} - 1$ ), with models with a  $D < 0$  assigned a weight of zero (Breiner et al. 2015). Somer's D was calculated using training presences and absences. In some cases, all the sub-models for a particular replicate were assigned a weight of zero. In these cases, we considered the replicate 'unconverged.'

#### Hierarchical models of species communities (HMSC)

HMSC models each species' response to the environment as per a GLM, albeit with joint priors on latent factors shared across species. To implement the model, we used the 'Hmsc' R package ([www.r-project.org](http://www.r-project.org), Tikhonov et al. 2021) with 20 000 iterations after a burn-in of 2000 with one latent factor and default priors (Ovaskainen and Abrego 2020, Chapter

8; Tikhonov et al. 2020). We did not use the option to allow the model to estimate interspecific associations (i.e. species-by-species covariances; therefore, a 'joint' species distribution model; Pollock et al. 2014). We ran HMSC for all sample sizes. For  $N \leq 4$ , models only included linear terms, while for  $N \geq 8$  models included both linear and quadratic terms. We considered a model to have converged if the Gelman–Rubin statistic  $\check{R}$  was  $< 1.2$  (Gelman and Rubin 1992).

#### Species archetype models (SAMs)

We used species archetype models (SAMs), another type of MSDM, which probabilistically assign each species to an 'archetype' category based on how it responds to the environment (Hui et al. 2013). Species in an archetype respond in the same general fashion to variables (e.g. linearly), but are allowed to have different coefficients (Hui et al. 2013). We implemented SAMs using the 'ecomix' R package ([www.r-project.org](http://www.r-project.org), Wooley et al. 2022). Based on preliminary

modeling of the real species (without the virtual species), we used six species archetypes. In some cases, the expectation/conditional maximization algorithm used by ‘ecomix’ was unable to settle on the maximum likelihood (cf. Norberg et al. 2019). In these cases, we considered the replicate unkonverged. We trained SAMs for all sample sizes.

## Evaluating model performance

To evaluate the ability of models to discriminate between sites where the species was present and absent, we calculated the area under the receiver operator curve (AUC) and Tjur’s  $R^2$  (the average model prediction at presence sites minus the average prediction at absence sites; Tjur 2009). AUC indicates the probability a randomly chosen occurrence has a higher model prediction than a random absence (Mason and Graham 2002), and Tjur’s  $R^2$  the ease of discriminating between presences and absences sites given model predictions.

To evaluate calibration accuracy, we measured the match between the species’ true response curves and estimated responses from the SDMs. For PC1 and PC3 (the environmental covariates we used to generate the simulated species), we used the *compareResponse* function in the ‘enmSdmX’ R package ([www.r-project.org](http://www.r-project.org), Smith 2022) to calculate the rank correlation between the actual habitat suitability and the models’ predicted response curves across the environmental range of all sites, while holding other variables at the mean across the species’ test occurrences. Models that are better able to recover the species response curve should have higher rank correlation. For PC2, which was not used to generate the simulated species, we calculated the standard deviation of the models’ predicted response curves, as PC2 was varied from the minimum to the maximum value across all plots while values of PC1 and PC3 were held constant at the mean value across the species’ occurrences. Well-calibrated models should predict the response along PC2 to have a standard deviation of zero, indicating a ‘flat’ response to this variable. We also evaluated overall model calibration using the root-mean square error (RMSE).

We did not conduct statistical hypothesis tests because we know, a priori, that differences exist between species’ niches and sample sizes (White et al. 2014). Rather, we focused on effect size differences by examining overlap between the inner 80th percent of values of each test statistic (AUC, Tjur’s  $R^2$ , RMSE). All analyses were conducted in R ver. 4.2.0 ([www.r-project.org](http://www.r-project.org)). Code for running all analyses is available at [www.github.com](http://www.github.com).

## Results

### Model convergence

GLMs and ESM-P always converged, and ESM-L almost always converged. In contrast, both MSDMs (HMSC and SAM) converged in roughly 75% of all cases regardless of sample size (Fig. 2). Unkonverged models were not considered in further analyses.

### Model discrimination

For all niche types, at small sample sizes ( $N \leq 4$ ), ESM-L had the highest median AUC values and were dependably better than random (Fig. 3a). HMSC had the second-highest AUC, followed by SAMs, which often had models that did no better than or even worse than random. For larger sample sizes ( $N \geq 8$ ), niche type had a greater effect on which models did best. Generally, GLMs and ESM-P had the highest AUC values, but these were often followed closely by HMSC as sample size increased. At  $N \geq 8$ , ESM-L did poorly relative to the other algorithms, except for the marginal generalist, where it was among the best models. SAMs had the lowest median AUC values for all niches and sample sizes. For a given sample size, the maximum AUC was higher for the two specialist species than for the two generalist species.

Although there was very little distinction between Tjur’s  $R^2$  at small sample sizes ( $N \leq 4$ ), Tjur’s  $R^2$  indicated some further differences between models where AUC did not (Fig. 3b). For  $N \geq 8$  for all species, GLMs had the highest

