No optimal spatial filtering distance for mitigating sampling bias in ecological niche models

Quentin Lamboley, Yoan Fourcade

Received: 30 October 2023 | Revised: 9 April 2024 | Accepted: 15 April 2024

DOI: 10.1111/jbi.14854

Journal of Biogeography

WILEY

METHOD

No optimal spatial filtering distance for mitigating sampling bias in ecological niche models

Quentin Lamboley | Yoan Fourcade

Univ Paris-Est Creteil, Sorbonne Université, Université Paris-Cité, CNRS, IRD, INRAE, Institut d'Écologie et des Sciences de l'Environnement, IEES, Créteil, France

Correspondence
Yoan Fourcade, Univ Paris Est Creteil, Sorbonne Université, Université Paris-Cité, CNRS, IRD, INRAE, Institut d'Écologie et des Sciences de l'Environnement, IEES, F-94010 Créteil, France.
Email: yoan.fourcade@u-pec.fr

Abstract

Aim: The continuous development of statistical tools applied to ecology has contributed to great advances for modelling species' niches and distributions from opportunistic observations. However, as these observations are subject to biases caused by spatial variation in sampling effort, ecological niche models (ENMs) are also frequently biased. Among several bias correction methods that have been proposed, spatial filtering—imposing a minimum distance between occurrences—is widely used, yet lacks clear guidelines for choosing the filtering distance. Here, we aimed to explore the impact of spatial filtering distances on the performance of ENMs.

Location: Europe.

Taxon: Virtual species.

Methods: We applied ENMs to two virtual species with contrasting levels of specialisation, across a spectrum of modelling conditions, bias types and sample sizes.

Results: Models applied to the specialist species had on average a lower performance than those applied to the generalist species. Using a biased sample reduced model performance, especially when the bias was strong, and when the sample size was large. In many cases, spatial filtering failed to improve model performance or even reduced it. We did find an improvement for the generalist species modelled with large and strongly biased datasets. However, there was no optimal filtering distance, as this improvement was linearly and positively associated with filtering distance. Moreover, because the initial bias was strong and the filtered dataset became very small, the resulting models had only very low accuracy.

Main Conclusions: Our results suggest that there is no optimal filtering distance for dealing with sampling bias in ENMs, and that spatial filtering never improves model performance enough to draw accurate predictions. We therefore recommend spatial filtering to be employed cautiously, only when enough data are available, and bearing in mind that its effectiveness remains highly uncertain.

KEYWORDS
bias, ecological niche, MaxEnt, spatial filtering, spatial thinning, species distribution modelling, sub-sampling

1 | INTRODUCTION

Most of the occurrence data that compose biodiversity databases such as the Global Biodiversity Information Facility (GBIF) originate from opportunistic observations collected as part of citizen science projects or from monitoring programmes (Edwards, 2004). The density of these data is associated with geographical and temporal patterns of sampling effort, which are linked to accessibility, season, or human population density (Bowler et al., 2022; Correia et al., 2019; Mair & Ruete, 2016). We often observe a scattering of occurrence data towards easily accessible areas that does not correspond to a higher population density. On the contrary, an absence of occurrences may reflect the lack of sampling effort instead of the absence of the species. As a result, occurrence datasets are frequently biased and characterised by a geographical distribution that does not fully reflect the true distribution of species (Beck et al., 2014; Daru & Rodriguez, 2023; Garcia-Rosello et al., 2023; Hughes et al., 2021).

These—potentially biased—datasets are frequently used in a variety of methods known as species distribution models (SDMs) or ecological niche models (ENMs), whose principle is to link occurrence data and environmental variables to map species' habitat suitability (Elith & Leathwick, 2009). These approaches have become very common in the fields of biogeography, climate-change ecology or invasion biology, and are implemented through various statistical tools, from simple generalised linear models to advanced machine-learning algorithms such as the popular MaxEnt (Elith et al., 2006, 2010). If the initial datasets are biased, models tend to overestimate the importance of environmental values in oversampled areas, resulting in distorted predictions of species distributions (Baker et al., 2022; Beck et al., 2014; Bystriakova et al., 2012; Guillera-Arroita et al., 2015).

A number of approaches have been proposed to address the problem of sampling bias in occurrence data in the context of ENMs. Among them, some methods focus on manipulating the background data, which are often used in lieu of true absences when only species' sightings (i.e., presence data) are available. The general principle consists in sampling the species' accessible area (Barve et al., 2011) with the same bias as in the occurrence dataset. This can be achieved for instance by using background thickening at the proximity of species records (Vollering et al., 2019), or using occurrences from sister species (supposedly collected with the same bias) as a so-called target-group background (Barber et al., 2021; Phillips et al., 2009; Ranc et al., 2017). Another possibility, that is offered in the MaxEnt method, is to incorporate a bias grid which weights occurrences by sampling effort (Dudík et al., 2007; Elith et al., 2010; Merow et al., 2013). The performance of background manipulation methods, however, is dependent on a good knowledge of the structure of the bias, which is rare in real-case studies (Baker et al., 2024).

A simple alternative is to manipulate the occurrence records. In this case, the most common strategy is the spatial filtering or thinning of input occurrence data (Aiello-Lammens et al., 2015; Boria et al., 2014; Fourcade et al., 2014; Galante et al., 2018). It consists in sub-sampling the data by enforcing a minimum distance between two occurrences. The removal of points that are located in close proximity to one another ensures that the spatial density of occurrence data is homogenised. Thus, the principal of spatial filtering involves removing information in a targeted way so that it can improve the performance of models. An alternative approach is to filter occurrences in the environmental space instead of the geographical space, to eliminate data aggregates in similar environmental conditions (Castellanos et al., 2019; Varela et al., 2014). In several studies that tested methods of sampling bias correction, spatial filtering appeared to be able to mitigate, at least to some extent, the effect of bias on the resulting models (Boria et al., 2014; Fourcade et al., 2014; Inman et al., 2021; Kramer-Schadt et al., 2013). Due to the positive outcomes observed in these studies, and because it is easy to implement (e.g. Aiello-Lammens et al., 2015), spatial filtering has emerged as one of the most frequently used correction method for dealing with datasets suspected to be biased.

Currently, though, the choice of an appropriate filtering distance remains an open question. The optimal distance is likely to depend on the species' ecology, the environmental heterogeneity, and the patterns of sampling scheme, information that is often unavailable (Aiello-Lammens et al., 2015). Therefore, in practice, the filtering distance is generally chosen by the modeller based on a visual evaluation of the spatial smoothing of occurrences. Modelling frameworks usually retain a single occurrence record within each grid cell of the environmental rasters used for modelling (e.g. Phillips et al., 2006), effectively removing data aggregation at the resolution of the input variables. However, although it avoids pseudoreplication, it should not be viewed as an effective method of bias correction. While these approaches may succeed in correcting sampling bias, they lack objectivity and reproducibility. The absence of guidelines for choosing an optimal filtering distance that would balance information loss and mitigation of sampling bias presents a significant knowledge gap that hinders the reproducible implementation of ENMs for species with biased input data.

Since this 'ideal' distance, if it exists, is certainly context-dependent, there is a need for developing a set of recommendations for selecting filtering distance in different scenarios of bias or varying species ecologies. The novel objective of this study is therefore to explore, using virtual species, the efficacy of different filtering distances in recovering an unbiased distribution under different modelling, bias, ecology and sampling conditions. Through simulations of species with different levels of specialisation, sampled with various effort and different bias intensities, we assessed how ENMs trained from biased occurrence data, and corrected using distinct spatial filtering distances, can generate predictions that align with the original species niche. By employing a virtual ecologist approach, this study aims to offer for the first time an evaluation of the most suitable filtering distance for correcting sampling bias across contrasting ecological contexts.

2 | MATERIALS AND METHODS

2.1 | Acquisition of climatic variables

For this work, which was carried out entirely using the R software (R Core Team, 2020), we first compiled various climatic data downloaded from the WorldClim database (http://www.worldclim.org). Each climate variable corresponds here to a raster of 2.5 arc minutes resolution (approximately 4.63 km at the equator), that we cropped at the extent of Europe. We chose the variables bio1 and bio12, representing mean annual temperature and annual precipitation, respectively, to define the virtual species' niche. In addition, we also obtained six additional variables that were used for fitting MaxEnt models (see Section 2.4): temperature seasonality (bio4), maximum temperature of the warmest month (bio5), minimum temperature of the coldest month (bio6), precipitation of the wettest month (bio13), precipitation of the driest month (bio14), and precipitation seasonality (bio15).

2.2 | Creation of virtual species

To obtain reference species for which their true niche is known, we simulated two terrestrial European species, defined solely in terms of their response to two climatic variables (bio1 and bio12) (see Bazzichetto et al., 2023 for a similar approach). Two types of virtual species were established via the *virtualspecies* R package (Leroy et al., 2016): a generalist species and a specialist species, defined as such by their response curves (Figure 1).

The response function of the generalist species to temperature (bio1) followed a Gaussian distribution with a mean value (i.e., the optimal temperature conditions) approaching the European mean temperature, that is, 10°C, and a standard deviation of 10°C. The response function for precipitation (bio12) was logistic, with an inflection point at 800 mm per year, once again close to the European average, and a parameter *alpha* of −125, which conditions the smoothness of the transition between low and high suitability (Figure 1).

The response functions of the specialist species to climatic variables were identical to the previous one, but with a standard deviation of only 2°C for temperature and an alpha value equals to −25 for the logistic response to precipitation (Figure 1). As a result, the specialist species as a narrower tolerance to temperature and its suitability declines rapidly when confronted to dry conditions. It also results in a smaller range and a lower prevalence.

A raster of environmental suitability was obtained for each species by summing the responses to each variable. This was then converted into a presence–absence raster using suitability values as sampling probabilities in each grid cell (Figure 1), thus avoiding defining a fixed threshold to convert suitability into presence–absence.

2.3 | Biased sampling and spatial filtering

We simulated the process of occurrence sampling by randomly drawing 20, 200 and 2000 grid cells corresponding to 'presences' from the presence–absence rasters, as well as the same number of absences. In addition to a true random sampling, we created biased samples simulating real situations in which the sampling was not representative of the actual distribution of the species.

