Received: 10 February 2023 | Revised: 1 July 2023 | Accepted: 21 August 2023

DOI: 10.1002/ecs2.4703

ECOSPHERE
AN ESA OPEN ACCESS JOURNAL

ARTICLE
Methods, Tools, and Technologies

Thinning occurrence points does not improve species distribution model performance

Cleber Ten Caten | **Tad Dallas**

Department of Biological Sciences, University of South Carolina, Columbia, South Carolina, USA

Correspondence
Cleber Ten Caten
Email: clebertencaten@gmail.com

Funding information
National Science Foundation, Grant/Award Number: NSF-DEB-2017826

Handling Editor: Manuel Lerdau

Abstract

Spatial biases are an intrinsic feature of occurrence data used in species distribution models (SDMs). Thinning species occurrences, where records close in the geographic or environmental space are removed from the modeling procedure, is an approach often used to address these biases. However, thinning occurrence data can also negatively affect SDM performance, given that the benefits of removing spatial biases might be outweighed by the detrimental effects of data loss caused by this approach. We used real and virtual species to evaluate how spatial and environmental thinning affected different performance metrics of four SDM methods. The occurrence data of virtual species were sampled randomly, evenly spaced, and clustered in the geographic space to simulate different types of spatial biases, and several spatial and environmental thinning distances were used to thin the occurrence data. Null datasets were also generated for each thinning distance where we randomly removed the same number of occurrences by a thinning distance and compared the results of the thinned and null datasets. We found that spatially or environmentally thinned occurrence data is no better than randomly removing them, given that thinned datasets performed similarly to null datasets. Specifically, spatial and environmental thinning led to a general decrease in model performances across all SDM methods. These results were observed for real and virtual species, were positively associated with thinning distance, and were consistent across the different types of spatial biases. Our results suggest that thinning occurrence data usually fails to improve SDM performance and that the use of thinning approaches when modeling species distributions should be considered carefully.

KEYWORDS
environmental thinning, model performance, spatial bias, spatial thinning, species distribution models

INTRODUCTION

A species geographic distribution expresses its ecology and evolutionary history (Brown, 1995; Gaston, 2003) where abiotic conditions, biotic factors, and dispersal ability determine the areas a species can occupy (Soberón & Peterson, 2005). Having a refined knowledge of species distributions is essential as it is a key variable

for assessing biogeographical and macroecological patterns (Herkt et al., 2017; Hortal et al., 2015) and to describe the conservation status and extinction risk of a species (Cardillo et al., 2008; Lee & Jetz, 2011). However, currently there is an incomplete knowledge of the distribution of several species, a shortcoming called the *Wallacean shortfall* (Hortal et al., 2015; Lomolino, 2004; Whittaker et al., 2005). The ongoing development of species occurrence databases and species distribution models (SDMs) has improved the ability to estimate species geographic ranges, partially addressing this Wallacean shortfall (Jetz et al., 2012; Peterson et al., 2011; Terribile et al., 2018). Nevertheless, using SDMs to estimate species geographic distribution might be challenging as several factors, such as the quality of the data available, affect the performance and predictive ability of these models (Santini et al., 2020).

The Wallacean shortfall is partially driven by spatial biases present during the sampling process such as sampling locations based on proximity to universities and local accessibility (Moerman & Estabrook, 2006; Oliveira et al., 2016; Sousa-Baena et al., 2014; Vale & Jenkins, 2012). This leads species occurrence points (i.e., localities where the species was recorded) to be sampled in a nonrandom subset of areas from which the species could occupy. Consequently, sampled occurrence points tend to be spatially clustered, a problem that is often observed in online databases (Beck et al., 2014; Inman et al., 2021). These clustered points have the potential to disproportionately represent the environmental conditions of the most sampled regions rather than representing the set of suitable environmental conditions for a species (Anderson & Gonzalez, 2011; Kadmon et al., 2004), which can affect SDM performance (Beck et al., 2014; Veloz, 2009). This occurs because the occurrence points used for training and testing the SDMs might be spatially adjacent, which can lead to higher (inflated) model performances than expected (Bahn & McGill, 2013; Radosavljevic & Anderson, 2014).

Although spatial block validation can be used to obtain non-inflated SDM performances (Bahn & McGill, 2013; Radosavljevic & Anderson, 2014; Roberts et al., 2017), it does not remove the bias present in species occurrence data. Background manipulation methods can also be used to address bias in occurrence data, but this approach requires knowledge of the sampling effort for a species, which is rarely available, or a large observation dataset to estimate the biased sampling effort (Inman et al., 2021; Phillips et al., 2009; Ranc et al., 2017). Alternatively, thinning approaches are perhaps the most commonly used methods that have been developed to remove spatial biases found in occurrence data (Aiello-Lammens et al., 2015; Anderson & Raza, 2010; Hidalgo-Mihart et al., 2004; Varela et al., 2014; Veloz, 2009). Occurrence points can be thinned based on the geographical (Aiello-Lammens et al., 2015; Veloz, 2009) or environmental (Varela et al., 2014) distances between them such that only the most spatially or environmentally unique occurrences are kept and used in the modeling process. The effects of spatial thinning on model performance are unclear (Beck et al., 2014; Boria et al., 2014; Castellanos et al., 2019; Varela et al., 2014) and dependent on the ecological characteristic of the species (Baker et al., 2022; Steen et al., 2021). For example, spatial thinning might not be appropriate for species that have spatially clustered occurrences because it can lead to an undesirable loss of data that negatively affects model performance (Varela et al., 2014). Consequently, environmental thinning has been suggested to be a superior alternative in these situations, given that occurrence points are filtered in the environmental space in this approach, which can reduce the data loss experienced by species with spatially clustered occurrences (Castellanos et al., 2019; Varela et al., 2014). However, regardless of the type of thinning that is used, a challenge during the thinning process is choosing the distance to filter the dataset (Castellanos et al., 2019) as using larger thinning distances will inherently remove more occurrence points from the initial dataset. Given that the number of occurrence points has a larger impact on model performance than spatial bias (Gaul et al., 2020), the potential benefits of thinning might be lost if considerably fewer occurrence points are left for the modeling procedure (Steen et al., 2021).

Here, we examine the effects of spatial and environmental thinning occurrence points on SDM performance for real and virtual species (Figure 1). For virtual species, we simulated species that had occurrence data sampled randomly, clustered, and evenly spaced (i.e., a similar distance between the sampled occurrences) in the geographic space. Our goal was to evaluate whether the effectiveness of thinning is dependent on the types of spatial bias observed in occurrence points. We thinned occurrence points considering different thinning distances and used four modeling methods to model the species distributions and four performance metrics to evaluate the models. We compared the results of these models with those of null models where we randomly removed species occurrence points from the modeling procedure. Spatial and environmental thinning decreased model performance for real and virtual species and performed no better than null models, suggesting that thinning approaches have limited benefits for SDMs.

