Received: 21 April 2020 | Revised: 8 October 2020 | Accepted: 19 November 2020

DOI: 10.1111/geb.13240

RESEARCH PAPER

Global Ecology and Biogeography | A Journal of Macroecology | WILEY

Using centroids of spatial units in ecological niche modelling: Effects on model performance in the context of environmental data grain size

Yanchao Cheng[1] | Nils Benjamin Tjaden[1] | Anja Jaeschke[1] | Stephanie Margarete Thomas[1,2] | Carl Beierkuhnlein[1,2]

[1]Department of Biogeography, University of Bayreuth, Bayreuth, Germany

[2]BayCEER, Bayreuth Center for Ecology and Environmental Research, Bayreuth, Germany

Correspondence
Yanchao Cheng, Department of Biogeography, University of Bayreuth, Universitätsstr. 30, 95447, Bayreuth, Germany.
Email: yanchao1.cheng@uni-bayreuth.de

Funding information
Bayerisches Staatsministerium für Gesundheit und Pflege, Grant/Award Number: TKP 01KPB-73560; Bayerisches Staatsministerium für Umwelt und Verbraucherschutz, Grant/Award Number: TKP 01KPB-73560; China Scholarship Council, Grant/Award Number: 201506040059

Editor: Pedro Peres-Neto

Abstract

Aim: Ecological niche models (ENMs) typically require point locations of species' occurrence as input data. Where exact locations are not available, geographical centroids of the respective administrational spatial units (ASUs) are often used as a substitute. We investigated how the use of ASU centroids in ENMs affects model performance, what role the size of ASUs plays, and what effects different grain sizes of explanatory variables have.

Location: Europe.

Major taxa studied: Virtual species.

Methods: We set up a two-factorial study design with artificial ASUs of three different sizes and environmental data of four commonly used grain sizes, repeated over three study regions. To control other factors that may affect ENM performance, we created a virtual species with a known response to environmental variables, precise and even sampling and a known spatial distribution. We ran a series of Maxent models for the virtual species based on centroids and precise occurrence locations under varying ASU and grain sizes.

Results: The use of ASU centroids introduces a value frequency mismatch of the explanatory variables between centroids and true occurrence locations, and it has a negative effect on ENM performance. Value frequency mismatch, negative effect on ENM performance and over-prediction of the species' range all increase with ASU size. The effect of grain size of environmental data, on the contrary, was small in comparison.

Main conclusions: ENMs built upon ASU centroids can suffer considerably from the introduced error. For ASUs that are sufficiently small or show low spatial heterogeneity of explanatory variables, ASU centroids can still be a viable and convenient surrogate for precise occurrence locations. When possible, however, central tendency values (median, mean) that represent the whole ASU rather than just a single point location need to be considered.

KEYWORDS
administrative spatial unit, centroid, county, ecological niche model, grain size, Maxent, spatial heterogeneity, species distribution model, virtual species

1 | INTRODUCTION

Ecological niche models (ENMs), based on niche theory, are widely used in many fields such as invasion and conservation ecology, biogeography, as well as epidemiology (Elith & Leathwick, 2009; Escobar & Craft, 2016; Liu et al., 2018; Peterson, 2014). They are often employed to estimate the spatial distribution of certain species (Elith et al., 2006, 2011; Elith & Leathwick, 2009) or diseases (Tjaden et al., 2018), and thus also known as 'species distribution models'. ENMs typically use geographical occurrence locations of the target species as input data. These locations are then related to a series of explanatory variables (spatial raster data describing the environmental and/or socio-economic conditions in the study area), forming a correlative model of the species' environmental niche. This model can be projected onto regions where the presence–absence state of the species is unknown, resulting in a map showing probability of presence or environmental suitability.

By default, ENMs assume that the whole study area was sampled consistently with precisely recorded geographical locations of occurrence, and that the selected explanatory variables can represent the species well (Yackulic et al., 2013). In practice, however, sampling bias is inevitable, especially when the input data have to be assembled from different sources that are based on different sampling methods (Liu et al., 2018; Lobo & Tognelli, 2011; Qiao et al., 2017; Stolar & Nielsen, 2015). While the sampling bias caused by *uneven* sampling can be reduced by rarifying or filtering the occurrence records (Castellanos et al., 2019; Gábor et al., 2020; Kramer-Schadt et al., 2013), *imprecisely* recorded occurrence locations are very difficult (although not entirely impossible) to correct (Hefley et al., 2017). In certain cases, for example, when using citizen-science databases or local monitoring systems, the occurrence locations of the species may be of a coarse precision or only available at municipal or county level (i.e., related to geographical surfaces of differing sizes). For epidemiological data, such missing spatial precision ensures information privacy.

Internet databases like the Global Biodiversity Information Facility (GBIF; http://www.gbif.org) gather and compile species occurrence data from different sources (scientific, governmental, citizen-science) across a tremendous geographical extent and across national boundaries. However, the precision of the occurrence locations in this kind of database is not always sufficient (Liu et al., 2018), and the record precision differs considerably depending on how the occurrence records were collected and processed (in certain cases, the precision might even be unknown) (Collins et al., 2017). In other databases, only very coarse administrational level information is available, that is, instead of geographical locations, the occurrence records are assigned to counties or postal regions. For instance, the European Centre for Disease Prevention and Control (ECDC) and the European Food Safety Authority (EFSA) maintain a joint collection of occurrence records of epidemiologically relevant mosquito, tick and sand fly species in their VectorNet database. This highly relevant database covers the entire European Union and adjoining countries, but only maps showing local administrational units are publicly available (https://www.ecdc.europa.eu/en/disease-vectors/surveillance-and-disease-data). Similarly, occurrences of species are often reported as inventories of protected areas that can differ considerably in size. For the sake of simplicity, hereafter we will refer to all kinds of administrational areas as 'administrative spatial units' (ASUs).