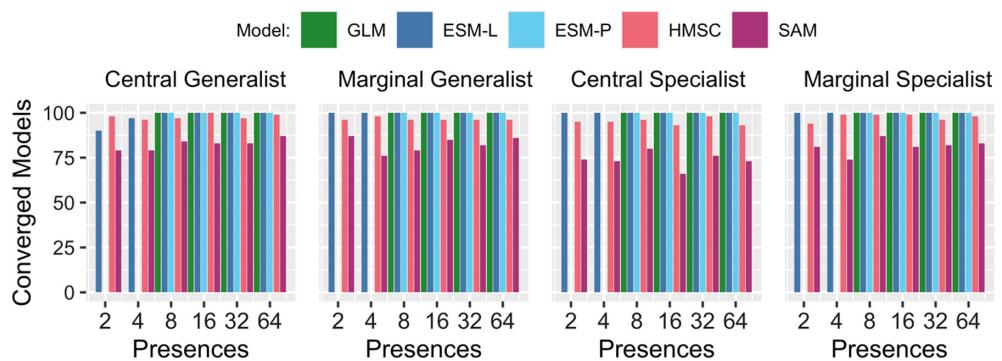


Figure 2. Number of converged models as a function of species’ niche breadth and position, and number of training presences. Ensembles of small models with only linear terms (ESM-L) and ESL with polynomial terms (ESM-P) were considered to be converged if at least one sub-model had a nonzero weight. Hierarchical models of species communities (HMSC) were considered converged if all values of their Gelman–Rubin statistic  $\bar{R}$  were  $< 1.2$ . Generalized linear models (GLMs) and species archetype models (SAMs) were considered converged if the maximum likelihood estimation converged.

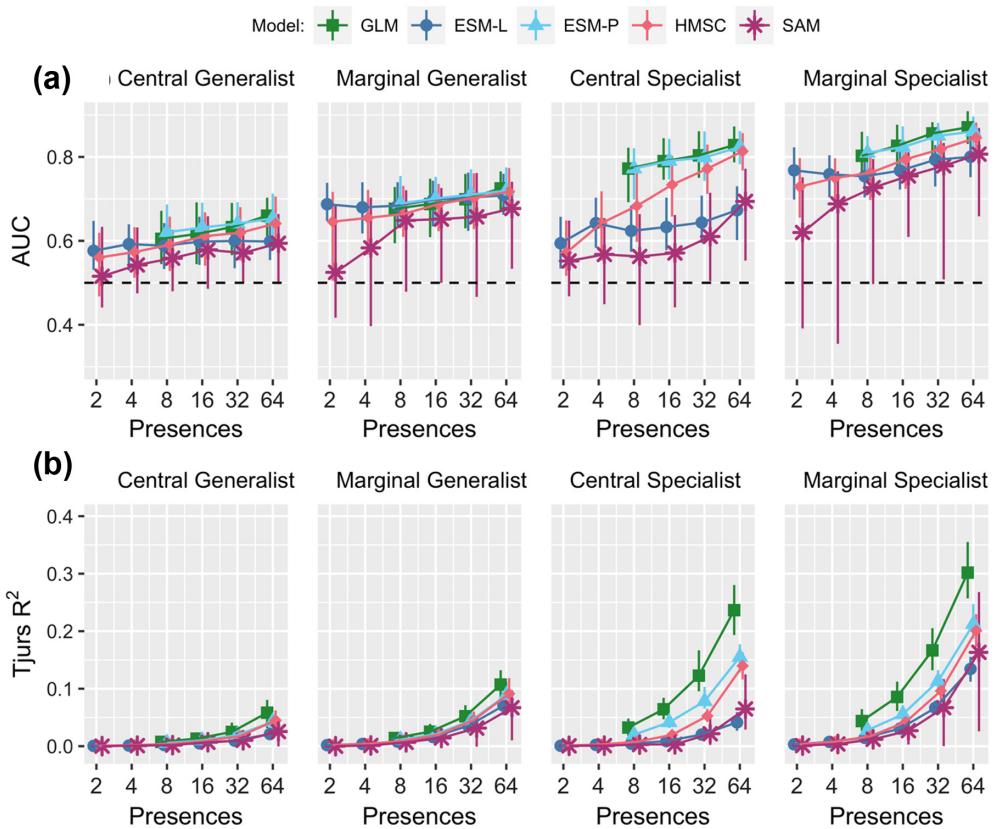


Figure 3. Measures of model discrimination (a) AUC and (b) Tjur's  $R^2$  as a function of number of presences, and species' niche breadth and position. Provided sample size was adequate, each species was modeled with generalized linear models (GLMs), ensembles of small models with linear terms (ESM-L), ensembles of small models with polynomial terms (ESM-P), hierarchical models of species communities (HMSC), and species archetype models (SAMs). Trend lines are included to make it easier to discern the relative position of models. Error bars span the 10th and 90th percentiles of AUC or Tjur's  $R^2$  across converged models.

Tjur's  $R^2$  (i.e. model predictions more clearly separated presences from absences). For the generalist species, there was little difference in Tjur's  $R^2$  between the other models, whereas for the specialists, ESM-P had the second-best values, followed by HMSC.

For all models and all species, RMSE improved (decreased) with sample size (Supporting information), but otherwise displayed trends within and across species qualitatively similar to AUC and Tjur's  $R^2$ .

## Model calibration

For ease of interpretation, we classified the correlation between model predictions and true environmental suitability along PC1 and PC3 as 'positive' if it was  $> 0.33$ , 'neutral' if between  $0.33$  and  $-0.33$ , and 'negative' if  $< -0.33$ . Across models and species, calibration with PC3 was relatively high (Fig. 4b), so we focus on here PC1 (Fig. 4a). Of the three algorithms used for  $N \leq 4$  (HMSC, SAMs, and ESM-L), HMSC had the lowest rate of mis-calibrated models, though correlations between true and predicted suitability were still often negative, especially for the two marginal species. For  $N \geq 8$ , across all niche types GLMs and ESM-P had the lowest rates of mis-calibrated models. HMSC also had few

mis-calibrated models for the two central niche types. For PC2, which was not used to generate the niche, at  $N \leq 4$  ESM-L had lower variation in response than HMSC (indicating that ESM-L tended not to mistakenly identify PC2 as important). For  $N \geq 8$ , ESM-L also had the lowest rate of overfitting to PC2, while GLMs had the worst.