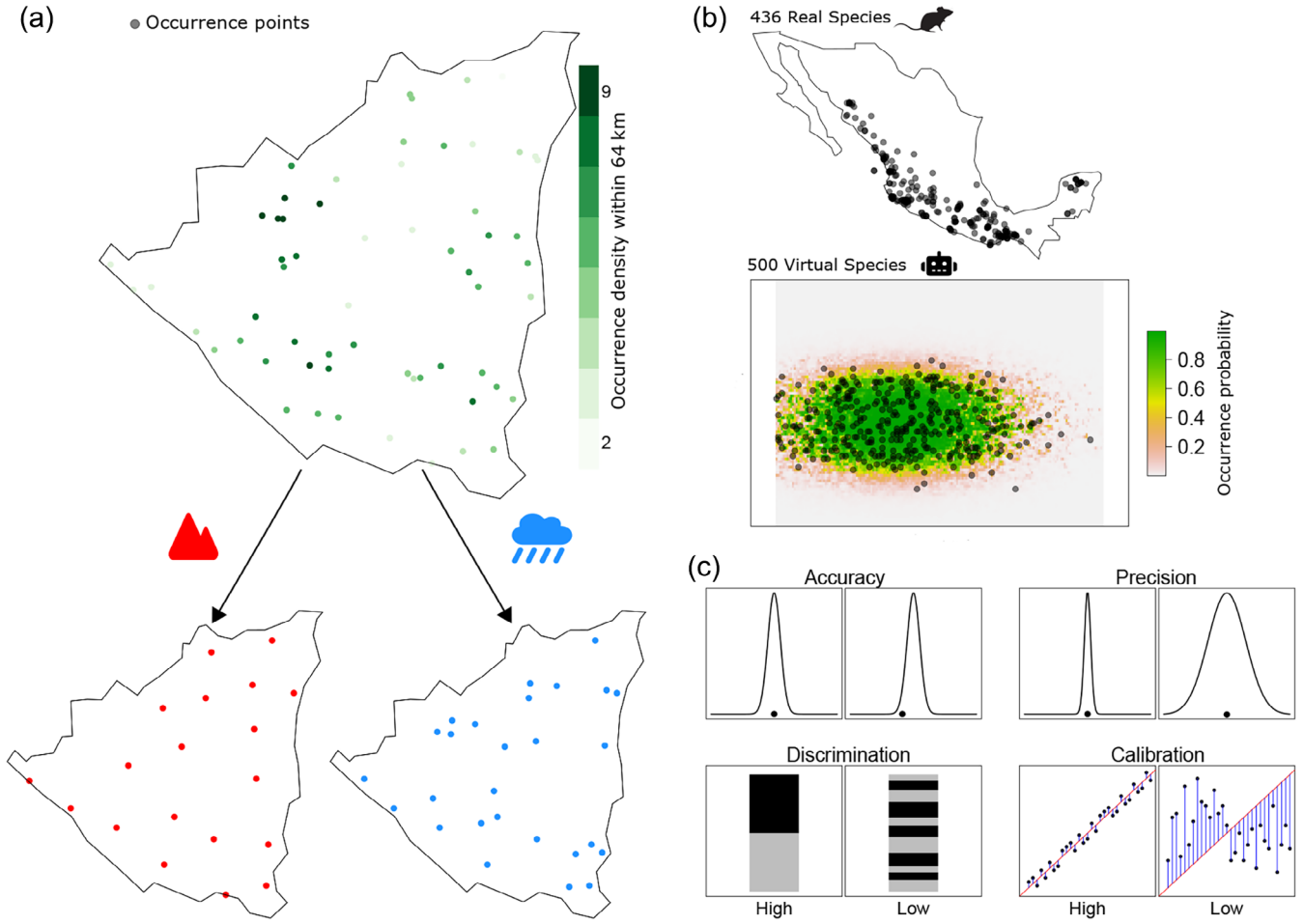

FIGURE 1 An example of how spatial (red mountain) and environmental (blue cloud) thinning can lead to a selection of different sets of occurrence points that are used to model the species distributions. We obtained data for 436 real species and we simulated 500 virtual species with different types of spatial biases to assess how thinning affects alternative metrics of model performance. For all model performance measures, the right panel indicates poor performance and the left panel indicates good performance. Models with high accuracy will have their predicted values close to the observed values; models with high precision will have their predicted values close to each other (i.e., not spread); models with high discrimination will be able to efficiently separate occurrences (black bars) from pseudo-absences (gray bars); and models with high calibration will have the predicted suitability of a bin close to the observed suitability of that bin.

METHODS

Simulating landscapes

We simulated 500 landscapes where 500 virtual species would occur. The virtual landscapes were modeled as a raster with 90 × 180 cells where two gradients, one horizontal and another vertical, shaped the species virtual environment (Figure 1). The gradients were modeled following a Gaussian distribution (Dallas & Santini, 2020) where the environmental values present in the landscape ranged from 0 to 2.5. A Gaussian random field was added to the modeled landscapes in order to simulate a more realistic spatially autocorrelated environment. We randomly selected the strength of the Gaussian random field (ranging from 0.1 to 1.5) to make each virtual landscape a unique set of environmental conditions for a species. The strength of the Gaussian random field added to the landscapes is positively related to how heterogenous the environmental conditions are in the landscape (Appendix S1: Figure S1). The maximum strength of 1.5 in the Gaussian random field was chosen because, at this level, the landscape is highly heterogenous, but it still maintains its spatially autocorrelated nature (see Appendix S1 for further information).

Simulating species distributions

Species environmental niches, and distributions within these virtual landscapes, were simulated with the `virtualspecies` R package (Leroy et al., 2016).

Species niches were modeled following a Gaussian response to the virtual environment as it has been shown that many species exhibit this response to environmental and ecological gradients (Boucher-Lalonde et al., 2014; Oksanen & Minchin, 2002), and it is an assumption that studies using virtual species usually adopt (van Proosdij et al., 2016; Varela et al., 2014). The optimum environmental value for each species was randomly chosen, and it varied from 1 to 2.2 in each environmental layer with a SD around it ranging from 0.05 to 0.6. We chose this wide range of optimum values (1–2.2) for the two environmental gradients because our goal was to simulate species that are ecologically different regarding the conditions required for their occurrence. Thus, these optimum values allowed us to simulate species that are capable of occupying all the parts of our virtual landscape. The SD around the optimum value was chosen to range between 0.05 and 0.6 because this allowed us to simulate specialist and generalist species when small or large SDs were selected, respectively. We selected the maximum SD of 0.6 because values above that would make the species unrealistically generalist where it would be able to occupy nearly the entire landscape. A logistic transformation of the species environmental suitability defined the chance of a species occurrence being sampled in a particular cell of the landscape.

We randomly sampled 6–640 occurrence points to model the distribution of each virtual species. This covers a realistic range of sample sizes that are commonly found when modeling species distributions as rare species often have few sampled occurrence points whereas other species might be better sampled and have more occurrence points available for modeling. We chose a maximum of 640 sampled occurrence points because models perform well with this sample size and performance would likely remain the same if larger sample sizes were used. Three sets of occurrence points were obtained for each species following three types of spatial biases (i.e., random, clustered, and evenly spaced).

Clustering and dispersion of sampled occurrences

First, we simulated an instance where species occurrence points are sampled randomly across the geographic space. In this case, there is no intrinsic sampling bias driving the sampling process. In the second case, we simulated a situation where there is spatial bias in the sampling of occurrence points, and the occurrence points sampled are spatially clustered in the geographic space. Such situation might occur when accessible locations are sampled more often (Botts et al., 2011; Mair & Ruete, 2016). To simulate this situation, we first randomly selected five initial points that represented five different clusters of occurrence data sampled. Next, we used the distm function from the `geosphere` package (Hijmans, 2019) and calculated the distance between the initial points and all other occurrence points. The 50 closest points to each cluster were selected, and a point was sampled from it, with closer points having greater probabilities of being sampled. This process was repeated for each cluster of occurrence points until we reached the desired number of occurrences for each species. In the last case, we simulated a case where the sampled points are evenly spaced from each other. To achieve this goal, we created a grid that covered the entire landscape where the number of cells in the grid would be equal to the number of occurrences desired to be sampled. Occurrence points were sampled from this grid such that points that were closer to the centroid of each cell in the grid had a greater probability of being sampled. All of the data manipulation in this step was done using the `sf` package (Pebesma, 2018).