We defined first a bias (referred to as 'accessibility' later on) using a raster of road density (from highways to local roads) obtained from the Global Roads Inventory Project (Meijer et al., 2018) and hosted by the Food and Agriculture Organisation of the United Nations (https://data.apps.fao.org). We simulated different bias intensities by raising the density values to variable exponents (^1, ^2, ^5 and ^10). These rasters were used as sampling weights, resulting in species samples that were biased in proportion to road density, which is typical of many real datasets that feature higher sampling effort at the proximity of roads because of better accessibility (Hughes et al., 2021; Phillips et al., 2009).

We defined a second bias (referred to as 'niche truncation' later on) where sampling was restricted to a portion of the virtual species' climatic niche. To do so, we discarded regions located in the 25% (>12°C and >652 mm), 50% (>7.5°C and >558 mm) and 75% (>2.5°C and >368 mm) highest values of climate conditions. Such niche truncation happens when there is a large imbalance between regions in terms of sampling effort or data availability (e.g. Fourcade et al., 2013).

For each replicate, we tested filtering distances from 0 to 1000 km, with steps of 10 km between 0 and 200 km, and then steps of 50 km from 200 to 1000 km. Spatial filtering was implemented using the *spThin* R package (Aiello-Lammens et al., 2015), which randomly removes data until it returns the maximum possible number of occurrences for a given filtering distance. Since the process involves removing data at random, we repeated the filtering algorithm and all subsequent analyses 10 times to establish the minimum and maximum outcomes for each distance tested.

2.4 | Ecological niche modelling

We used two complementary methods to measure the ability of different strategies of spatial filtering to help modelling the niche and distribution of species under different bias conditions. First, we fitted ENMs using logistic regressions Generalised Linear Models (GLM), based on presences and absences, sampled in equal number. To do so, we modelled species' probability of presence as a function of temperature (bio1) and precipitation (bio12), using a quadratic response function for the temperature variable and a linear response function for the precipitation variable, mimicking the processes that defined the virtual species' niches. Therefore, we simulated here a case in which the true variables and responses are known from the modeller, and absence data are available.

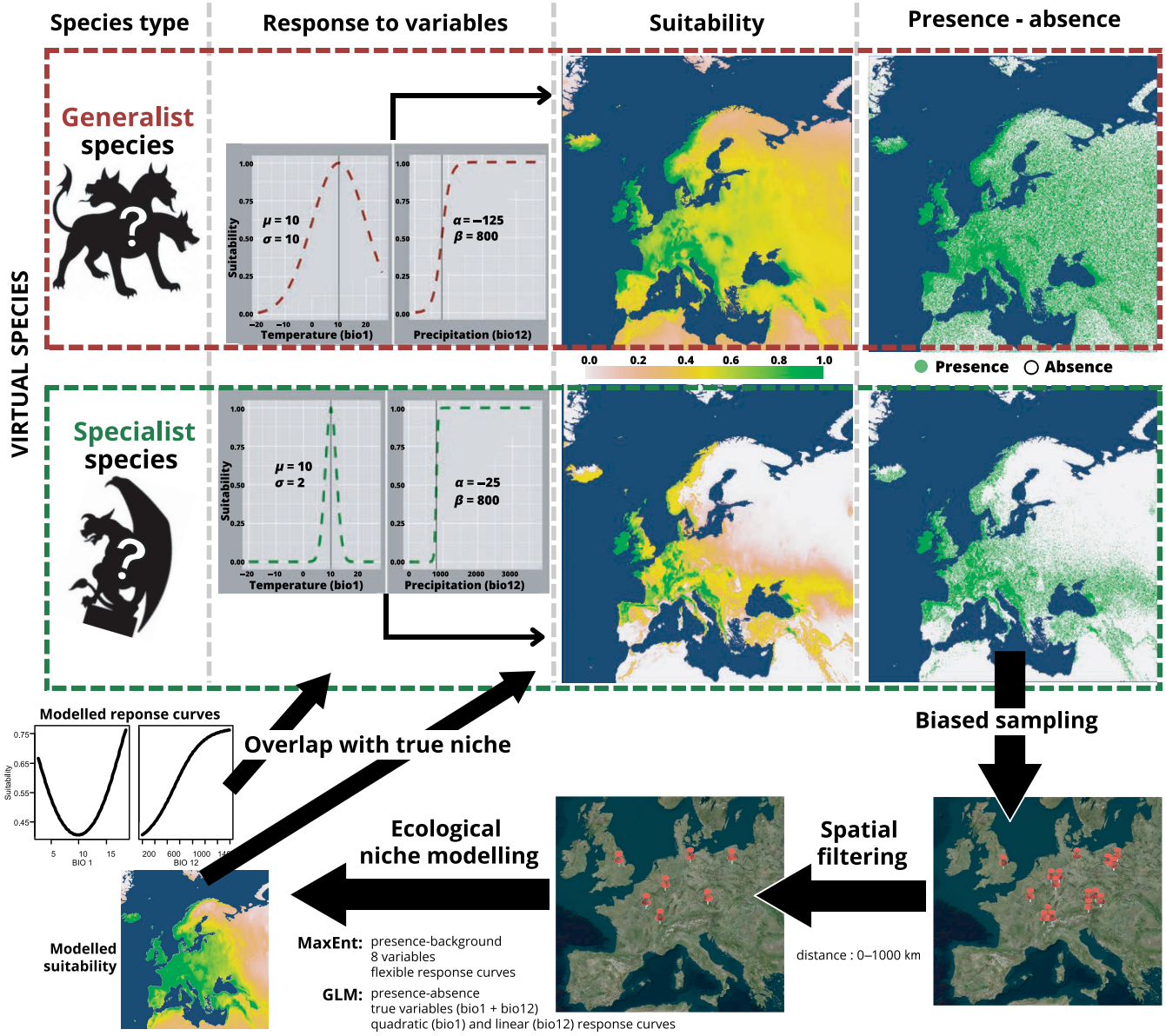

FIGURE 1 Framework used in this study: creation of virtual species by defining their responses to two climate variables, from which are derived a raster of suitability and a raster of presence–absence; biased sampling of presence and absences; application of a spatial filtering approach to reduce the bias (36 filtering distances up to 1000 km); ecological niche modelling to produce modelled response curves and suitability maps, which are then compared with the 'true' species' niche and distribution. Niche models from biased datasets were also compared with models produced from unbiased (i.e. random) samples to evaluate the ability of spatial filtering to mitigate the effect of sampling bias.

This case is, however, very rare in practice. Most often, only presence data are available (occurrences) and the response functions and variables involved would be unknown. Secondly, we thus fitted maximum entropy models (MaxEnt) (Phillips et al., 2006) with 10,000 background points randomly sampled across Europe and eight climate variables as predictors: mean annual temperature (bio1), temperature seasonality (bio4), maximum temperature of the warmest month (bio5), minimum temperature of the coldest month (bio6), annual precipitation (bio12), precipitation of the wettest month (bio13), precipitation of the driest month (bio14) and precipitation seasonality (bio15). Here, models were fitted with the 'maxnet' R package (Phillips et al., 2017), and we restricted feature classes to 'linear', 'quadratic' and 'hinge' and used a regularisation parameter of 2 to avoid overfitting. To check whether the different number of variables used in the MaxEnt and GLM models could influence our result, we also fitted MaxEnt models (six replicates) using the same two variables bio1 and bio12 that were used in the GLM and to create the virtual niche.

For both GLMs and MaxEnt models, we projected the modelled niche into the European climatic space, producing a map of predicted species' suitability ranging from 0 to 1.

TABLE 1 Performance of the unbiased models (trained with occurrences [MaxEnt] or presence–absence data [GLM] randomly sampled, that is, with no sampling bias) at modelling species' true distributions and response curves with temperature (bio1) and precipitations (bio12) variables.

		Similarity with true distribution		Similarity with true response curve	
Virtual species type	Number of sampled points	MaxEnt	GLM	Temperature	Precipitations
Generalist	20	0.913	0.864	0.763	0.702
	200	0.950	0.925	0.762	0.702
	2000	0.958	0.925	0.763	0.702
Specialist	20	0.625	0.628	0.767	0.702
	200	0.660	0.723	0.766	0.531
	2000	0.683	0.764	0.765	0.574

Note: Model's ability to recover species' distribution was evaluated with the Schoener's *D* niche overlap index between modelled and true suitability maps, and model's ability to recover response curves was assessed by directly comparing the true response curves with the modelled responses obtained from the GLMs (see Section 2.5; Figure S1); both indices range from 0 (totally different) to 1 (perfectly similar).

2.5 | Evaluating unbiased, biased and corrected niche models

Because we used virtual species, we could directly compare niche model outputs to the true species' distributions without relying on model evaluation metrics. First, we assessed how well niche models trained with unbiased data were able to predict species' environmental suitability by comparing modelled suitability maps with their true suitability, using the Schoener's *D* index ($D_{\text{unbiased vs. true}}$). This index of niche overlap, which lies between 0 and 1, represents the similarity between two rasters and is recommended to compare predictions of niche models (Rödder & Engler, 2011).

Then, we used each of these unbiased models (two types of species × three sampling efforts × two modelling methods = 12 models) as reference benchmarks against which we compared the suitability maps produced from biased datasets and corrected with various distances of spatial filtering (0 corresponding to uncorrected datasets). We computed Schoener's *D* between the outputs of niche models trained with biased (corrected) datasets and the true suitability ($D_{\text{biased vs. true}}$), and reported the percent difference between this value and the $D_{\text{unbiased vs. true}}$ value previously obtained for the same type of model (i.e., same species, sampling effort and modelling method). Thus, the obtained estimate of model performance is <0 when the models trained from biased datasets are less performant at recovering the true distribution than the same models trained from unbiased datasets. It reaches 0 when the spatial filtering has perfectly corrected for sample bias.