When occurrence records are available at the level of entire ASUs only, the geographical centroids of the ASU are often used in ENMs as a substitute for precise point locations (Collins et al., 2017; Park & Davis, 2017). As mentioned above, the ENMs allocate explanatory variables' values at the respective geographical occurrence locations and form a correlative model of the species' environmental niche. The use of centroid locations introduces geographical distance between the true (but unknown) occurrence locations and the geographical centroids representing them. This induces a mismatch in the values of explanatory variables (Figure 1a): it is very unlikely (although not impossible) for the true geographical location of the observed record to exactly match the environmental conditions at the centroid location (Figure 1b). This means that between each pair of true location and geographical centroid, there is likely a mismatch in values of explanatory variables. It can consequently be expected that substituting geographical centroids for true occurrence locations also leads to a change in the overall frequency of values of explanatory variables. The correlative model, built with the shifted values, will further lead to a biased prediction for the species' distribution, and will probably lead to over-prediction.

While finding a substitute for a geographically unknown occurrence location, drawing the geographical centroid of the ASU minimizes the largest possible spatial distance between the substitute location and the unknown true location (Figure 1b1). However, this does not necessarily minimize the difference in environmental conditions (i.e., values of explanatory variables) at the two locations. In fact, it is entirely possible that among all possible locations within an ASU, the environmental conditions at the ASU centroid are the worst possible substitute for the conditions at the true location—especially in areas where spatial heterogeneity is high. Approaching the substitute from another angle (Figure 1b2), using a central tendency value (median, mean) of each explanatory variable across the entire ASU has been presented as a better option (Park & Davis, 2017). Instead of minimizing the largest possible spatial distance, central tendency values minimize the largest value mismatch directly. The boxplot in Figure 1b2 illustrates this on the basis of the median: when using the median as the substitute, very likely (97% in this example) the largest potential value mismatch is half of the range between the two bars. Possibly (50%) the value of the occurrence location falls in the box. In this case, the largest potential value mismatch is even smaller. It is obvious that using central tendency values reduces

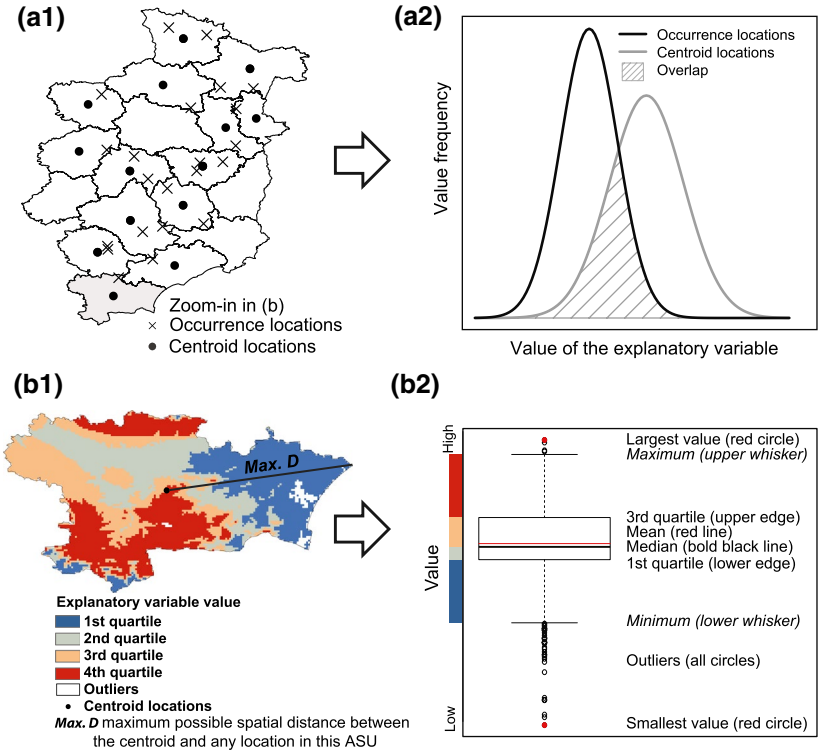

FIGURE 1 (a) Value frequency mismatch of an explanatory variable resulting from the use of administrational spatial unit (ASU) centroids. (a1) is a group of ASUs (here: counties) with true occurrence locations and respective centroid locations. Note that for each ASU only one centroid location will be kept, as there exists only one. (a2) is the value frequency curve mismatch between these two groups of locations, concerning an explanatory variable. (a1) and (a2) illustrate our hypotheses on how geographical distance between occurrence locations and ASU centroids leads to value frequency mismatch. (b) Zooming in to each ASU, for a single pair of occurrence location and centroid location. (b1) shows that using ASU centroids minimizes the largest potential spatial distance (thick black line) between the centroid location (black dot) and any possible unknown occurrence location in the given ASU. However, this does not mean that the difference in values between those two points is minimized. (b2) shows the variation of values of an explanatory variable across all the grid cells within the ASU. More than 97% of values fall into the range between the whiskers, and 50% of the values fall into the rectangular box

the possibility of introducing extreme values. Not surprisingly, it has been shown that the central tendency values outperform the variable values at the centroids (Collins et al., 2017; Park & Davis, 2017). Despite this, however, ASU centroids are still widely being used (e.g., Evans et al., 2010; Fois et al., 2018; Gao & Cao, 2019; Johnson et al., 2017; Quiner & Nakazawa, 2017). It is thus worth further investigating if and under what circumstances the much simpler approach of using ASU centroids can lead to sound results.