## Discussion

We evaluated the ability of two promising modeling frameworks – MSDMs and ESMs – for modeling species with extremely low numbers of occurrences. We hypothesized that when the number of occurrences was very low, species with 'central' niches that closely resembled those of others would benefit from MSDMs because they employ statistical borrowing of strength. We also hypothesized that as sample size increased and the species' niche became more 'marginal' to the community, the single-species approach based on ESMs would become better because the species' niche was more distinct from the background environment and there would be enough data to inform simple models, yet not enough to correctly parameterize a single-species model like a GLM (Lomba et al. 2010, Smith et al. 2013, Breiner et al. 2015, Soltan and Saf

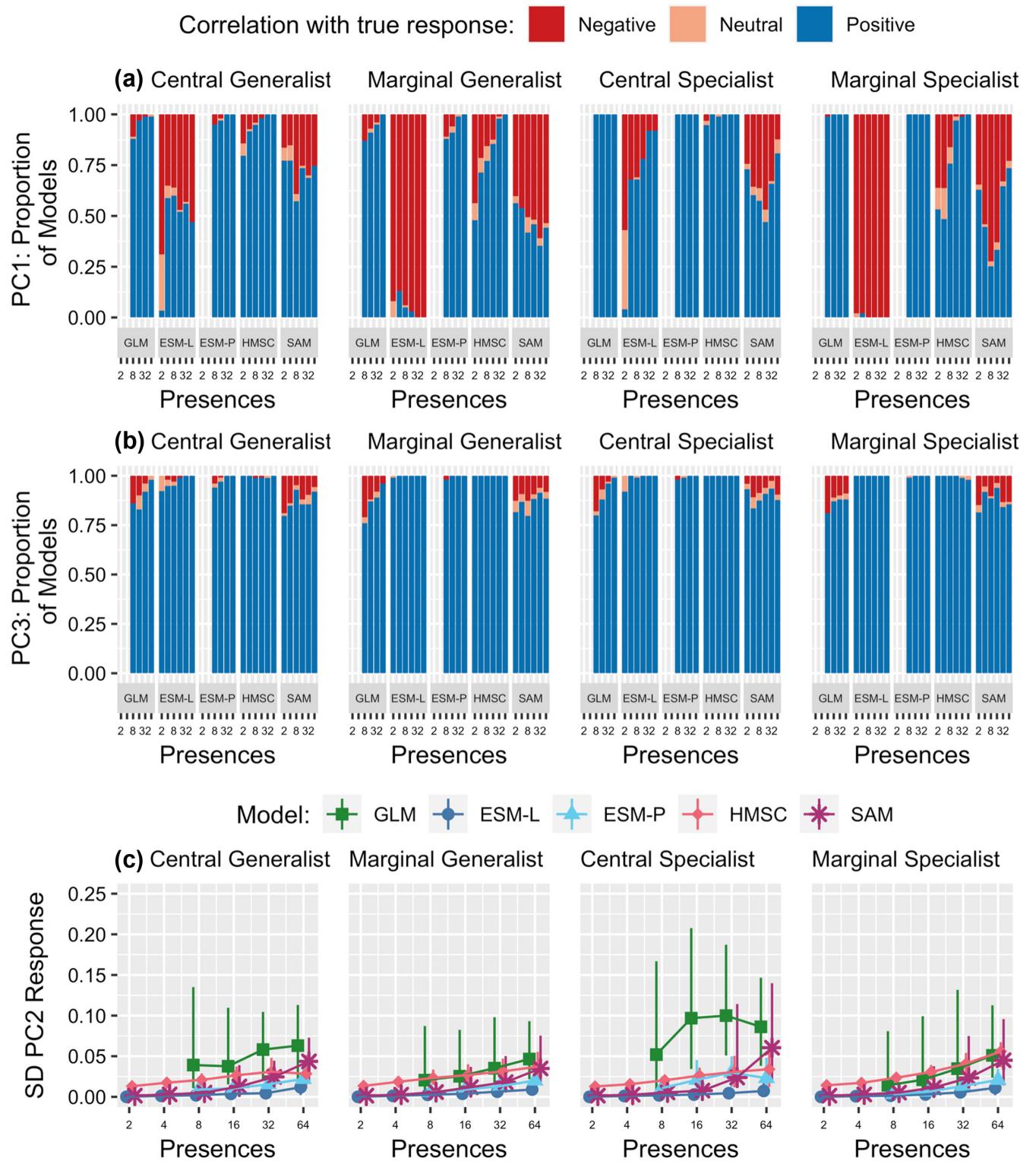


Figure 4. Ability of models to correctly calibrate predictions with the true probability of presence as a function of niche breadth and position, number of presences, and model algorithm. (a) and (b) Proportion of models where the rank correlation between the real and predicted response curve for PC1 and PC3 was > 0.33 (positive), between -0.33 and 0.33 (neutral) or < -0.33 (negative). (c) Standard deviation of the predicted response curve for PC2, which was provided to models but not used to create niches. Values closer to zero indicate better calibration for PC2, which had no influence on the species. Error bars span the 10th and 90th percentiles across converged models.

2017, Connor et al. 2018). We found support for our predictions only when the modeling goal was related to calibration accuracy, but not discrimination (Fig. 3 versus 4). Indeed, for calibration tasks where the species' niche was centered on the community mean, the HMSC multi-species distribution model was best at small sample sizes when the goal was to calibrate model responses. As sample size increased and niches were removed to a marginal position away from other species, ESMs with polynomial terms (ESM-P) were better calibrated. In contrast, for discrimination tasks, single-species modeling algorithms (ESMs or GLMs) were nearly always best regardless of niche position and sample size (Fig. 3). In sum, the

best algorithm depends on the modeling task, how sample size influences model complexity and bias-variance tradeoffs, and niche characteristics (summarized in Fig. 5).