In addition, we compared the response curves obtained from the GLMs to the initial functions of each climate variable as defined in the virtual species. After every GLMs, we retrieved the regression coefficients obtained for each of the two climate variables and used them to plot their response curve normalised between 0 and 1. Normalising allowed us to compare each modelled curve with the initial curves of temperature and precipitation variables we used to create the two virtual species. At each integer value of temperature and precipitation, we calculated a proportion of similarity between the predicted and true suitability, which we averaged to produce a new similarity index ranging between 0 and 1. We reported the percent difference between this index as produced by models in the absence of bias and for models fitted with biased (and corrected) datasets, for each replicate. The procedure is illustrated in Figure S1.

3 | RESULTS

3.1 | Performance in recovering species' true distributions

3.1.1 | Unbiased models

For both modelling algorithms, in the absence of bias and for both species, the value of $D_{\text{unbiased vs. true}}$, that is, the similarity between the modelled and true distributions, increases with sampling effort (here the number of data points used), the maximum performance being systematically achieved for 2000 occurrences (Table 1). However, niche models obtained for the specialist species were less effective at recovering the true species' niches (min $D_{\text{unbiased vs. true}}$ = 0.625; max $D_{\text{unbiased vs. true}}$ = 0.764) than those obtained for the generalist species (min $D_{\text{unbiased vs. true}}$ = 0.864; max $D_{\text{unbiased vs. true}}$ = 0.958). MaxEnt models appeared to obtain a better performance for the generalist species, while GLMs performed better for the specialist species.

3.1.2 | Biased and uncorrected models

Biased sampling resulted, for both accessibility and niche truncation biases, in a reduction of the performance of niche models applied to the generalist species (see Figure 2 with spatial filtering distance = 0). The reduction in performance was stronger when the

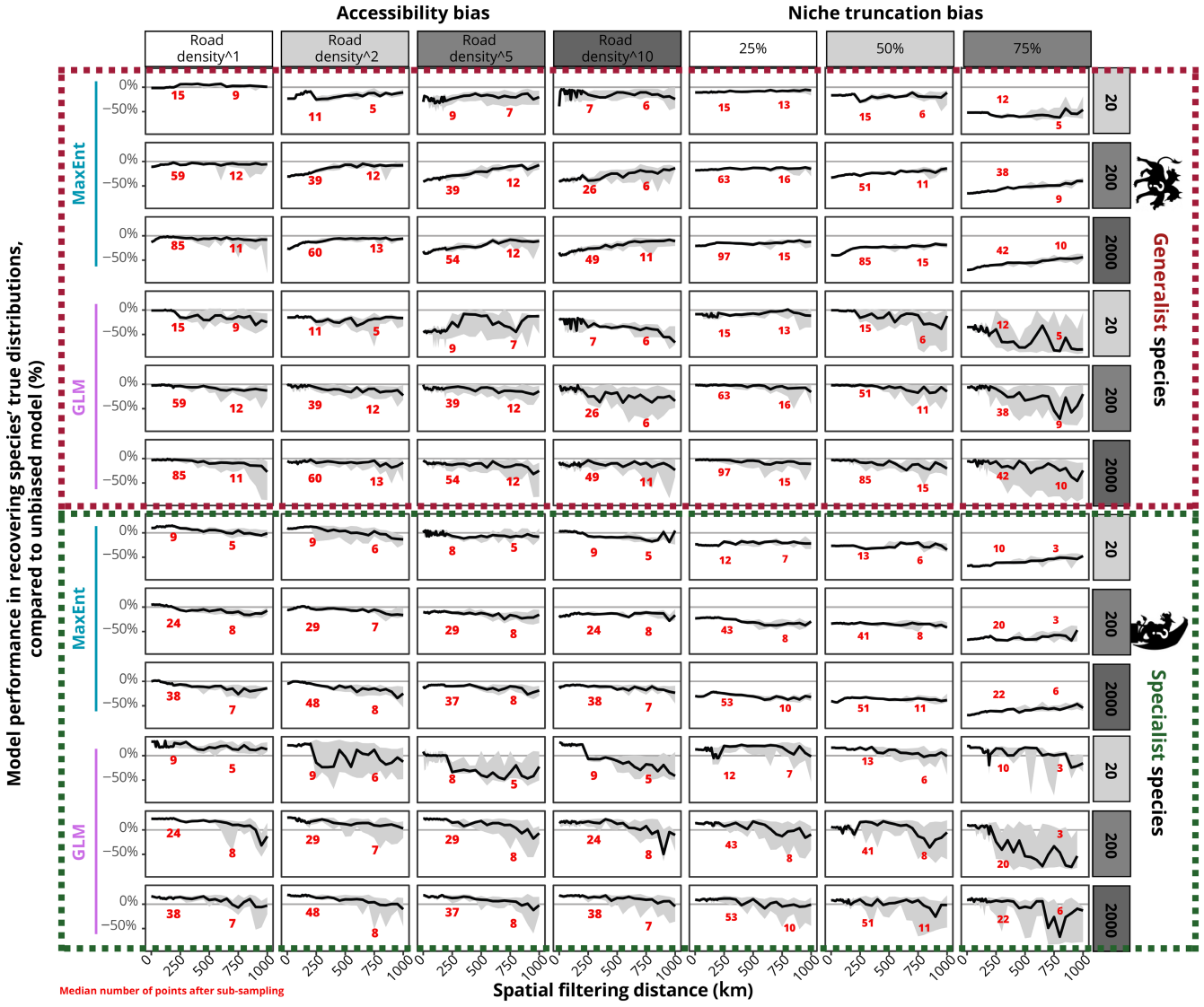

FIGURE 2 Performance of niche models trained with biased datasets in recovering the true distribution of virtual species, corrected with various spatial filtering distances (models with no correction correspond to a filtering distance = 0). Results are shown for various modelling methods (presence-background MaxEnt and presence–absence GLM), different bias types (accessibility derived from road density and niche truncation) at different intensities, and different sample sizes (20, 200 or 2000 species occurrences sampled in the virtual species' niche). Model performance is expressed as a percent difference between models computed from the biased datasets and from the corresponding models fitted with unbiased (i.e. randomly sampled) datasets, and is based on Schoener's *D* niche overlap with the true suitability maps of virtual species; hence, values <0 show models that perform worse than the unbiased models. Since the spatial filtering approach involves random sampling, it is repeated 10 times; results show the median values as a solid black line, along the minimum and maximum values as grey ribbons. The number of presence points available for modelling after spatial filtering is presented for two filtering distances (250 and 750 km) as red numbers (Figure S2).

MaxEnt method was used rather than GLMs, and with higher bias intensities and larger number of samples.

For the specialist species, we also observed that MaxEnt models trained with biased data showed a decrease in performance, except in a few cases (accessibility bias, low bias and few samples) where model performance increased with the bias. Unexpectedly, GLMs trained on the specialist species with biased and uncorrected datasets had always a better performance than the same models trained using an unbiased sample (Figure 2).

3.1.3 | Biased and corrected models

In all cases, increasing the filtering distance reduced the number of occurrences available for modelling. This reduction is proportional to the initial sampling effort, and led to roughly the same number of remaining occurrences for a given filtering distance, regardless of the number of initial sampling points (Figure S2). For the strongest biases, it means that large spatial filtering distances resulted in less than 10 occurrences (Figure S2; Figure 2).

Spatial filtering almost never improved the performance of GLMs (Figure S3). On the contrary, it often decreased the ability of models to recover the original species distributions compared to the corresponding unbiased models. This reduction in model performance was stronger with increasing filtering distances (Figure 2).

The effect of spatial filtering on MaxEnt models was more variable depending on species type and bias type/intensity. When applied to the generalist species, spatial filtering resulted in improved model performance (Figure S3), especially for strong biases and large datasets (200 or 2000 occurrences). We did not observe an optimal filtering distance, as model improvement increased roughly linearly with increasing distance (Figure 2), although it reached a plateau for some cases of intermediate bias intensity (e.g. second level of accessibility bias with 2000 occurrences). Spatial filtering had little impact on model performance for small datasets biased with low intensity.

For the specialist species, applying spatial filtering resulted in a decrease in the performance of MaxEnt models, which was stronger with increasing spatial filtering distance. For the niche truncation bias with the highest intensity (i.e. when 75% of the species' niche is unobserved), increasing spatial filtering distance contributed to increase model performance with no observable plateau (Figure 2). In all cases where spatial filtering helped in bias correction, model performance remained lower than that with the unbiased dataset.

3.2 | Performance in recovering species' true response curves

Using an unbiased dataset, the similarity between response curves obtained by the GLMs and the original responses differed only little between species and sampling efforts. Still, we observed a clear decrease for the specialist species modelled with 200 and 2000 occurrences, for the precipitation variable only (Table 1).

When presence–absence data were sampled with a bias, and not corrected by spatial filtering, the temperature response curves were only little affected by the bias (Figure 3). Accordingly, spatial filtering did not contribute to improvements in models' ability to recover the true response, and occasionally made it worse.

In contrast, sampling bias made the modelled responses to precipitations sometimes closer (e.g. specialist species with strong biases and large sample sizes) or further away (generalist species sampled with the 25% niche truncation and low sample size) from the species' real response curve than unbiased models (Figure 3). The effect of spatial filtering on models performance in recovering the true response curves to precipitation was very uncertain, with seemingly random improvements or performance loss depending on filtering distance, or even within a given distance (Figure 3).

4 | DISCUSSION

4.1 | How do unbiased ENMs perform on different datasets?

Beyond the main objective of this study, which was to evaluate the effect of spatial filtering distances on sampling bias mitigation, our virtual ecologist approach (Figure 1) yielded valuable insights into the ability of ENMs to predict species distributions. We observed that unbiased models applied to the generalist species consistently exhibited higher performance than those fitted to the specialist species, regardless of the method used (Table 1). It suggests that species with a high degree of specialisation are more difficult to model, which contrasts with prior research that established on the contrary that niche models applied to habitat specialists were more accurate (Hallman & Robinson, 2020; McCune et al., 2020; McPherson & Jetz, 2007; Tessarolo et al., 2021), although the opposite has also been shown (Inman et al., 2021; Soultan & Safi, 2017). This may be explained by the fact that these earlier studies did not make a direct comparison of model outputs against the species' known niche. There are also theoretical underpinnings that may explain why we failed to reach the same level of performance for specialists compared with generalist species. Indeed, ENMs benefit from using background or absence data sampled in proximity to species presence, ideally within the species' accessible area (Barve et al., 2011). Here, we sampled instead the entire European region for both generalist and specialist species, which may be less appropriate for species whose distribution is limited to restricted areas (Araújo & Guisan, 2006).