Here, we investigate how the application of geographical centroids affects the ENM results, a factor that has not received much attention so far. Of course, ENMs are also affected by a series of other factors. This includes the selection of explanatory variables, the specific modelling algorithm, model settings and the spatial resolution or pixel size of explanatory variables, from here on referred to as 'grain size' (Connor et al., 2018; Fourcade et al., 2018; García-Callejas & Araújo, 2016; Moudrý & Simova, 2012; Nezer et al., 2017; Record et al., 2018; Warren & Seifert, 2011; Yates et al., 2018). To best control these factors, we generated a virtual species with true, known occurrence locations and a known response to a fixed set of variables. Similar to the use of ASU centroids, the choice of grain size of explanatory variables may also cause a mismatch of explanatory variable value. For a given location, the explanatory variable value may differ at different grain sizes, though the difference is in general small due to spatial autocorrelation. Hence, we included a series of commonly used grain sizes of explanatory variables. Focusing on the bias resulting from the use of ASU centroids, we hypothesize that: (a) using ASU centroids as substitutes for true occurrence locations leads to a value frequency mismatch of explanatory variables between the true locations and the centroids. An increased spatial heterogeneity within the ASU elevates that mismatch, assuming that larger ASUs tend to have higher spatial heterogeneity. (b) When using ASU centroids, the ENM's performance decreases with increasing ASU size. (c) The size of the ASUs affects the model performance more than the grain size of explanatory variables. (d) The use of ASU centroids leads to an over-estimation of the modelled species' distributional range.

2 | MATERIAL AND METHODS

To investigate how much ASU size and grain size affect ENM performance, a two-factorial design with three replicates was applied

(Figure 2). For this, a virtual species was generated based on three explanatory variables across Europe (see below for details). According to the presence–absence map of this virtual species, three squared regions (sized 5° × 5°) were selected within which the virtual species occupancy was about 50%. For each squared region, pseudo-ASUs of three different sizes were constructed by dividing it evenly into 25 (large), 100 (medium) and 400 (small) squares (Figure 2). The size range of these pseudo-ASUs corresponds to those of low-level administrational units across Europe. There, the Nomenclature of Territorial Units for Statistics (NUTS, https://ec.europa.eu/eurostat/web/nuts/background) consists of three levels (NUTS 1–3). NUTS 3 is for small regions with a population size threshold of 150,000–800,000. The average area of the NUTS 3 units is about 7,000 km^2. The large pseudo-ASU size applied in this study is about 8,000 km^2, which can be treated as an equivalent of the average NUTS 3 administrative units across Europe. The medium pseudo-ASU size is about 2,000 km^2, and the small pseudo-ASU size is about 500 km^2.

The hypotheses were first tested with these artificial pseudo-ASUs from the three rectangular regions (with three pseudo-ASU sizes; Figure 2). Afterwards, data from real countries (Germany and France) with irregular ASU size and shape were used to confirm the previous results in a real-life environment (Supporting Information Appendix S1, Figure S1.1). The varying pseudo-ASU sizes for the regions were applied to detect the general trend of centroid-arisen bias. France and Germany were chosen as test cases as they are of very different NUTS 3 ASU sizes. For France, the average area of NUTS 3 ASUs is about 6,000 km^2. For Germany, it is 1,200 km^2. As the NUTS 3 ASU size of Germany is much smaller than that of France, we expected the ENM models based on Germany's NUTS 3 centroids to outperform the ones based on French NUTS 3 centroids.

For the rectangular regions as well as France and Germany, a series of commonly used grain sizes was taken into consideration (four grain sizes 0.5, 2.5, 5, 10 arc-min, roughly equivalent to 0.5, 10, 40 and 200 km^2, respectively), because the grain size (raster resolution) of explanatory variables in ENMs also affects model performance (Connor et al., 2018; Guisan et al., 2007; Lauzeral et al., 2013; Manzoor et al., 2018). It is necessary to view both ASU size and grain size as factors that affect models on similar special scales.