### Modeling goal

Determining which model type to use depends on whether the focus of the modeling effort is in understanding the fundamental niche or the geographic distribution of a species (Fig. 5). Ideally, models would be able to reproduce both a species' geographic distribution and environmental responses reliably, but often there is a tradeoff between model discrimination and

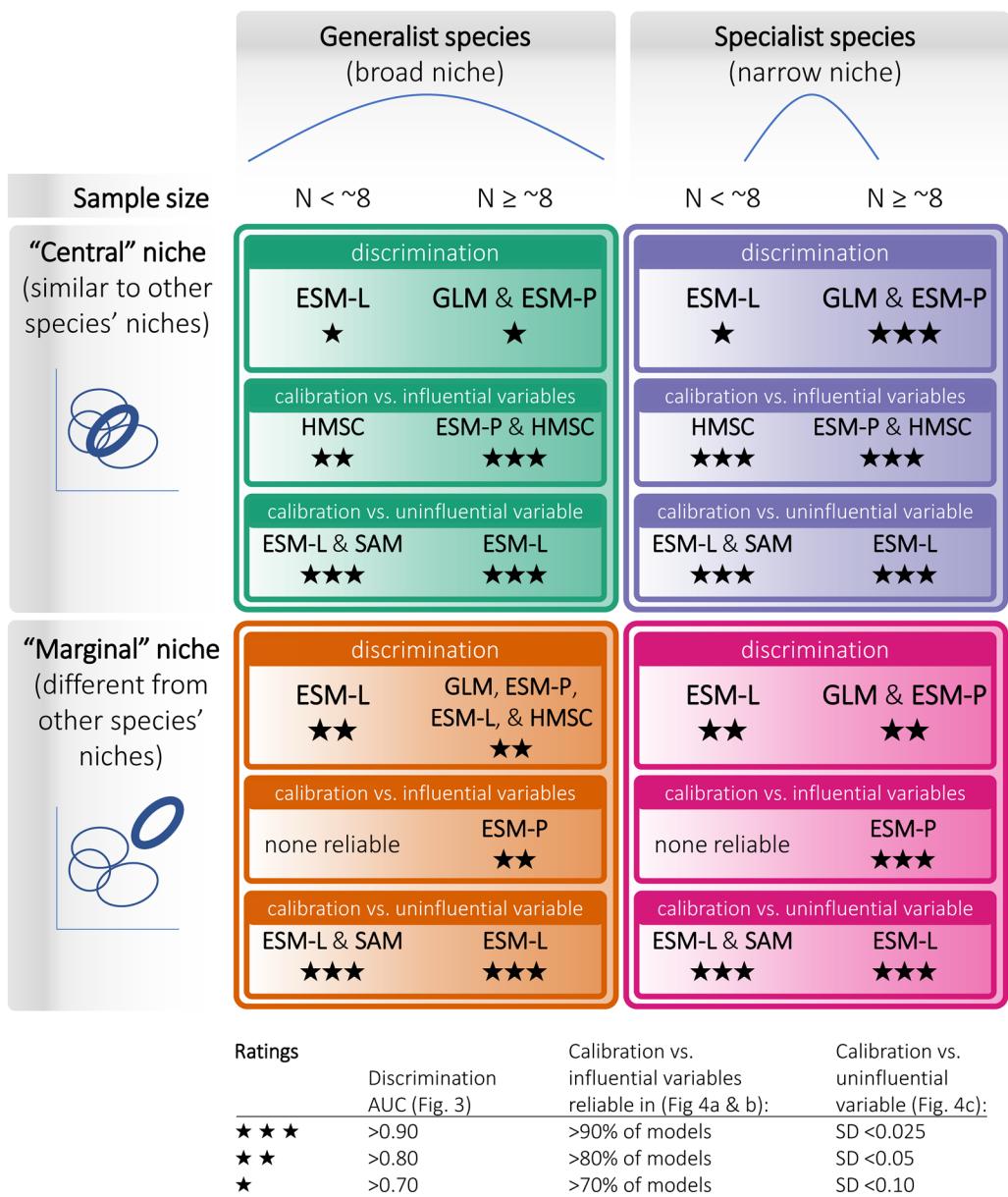


Figure 5. Summary of results. For each combination of modeling aim (discrimination versus calibration), niche breadth (narrow versus broad), niche position (similar to other species versus peripheral), and number of occurrences (N), we note the models that performed the best based on AUC and Tjur's R<sup>2</sup> (Fig. 3), and calibration with influential and non-influential variables (Fig. 4). To break rough ties between models in discrimination, we used Tjur's R<sup>2</sup>.

calibration (Jiménez-Valverde et al. 2013). If the focus is on estimating the geographic range, ESMs or GLMs would be a good choice as they had better discrimination metrics than the other models across all sample sizes (Fig. 3). However, if identifying niche relationships is the focus, the best algorithm depends on niche characteristics and sample size, and on whether the focus is on identifying influential or uninfluent variables (Fig. 4). When the task was to calibrate model responses to influential variables (i.e. PC1 and PC3), HMSC had the lowest rate of mis-calibration for species with few occurrences and niches centered on the community mean. HMSC was also well calibrated for these same species at larger sample sizes, as was ESM-P. For species with niches marginal to the community, no algorithm was well calibrated when sample size was low, though ESM-P had better success at larger sample sizes (Fig. 4). When the task was to calibrate responses to the uninfluent variable (PC2), ESM-L and SAMs did well at small sample sizes, and ESM-L at larger sizes.