Empirical species distributions

To model the distribution of real species, we sampled 500 mammal species from the Global Biodiversity Information Facility (GBIF) using the `rgbif` package (Chamberlain et al., 2021). We removed 61 species that had less than five sampled occurrence points from the modeling procedure because it is challenging to reliably model the distribution of such species, and we also removed 3 species that do not occur in the Americas. We obtained the 19 bioclimatic variables available in the BioClim database (Hijmans et al., 2005) to model the species distribution. These variables represent different facets of temperature and precipitation patterns that have been recorded from weather stations across the globe between 1960 and 1990 (Hijmans et al., 2005). The bioclimatic variables were obtained at a resolution of 10 arcminutes (i.e., ≈18 km^2) covering the Americas. A principal components analysis (PCA) was performed on these variables, and the first five axes explained approximately 95% of the variance of the data and were used to model the species distribution.

Spatial and environmental thinning

Spatial thinning for real species was done using the thin function from the `spThin` package (Aiello-Lammens et al., 2015). In this function, a pairwise distance matrix between all occurrence points is calculated based on a chosen distance, and a neighbor point of the observation with the most neighbors is removed. This process is repeated until there are no more points with neighbors

based on the chosen distance. For each species, we thinned occurrence points with a distance of 16, 32, 64, and 128 km. These values were chosen because our environmental values were obtained at the spatial resolution of 16 km^2. For the virtual species, we spatially thinned the occurrence points following a similar framework. Virtual species only have one occurrence point per cell (i.e., the centroid of the cell). We used the select.window function from the `CommEcol` package (Melo, 2019) and removed neighbor points to the focal occurrence points based on a chosen distance. We chose the distances of 2, 4, 8, and 16. In the distance of 2, all eight neighbor cells to a focal point are selected and their points are removed. Similarly, with a distance of 4, 16 of neighbor cells are selected and their points are removed; with a distance of 8, 32 neighbor cells are selected and their points are removed; with a distance of 16, 64 neighbor cells are selected and their points are removed. Thus, thinning distance doubles for every increase in the distance used for virtual species, which is analogous to how we thinned the occurrence points of real species.

Environmental thinning virtual and real species occurrence points was done using the method developed by Varela et al. (2014). In this framework, bins are created based on a chosen environmental distance and then only one point that falls within an environmental bin is selected for the modeling. Thus, more points are discarded when larger environmental distances are used. For real species, environmental thinning was based on the two PCA axes used to model their distribution, while for virtual species, it was based on the two virtual climatic conditions. The environmental intervals we used for both real and virtual species were 0.05, 0.1, 0.2, and 0.4. We used the envSample function to environmentally thin our data (Varela et al., 2014).

Comparing thinned and null models

Since thinning occurrence points decreases the number of points available to model the species distribution and this alone can negatively affect SDM performance (Loiselle et al., 2008; Tessarolo et al., 2014), we also generated null datasets for all spatial and environmental thinning distances we used. In the null dataset, we randomly removed the same number of occurrence points that were removed by a specific spatial or environmental thinning distance for a species. Thus, a similar performance between thinned and null datasets would suggest that the benefits of removing spatial bias through thinning approaches are countered by the loss of occurrence points that are important to model the species distribution.

Modeling species distributions

We evaluated the effects that spatial and environmental thinning have on the performance of MaxEnt, support vector machine (SVM), generalized linear models (GLMs), and generalized additive models (GAMs) modeling methods. We chose these modeling methods because they are commonly used methods that utilize presence/pseudo-absence data. Specifically, MaxEnt is a widely used machine learning method that generally performs well (Elith et al., 2010; Santini et al., 2020). SVM is also a machine learning method that performs well for species with limited occurrence data (Drake et al., 2006; Schölkopf et al., 2001; Tax & Duin, 2004). GLM is a simple regression-based modeling method (Loyola, 2012; Nelder & Wedderburn, 1972), and GAM is an extension of GLM that allows some predictors to be modeled non-parametrically (Guisan et al., 2002). MaxEnt models were run with the maxent function from the R `dismo` package (Hijmans et al., 2017); SVM models were run with the svm function from the R `kernlab` package (Karatzoglou et al., 2019); GLM models were run using the glm function from the R `stats` package (R Core Team, 2018); and GAM models were run using the gam function from the R `mgcv` package (Wood, 2017).

As class imbalances in the number of occurrence/pseudo-absence points can affect model intercomparisons (Liu et al., 2005; McPherson et al., 2004), we kept prevalence at 0.5, where 50% of the data are occurrences and 50% are pseudo-absences (Iturbide et al., 2015; Senay et al., 2013). Pseudo-absences were randomly sampled for virtual and real species, sampling from a 200-km circular buffer for real species and from the entire landscape for virtual species. Next, we applied a cross-validation procedure where we randomly split our occurrence data into two datasets: 75% of the data were used to train the models and the other 25% of the data were used to test the models, repeating the sampling process 20 times. Since we only have one dataset for the real species, we used that same dataset in each of the 20 sampling processes. For the virtual species, we have 20 unique datasets, where the number of occurrences and pseudo-absences was the same in each dataset, but different occurrences and pseudo-absences were sampled in each dataset. Similar patterns of model performance were observed for virtual species when the same and unique datasets were used to model the species distribution in each sampling process (Appendix S1: Figure S2). Thus, using the same and unique datasets during the cross-validation procedure does not affect the estimation of the virtual species distributions. Virtual species SDMs were evaluated with independent data (i.e., data that were not sampled to model the species distributions).

Assessing model performance

We evaluated SDM performance with regard to accuracy, calibration, discrimination power (hereafter discrimination), and precision (Norberg et al., 2019). Accuracy measures the agreement between the model predicted values and the observed values (i.e., how close the predicted values are to the observed values). We measured accuracy as the absolute difference between the predicted (varying from 0 to 1) and observed (0 for pseudo-absence and 1 for occurrence) values. Calibration was assessed as the statistical accuracy between the predicted and observed values. As a measure of calibration, we calculated the root-mean-square error (RMSE) between the predicted and observed values in 10 probability bins (see methods section in Appendix S1 and Appendix S1: Figure S3 for details). Discrimination evaluates how well the predictive values of a model can differ between occurrences and pseudo-absences. We used the area under the receiver operating characteristic curve (AUC) (Fielding & Bell, 1997) to evaluate discrimination. At last, precision measures the breadth of the predictive distribution of the models. As a measure of precision, we calculated the square root of the product of the probability of species occurrence in a cell times the probability of the species absence in that same cell. All the performance metrics were averaged across species. To facilitate the interpretation of the results, we reversed the signs of the performance metrics when applicable, such that higher performance values always indicated a higher discrimination, calibration, accuracy, or precision of the models.

We used a linear mixed-effects model (LMM) to evaluate whether thinning occurrence points affected SDM performances. In this framework, model performance was used as the response variable and spatial and environmental thinning distances were used as fixed effects and species was used as a random effect. The non-filtered dataset was used as the baseline case to evaluate the effects of thinning occurrence points on model performance. We present the effects of thinning on MaxEnt discrimination power (AUC) in the *Results* section except where noted otherwise.

RESULTS

General effects of thinning occurrence points

Thinning occurrence points consistently decreased model performance irrespective of spatial bias for real and virtual species (Figure 2, Table 1). Using larger spatial and environmental thinning distances resulted in fewer occurrence points being available for modeling (Appendix S1: Figure S4), and worse model performances were observed in these cases. Spatially and environmentally thinning species occurrence points was functionally the same as removing occurrences randomly, given that model performances were similar between thinned and null datasets (i.e., error bars overlapping zero; Figure 3).

Thinning real and virtual species occurrences

Spatial thinning had stronger negative effects on clustered occurrences relative to random or evenly spaced occurrences (Figure 2). Specifically, even the smallest spatial distances caused significant decreases in model performance of real and virtual species with clustered occurrences (Table 1). Alternatively, virtual species with random and evenly spaced occurrences only experienced significant decreases in model performance when the two largest spatial distances were used (Table 1), suggesting that these species are less susceptible to the negative effects of spatial thinning.