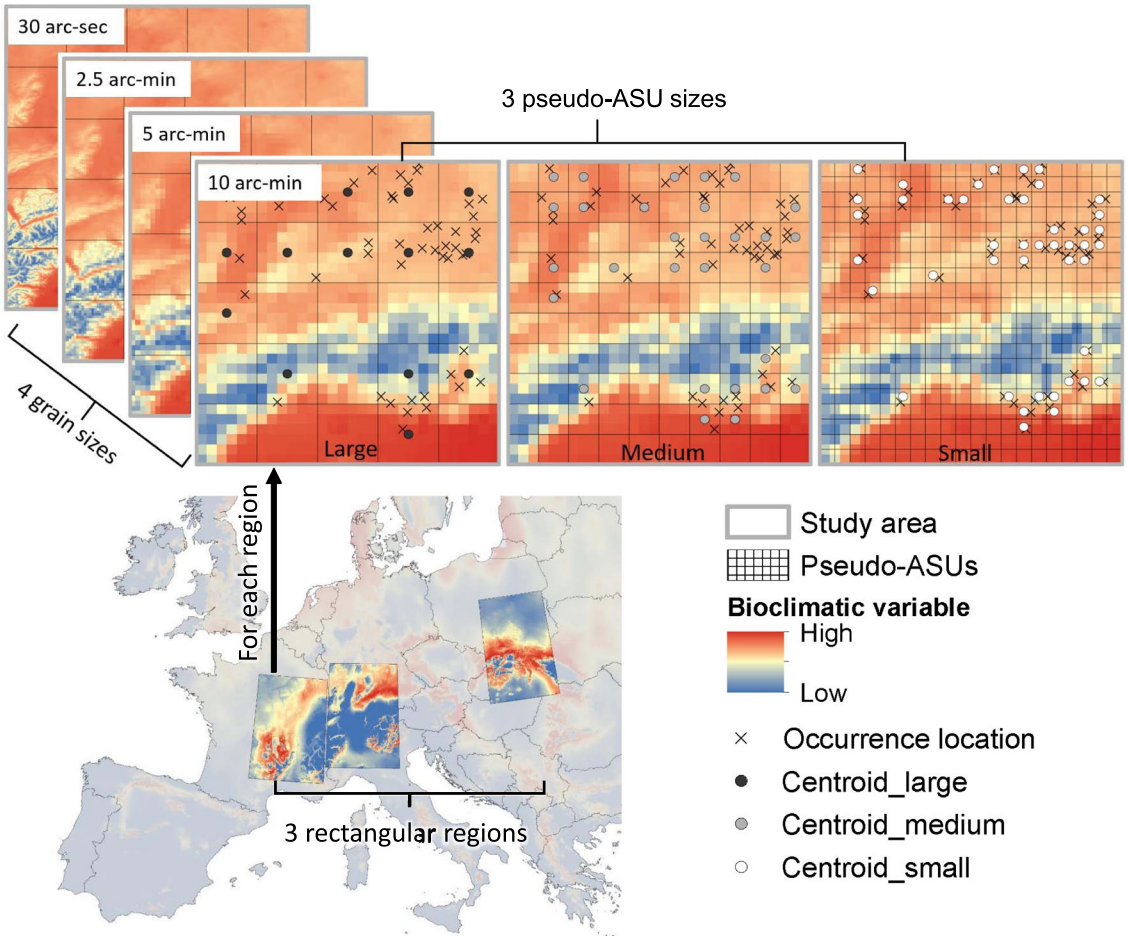

FIGURE 2 The two-factorial study design with three replicates. Three rectangular regions [with varying grain size and pseudo-administrational spatial unit (pseudo-ASU) size] were used to assess the general trend of bias resulting from the use of ASU centroids together with varying grain size. For each region, 200 random locations were drawn to keep the sampling effort even, and only locations where the virtual species occurs were kept (black crosses). The whole setup was repeated with the three bioclimatic variables that the virtual species was generated with (see Material and methods)

2.1 | Virtual species

A virtual species was generated using the 'virtualspecies' package version 1.4.2 (Leroy et al., 2016) in R 3.4.2 (R Core Team, 2015), with a spatial resolution of 0.5 arc-min (for details on explanatory variables see section 'Explanatory variables'). Virtual species generation can be understood as defining a niche of a virtual species by limiting the determining environmental variables, setting the response to the variables, and setting prevalence or tolerance levels. By applying a virtual species, (a) the exact explanatory variables are known, (b) the occurrence locations are precise, (c) the whole study area is evenly sampled without sampling bias, (d) the true spatial distribution probability and presence–absence maps are available. These advantages make virtual species an ideal tool for testing our hypotheses.

To generate a virtual species, a certain number of explanatory variables is needed, as well as parameters such as the response of the virtual species to each variable. Depending on the parameters and the presence–absence conversion method applied, different species distribution patterns can be achieved. The spatial distribution of the virtual species, both the logistic distribution map and presence–absence map, can be exported as raster files for further use. A dataset of presence–absence or presence-only locations can be generated in order to simulate real-world sampling of occurrence records in the field. In this study, 200 random locations were drawn for each rectangular region to simulate sampling locations of an unbiased field campaign. By allocating the same number of sampling locations, the simulated sampling effort for each region is the same, thus the model performance is comparable across the regions. Locations where the species was recorded as 'absent' were discarded and the remaining presence locations (*c*. 60 per region) were used to build the ENMs. For more details about the virtual species in this study see Supporting Information Appendix S2.

2.2 | Explanatory variables

To keep the virtual species simple, only bioclimatic variables were taken into account. For this, the standard set of 19 bioclimatic variables was acquired from https://www.worldclim.org (Fick & Hijmans, 2017), with grain sizes of 0.5, 2.5, 5 and 10 arc-min. Three bioclimatic variables were selected according to the following criteria: (a) the set must include both hydrological and thermal factors, which are essential to most life-forms; (b) the variables should not be closely related to each other [De Marco & Nóbrega, 2018; i.e., |Pearson's r| > .7 (see Supporting Information Appendix S3, Table S3.1), calculated with the European extent (Supporting Information Appendix S2, Figure S2.2) of the virtual species]. As a consequence, three bioclimatic variables, namely annual mean temperature (Bio 1), annual precipitation (Bio 12), and precipitation seasonality (Bio 15), were chosen (for details see Supporting Information Appendix S2, Figures S2.3 and S2.4).

2.3 | Value frequency mismatch in explanatory variables due to the use of ASU centroids

To visualize the value frequency mismatch, the three explanatory variables' values were extracted at the centroid locations of the three different pseudo-ASU sizes separately per region. This was repeated for the four different grain sizes (Figure 2). The value frequencies of the explanatory variables were then described through a kernel density curve (which can be understood as the response curve of the virtual species to the variable) using the R package 'caTools' version 1.17.1.2 (Tuszynski, 2014). Relative overlap of the curve for the centroid locations with the corresponding curve for the original occurrence records was calculated using 'caTools'. Mismatch between these curves was then calculated as *1 − overlap*, so that a mismatch of 0 means identical curves and 1 means no overlap at all.

The spatial heterogeneity of each explanatory variable was assessed by calculating its standard deviation within the respective pseudo-ASU and for the respective grain size, using the 'raster' package version 2.6.7 (Hijmans, 2019) in R. The spatial heterogeneity was expected to increase with the size of pseudo-ASUs. For each explanatory variable, a linear regression was applied to describe the correlation between frequency curve mismatch and spatial heterogeneity.