It appeared that a larger number of sampling points led to more accurate predictions of species' distributions. Such pattern has been shown before (Hallman & Robinson, 2020), and is intuitive because larger sampling means greater knowledge of the ecological niche of the species studied (Boyd et al., 2023). However, it has often been suggested that small sample sizes—as low as ca. 10 occurrences—were enough to reach good model performance, and that increasing sample size beyond a few dozens or hundreds was unnecessary, especially for specialist species (Boria & Blois, 2018; Stockwell & Peterson, 2002; van Proosdij et al., 2016). Our results show a clear increase in model performance between unbiased models fitted with 20 and 200 occurrences, less so between 200 and 2000 occurrences (Table 1), suggesting both that a couple dozen occurrences may be too few, and that increasing sample size up to thousands of data points may be useless.

4.2 | What is the effect of sampling bias on the performance of ENMs?

Gaul et al. (2020) showed that sample size was a more important predictor of ENM accuracy than the spatial bias in the training data. Here, we show that these factors interact, leading to a larger

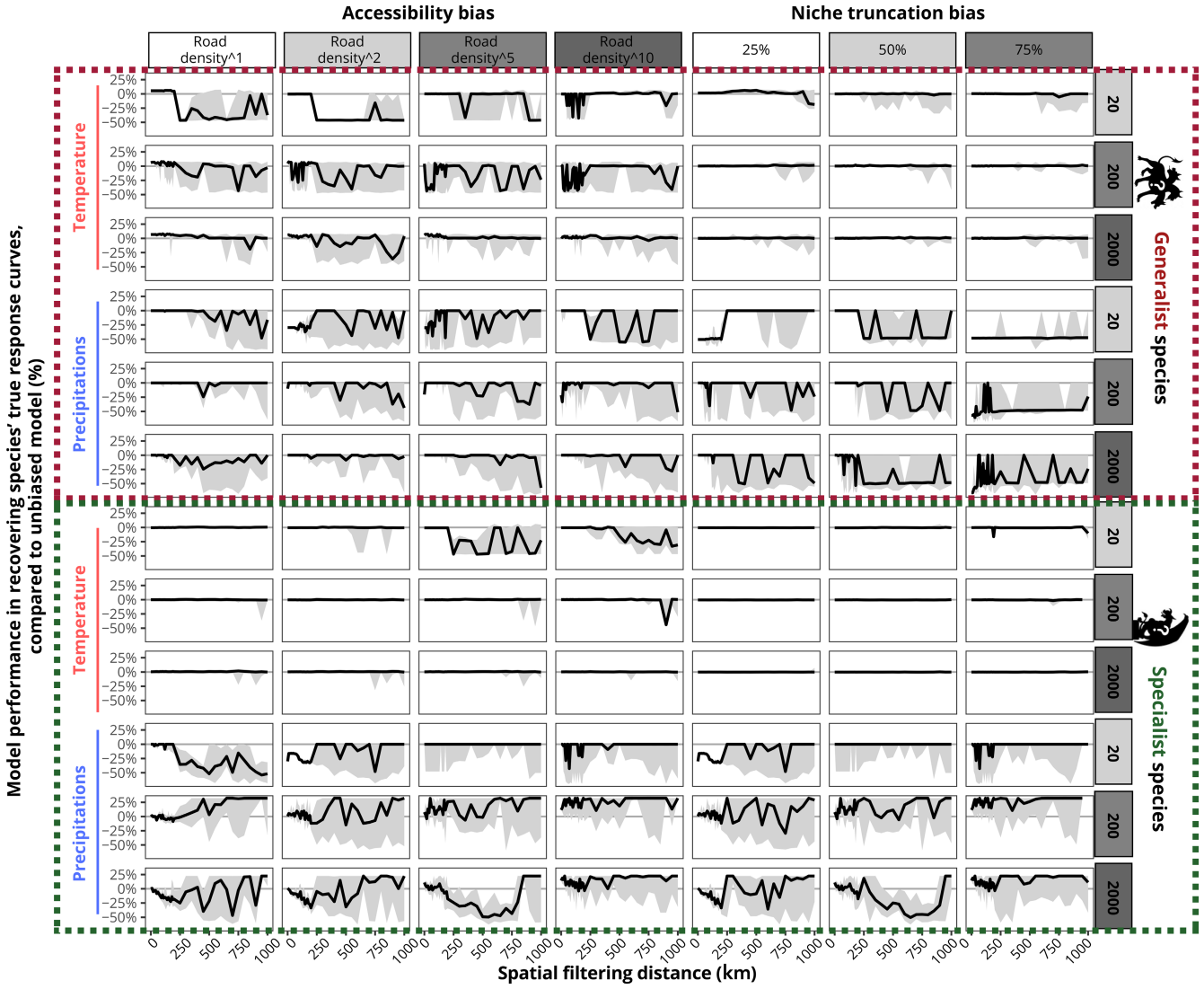

FIGURE 3 Performance of generalised linear models trained with biased datasets in recovering the true response curves of virtual species, corrected with various spatial filtering distances. Model performance is expressed as a percent difference between models computed from the biased and unbiased datasets, and is based on a comparison between the modelled and true response curves, for the temperature and precipitations variables (see Section 2.5; Figure S1). See Figure 2 for full legend details.

reduction in model performance when large datasets are biased, compared to smaller datasets biased in the same way (Figure 2). A similar result was observed by Bean et al. (2012), who suggested that small samples, even if they were biased, prevented overfitting and thus allowed models to extrapolate species suitability beyond the sampled area. In this regard, spatial filtering, which reduces the amount of data while smoothing them in the geographical spaces, may provide an interesting solution for mitigating the effect of sampling bias in large datasets.

It is remarkable that some biased datasets led to predictions of species distribution that were closer to the true distribution (as defined by our virtual species) than the same type of model fitted to an unbiased dataset. This situation was found almost exclusively for the specialist species modelled using a GLM, here fitted with presence–absence data and the true variables that constrained the species' niche. This outcome appears counterintuitive at first glance. However, for a highly specialised species, a concentration of samples with 100% location accuracy (Naimi et al., 2014)—since they are sampled from the true presence–absence raster—in a small region may be more effective at representing the subtle details of the niche than a random sample distributed across a large area (Araújo & Guisan, 2006).

We observed that ENMs fitted with the MaxEnt method were more affected by the bias than the GLMs (they were also less performant in modelling the specialist species with unbiased data, see Table 1). This could be because of the fact that MaxEnt models used presence-background data, complex response functions and a larger set of climate variables (including irrelevant ones), contrary to our GLMs that were fitted with parameters that were as close as possible to the true species' niche (Brotons et al., 2004). However, results obtained for MaxEnt models fitted with the same two variables as GLMs were highly correlated with those obtained for the

model with eight variables, suggesting that variable choice did not cause these differences (Figure S4). Modellers are usually unaware of the important variables, must accommodate presence-only data, and thus rely on flexible machine-learning methods such as MaxEnt, but also random forest or boosted regression trees algorithms (Elith et al., 2006). Therefore, although our simulations conducted using GLMs are important from a theoretical point of view, we consider that they provide little insights into the effect of bias and spatial filtering for real-life applications.

4.3 | At what scale does spatial filtering improve the performance of biased ENMs?

Our spatial filtering approach proved to get vastly different effectiveness depending on the modelling settings (Figure 2; Figure S3). With the GLM algorithm, which was already little impacted by the bias, sub-sampling the data appeared to only reduce the accuracy of the models, or at best to have no effect on model performance. The results obtained with the MaxEnt algorithm, on the other hand, showed that when the bias is strong and the sample size is large, spatial filtering can help improving the model. Spatial filtering has been tested previously with MaxEnt models, because of their prevalence in the ENM literature, and was frequently found to be effective in mitigating sampling bias in the input data (e.g. Boria et al., 2014; Fourcade et al., 2014; Kramer-Schadt et al., 2013; Radosavljevic & Anderson, 2014, but see Ten Caten & Dallas, 2023). Under which circumstances it performs best remained an open question. Using a fixed filtering distance of 15 km, Inman et al. (2021) found, like us, that sampling bias correction was positively correlated with the intensity of the underlying bias and that it performed better on generalist species. While it was suggested that spatial filtering could be appropriate for dealing with sampling bias in small datasets (e.g. Galante et al., 2018), our simulations using virtual species suggest that this strategy is mostly effective for large sample sizes. Similarly, Kramer-Schadt et al. (2013) recommended spatial filtering to be employed for large datasets only, as the approach necessarily removes part of the dataset.

The novelty of our study was that we explored a range of spatial filtering distances to determine the best distance to thin a biased occurrence dataset in different ecological contexts. Our hypothesis was that model performance would exhibit an initial improvement as the filtering distance increased, followed by a subsequent decline once the loss of information outweighed the benefits of bias mitigation, essentially observing a bell-shaped curve with an optimal filtering distance at its top. Unexpectedly, none of our simulation parameters produced this pattern. Instead, instances where spatial filtering effectively aided in mitigating sampling bias demonstrated a predominantly linear and positive relationship between filtering distance and model performance improvement. This enhancement in performance was still insufficient to completely counterbalance the impact of bias. Therefore, our findings not only revealed that no filtering distance succeeded in producing a model as good as that produced with unbiased data, but also highlighted that the accuracy of the biased model could still be improved up to a filtering distance of 1000 km. However, very few (less than 10) occurrences remained at the large filtering distances that led to the highest improvements. Such a small dataset is most likely not enough to produce accurate models of species distributions. Across biases and species types, MaxEnt models fitted to data filtered with a distance of 1000 km had a mean overlap with the true suitability ($D_{\text{biased vs. true}}$) of 0.66 (SD = 0.16) only (Figure S5), denoting predictions that were quite far from reality, even though they constitute improvements compared to the biased models.