For the same species and sample size, specific algorithms tended to do well in one task but poorly at the others (Fig. 5). We believe that these alternating ‘successes’ and ‘failures’ may arise from the configuration of niches and environmental space in our data (Fig. 1). We generated virtual species such that they had unimodal (i.e. quadratic) responses to PC1 and PC3. Ostensibly, a model would need to be able to recreate a unimodal response to accurately recreate these trends. In our data there are sites in environmental space with values of PC1 on either side of each simulated species’ niche (Fig. 1). Hence, to recreate a response to PC1, a model needs to have the capacity to create a unimodal response. ESM-L, for example, were constructed only with linear terms (Breiner et al. 2015), so they were poorly calibrated against PC1 (Fig. 4a). In contrast, even though the true response to PC3 is also unimodal, the niches (and indeed, those of most of the real species) lie very close to the edge of available environmental space. Hence, a model needs only utilize a linear term to recreate most of the response to PC3. Thus, ESM-L was well calibrated against PC3 even though the model terms are not quadratic. In summary, model calibration accuracy is affected not only by the underlying functional form of the species’ response to its environment and model flexibility (Merow et al. 2014), but also by the position of niches in environmental space.

## Sample size

Sample size strongly influences the bias–variance tradeoff (Hastie et al. 2009), wherein models with many parameters run the risk of overfitting the data (high bias, with poor out-of-sample accuracy), while models with few are more prone to underfit the data (high variance, with increased sensitivity of coefficients to small changes in data; Hastie et al. 2009, James et al. 2013). Using single-species frequentist-based modeling like GLMs, only two parameters (without variance) can be estimated from only two presences, resulting in both overfitting and high bias. MSDMs and ESMs attempt to address this issue in different ways. In multi-species frameworks, borrowing of strength can make the focal species’

sample size less relevant. Indeed, for some kinds of multi-species models, it is possible for the number of parameters being estimated to exceed the number of data points for a given species (Gelfand and Sahu 1999, Ovaskainen and Soininen 2011). Likewise, ESMs attempt to overcome the problem of small samples by averaging across simple models, which are less likely to be overfit (Lomba et al. 2010). Nonetheless, we found that GLMs did as well as any other algorithm when sample size was sufficient to implement them (although GLMs also often incorrectly identified PC2 as influential). Further improvements in modeling rare species could be gained by using informative priors (Gelfand and Sahu 1999), perhaps based on experiments or information from closely related species (Qiao et al. 2017, Smith et al. 2019, van der Veen et al. 2021).

## Niche breadth and similarity to other species

We posit that most ‘rare’ species of conservation interest are likely either inhabitants of uncommon environments (marginal specialists; Gregory and Gaston 2000, Albrecht and Long 2019, Díaz et al. 2020, but see Morueta-Holme et al. 2013) or were once common but have suffered from reduction due to anthropogenic factors (central generalists, Channell and Lomolino 2000). If this is the case, our results are widely germane to species of conservation concern. We found that species with narrow niches (especially those near the margin of available environmental space) yielded the most accurate models (Fig. 3, 4). Hence, application of methods explored here can assist in highlighting promising areas to search for or establish new populations, or for estimating vulnerability to climate change. In contrast, generalists that have suffered recent decline will be challenging to model, especially as their current distributions may underrepresent their environmental tolerances (Faurby and Araújo 2018). Species with marginal niches were also easier to model than their central counterparts. However, species with niches on the periphery of available environmental space may also be prone to having their fundamental niche truncated (Fig. 1). In these cases, extrapolation of models ‘beyond’ the region of truncation will be risky (Feeley and Silman 2010, Anderson 2013).

## Borrowing of strength versus being misled by the crowd

The two MSDMs in our analysis excelled at very different tasks (Fig. 5; Norberg et al. 2019). These differences likely arise from how the models approach estimation (i.e. HMSC: Bayesian with shared priors; SAMs: mixture models). Rare species modeled with HMSC do seem to benefit from borrowing of strength for calibration tasks when their sample size is adequate. In contrast, SAMs only excelled at correctly identifying the lack of response to PC2, the uninfluent predictor. However, the success of SAMs in this situation may arise from roughly equal assignment of the rare species to each of the six archetypes, resulting in an averaged ‘flat’ response. Hence, small sample size may dilute signal enough

that borrowing of strength can actually mislead the model. Borrowing of strength alone is not sufficient to ensure success in modeling rare species.

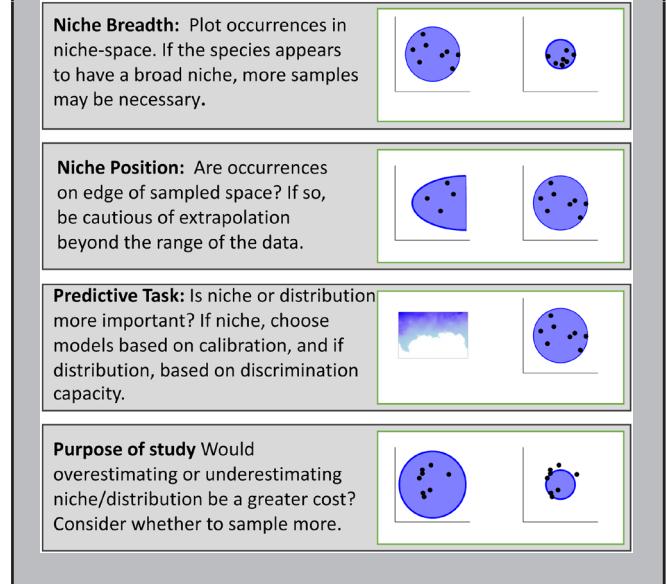
## Ways forward

Our work points the way toward further potential improvements for modeling rare species. We explored just two versions of multi-species models (Ovaskainen and Abrego 2020), various forms of which can account for the underlying spatial structure of the sampling locations (Ver Hoef et al. 2018), phylogenetic relationships and traits (Ovaskainen et al. 2017, Morales-Castilla et al. 2017, Smith et al. 2019, Tikhonov et al. 2020), and species-by-species associations (Pollock et al. 2014, Ovaskainen et al. 2016). However, including additional data and model features can be demanding of field and computational time. In our case, for example, preliminary test runs using HMSC with latent factors for spatial autocorrelation required approximately eight days of run time per model for a single data instance, so we opted not to use them in our study (cf. Velásquez-Tibatá et al. 2016). However, it is possible that model performance (and run time) could be improved by using a subset of the total community with species that share functional traits or habitats (Pacifici et al. 2014). We also created our virtual species such that it did not interact with other species. Some algorithms, like HMSC, can estimate positive and negative interspecific associations independently of shared environmental preferences (Ovaskainen et al. 2016). It is possible that a model accounting for associations (Wilkinson et al. 2019) would perform better than single-species models in these circumstances. We also used a presence–absence dataset, assuming perfect detection with no sampling bias, and that all species only responded to two variables. As a result of these considerations, we do not interpret our apparent success at extremely low sample size to be indicative of what can be achieved using real-world data. Rather, the minimum sample size necessary for generation of robust models using real-world data will surely be much higher than those explored here.