2.4 | Ecological niche model

Maxent, an ENM algorithm widely used and known for its good performance with small occurrence location datasets (Baldwin, 2009; Elith et al., 2006; Hernandez et al., 2006), was chosen in this study. To make the models comparable, the settings were kept the same for all runs, for the three rectangular patches, Germany and France. Default model settings were applied (with 10,000 background locations), with 10 replicates. Instead of commonly used methods such as the true skill statistic (TSS) or the area under the curve (AUC) of the receiver operating characteristic, model performance was assessed using Spearman's rank correlation coefficient (Spearman's rho). Spearman's rho is obviously a better choice than AUC, and compared to TSS, Spearman's rho has the advantage of being threshold-independent. As the true spatial distribution probability map for the virtual species is available, Spearman's rho for the correlation between the environmental suitability predicted by the model and the true probability of presence can easily be achieved [via R package 'pspearman' (Savicky, 2014)]. In this case, Spearman's rho can range from 0 (no correlation) to 1 (perfect linear positive correlation, the compared models are identical).

To calculate Spearman's rho, the larger grain-sized (2.5, 5 and 10 arc-min) outputs from Maxent models were resampled to 0.5 arc-min resolution using the 'nearest neighbour' method [R package 'raster' (Hijmans, 2019)]. Essentially, this means cutting the large raster cells into smaller ones, while keeping the original values without interpolation or loss of information. The model results were then compared with the true (distribution) probability map of the

virtual species (the virtual species was generated with 0.5 arc-min resolution; for more details see above and Supporting Information Appendix S2).

2.5 | Calculation of over-prediction ratio

The model results were transformed into binary presence–absence maps according to the thresholds: maximum training sensitivity plus specificity logistic threshold (MaxSSS; Liu et al., 2005, 2016), equal training sensitivity and specificity logistic threshold (eqSS; Liu et al., 2005; Nenzen & Araujo, 2011) and 10 percentile training presence logistic threshold (10 percentile; Pearson et al., 2004). Over-prediction was then calculated as the ratio of raster cells classified as 'presence' in the model output versus the original virtual species [i.e., $(\text{Presence}_{\text{modelled}} - \text{Presence}_{\text{original}})/\text{Presence}_{\text{original}}$]. Here, a value of 0 suggests that the distributional range predicted by the model has the same size as the one defined in the virtual species. Values larger or smaller than 0 mean that the predicted range is larger (over-prediction) or smaller (under-prediction) than that of the original species, respectively.

3 | RESULTS

3.1 | The larger the pseudo-ASU size, the larger the value frequency mismatch of the explanatory variables

Using ASU centroids resulted in a mismatch of value frequency curves of the explanatory variables. This mismatch increases with the spatial heterogeneity within the respective ASU. A statistically significant positive relationship between variable mismatch and spatial heterogeneity was revealed through linear regression analysis for all three bioclimatic variables (Figure 3, top; Bio1: $p < .01$, $R^2 = .16$; Bio12: $p < .05$, $R^2 = .11$; Bio15: $p << .001$, $R^2 = .56$). The overall spatial heterogeneity of an ASU increases with its size (Figure 3, bottom).

3.2 | For ENM performance, ASU size matters more than the grain size

Ecological niche models built with the original occurrence locations showed strong correlations of predicted environmental suitability