In light of our findings, it becomes evident that the strategy of spatially filtering biased occurrence data is not universally as successful as previous studies suggested (Boria et al., 2014; Fourcade et al., 2014; Inman et al., 2021; Kramer-Schadt et al., 2013; Radosavljevic & Anderson, 2014). We demonstrated that this approach leads to a reduction in model performance when the model already incorporates a substantial amount of data and knowledge about the species being modelled. This includes absence data, insights into crucial variables, and an understanding of the shape of their responses, such as in the GLMs we simulated. In a more realistic application of ENM methods, that is, using presence-background data and a flexible algorithm fitted with multiple variables, we found that spatial filtering improves model performance in cases where the bias is so strong that the resulting models may be of little use even after correction. Recently, Ten Caten and Dallas (2023) used real data and simulations to test filtering distances up to 128 km, and reached the same conclusion that 'thinning occurrence points does not improve SDM performance'.

Our study does not prove the existence of an optimal filtering distance that could definitely solve the problem of sampling bias in ENMs. Consequently, despite the current prevalence of spatial filtering in many modelling routines, as evidenced by its implementation in various ENM workflows (e.g. Dobson et al., 2023; Kass et al., 2018; Velazco et al., 2022), its ability to effectively address sampling bias remains uncertain. Failure to improve model performance is especially clear for specialist species sampled with low intensity, where spatial filtering sometimes decreases the ability of MaxEnt models to predict species distributions (see Ten Caten & Dallas, 2023 for a similar conclusion). In this case, any attempt to filter the input data runs the risk of removing key data points that were crucial for modelling the species' niche. It may be then advisable to switch instead to methods of background manipulation that do not contribute to information loss (Barber et al., 2021; Dubos et al., 2022; Phillips et al., 2009; Ranc et al., 2017; Vollering et al., 2019). Studies that tested environmental filtering along with spatial filtering also concluded that filtering data in the environmental space could lead to better performance (Varela et al., 2014) and less information loss (Castellanos et al., 2019).

All the aspects discussed so far concern the capacity of ENMs, unbiased, biased and corrected to model the true distribution of

habitat suitability of virtual species. In addition, we aimed to complement this analysis with an assessment of model performance at recovering the shape of response curves, such as in Bazzichetto et al. (2023) or Inman et al. (2021). We carried out this analysis for the GLMs only, since they were calibrated in such a way that they could directly model the true response of both variables involved in the definition of virtual species (although regression on the temperature variable was not carried out with the original function but simply by a quadratic approximation). Surprisingly, we found little correlation between GLMs' ability to model response curves and to predict species distributions (Figure S6). Several methods of comparing the modelled and true response curves have been tried, such as by comparing the area under the curves, which gave similar results as the method we used (calculating the average distance between points on the curves). Given that there was no clear pattern of model improvement across spatial filtering distances, it is once again impossible to use this approach to recommend an optimal filtering distance.

5 | CONCLUSION

Despite two decades of rapid methodological advances, modelling ecological niches remains a challenge for biogeographers. This is made even more difficult by the existence of numerous correlates of sampling effort, such as accessibility or population densities, which generate biases in the available data. In the current context of global changes and biodiversity crisis, it is crucial to be able to handle this bias, particularly when models are used for the purpose of delineating protected areas or managing threatened species. In this study, we aimed to identify the optimal distance for filtering biased occurrence data in different contexts, a strategy that is frequently employed despite the absence of guidelines to select that distance. Clearly, we failed in this regard. We are confident that although there is always ground for improvement, our methodology was robust and adequate to find this optimal filtering distance. Instead, our results suggest that such an optimal filtering distance may not actually exist. The spatial filtering approach appears to yield little benefit when the initial bias is low, and it struggles to sufficiently mitigate strong sampling biases. Still, we highlighted an apparent interaction between species traits (here climate specialisation), the strength of the bias and sample size in the ability of spatial filtering to correct for sampling bias. In light of these results, we recommend spatial filtering to be employed—carefully—only when enough data are available, and to explore alternative options of sampling bias correction for small sample sizes.

ACKNOWLEDGEMENTS

We thank M. Nicolas Dubos and an anonymous reviewer for their helpful comments. The authors did not receive any funding for conducting this study. Being based on virtual data, no permit was required to conduct this work.

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflict of interest.

DATA AVAILABILITY STATEMENT

The results obtained from the simulations are hosted in the Figshare repository (https://doi.org/10.6084/m9.figshare.24032196).

ORCID

Quentin Lamboley https://orcid.org/0009-0004-1382-1382
Yoan Fourcade https://orcid.org/0000-0003-3820-946X

REFERENCES

Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., & Anderson, R. P. (2015). spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, *38*(5), 541–545. https://doi.org/10.1111/ecog.01132

Araújo, M. B., & Guisan, A. (2006). Five (or so) challenges for species distribution modelling. *Journal of Biogeography*, *33*(10), 1677–1688. https://doi.org/10.1111/j.1365-2699.2006.01584.x

Baker, D. J., Maclean, I. M. D., & Gaston, K. J. (2024). Effective strategies for correcting spatial sampling bias in species distribution models without independent test data. *Diversity and Distributions*, *30*(3), e13802. https://doi.org/10.1111/ddi.13802

Baker, D. J., Maclean, I. M. D., Goodall, M., & Gaston, K. J. (2022). Correlations between spatial sampling biases and environmental niches affect species distribution models. *Global Ecology and Biogeography*, *31*(6), 1038–1050. https://doi.org/10.1111/geb.13491

Barber, R. A., Ball, S. G., Morris, R. K. A., & Gilbert, F. (2021). Target-group backgrounds prove effective at correcting sampling bias in Maxent models. *Diversity and Distributions*, *28*(1), 128–141. https://doi.org/10.1111/ddi.13442

Barve, N., Barve, V., Jimenez-Valverde, A., Lira-Noriega, A., Maher, S. P., Peterson, A. T., Soberón, J., & Villalobos, F. (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, *222*(11), 1810–1819. https://doi.org/10.1016/j.ecolmodel.2011.02.011

Bazzichetto, M., Lenoir, J., Da Re, D., Tordoni, E., Rocchini, D., Malavasi, M., Barták, V., & Sperandii, M. G. (2023). Sampling strategy matters to accurately estimate response curves' parameters in species distribution models. *Global Ecology and Biogeography*, *32*, 1717–1729. https://doi.org/10.1111/geb.13725

Bean, W. T., Stafford, R., & Brashares, J. S. (2012). The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. *Ecography*, *35*, 250–258. https://doi.org/10.1111/j.1600-0587.2011.06545.x

Beck, J., Böller, M., Erhardt, A., & Schwanghart, W. (2014). Spatial bias in the GBIF database and its effect on modelling species' geographic distributions. *Ecological Informatics*, *19*, 10–15. https://doi.org/10.1016/j.ecoinf.2013.11.002

Boria, R. A., & Blois, J. L. (2018). The effect of large sample sizes on ecological niche models: Analysis using a North American rodent, *Peromyscus maniculatus*. *Ecological Modelling*, *386*, 83–88. https://doi.org/10.1016/j.ecolmodel.2018.08.013

Boria, R. A., Olson, L. E., Goodman, S. M., & Anderson, R. P. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, *275*, 73–77. https://doi.org/10.1016/j.ecolmodel.2013.12.012

Bowler, D. E., Callaghan, C. T., Bhandari, N., Henle, K., Benjamin Barth, M., Koppitz, C., Klenke, R., Winter, M., Jansen, F., Bruelheide, H., & Bonn, A. (2022). Temporal trends in the spatial bias of species

occurrence records. *Ecography*, *8*, e06219. https://doi.org/10.1111/ecog.06219

Boyd, R. J., Harvey, M., Roy, D. B., Barber, T., Haysom, K. A., Macadam, C. R., Morris, R. K. A., Palmer, C., Palmer, S., Preston, C. D., Taylor, P., Ward, R., Ball, S. G., & Pescott, O. L. (2023). Causal inference and large-scale expert validation shed light on the drivers of SDM accuracy and variance. *Diversity and Distributions*, *29*(6), 774–784. https://doi.org/10.1111/ddi.13698

Brotons, L., Thuiller, W., Araújo, M. B., & Hirzel, A. H. (2004). Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. *Ecography*, *4*, 437–448.

Bystriakova, N., Peregrym, M., Erkens, R. H. J., Bezsmertna, O., & Schneider, H. (2012). Sampling bias in geographic and environmental space and its effect on the predictive power of species distribution models. *Systematics and Biodiversity*, *10*(3), 1–11. https://doi.org/10.1080/14772000.2012.705357

Castellanos, A. A., Huntley, J. W., Voelker, G., & Lawing, A. M. (2019). Environmental filtering improves ecological niche models across multiple scales. *Methods in Ecology and Evolution*, *10*(4), 481–492. https://doi.org/10.1111/2041-210X.13142

Correia, R. A., Ruete, A., Stropp, J., Malhado, A. C. M., dos Santos, J. W., Lessa, T., Alves, J. A., & Ladle, R. J. (2019). Using ignorance scores to explore biodiversity recording effort for multiple taxa in the Caatinga. *Ecological Indicators*, *106*, 105539. https://doi.org/10.1016/j.ecolind.2019.105539

Daru, B. H., & Rodriguez, J. (2023). Mass production of unvouchered records fails to represent global biodiversity patterns. *Nature Ecology & Evolution*, *7*, 816–831. https://doi.org/10.1038/s41559-023-02047-3

Dobson, R., Challinor, A. J., Cheke, R. A., Jennings, S., Willis, S. G., & Dallimer, M. (2023). dynamicSDM: An R package for species geographical distribution and abundance modelling at high spatiotemporal resolution. *Methods in Ecology and Evolution*, *14*(5), 1190–1199. https://doi.org/10.1111/2041-210X.14101

Dubos, N., Préau, C., Lenormand, M., Papuga, G., Monsarrat, S., Denelle, P., Louarn, M. L., Heremans, S., May, R., Roche, P., & Luque, S. (2022). Assessing the effect of sample bias correction in species distribution models. *Ecological Indicators*, *145*, 109487. https://doi.org/10.1016/j.ecolind.2022.109487

Dudík, M., Phillips, S. J., & Schapire, R. E. (2007). Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *Journal of Machine Learning Research*, *8*, 1217–1260.