## Conclusions

In sum, the best modeling algorithm for a given task (discrimination versus calibration) depends on the species' niche breadth, similarity to other species' niches, and sample size (Fig. 5). Neither borrowing of strength inherent in multi-species modeling approaches nor ensembles of small models is a panacea to the problem of modeling extremely rare species: sometimes simpler methods are best (Jeliazkov et al. 2022). Before deciding on the appropriate modeling approach, researchers should consider what they already know about their species of interest (i.e. niche characteristics) and the goal of their modeling endeavor (Box 1; Jeliazkov et al. 2022). Plotting presences in environmental space can help determine whether a species will be easier to model (a specialist) or more challenging (a generalist). Additionally, comparing the

### Box 1. Considerations for modeling rare species.



location of the presences in niche space to the accessible niche space will help identify whether the niche is marginal to available environmental space or more interior. Taken together, these considerations can help identify which modeling algorithms are fit for the desired task.

**Acknowledgements** – We would like to thank the Global Change and Conservation lab group at the Missouri Botanical Garden, Saint Louis, MO for helpful comments on the manuscript.

**Funding** – This project was made possible in part by US Geological Survey grant no. G21AC10157-00 (FAIN G21AC10157) and the Alan Graham Fund in Global Change.

## Author contributions

**Kelley D. Erickson:** Conceptualization (equal); Formal analysis (lead); Methodology (equal); Software (lead); Visualization (lead); Writing – original draft (lead). **Adam B. Smith:** Conceptualization (equal); Methodology (equal); Supervision (lead); Writing – review and editing (equal).

## Transparent peer review

The peer review history for this article is available at <https://publons.com/publon/10.1111/ecog.06500>.

## Data availability statement

Code for running all analyses is available at [https://github.com/kerickson22/SDMs\\_for\\_rare\\_species\\_modeling](https://github.com/kerickson22/SDMs_for_rare_species_modeling), and data for the real species from <https://zenodo.org/record/2637812> (Norberg et al. 2019).

## Supporting information

The Supporting information associated with this article is available with the online version.