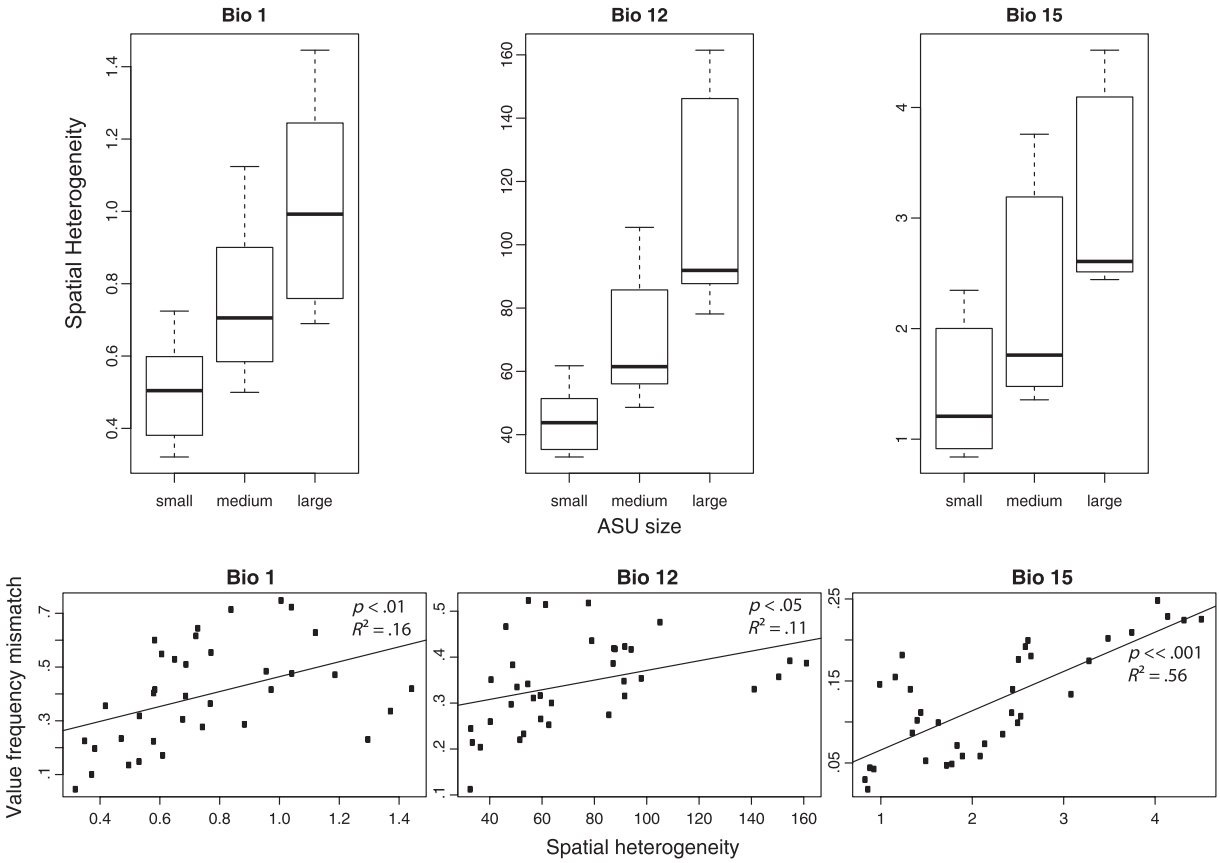

FIGURE 3 Top: the spatial heterogeneity of each explanatory variable (Bio 1: annual mean temperature, Bio 12: annual precipitation, Bio 15: precipitation seasonality) increases with increasing centroid region size. Bottom: value frequency curve mismatch for explanatory variables (Bio 1, Bio 12, Bio 15) at occurrence locations versus centroid locations increases with elevated spatial heterogeneity. Each black dot represents one pair of comparisons of the recorded variable values between centroid locations and real locations. For each variable, the spatial heterogeneity is measured by the standard deviation of that variable within the individual centroid regions. ASU = administrational spatial unit.

with the true distribution of the virtual species, suggesting good model performance (Figure 4a). For these ENMs, the model performance decreased with increasing grain size (see Supporting Information Appendix S4, Figure S4.5). For those ENMs built with centroids, a Kruskal–Wallis rank sum test with multiple comparison post-hoc test revealed statistically significant ($p < .05$) effects of ASU size on model performance. There is a clear trend of model performance decreasing with increasing ASU size (Figure 4a). No clear model performance pattern was observed for the different grain sizes (Figure 4b). A direct comparison using two-way ANOVA (see Supporting Information Appendix S4, Table S4.2; only including the ENMs built with centroids, i.e., the three grey boxes in Figure 4a) reveals that while ASU size can explain more than half (52%) of the variability in model performance ($f(2) = 16.414$, $p << .001$); grain size appears to have almost no effect ($f(3) = 0.876$, $p = .467$); the interaction between grain size and ASU size shows no significance ($f(6) = 0.561$, $p = .757$), either.

3.3 | Over-prediction of species spatial distribution due to the use of centroid data

Almost all ENM runs in this study, including those performed with true occurrence locations, over-predicted the virtual species' occurrence. Based on the MaxSSS threshold, over-prediction tends to be stronger with increasing ASU size (Figure 5). However, this increase is statistically significant only for the large ASUs ($p < .01$ based on ANOVA followed by a Tukey honest significant difference post-hoc test). This is consistent with the results obtained from the eqSS and 10 percentile thresholds (Supporting Information Appendix S4, Figure S4.6).

3.4 | Real-world application example using French and German ASU centroids

The ENM built with centroids of NUTS 3 administrative units in Germany outperforms the model for France, with Spearman's rho value of .818 and .790, respectively. For the ENM built with true occurrence locations, Spearman's rho for Germany and France is .894 and .924, respectively (Table 1). When occurrence locations are available, the model performance decreases with increasing grain size. However, when only centroid locations are available, fine grain size was not always the best. For France, the ENM with 2.5 arc-min had the best model performance (Table 1). This is in accordance with the pattern shown in Figure 4: ASU size matters more than grain size.

4 | DISCUSSION

In this study, we looked into the mechanism of how the use of centroids affects ENM performance. Though there have been studies focusing solely on centroid size (Collins et al., 2017; Park & Davis, 2017) or grain size (Connor et al., 2018; Lauzeral et al., 2013; Manzoor et al., 2018), we investigated and compared how much ASU size and grain size affect ENM performance. Our results confirm that, in general, larger ASUs have higher spatial heterogeneity, and higher spatial heterogeneity is associated with higher value frequency mismatch of

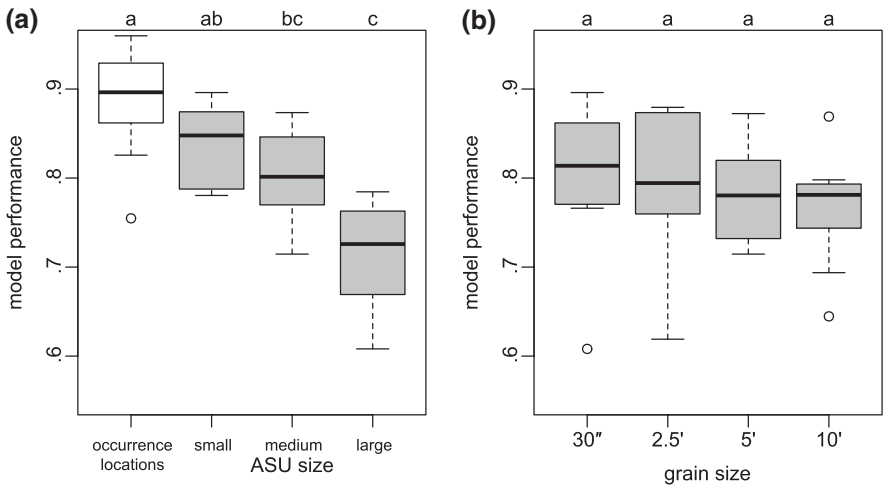

FIGURE 4 Administrational spatial unit (ASU) size affects model performance more than grain size. (a) Grey: performance of the ecological niche model (ENM) at different ASU sizes. White, for reference: ENM performance when using true occurrence locations. Lower case letters above the boxes indicate differences between groups as indicated by a Kruskal–Wallis rank sum test with multiple comparison post-hoc test. (b) Performance of the ENM at different grain sizes (i.e., spatial resolution of environmental data). Lower case letters above the boxes indicate that an ANOVA followed by a Tukey honest significant difference post-hoc test revealed no statistically significant differences between any of the groups. Grey boxes in (a) and (b) refer to the same set of models. Model performance was assessed through the correlation coefficient (Spearman's rho) between the environmental suitability predicted by the ecological niche model and the true species distribution defined for the virtual species. Model performance ranges from 0 to 1. The larger the value, the better the model performance.