Edwards, J. L. (2004). Research and societal benefits of the Global Biodiversity Information Facility. *Bioscience*, *54*(6), 485–486. https://doi.org/10.1641/0006-3568(2004)054[0486:RASBOT]2.0.CO;2

Elith, J., Graham, C. H., Anderson, R. P., Dudík, M., Ferrier, S., Guisan, A., Hijmans, R. J., Huettmann, F., Leathwick, J. R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B. A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, J. M., Townsend Peterson, A., … Zimmermann, N. E. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*(2), 129–151. https://doi.org/10.1111/j.2006.0906-7590.04596.x

Elith, J., Kearney, M., & Phillips, S. J. (2010). The art of modelling range-shifting species. *Methods in Ecology and Evolution*, *1*(4), 330–342. https://doi.org/10.1111/j.2041-210X.2010.00036.x

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, *40*(1), 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Fourcade, Y., Engler, J. O., Besnard, A. G., Rödder, D., & Secondi, J. (2013). Confronting expert-based and modelled distributions for species with uncertain conservation status: A case study from the corncrake (*Crex crex*). *Biological Conservation*, *167*, 161–171. https://doi.org/10.1016/j.biocon.2013.08.009

Fourcade, Y., Engler, J. O., Rödder, D., & Secondi, J. (2014). Mapping species distributions with MAXENT using a geographically biased sample of presence data: A performance assessment of methods for correcting sampling bias. *PLoS One*, *9*(5), e97122. https://doi.org/10.1371/journal.pone.0097122

Galante, P. J., Alade, B., Muscarella, R., Jansa, S. A., Goodman, S. M., & Anderson, R. P. (2018). The challenge of modeling niches and distributions for data-poor species: A comprehensive approach to model complexity. *Ecography*, *41*(5), 726–736. https://doi.org/10.1111/ecog.02909

Garcia-Rosello, E., Gonzalez-Dacosta, J., Guisande, C., & Lobo, J. M. (2023). GBIF falls short of providing a representative picture of the global distribution of insects. *Systematic Entomology*, *48*, 489–497. https://doi.org/10.1111/syen.12589

Gaul, W., Sadykova, D., White, H. J., Leon-Sanchez, L., Caplat, P., Emmerson, M. C., & Yearsley, J. M. (2020). Data quantity is more important than its spatial bias for predictive species distribution modelling. *PeerJ*, *8*, e10411. https://doi.org/10.7717/peerj.10411

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

Hallman, T. A., & Robinson, W. D. (2020). Deciphering ecology from statistical artefacts: Competing influence of sample size, prevalence and habitat specialization on species distribution models and how small evaluation datasets can inflate metrics of performance. *Diversity and Distributions*, *26*(3), 315–328. https://doi.org/10.1111/ddi.13030

Hughes, A. C., Orr, M. C., Ma, K., Costello, M. J., Waller, J., Provoost, P., Yang, Q., Zhu, C., & Qiao, H. (2021). Sampling biases shape our view of the natural world. *Ecography*, *44*(9), 1259–1269. https://doi.org/10.1111/ecog.05926

Inman, R., Franklin, J., Esque, T., & Nussear, K. (2021). Comparing sample bias correction methods for species distribution modeling using virtual species. *Ecosphere*, *12*(3), e03422. https://doi.org/10.1002/ecs2.3422

Kass, J., Vilela, B., Aiello-Lammens, M., Muscarella, R., Merow, C., & Anderson, R. P. (2018). Wallace: A flexible platform for reproducible modeling of species niches and distributions built for community expansion. *Methods in Ecology and Evolution*, *9*, 1151–1156. https://doi.org/10.1111/2041-210x.12945

Kramer-Schadt, S., Niedballa, J., Pilgrim, J. D., Schröder, B., Lindenborn, J., Reinfelder, V., Stillfried, M., Heckmann, I., Scharf, A. K., Augeri, D. M., Cheyne, S. M., Hearn, A. J., Ross, J., Macdonald, D. W., Mathai, J., Eaton, J., Marshall, A. J., Semiadi, G., Rustam, R., … Wilting, A. (2013). The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions*, *19*(11), 1366–1379. https://doi.org/10.1111/ddi.12096

Leroy, B., Meynard, C. N., Bellard, C., & Courchamp, F. (2016). virtualspecies, an R package to generate virtual species distributions. *Ecography*, *39*, 599–607. https://doi.org/10.1111/ecog.01388

Mair, L., & Ruete, A. (2016). Explaining spatial variation in the recording effort of citizen science data across multiple taxa. *PLoS One*, *11*(1), e0147796. https://doi.org/10.1371/journal.pone.0147796

McCune, J. L., Rosner-Katz, H., Bennett, J. R., Schuster, R., & Kharouba, H. M. (2020). Do traits of plant species predict the efficacy of species distribution models for finding new occurrences? *Ecology and Evolution*, *10*(11), 5001–5014. https://doi.org/10.1002/ece3.6254

McPherson, J. M., & Jetz, W. (2007). Effects of species' ecology on the accuracy of distribution models. *Ecography*, *30*(1), 135–151. https://doi.org/10.1111/j.2006.0906-7590.04823.x

Meijer, J. R., Huijbregts, M. A. J., Schotten, K. C. G. J., & Schipper, A. M. (2018). Global patterns of current and future road infrastructure. *Environmental Research Letters*, *13*(6), 064006. https://doi.org/10.1088/1748-9326/aabd42

Merow, C., Smith, M. J., & Silander, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: What it does, and why inputs and settings matter. *Ecography*, *36*, 1058–1069. https://doi.org/10.1111/j.1600-0587.2013.07872.x

Naimi, B., Hamm, N. A. S., Groen, T. A., Skidmore, A. K., & Toxopeus, A. G. (2014). Where is positional uncertainty a problem for species distribution modelling? *Ecography*, *37*(2), 191–203. https://doi.org/10.1111/j.1600-0587.2013.00205.x

Phillips, S. J., Anderson, R. P., Dudík, M., Schapire, R. E., & Blair, M. E. (2017). Opening the black box: An open-source release of Maxent. *Ecography*, *40*, 887–893. https://doi.org/10.1111/ecog.03049

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190*(3–4), 231–259. https://doi.org/10.1016/j.ecolmodel.2005.03.026

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J. R., & Ferrier, S. (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. (2020). *R: A language and environment for statistical computing* [computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

Radosavljevic, A., & Anderson, R. P. (2014). Making better Maxent models of species distributions: Complexity, overfitting and evaluation. *Journal of Biogeography*, *41*(4), 629–643. https://doi.org/10.1111/jbi.12227

Ranc, N., Santini, L., Rondinini, C., Boitani, L., Poitevin, F., Angerbjörn, A., & Maiorano, L. (2017). Performance tradeoffs in target-group bias correction for species distribution models. *Ecography*, *40*(9), 1076–1087. https://doi.org/10.1111/ecog.02414

Rödder, D., & Engler, J. O. (2011). Quantitative metrics of overlaps in Grinnellian niches: Advances and possible drawbacks. *Global Ecology and Biogeography*, *20*(6), 915–927. https://doi.org/10.1111/j.1466-8238.2011.00659.x

Soultan, A., & Safi, K. (2017). The interplay of various sources of noise on reliability of species distribution models hinges on ecological specialisation. *PLoS One*, *12*(11), e0187906. https://doi.org/10.1371/journal.pone.0187906

Stockwell, D. R. B., & Peterson, A. T. (2002). Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, *148*(1), 1–13. https://doi.org/10.1016/S0304-3800(01)00388-X

Ten Caten, C., & Dallas, T. (2023). Thinning occurrence points does not improve species distribution model performance. *Ecosphere*, *14*(12), e4703. https://doi.org/10.1002/ecs2.4703

Tessarolo, G., Lobo, J. M., Rangel, T. F., & Hortal, J. (2021). High uncertainty in the effects of data characteristics on the performance of species distribution models. *Ecological Indicators*, *121*, 107147. https://doi.org/10.1016/j.ecolind.2020.107147

van Proosdij, A. S. J., Sosef, M. S. M., Wieringa, J. J., & Raes, N. (2016). Minimum required number of specimen records to develop accurate species distribution models. *Ecography*, *542–552*, 542–552. https://doi.org/10.1111/ecog.01509

Varela, S., Anderson, R. P., García-Valdés, R., & Fernández-González, F. (2014). Environmental filters reduce the effects of sampling bias and improve predictions of ecological niche models. *Ecography*, *37*(11), 1084–1091. https://doi.org/10.1111/j.1600-0587.2013.00441.x

Velazco, S. J. E., Rose, M. B., de Andrade, A. F. A., Minoli, I., & Franklin, J. (2022). flexsdm: An r package for supporting a comprehensive and flexible species distribution modelling workflow. *Methods in Ecology and Evolution*, *13*(8), 1661–1669. https://doi.org/10.1111/2041-210X.13874

Vollering, J., Halvorsen, R., Auestad, I., & Rydgren, K. (2019). Bunching up the background betters bias in species distribution models. *Ecography*, *42*(10), 1717–1727. https://doi.org/10.1111/ecog.04503

BIOSKETCHES

Quentin Lamboley is an MSc student who specialises in modelling approaches applied to ecology.

Yoan Fourcade is an associate professor at the University Paris-Est Créteil and at the Institute of Ecology and Environmental Sciences of Paris, who is interested in studying large scale patterns of biodiversity, including by using modelling methods for predicting future biodiversity changes.

Author Contributions: QL performed the analyses and wrote the first draft of the manuscript. YF conceived and supervised the study. Both authors contributed to the final version of the manuscript.