Environmental thinning decreased model performance similarly across real and all virtual species (Figure 2). In general, environmentally thinning occurrence points using the smallest environmental distance did not affect model performance, but model performance quickly deteriorated for all species as the environmental distance used increased (Table 1). Thus, environmental thinning had more consistent negative effects on SDM performance than spatial thinning.

Effects of thinning on different performance metrics and modeling methods

In general, discrimination, accuracy, calibration, and precision were positively correlated to each other (Figure 4), suggesting a general agreement between the performance metrics we used to evaluate the SDMs. As with discrimination, thinning occurrence points also decreased models accuracy (Appendix S1: Figure S5), precision (Appendix S1: Figure S6), and calibration (Appendix S1: Figure S7), indicating that models are in general making more incorrect predictions when occurrence points are thinned.

Overall, GLM tended to have the worst performance across the modeling methods whereas GAM, SVM, and MaxEnt had superior performances (see Appendix S1). Although thinning had consistent negative effects on the performance of the four modeling methods we assessed,

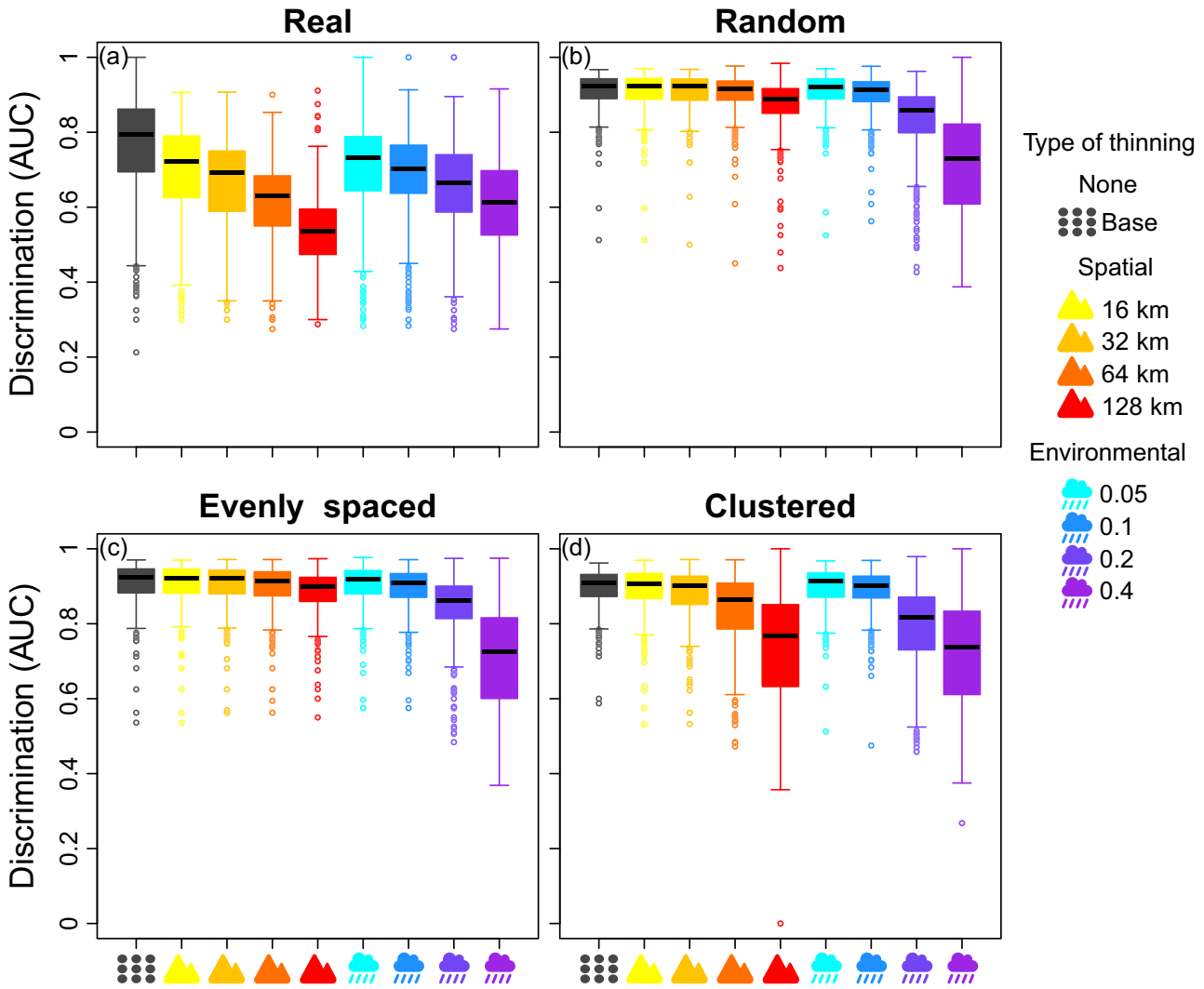

FIGURE 2 The effects of spatial and environmental thinning on MaxEnt discrimination (area under the curve, AUC) for real species (a), and virtual species with random (b), evenly spaced (c), and clustered (d) sampled occurrence points. While spatial thinning had a stronger negative effect on real species and virtual species with clustered occurrences, environmental thinning similarly negatively affected all species. Larger thinning distances consistently led to worse performances for both spatial and environmental thinning.

some methods were more affected by thinning than others. For example, SVM (Appendix S1: Figures S8–S11) and GLM (Appendix S1: Figures S12–S15) had relatively similar model performances for the virtual species when the non-thinned and thinned datasets were used. On the other hand, MaxEnt (Figure 2, Appendix S1: Figures S5–S7) and GAM (Appendix S1: Figures S16–S19) had more significant decreases in model performance when occurrences were thinned. This result suggests that the number of occurrence points is more important to efficiently model species distributions for some modeling methods (i.e., MaxEnt and GAM) than for others (i.e., SVM and GLM). The only improvement in model performance caused by thinning was observed when precision was the performance metric for GLM (Appendix S1: Figure S14) and GAM (Appendix S1: Figure S18). The fact that improvement in model performance was not observed by any other performance metric and modeling method we considered shows the general ineffectiveness of thinning approaches to address spatial bias (see Appendix S1 for details).

DISCUSSION

In general, we found that spatial and environmental thinning species occurrence points led to a consistent decrease in SDM performance. These reductions in performance were constant across real and virtual species, performance metrics, modeling methods, and were positively associated with thinning distance. Real species and virtual species with clustered points were particularly susceptible to spatial thinning, indicating that this

TABLE 1 Fitted linear mixed-effects models showing the negative effects of thinning on species distribution model discrimination (area under the curve) for real and virtual species with random, evenly spaced, and clustered spatial bias in sampled occurrence points.