explanatory variables between the true locations and centroids. When using ASU centroids, larger ASUs also lead to a larger decrease of ENM performance. Compared with ASU size, grain size does not affect ENM performance as much. The use of ASU centroids leads to over-prediction of the modelled species' distribution range, and the over-prediction ratio shows a tendency to increase with ASU size.

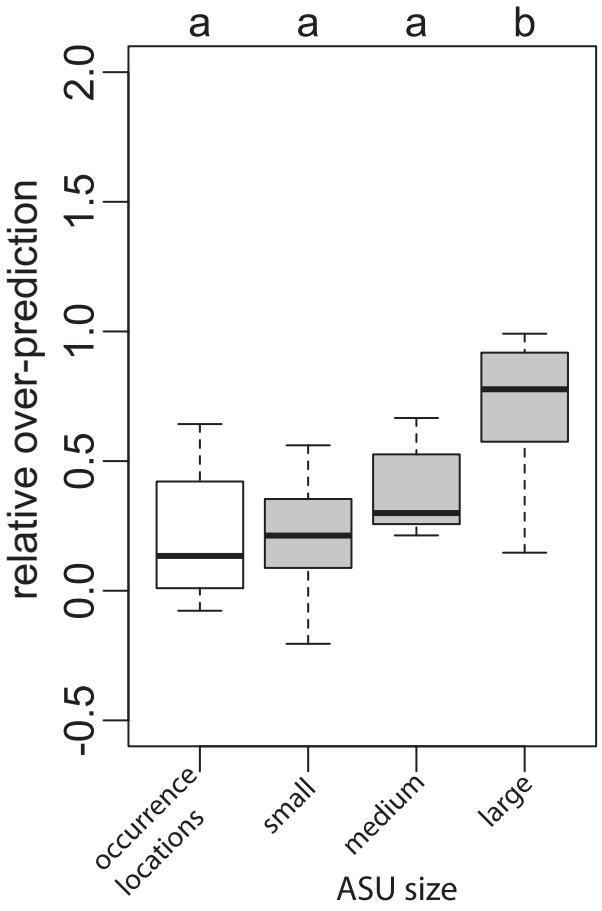

FIGURE 5 Relative over-prediction ratio [ecological niche model (ENM) results versus 'true' virtual species occurrence] for models built on point locations (white) as well as centroids of differently sized regions (grey). Lower case letters above the boxes indicate differences between groups as indicated by an ANOVA followed by a Tukey honest significant difference post-hoc test. ASU = administrational spatial unit

Using centroids of ASUs as a substitute for true occurrence records in ENMs introduces errors. Spatial distance between the centroids and the true, unknown occurrence locations leads to a mismatch of explanatory variables' values at these locations. These mismatched values at the centroid locations lead to a mismatch of explanatory variables' value frequency curves, which further results in a mismatch between the projected niche and the true niche. Our results show that the absolute size of ASUs affects the value frequency mismatch between true locations and centroids. How strong this effect is, ultimately depends on the explanatory variables' spatial heterogeneity (here: standard deviation) within the ASUs. Park and Davis (2017) found that spatial heterogeneity in climatic variables was mainly governed by the heterogeneity in topography in the US. Their findings that ASU (county) size only had minimal effects on spatial heterogeneity does not contradict our results, due to their different, non-nested study design. The absolute ASU size or grain size alone cannot determine how much the explanatory variables' values mismatch with the values at the true occurrence locations, but in general, larger ASU size leads to larger value mismatch (Figure 3).

Similarly, coarser grain size leads to a deterioration in model performance (Supporting Information Appendix S4, Figure S4.5). This is in line with previous findings (Connor et al., 2018; Guisan et al., 2007; Manzoor et al., 2018). However, although the grain size does affect model performance, its effect was found to be small compared that of the ASU size (Figure 4 and Supporting Information Appendix S4, Table S4.2). When centroids are drawn from ASUs with large extent, the ENM's performance can hardly be improved by using a fine grain size (Figure 4). However, when using centroids drawn from a small extent or using the true occurrence locations, a fine grain size is preferable (Table 1). Note that the 'small' pseudo-ASU size used in this study is 0.25° × 0.25° (*c.* 400 km^2). This is not much larger than the coarsest grain size (*c.* 170 km^2) in this study, which corresponds to the resolution of some commonly used environmental datasets [e.g., E-OBS (Cornes et al., 2018)]. For ASUs of this size, the value mismatch between the centroids and true locations is very small, and the effect of ASU size on the value mismatch cannot be distinguished from that introduced by a large grain size.