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: Lamboley, Q., & Fourcade, Y. (2024). No optimal spatial filtering distance for mitigating sampling bias in ecological niche models. *Journal of Biogeography*, *00*, 1–12. https://doi.org/10.1111/jbi.14854

Supporting information for:

No optimal spatial filtering distance for mitigating sampling bias in ecological niche models

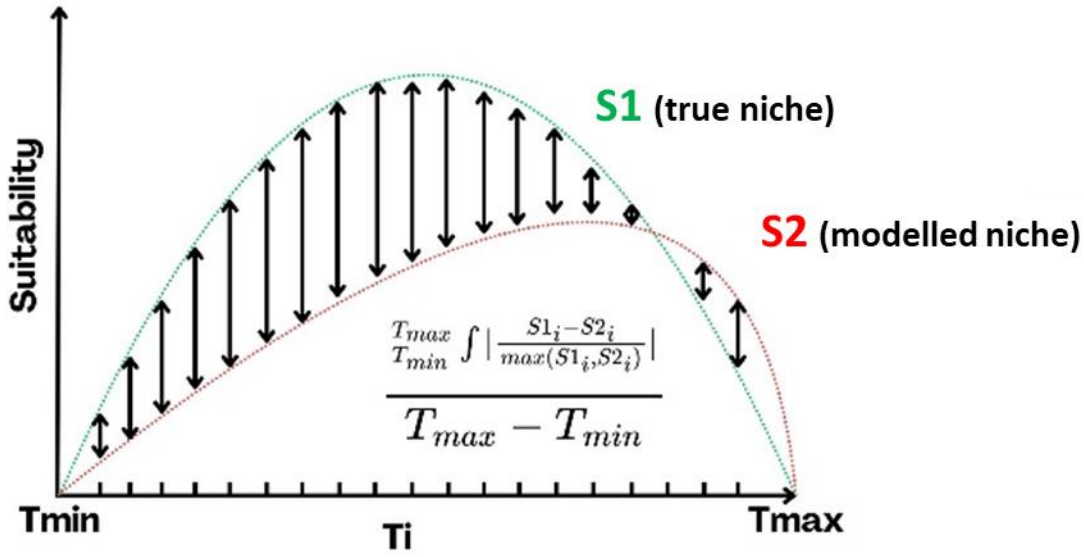

Figure S1: Illustration representing the calculation of GLMs' performance in recovering the true response curve. The distance between the true and modelled suitability is extracted along a gradient of temperature (or precipitation). At each integer value of the variable, the proportional difference between the predicted and true suitability is computed, and later averaged to produce a similarity index ranging between 0 and 1.

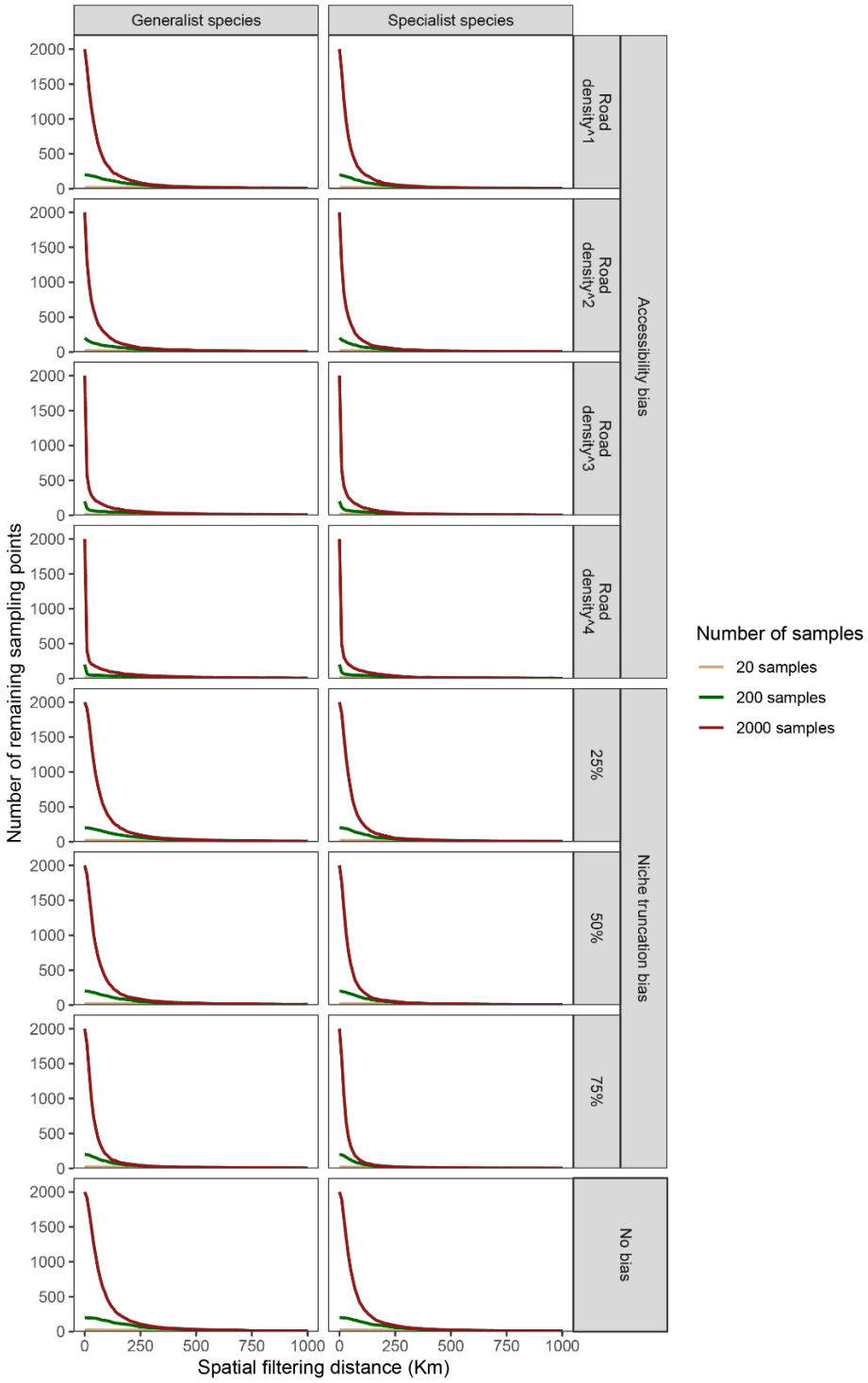

Figure S2: Median number of sampling points remaining after spatial filtering, for each species, each initial sampling size and each bias type and intensity (including the unbiased datasets), and for increasing filtering distances.

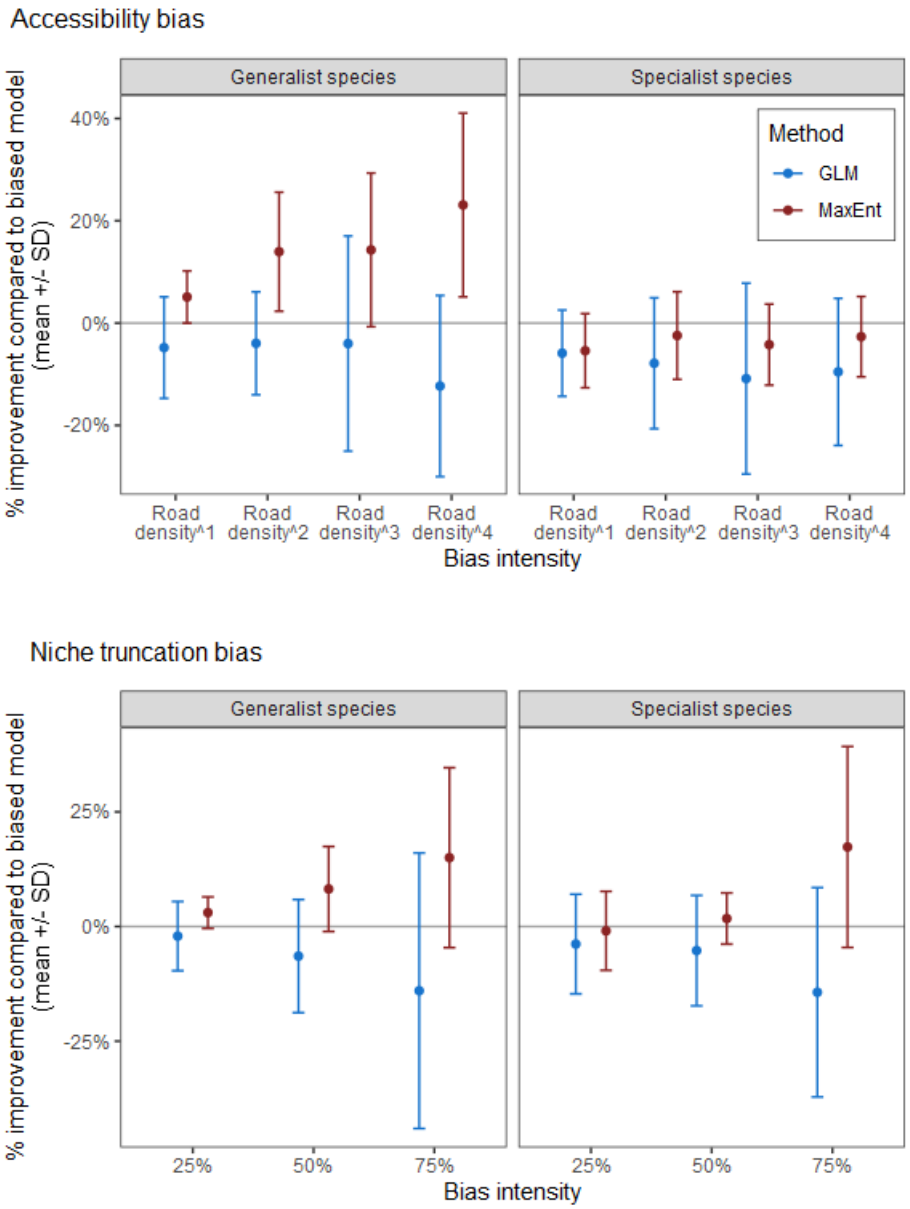

Figure S3: Mean (± SD) percent improvement in model performance at recovering the true species distribution, compared to the biased model, for models fitted with spatially filtered datasets, across all filtering distances. Results are presented for each species and each bias type and intensity.