## References

- Albrecht, M. A. and Long, Q. G. 2019. Habitat suitability and herbivores determine reintroduction success of an endangered legume. – *Plant Divers.* 41: 109–117.
- Anderson, R. P. 2013. A framework for using niche models to estimate impacts of climate change on species distributions. – *Ann. N. Y. Acad. Sci.* 1297: 8–28.
- Breiner, F. T., Guisan, A., Bergamini, A. and Nobis, M. P. 2015. Overcoming limitations of modeling rare species by using ensembles of small models. – *Methods Ecol. Evol.* 6: 1210–1218.
- Brun, P., Thuiller, W., Chauvier, Y., Pellissier, L., Wiest, R. O., Wang, Z. and Zimmermann, N. E. 2020. Model complexity affects species distribution projections under climate change. – *J. Biogeogr.* 47: 130–142.
- Butler, L. and Sanderson, R. A. 2022. National-scale predictions of plant assemblages via community distribution models: leveraging published data to guide future surveys. – *J. Appl. Ecol.* 59: 1559–1571.
- Carroll, C., Dunk, J. R. and Moilanen, A. 2010. Optimizing resiliency of reserve networks to climate change: multispecies conservation planning in the Pacific Northwest, USA. – *Global Change Biol.* 16: 891–904.
- Channell, R. and Lomolino, M. V. 2000. Trajectories to extinction: spatial dynamics of the contraction of geographical ranges. – *J. Biogeogr.* 27: 169–179.
- Clark, J. S., Gelfand, A. E., Woodall, C. W. and Zhu, K. 2014. More than the sum of the parts: forest climate response from joint species distribution models. – *Ecol. Appl.* 24: 990–999.
- Connor, T., Hull, V., Viña, A., Shortridge, A., Tang, Y., Zhang, J., Wang, F. and Liu, J. 2018. Effects of grain size and niche breadth on species distribution modeling. – *Ecography* 41: 1270–1282.
- D'Amen, M., Rahbek, C., Zimmermann, N. E. and Guisan, A. 2017. Spatial predictions at the community level: from current approaches to future frameworks: methods for community-level spatial predictions. – *Biol. Rev.* 92: 169–187.
- Díaz, D. M. V., Blundo, C., Cayola, L., Fuentes, A. F., Malizia, L. R. and Myers, J. A. 2020. Untangling the importance of niche breadth and niche position as drivers of tree species abundance and occupancy across biogeographic regions. – *Global Ecol. Biogeogr.* 29: 1542–1553.
- Elith, J. and Leathwick, J. R. 2009. Species distribution models: ecological explanation and prediction across space and time. – *Annu. Rev. Ecol. Evol. Syst.* 40: 677–697.
- Faúby, S. and Araújo, M. B. 2018. Anthropogenic range contractions bias species climate change forecasts. – *Nat. Clim. Change* 8: 252–256.
- Feeley, K. J. and Silman, M. R. 2010. Biotic attrition from tropical forests correcting for truncated temperature niches. – *Global Change Biol.* 16: 1830–1836.
- Gelfand, A. E. and Sahu, S. K. 1999. Identifiability, improper priors and Gibbs sampling for generalized linear models. – *J. Am. Stat. Assoc.* 94: 247–253.
- Gelman, A. and Rubin, D. B. 1992. Inference from iterative simulation using multiple sequences. – *Stat. Sci.* 7: 457–511.
- Gregory, R. D. and Gaston, K. J. 2000. Explanations of commonness and rarity in British breeding birds: separating resource use and resource availability. – *Oikos* 88: 515–526.
- Guillera-Arroita, G., Lahoz-Monfort, J. J., Elith, J., Gordon, A., Kujala, H., Lentini, P. E., McCarthy, M. A., Tingley, R. and Wintle, B. A. 2015. Is my species distribution model fit for purpose? Matching data and models to applications: matching distribution models to applications. – *Global Ecol. Biogeogr.* 24: 276–292.
- Guisan, A., Broennimann, O., Engler, R., Vust, M., Yoccoz, N. G., Lehman, A. and Zimmermann, N. E. 2006. Using niche-based models to improve the sampling of rare species. – *Conserv. Biol.* 20: 501–511.
- Guisan, A. et al. 2013. Predicting species distributions for conservation decisions. – *Ecol. Lett.* 16: 1424–1435.
- Hastie, T. R., Tibshirani, R. and Friedman, J. 2009. The elements of statistical learning: data mining, inference and prediction. – Springer.
- Hernandez, P. A., Graham, C. H., Master, L. L. and Albert, D. L. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. – *Ecography* 29: 773–785.
- Hui, F. K. C., Warton, D. I., Foster, S. D. and Dunstan, P. K. 2013. To mix or not to mix: comparing the predictive performance of mixture models vs separate species distribution models. – *Ecology* 94: 1913–1919.
- James, G., Witten, D., Hastie, T. and Tibshirani, R. 2013. An introduction to statistical learning with applications in R. – Springer.
- Jeliazkov, A., Gavish, Y., Marsh, C. J., Geschke, J., Brummit, N., Rocchini, D., Haase, P., Kunin, W. E. and Henle, K. 2022. Sampling and modeling rare species: conceptual guidelines for the neglected majority. – *Global Change Biol.* 28: 3754–3777.
- Jiménez-Valverde, A., Acevedo, P., Barbosa, A. M., Lobo, J. M. and Real, R. 2013. Discrimination capacity in species distribution models depends on the representativeness of the environmental domain: discrimination is context dependent. – *Global Ecol. Biogeogr.* 22: 508–516.
- Le Lay, G., Franc, E. and Guisan, A. 2010. Prospective sampling based on model ensembles improves the detection of rare species. – *Ecography* 33: 1015–1027.
- Lomba, A., Pellissier, L., Randin, C., Vicente, J., Horondo, J. and Guisan, A. 2010. Overcoming the rare species modeling paradox: a novel hierarchical framework applied to an Iberian endemic plant. – *Biol. Conserv.* 143: 2647–2657.
- Mason, S. J. and Graham, N. E. 2002. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: statistical significance and interpretation. – *Q. J. R. Meteorol. Soc.* 128: 2145–2166.
- Merow, C., Smith, M. J., Edwards Jr, T. C., Guisan, A., McMahon, S. M., Normand, S., Thuiller, W., Wüest, R. O., Zimmermann, N. E. and Elith, J. 2014. What do we gain from simplicity versus complexity in species distribution models? – *Ecography* 37: 1267–1281.
- Meynard, C. N. and Kaplan, D. M. 2013. Using virtual species to study species distributions and model performance. – *J. Biogeogr.* 40: 1–8.
- Meynard, C. N., Leroy, B. and Kaplan, D. M. 2019. Testing methods in species distribution modeling using virtual species: what have we learnt and what are we missing? – *Ecography* 42: 2021–2036.
- Morales-Castilla, I., Davies, T. J., Pearse, W. D. and Peres-Neto, P. 2017. Combining phylogeny and co-occurrence to improve single species distribution models. – *Global Ecol. Biogeogr.* 26: 740–752.
- Morueta-Holme, N. et al. 2013. Habitat area and climate stability determine geographical variation in plant species range sizes. – *Ecol. Lett.* 16: 1446–1454.