Species	Thinning	Estimate	SE	df	*T*	*p*
Real	S-16	−0.067	0.005	3444.2	−14.464	**<0.001**
	S-32	−0.099	0.005	3444.3	−21.416	**<0.001**
	S-64	−0.152	0.005	3444.5	−32.818	**<0.001**
	S-128	−0.229	0.005	3445.0	−49.289	**<0.001**
	E-10	−0.060	0.005	3443.9	−12.954	**<0.001**
	E-15	−0.076	0.005	3443.9	−16.490	**<0.001**
	E-20	−0.107	0.005	3444.3	−23.269	**<0.001**
	E-25	−0.158	0.005	3444.4	−34.147	**<0.001**
Random	S-16	−0.000	0.003	3992.0	−0.018	0.986
	S-32	−0.001	0.003	3992.0	−0.376	0.707
	S-64	−0.008	0.003	3992.0	−2.245	**0.025**
	S-128	−0.034	0.003	3992.0	−10.078	**<0.001**
	E-10	−0.002	0.003	3992.0	−0.498	0.619
	E-15	−0.009	0.003	3992.0	−2.704	**0.007**
	E-20	−0.076	0.003	3992.0	−22.641	**<0.001**
	E-25	−0.196	0.003	3992.0	−58.002	**<0.001**
Evenly spaced	S-16	−0.001	0.003	3992.0	−0.174	0.862
	S-32	−0.003	0.003	3992.0	−1.012	0.312
	S-64	−0.008	0.003	3992.0	−2.683	**0.007**
	S-128	−0.021	0.003	3992.0	−6.975	**<0.001**
	E-10	−0.004	0.003	3992.0	−1.296	0.195
	E-15	−0.014	0.003	3992.0	−4.697	**<0.001**
	E-20	−0.061	0.003	3992.0	−20.122	**<0.001**
	E-25	−0.197	0.003	3992.0	−65.381	**<0.001**
Clustered	S-16	−0.002	0.004	3992.0	−0.379	0.704
	S-32	−0.013	0.004	3992.0	−3.164	**0.002**
	S-64	−0.058	0.004	3992.0	−13.672	**<0.001**
	S-128	−0.155	0.004	3992.0	−36.414	**<0.001**
	E-10	0.003	0.004	3992.0	0.739	0.460
	E-15	−0.006	0.004	3992.0	−1.410	0.159
	E-20	−0.101	0.004	3992.0	−23.732	**<0.001**
	E-25	−0.175	0.004	3992.0	−41.116	**<0.001**

Note: Significant results ($p < 0.05$ appear in boldface). The performance of the models obtained with the non-thinned dataset was used to compare against the effects of spatial and environmental thinning. S-16, S-32, S-64, and S128 refer to spatial thinning using the thinning distances of 16, 32, 64, and 128 km, respectively, while E-10, E-15, E-20, and E-25 refer to environmental thinning using the thinning distances of 0.10, 0.15, 0.20, and 0.25, respectively.

approach might be inadequate for situations where its use is generally suggested (Beck et al., 2014; Boria et al., 2014). Environmental thinning had similar negative effects on model performance for all species, showing a general limitation of this approach to address biases in occurrence points. Finally, we showed that thinned models perform similarly to null models, suggesting that thinning occurrence points is not better than randomly removing them and further highlighting the inefficacy of thinning approaches in improving model performance. Taken together, our results show that using thinning approaches to remove spatial biases from species occurrence data often negatively affected SDM performance under different modeling scenarios.

Thinning occurrence points before modeling species distributions has the goal of removing sampling bias that

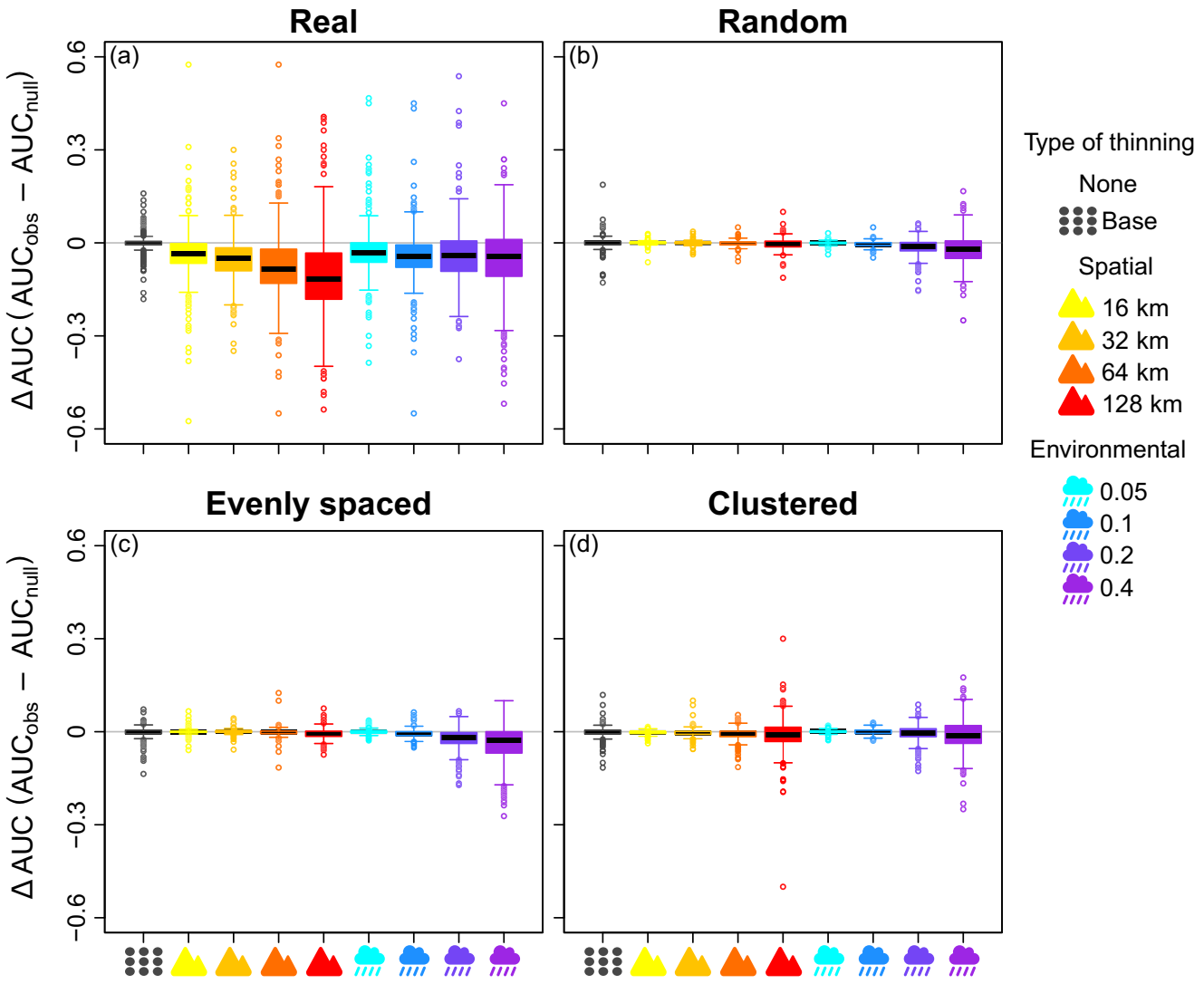

FIGURE 3 Differences in discrimination (area under the curve [AUC], ΔAUC) of models trained with thinned (AUC_{obs}) and null (AUC_{null}) data for real species (a) and virtual species with random (b), evenly spaced (c), and clustered (d) sampled occurrence points. Positive values indicate that thinned models perform better than null models while negative values suggest the opposite. ΔAUC values consistently overlapped 0, suggesting that spatially and environmentally thinning occurrence points often failed to improve model performance when compared with null expectations.

is often found in occurrence data (Aiello-Lammens et al., 2015; Varela et al., 2014). Although spatial and environmental thinning can improve model performance in individual cases (Boria et al., 2014; Varela et al., 2014), thinning can have mixed effects on model performance when it is assessed considering several species (Baker et al., 2022; Inman et al., 2021; Steen et al., 2021). We expand these results by showing that, when considering a broad ecological context with hundreds of real and virtual species, the effects of spatial and environmental thinning on model performance can often be negative and no different than null expectations. The decrease in model performance caused by spatial thinning is thought to be a consequence of the information loss caused by this approach (Varela et al., 2014). Environmental thinning is also likely leading to a similar degree of information loss given its stronger impacts on virtual species than spatial thinning. This general negative impact of thinning is not unexpected, given that the number of occurrence points can have a greater impact on model performance than spatial bias (Gaul et al., 2020), which also explains why using larger thinning distances led to worse model performances.