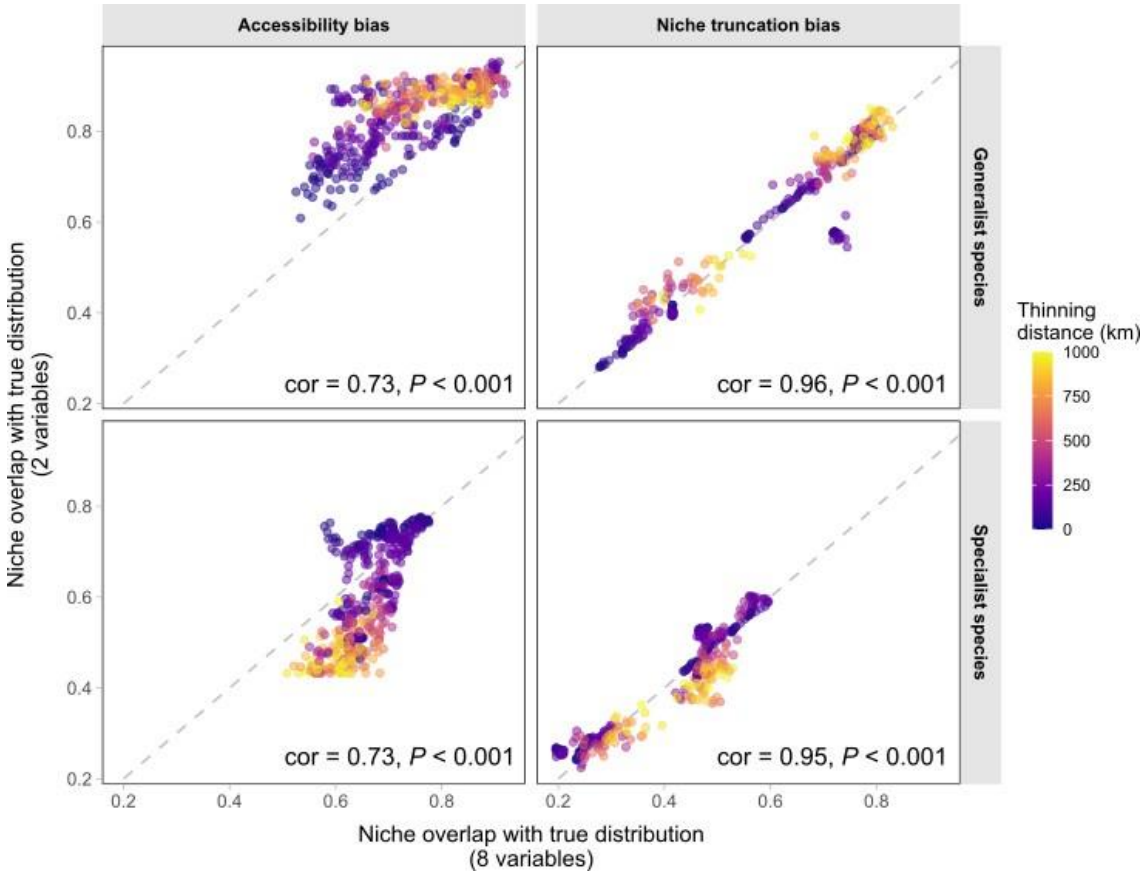

Figure S4: Scatterplot of the results, as expressed by Schoener's *D* index of overlap with the true distribution, obtained for 8 (x-axis) and 2 (y-axis) variables in the MaxEnt models.

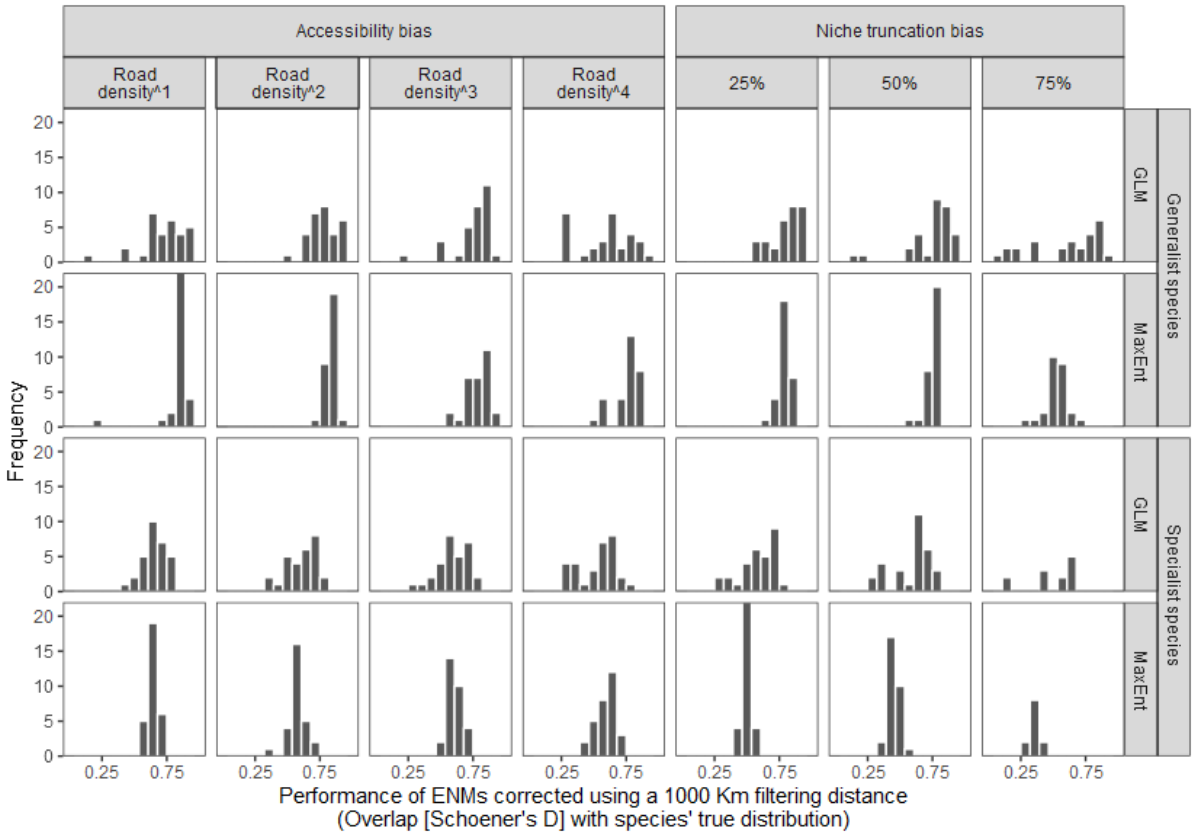

Figure S5: Distribution of model performance (based on Schoener's *D* overlap index with species' true distribution, $D_{\text{biased vs. true}}$) for ENMs fitted using a biased dataset corrected by using a 1000 Km filtering distance.

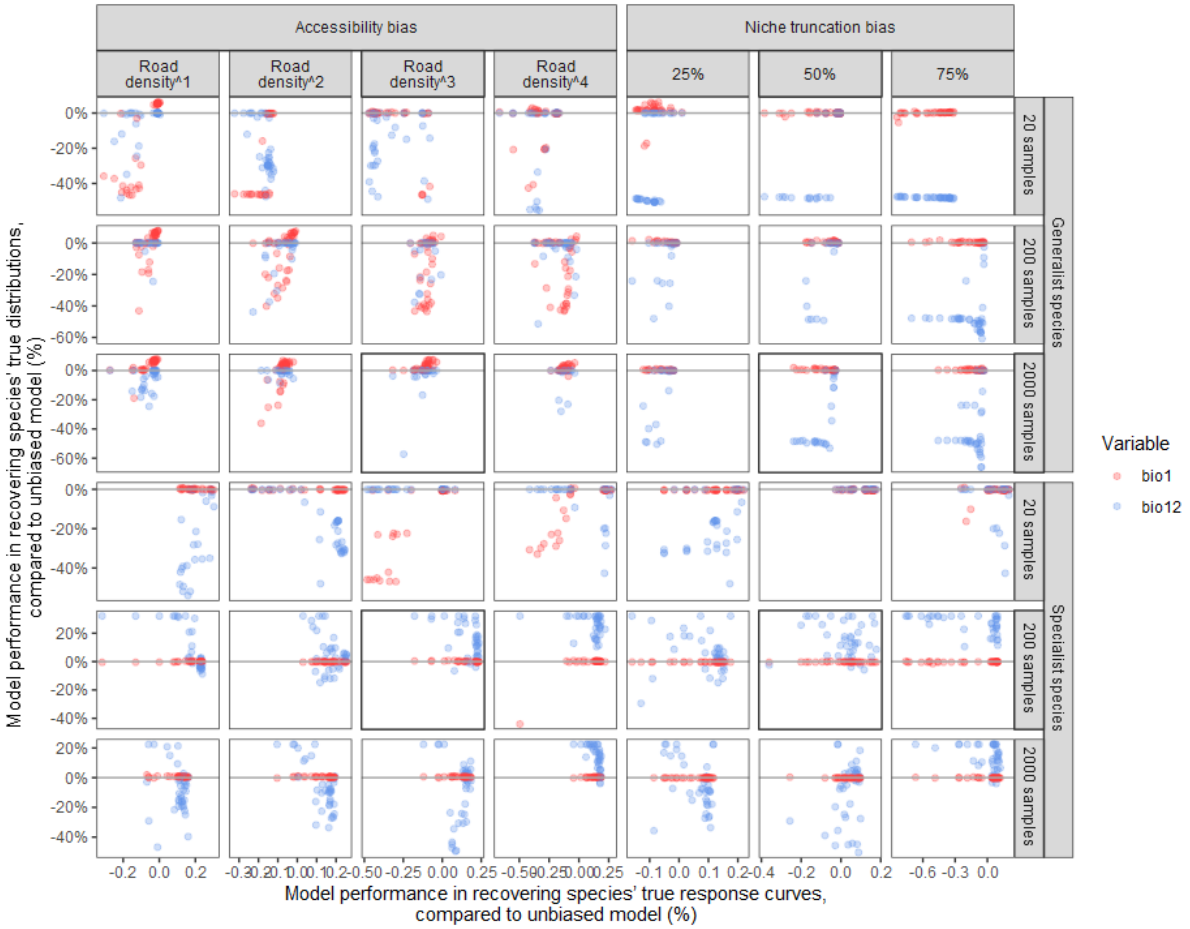

Figure S6: Relationship between the results presented in Figure 2 and 3 for the GLMs, i.e. model performance, expressed in percent compared to unbiased model, in recovering species' true distributions (y-axis) and response curves (x-axis).