- Norberg, A. et al. 2019. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. – *Ecol. Monogr.* 89: 1–24.
- Ovaskainen, O. and Soininen, J. 2011. Making more out of sparse data: hierarchical modeling of species communities. – *Ecology* 92: 289–295.
- Ovaskainen, O. and Abrego, N. 2020. Joint species distribution modeling with applications in R. – Cambridge Univ. Press.
- Ovaskainen, O., Hottola, J. and Siitonen, J. 2010. Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. – *Ecology* 91: 2514–2521.
- Ovaskainen, O., Roy, D. R., Fox, R. and Anderson, B. J. 2016. Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. – *Methods Ecol. Evol.* 7: 428–436.
- Ovaskainen, O., Tikhonov, G., Norberg, A., Blanchet, F. G., Duan, L., Dunson, D., Roslin, T. and Abrego, N. 2017. How to make more out of community data? A conceptual framework and its implementation as models and software. – *Ecol. Lett.* 20: 561–576.
- Pacifici, K., Zipkin, E. F., Collazo, J. A., Irizzary, J. I. and DeWan, A. 2014. Guidelines for a priori grouping of species in hierarchical community models. – *Ecol. Evol.* 4: 877–888.
- Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M. and Araújo, M. B. 2011. Ecological niches and geographic distributions. – Princeton Univ. Press.
- Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., Veski, P. A. and McCarthy, M. A. 2014. Understanding co-occurrence by modeling species simultaneously with a joint species distribution model (JSDM). – *Methods Ecol. Evol.* 5: 397–406.
- Qiao, H., Peterson, A. T., Ji, L. and Hu, J. 2017. Using data from related species to overcome spatial sampling bias and associated limitations in ecological niche modeling. – *Methods Ecol. Evol.* 8: 1804–1812.
- Rabinowitz, D. 1981. Seven forms of rarity. – In: Synge, H. (ed.), *The biological aspects of rare plant conservation*. Wiley, pp. 205–217.
- Santika, T. 2011. Assessing the effect of prevalence on the predictive performance of species distribution models using simulated data. – *Global Ecol. Biogeogr.* 20: 181–192.
- Santini, L., Benítez-López, A., Maiorano, L., Čengič, M. and Huijbregts, M. A. 2021. Assessing the reliability of species distribution projections in climate change research. – *Divers. Distrib.* 27: 1035–1050.
- Smith, A. 2022. enmSdmX: Tools for modeling niches and distributions of species, ver. 1.0.0. – <https://cran.r-project.org/package=enmSdmX>.
- Smith, A. B., Santos, M. J., Koo, M. S., Rowe, K. C., Rowe, K. M. C., Patton, J. L., Beissinger, S. and Moritz, C. 2013. Evaluation of species distribution models by resampling of sites surveyed a century ago by Joseph Grinnell. – *Ecography* 36: 1017–1031.
- Smith, A. B., Godsoe, W., Rodríguez-Sánchez, F., Wang, H-H. and Warren, D. 2019. Niche estimation above and below the species level. – *Trends Ecol. Evol.* 34: 260–273.
- Soultan, A. and Safi, K. 2017. The interplay of various sources of noise on reliability of species distribution models hinges on ecological specialization. – *PLoS One* 12: e0187906.
- Stockwell, D. R. B. and Peterson, A. T. 2002. Effects of sample size on accuracy of species distribution models. – *Ecol. Model.* 148: 1–13.
- Thibaud, E., Petitpierre, B., Broennimann, O., Davison, A. C. and Guisan, A. 2014. Measuring the relative effect of factors affecting species distribution model predictions. – *Methods Ecol. Evol.* 5: 947–955.
- Thuiller, B., Lavorel, S., Sykes, M. T. and Araújo, M. B. 2006. Using niche-based modeling to assess the impact of climate change on tree functional diversity in Europe. – *Divers. Distrib.* 12: 49–60.
- Tikhonov, G., Opedal, O. H., Abrego, N., Lehikoinen, A., de Jonge, M. M. J., Oksanen, J. and Ovaskainen, O. 2020. Joint species distribution modeling with the R-package Hmsc. – *Methods Ecol. Evol.* 11: 442–447.
- Tikhonov, G., Ovaskainen, O., Oksanen, J., de Jonge, M., Opedel, O. and Dallas, T. 2021. Hmsc: hierarchical model of species communities. – R package ver. 3.0-11. <https://CRAN.R-project.org/package=Hmsc>.
- Tjur, T. 2009. Coefficients of determination in logistic regression models – a new proposal: the coefficient of discrimination. – *Am. Stat.* 63: 366–372.
- van der Veen, B., Hui, F. K. C., Hovstad, K. A., Solbu, E. B. and O'Hara, R. B. 2021. Model-based ordination for species with unequal niche widths. – *Methods Ecol. Evol.* 12: 1288–1300.
- van Proosdij, A. S. J., Sosef, M. S. M., Wieringa, J. J. and Raes, N. 2016. Minimum required number of specimen records to develop accurate species distribution models. – *Ecography* 39: 542–552.
- Velásquez-Tibatá, J., Graham, C. G. and Munch, S. B. 2016. Using measurement error models to account for georeferencing error in species distribution models. – *Ecography* 39: 305–316.
- Ver Hoef, J. M., Peterson, E. E., Hooten, M. B., Hanks, E. M. and Fortin, M-J. 2018. Spatial autoregressive models for statistical inference from ecological data. – *Ecol. Monogr.* 88: 36–59.
- White, J. W., Rassweiler, A., Samhouri, J. F., Stier, A. C. and White, C. 2014. Ecologists should not use statistical significance tests to interpret simulation model results. – *Oikos* 123: 385–388.
- Wilkinson, D. P., Golding, N., Guillera-Arroita, G., Tingley, R. and McCarthy, M. A. 2019. A comparison of joint species distribution models for presence-absence data. – *Methods Ecol. Evol.* 10: 198–211.
- Winship, C. and Western, B. 2016. Multicollinearity and model misspecification. – *Sociol. Sci.* 3: 627–649.
- Wisz, M. S., Hijmans, R. J., Peterson, A. T., Graham, C. H., Guisan, A. and NCEAS Predicting Species Distributions Working Group. 2008. Effects of sample size on the performance of species distribution models. – *Divers. Distrib.* 14: 763–773.
- Wooley, S., Foster, S. and Dunstan, P. 2022. ecomix: Fitting finite mixture models to ecological data. – R package ver. 1.0.0. <https://github.com/skipontiam/ecomix>.
- Yoccoz, N. G. 2022. From design to analysis: a roadmap for predicting distributions of rare species. – *Global Change Biol.* 28: 3745–3747.
- Zhang, C., Chen, Y., Xu, B., Xue, Y. and Ren, Y. 2020. Improving prediction of rare species' distribution from community data. – *Sci. Rep.* 10: 12230.
- Zipkin, E. F., DeWan, A. and Royle, J. A. 2009. Impacts of forest fragmentation on species richness: a hierarchical approach to community modelling. – *J. Appl. Ecol.* 46: 815–822.
- Zizka, A., ter Steege, H., Pessoa, M. de C. R. and Antonello, A. 2018. Finding needles in the haystack: where to look for rare species in the American tropics. – *Ecography* 41: 321–330.