One of the most commonly used performance metrics to evaluate SDMs is AUC because it is a threshold-independent metric (Fourcade et al., 2018). However, relying on a single performance metric when assessing the effects of thinning on SDMs is problematic (Castellanos et al., 2019), and

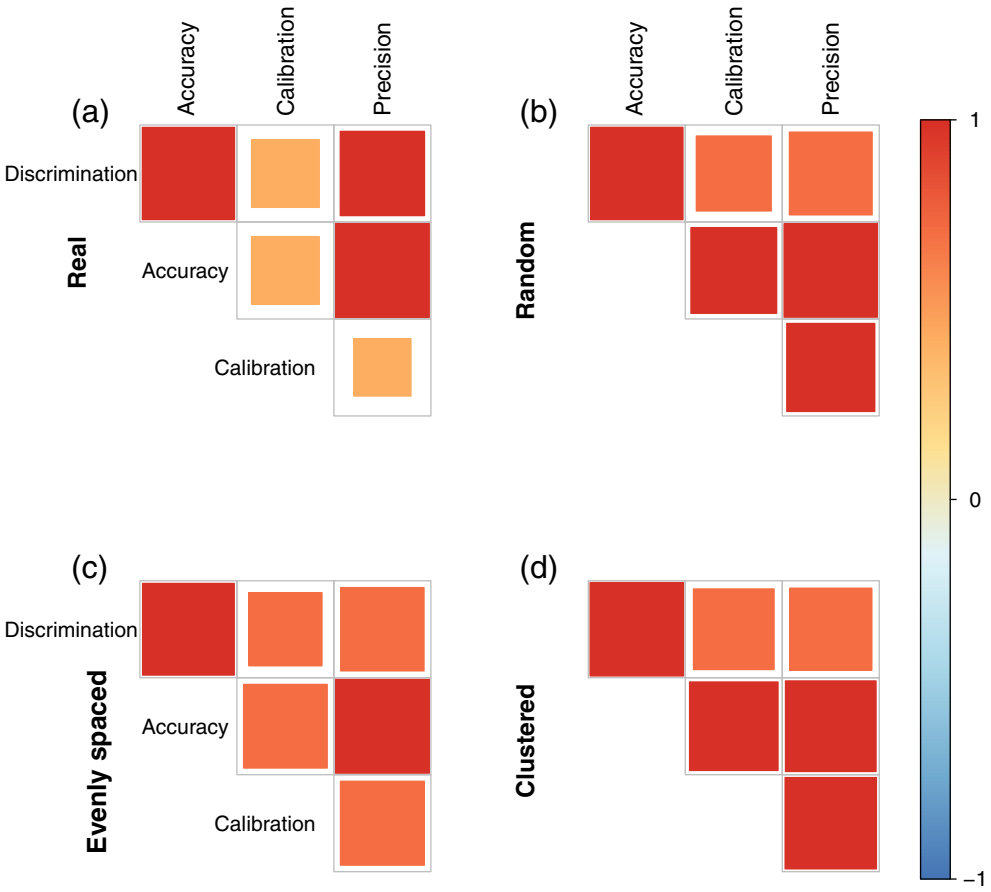

FIGURE 4 Correlation between the performance metrics for real species (a) and virtual species with random (b), evenly spaced (c), and clustered (d) sampled occurrence points. Discrimination, accuracy, calibration, and precision are generally positively correlated to each other, indicating that when a high performance is observed for one of the metrics, the other three also have a high performance.

using alternative metrics that capture different aspects of model effectiveness under different scenarios allows for a more complete assessment of model performance (Lotterhos et al., 2022; Norberg et al., 2019). We show that discrimination, accuracy, calibration, and precision are generally positively correlated to each other and are generally negatively affected by thinning. Specifically, thinning occurrence points not only decreased the ability of the models to discern different types of values (lower discrimination) but also led models to be more biased (lower accuracy) and less consistent (lower calibration) and precise (lower precision). Although thinning made some models (i.e., GLM and GAM) more precise, they also had lower accuracy, indicating that these models were more confident in making incorrect predictions. Thus, evaluating models considering single performance metrics should be avoided as it can potentially lead to misleading conclusions regarding the effects of thinning on SDMs.

Our results contribute to a growing body of evidence showing that spatially or environmentally thinning occurrence points might not necessarily improve SDM performance (Castellanos et al., 2019; Inman et al., 2021; Steen et al., 2021). Nonetheless, we recognize that our study has some limitations such as with regard to the types of spatial biases we simulated in our virtual species. For example, real species usually have occurrence points sampled in a biased manner (Oliveira et al., 2016; Sousa-Baena et al., 2014; Vale & Jenkins, 2012) that leads their occurrence points to be mostly spatially clustered (Beck et al., 2014; Inman et al., 2021) while the occurrence of species with random or evenly spaced occurrence points might be less common. Nevertheless, we show that the effects of thinning were similar across the three different types of spatial bias we considered, indicating that thinning has consistent negative effects on model performance regardless of the type of spatial bias found in the data that are used in SDMs. Additionally, more realistic effects of thinning on SDM performance might be obtained when thinning distances are chosen considering the region, spatial scale, and species being studied (Castellanos et al., 2019). Although we did not consider these factors when choosing thinning distances, the fact that we used a wide range of thinning distances and that

SDM performance consistently decreased across these different distances suggests that the potential positive effects of thinning on SDM performance are limited.

Evaluating the effects of thinning considering only real species is challenging because there is no knowledge of the "true" distribution of these species, which complicates the assessment of the models, a problem that is overcome with virtual species (Miller, 2014; Moudrý, 2015). On the other hand, assessing the effects of thinning considering only virtual species might also not be optimal, given that these species are usually modeled following simplistic assumptions that might not be observed in the real world, such as the lack of dispersal limitation or interspecific interactions that are important in controlling real species geographic distributions (Soberón & Peterson, 2005). Here, we use a set of real and virtual species to provide strong evidence that thinning occurrence points often decreases SDM performance. Although higher model performance does not necessarily mean more biologically realistic predictions (Fourcade et al., 2018; Godsoe, 2010), the fact that model performance consistently decreased in all circumstances considered in our study, and that these models performed no better than null models, highlights the limitation of thinning approaches to improve SDM performance. These negative effects of thinning might be particularly pronounced when rare species are considered, given that modeling their distributions is already a challenging task because of their limited data availability (Steen et al., 2021). Although thinning might not always be detrimental to SDM performance, our results suggest that its use should be considered carefully, given that it can easily lead to decreases in model performance.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support provided by the National Science Foundation (NSF-DEB-2017826) *Macrosystems Biology and NEON-Enabled Science* program.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data and R code (Ten Caten & Dallas, 2022) are available from Figshare: https://doi.org/10.6084/m9.figshare.21764129.

ORCID

Cleber Ten Caten https://orcid.org/0000-0003-3788-3508
Tad Dallas https://orcid.org/0000-0003-3328-9958

REFERENCES

Aiello-Lammens, M. E., R. A. Boria, A. Radosavljevic, B. Vilela, and R. P. Anderson. 2015. "Spthin: An r Package for Spatial Thinning of Species Occurrence Records for Use in Ecological Niche Models." *Ecography* 38(5): 541–45.

Anderson, R. P., and I. Gonzalez, Jr. 2011. "Species-Specific Tuning Increases Robustness to Sampling Bias in Models of Species Distributions: An Implementation with Maxent." *Ecological Modelling* 222(15): 2796–2811.

Anderson, R. P., and A. Raza. 2010. "The Effect of the Extent of the Study Region on gis Models of Species Geographic Distributions and Estimates of Niche Evolution: Preliminary Tests with Montane Rodents (Genus *Nephelomys*) in Venezuela." *Journal of Biogeography* 37(7): 1378–93.

Bahn, V., and B. J. McGill. 2013. "Testing the Predictive Performance of Distribution Models." *Oikos* 122(3): 321–331.

Baker, D. J., I. M. Maclean, M. Goodall, and K. J. Gaston. 2022. "Correlations between Spatial Sampling Biases and Environmental Niches Affect Species Distribution Models." *Global Ecology and Biogeography* 31: 1038–50.

Beck, J., M. Böller, A. Erhardt, and W. Schwanghart. 2014. "Spatial Bias in the gbif Database and Its Effect on Modeling Species' Geographic Distributions." *Ecological Informatics* 19: 10–15.

Boria, R. A., L. E. Olson, S. M. Goodman, and R. P. Anderson. 2014. "Spatial Filtering to Reduce Sampling Bias Can Improve the Performance of Ecological Niche Models." *Ecological Modelling* 275: 73–77.

Botts, E. A., B. F. Erasmus, and G. J. Alexander. 2011. "Geographic Sampling Bias in the South African Frog Atlas Project: Implications for Conservation Planning." *Biodiversity and Conservation* 20(1): 119–139.

Boucher-Lalonde, V., A. Morin, and D. J. Currie. 2014. "A Consistent Occupancy–Climate Relationship across Birds and Mammals of the Americas." *Oikos* 123(9): 1029–36.

Brown, J. H. 1995. *Macroecology*. Chicago, IL: University of Chicago Press.

Cardillo, M., G. M. Mace, J. L. Gittleman, K. E. Jones, J. Bielby, and A. Purvis. 2008. "The Predictability of Extinction: Biological and External Correlates of Decline in Mammals." *Proceedings of the Royal Society B: Biological Sciences* 275(1641): 1441–48.

Castellanos, A. A., J. W. Huntley, G. Voelker, and A. M. Lawing. 2019. "Environmental Filtering Improves Ecological Niche Models across Multiple Scales." *Methods in Ecology and Evolution* 10(4): 481–492.

Chamberlain, S., V. Barve, D. Mcglinn, D. Oldoni, P. Desmet, L. Geffert, and K. Ram. 2021. "rgbif: Interface to the Global Biodiversity Information Facility API." R Package Version 3.5.2.

Dallas, T. A., and L. Santini. 2020. "The Influence of Stochasticity, Landscape Structure and Species Traits on Abundant–Centre Relationships." *Ecography* 43(9): 1341–51.

Drake, J. M., C. Randin, and A. Guisan. 2006. "Modelling Ecological Niches with Support Vector Machines." *Journal of Applied Ecology* 43(3): 424–432.

Elith, J., M. Kearney, and S. Phillips. 2010. "The Art of Modelling Range-Shifting Species." *Methods in Ecology and Evolution* 1(4): 330–342.

Fielding, A. H., and J. F. Bell. 1997. "A Review of Methods for the Assessment of Prediction Errors in Conservation Presence/Absence Models." *Environmental Conservation* 24(1): 38–49.

Fourcade, Y., A. G. Besnard, and J. Secondi. 2018. "Paintings Predict the Distribution of Species, or the Challenge of Selecting Environmental Predictors and Evaluation Statistics." *Global Ecology and Biogeography* 27(2): 245–256.

Gaston, K. J. 2003. *The Structure and Dynamics of Geographic Ranges*. Oxford and New York: Oxford University Press.

Gaul, W., D. Sadykova, H. J. White, L. Leon-Sanchez, P. Caplat, M. C. Emmerson, and J. M. Yearsley. 2020. "Data Quantity Is More Important than Its Spatial Bias for Predictive Species Distribution Modelling." *PeerJ* 8: e10411.

Godsoe, W. 2010. "I Can't Define the Niche but I Know It When I See It: A Formal Link between Statistical Theory and the Ecological Niche." *Oikos* 119(1): 53–60.

Guisan, A., T. C. Edwards, Jr., and T. Hastie. 2002. "Generalized Linear and Generalized Additive Models in Studies of Species Distributions: Setting the Scene." *Ecological Modelling* 157(2–3): 89–100.

Herkt, K. M. B., A. K. Skidmore, and J. Fahr. 2017. "Macroecological Conclusions Based on Iucn Expert Maps: A Call for Caution." *Global Ecology and Biogeography* 26(8): 930–941.

Hidalgo-Mihart, M. G., L. Cantú-Salazar, A. González-Romero, and C. A. López-González. 2004. "Historical and Present Distribution of Coyote (*Canis latrans*) in Mexico and Central America." *Journal of Biogeography* 31(12): 2025–38.

Hijmans, R. J. 2019. "Geosphere: Spherical Trigonometry." R Package Version 1.5–10.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis. 2005. "Very High Resolution Interpolated Climate Surfaces for Global Land Areas." *International Journal of Climatology: A Journal of the Royal Meteorological Society* 25(15): 1965–78.

Hijmans, R. J., S. Phillips, J. Leathwick, J. Elith, and M. R. J. Hijmans. 2017. "Package 'dismo'." *Circles* 9(1): 1–68.

Hortal, J., F. de Bello, J. A. F. Diniz-Filho, T. M. Lewinsohn, J. M. Lobo, and R. J. Ladle. 2015. "Seven Shortfalls that Beset Large-Scale Knowledge of Biodiversity." *Annual Review of Ecology, Evolution, and Systematics* 46: 523–549.

Inman, R., J. Franklin, T. Esque, and K. Nussear. 2021. "Comparing Sample Bias Correction Methods for Species Distribution Modeling Using Virtual Species." *Ecosphere* 12(3): e03422.

Iturbide, M., J. Bedia, S. Herrera, O. del Hierro, M. Pinto, and J. M. Gutiérrez. 2015. "A Framework for Species Distribution Modelling with Improved Pseudo-Absence Generation." *Ecological Modelling* 312: 166–174.

Jetz, W., J. M. McPherson, and R. P. Guralnick. 2012. "Integrating Biodiversity Distribution Knowledge: Toward a Global Map of Life." *Trends in Ecology & Evolution* 27(3): 151–59.

Kadmon, R., O. Farber, and A. Danin. 2004. "Effect of Roadside Bias on the Accuracy of Predictive Maps Produced by Bioclimatic Models." *Ecological Applications* 14(2): 401–413.

Karatzoglou, A., A. Smola, K. Hornik, and M. A. Karatzoglou. 2019. "Package 'kernlab'." Technical Report, CRAN, 03 2016.

Lee, T. M., and W. Jetz. 2011. "Unravelling the Structure of Species Extinction Risk for Predictive Conservation Science." *Proceedings of the Royal Society B: Biological Sciences* 278(1710): 1329–38.

Leroy, B., C. N. Meynard, C. Bellard, and F. Courchamp. 2016. "Virtualspecies, an r Package to Generate Virtual Species Distributions." *Ecography* 39(6): 599–607.

Liu, C., P. M. Berry, T. P. Dawson, and R. G. Pearson. 2005. "Selecting Thresholds of Occurrence in the Prediction of Species Distributions." *Ecography* 28(3): 385–393.

Loiselle, B. A., P. M. Jørgensen, T. Consiglio, I. Jiménez, J. G. Blake, L. G. Lohmann, and O. M. Montiel. 2008. "Predicting Species Distributions from Herbarium Collections: Does Climate Bias in Collection Sampling Influence Model Outcomes?" *Journal of Biogeography* 35(1): 105–116.

Lomolino, M. V. 2004. *Conservation Biogeography. Frontiers of Biogeography: New Directions in the Geography of Nature*, 293. Sunderland, MA: Sinauer Associates, Inc.

Lotterhos, K. E., M. C. Fitzpatrick, and H. Blackmon. 2022. "Simulation Tests of Methods in Evolution, Ecology, and Systematics: Pitfalls, Progress, and Principles." *Annual Review of Ecology, Evolution, and Systematics* 53: 113–136.

Loyola, T. F. R. R. D. 2012. "Labeling Ecological Niche Models." *Natureza Conservação* 10: 119–126.

Mair, L., and A. Ruete. 2016. "Explaining Spatial Variation in the Recording Effort of Citizen Science Data across Multiple Taxa." *PLoS One* 11(1): e0147796.

McPherson, J. M., W. Jetz, and D. J. Rogers. 2004. "The Effects of species' Range Sizes on the Accuracy of Distribution Models: Ecological Phenomenon or Statistical Artefact?" *Journal of Applied Ecology* 41(5): 811–823.

Melo, A. S. 2019. "CommEcol: Community Ecology Analyses." R Package Version 1.7.0.

Miller, J. A. 2014. "Virtual Species Distribution Models: Using Simulated Data to Evaluate Aspects of Model Performance." *Progress in Physical Geography* 38(1): 117–128.

Moerman, D. E., and G. F. Estabrook. 2006. "The Botanist Effect: Counties with Maximal Species Richness Tend to be Home to Universities and Botanists." *Journal of Biogeography* 33(11): 1969–74.

Moudrỳ, V. 2015. "Modelling Species Distributions with Simulated Virtual Species." *Journal of Biogeography* 42(8): 1365–66.

Nelder, J. A., and R. W. Wedderburn. 1972. "Generalized Linear Models." *Journal of the Royal Statistical Society: Series A (General)* 135(3): 370–384.

Norberg, A., N. Abrego, F. G. Blanchet, F. R. Adler, B. J. Anderson, J. Anttila, M. B. Araújo, et al. 2019. "A Comprehensive Evaluation of Predictive Performance of 33 Species Distribution Models at Species and Community Levels." *Ecological Monographs* 89(3): e01370.

Oksanen, J., and P. R. Minchin. 2002. "Continuum Theory Revisited: What Shape Are Species Responses along Ecological Gradients?" *Ecological Modelling* 157(2–3): 119–129.

Oliveira, U., A. P. Paglia, A. D. Brescovit, C. J. de Carvalho, D. P. Silva, D. T. Rezende, F. S. F. Leite, et al. 2016. "The Strong Influence of Collection Bias on Biodiversity Knowledge Shortfalls of Brazilian Terrestrial Biodiversity." *Diversity and Distributions* 22(12): 1232–44.

Pebesma, E. 2018. "Simple Features for R: Standardized Support for Spatial Vector Data." *The R Journal* 10(1): 439–446.

Peterson, A. T., J. Soberón, R. G. Pearson, R. P. Anderson, E. Martnez-Meyer, M. Nakamura, and M. B. Araújo. 2011. *Ecological Niches and Geographic Distributions (MPB-49)*, Vol. 49. Princeton, NJ: Princeton University Press.

Phillips, S. J., M. Dudk, J. Elith, C. H. Graham, A. Lehmann, J. Leathwick, and S. Ferrier. 2009. "Sample Selection Bias and

Presence-Only Distribution Models: Implications for Background and Pseudo-Absence Data." *Ecological Applications* 19(1): 181–197.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Radosavljevic, A., and R. P. Anderson. 2014. "Making Better Maxent Models of Species Distributions: Complexity, Overfitting and Evaluation." *Journal of Biogeography* 41(4): 629–643.

Ranc, N., L. Santini, C. Rondinini, L. Boitani, F. Poitevin, A. Angerbjörn, and L. Maiorano. 2017. "Performance Tradeoffs in Target-Group Bias Correction for Species Distribution Models." *Ecography* 40(9): 1076–87.

Roberts, D. R., V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, et al. 2017. "Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure." *Ecography* 40(8): 913–929.

Santini, L., A. Bentez-López, M. Čengić, L. Maiorano, and M. A. Huijbregts. 2020. "Assessing the Reliability of Species Distribution Projections in Climate Change Research." BioRxiv.

Schölkopf, B., J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. 2001. "Estimating the Support of a High-Dimensional Distribution." *Neural Computation* 13(7): 1443–71.

Senay, S. D., S. P. Worner, and T. Ikeda. 2013. "Novel Three-Step Pseudo-Absence Selection Technique for Improved Species Distribution Modelling." *PLoS One* 8(8): e71218.

Soberón, J., and A. T. Peterson. 2005. "Interpretation of Models of Fundamental Ecological Niches and species' Distributional Areas." *Biodiversity Informatics* 2: 1–10.

Sousa-Baena, M. S., L. C. Garcia, and A. T. Peterson. 2014. "Completeness of Digital Accessible Knowledge of the Plants of Brazil and Priorities for Survey and Inventory." *Diversity and Distributions* 20(4): 369–381.

Steen, V. A., M. W. Tingley, P. W. Paton, and C. S. Elphick. 2021. "Spatial Thinning and Class Balancing: Key Choices Lead to Variation in the Performance of Species Distribution Models with Citizen Science Data." *Methods in Ecology and Evolution* 12(2): 216–226.

Tax, D. M., and R. P. Duin. 2004. "Support Vector Data Description." *Machine Learning* 54(1): 45–66.

Ten Caten, C., and T. Dallas. 2022. "Data and Code to Reproduce 'Thinning Presence Points Does Not Improve Species Distribution Model Performance'." Figshare. Dataset. https://doi.org/10.6084/m9.figshare.21764129.v3.

Terribile, L. C., D. T. Feitosa, M. G. Pires, P. C. R. de Almeida, G. de Oliveira, J. A. F. Diniz-Filho, and N. J. da Silva Jr. 2018. "Reducing Wallacean Shortfalls for the Coralsnakes of the *Micrurus lemniscatus* Species Complex: Present and Future Distributions under a Changing Climate." *PLoS One* 13(11): e0205164.

Tessarolo, G., T. F. Rangel, M. B. Araújo, and J. Hortal. 2014. "Uncertainty Associated with Survey Design in Species Distribution Models." *Diversity and Distributions* 20(11): 1258–69.

Vale, M. M., and C. N. Jenkins. 2012. "Across-Taxa Incongruence in Patterns of Collecting Bias." *Journal of Biogeography* 39(9): 1744–48.

van Proosdij, A. S., M. S. Sosef, J. J. Wieringa, and N. Raes. 2016. "Minimum Required Number of Specimen Records to Develop Accurate Species Distribution Models." *Ecography* 39(6): 542–552.

Varela, S., R. P. Anderson, R. Garca-Valdés, and F. Fernández-González. 2014. "Environmental Filters Reduce the Effects of Sampling Bias and Improve Predictions of Ecological Niche Models." *Ecography* 37(11): 1084–91.

Veloz, S. D. 2009. "Spatially Autocorrelated Sampling Falsely Inflates Measures of Accuracy for Presence-Only Niche Models." *Journal of Biogeography* 36(12): 2290–99.

Whittaker, R. J., M. B. Araújo, P. Jepson, R. J. Ladle, J. E. Watson, and K. J. Willis. 2005. "Conservation Biogeography: Assessment and Prospect." *Diversity and Distributions* 11(1): 3–23.

Wood, S. 2017. *Generalized Additive Models: An Introduction with R*, 2nd ed. Boca Raton, FL: Chapman and Hall/CRC.

SUPPORTING INFORMATION

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

How to cite this article: Ten Caten, Cleber, and Tad Dallas. 2023. "Thinning Occurrence Points Does Not Improve Species Distribution Model Performance." *Ecosphere* 14(12): e4703. https://doi.org/10.1002/ecs2.4703