The use of a virtual species in this study means that the environmental suitability and the presence–absence status of the species across the study region are known, and the species' occurrence records are precise. This makes the comparison between models based on true

Country	Grain size	Model performance (centroids)	Model performance (true occurrence locations)
France	10 arc-min	.725	.852
France	5 arc-min	.798	.927
France	2.5 arc-min	**.858**	.942
France	30 arc-sec	.779	**.974**
Germany	10 arc-min	.807	.837
Germany	5 arc-min	.812	.897
Germany	2.5 arc-min	.826	.914
Germany	30 arc-sec	**.828**	**.927**

TABLE 1 Model performance of the real-world examples using French and German administrational spatial units (ASUs). Performance of the models was assessed by calculating Spearmans' rho for the correlation of the predicted probability of presence with the known true probability of presence of the virtual species. Bold: best-performing models for centroid- and true location-based models for France and Germany

locations and models based on ASU centroids feasible and reliable. The difference between models based on observed point locations from field data versus centroid-based locations has previously been quantified (Collins et al., 2017). However, using observed locations, additional effects resulting from sampling bias or uncertainty cannot be excluded. Generating a virtual species and drawing precise geographical locations ensures that the observed differences between model results are due to the use of centroid locations itself and that they can be quantified. As the true environmental suitability of the virtual species is known, it can be used as a benchmark for the performance of the models by directly calculating Spearman's rho. This obviates the use of threshold-based performance measures such as TSS or the commonly used but controversial AUC (Allouche et al., 2006; Tjaden et al., 2018).

The value frequency mismatch was for the first time applied as an indicator of potential projected niche mismatch. This method calculates, for each explanatory variable, the value frequency mismatch between ASU centroids and occurrence locations. Compared with conventional statistical tests (such as ANOVA or Kruskal–Wallis test) that show whether a statistical difference between centroid-based and true location-based environmental data exists or not, this method focuses on quantifying the differences between the two. As ENMs typically process variables' values in a continuous way (for a given variable value at a single location, the ENMs calculate a probability of presence rather than a binary presence–absence value), the true/false information alone is clearly not sufficient to assess the effect of using ASU centroids. While test statistics fail to measure how much two groups of data differ, the continuous method makes it possible to capture the general trend of the value mismatch and compare it across groups. For instance, our results suggest that a larger ASU size leads to larger value frequency mismatch. If a binary-result-only method had been applied in this study, this general trend could not have been captured, or evaluated. It should be noted that the value frequency mismatch ranges from 0 to 100%, that is, though the value frequency mismatch increases with spatial heterogeneity within the ASU, this increase is not always linear. Here, a linear model was applied to show the general trend, but it should not be interpreted as an indication for how much the mismatch will be when larger spatial heterogeneity occurs.

As the virtual species was generated with the grain size of 0.5 arc-min, it is to be expected that the output from the combination of this grain size and original locations has the best performance. While generating the virtual species, we assumed that the grain size employed is the smallest unit in which the virtual species can survive, as an individual (similar to e.g., a deer, a bird) or a population (e.g., ants, bees). However, modelling with real species, it should be questioned which grain size should be utilized. It has been suggested to use a grain size smaller than 1 km when possible, especially for habitat specialists (Manzoor et al., 2018). Nevertheless, the question of choosing optimal grain size needs to be further investigated.

It is not surprising that centroids of ASUs are chosen as a substitute when no precise occurrence locations are available. After all, the changes in the modelling workflow required by this approach are minimal compared with the calculation of central tendency measures across ASUs. Using the geographical centroid of an ASU minimizes the largest possible spatial distance between the centroid and any (unknown) occurrence location within the ASU (Figure 1b1). Considering spatial autocorrelation, the value mismatch between the unknown true location and the chosen substitute (centroid) has a chance of being limited as well. However, central tendency values (e.g., mean or median) of the variable value within the ASU are a better alternative for minimizing the value mismatch, as they limit the potential value distance directly rather than indirectly through spatial distance (Figure 1b2). Although the value mismatch cannot be eliminated completely, using central tendency values lowers the probability of introducing extreme values or outliers. In accordance with previous studies (Park & Davis, 2017), we thus suggest the use of central tendency values as substitutes wherever possible. However, when the ASU size is very small (e.g., 0.25° × 0.25°) or the environment within the ASU is very homogenous, geographical centroids can work as good substitutes. In this case, it is worth comparing the variables' values from the centroid locations with those from available occurrence locations. If the centroids lead to value outliers, central tendency values should be used instead or the outliers discarded. Of course, this assumes that only a fraction of the available records consists of un-precisely recorded location data.

When using a limited number of ASU centroids *in addition* to precise locations, there are two critical aspects that need to be considered. First, these coarse centroids are typically assigned the same weight as the precise location by the ENMs, which may result in a distorted model of the spatial distribution of the species. This could be ameliorated by down-weighting centroid locations, provided that the chosen modelling algorithm allows for that. Second, for each ASU, only one centroid record will be kept by the ENMs. As a consequence, regions with only one record are treated the same as regions with several records, and the abundance information (if available at all) is neglected. It has been shown that a mixture of precise and centroid-based data would be more robust against the issues demonstrated in this work (Collins et al., 2017), but to what degree and how the ratio of the two data types affect the robustness needs to be clarified in future studies.

It should be noted that the insights gained here only apply to models built with continuous variables. Categorical predictors like land use classes typically show sharp edges on the map, so that even small differences in the spatial location of occurrence records can lead to dramatically different values being assigned to them. Thus, it seems reasonable to assume that models built with categorical variables would be affected more strongly by ASU size, but further investigations are needed to confirm this. Similarly, all our analyses were conducted at an intermediate (sub-continental) spatial scale, as this is the scale where ASU centroids are most likely to be used. Further research is needed to verify whether the conclusions drawn from this can be transferred to coarser or finer scales.

5 | CONCLUSIONS

Whether ASU centroids can be a viable surrogate for precise occurrence locations depends on the ASUs' sizes and how heterogeneous

they are in terms of environmental explanatory variables. For instance, in the northern German flatlands, where ASUs are small and the environment comparably homogenous, the use of centroid locations is much less of a problem than in the alpine regions of France, where ASUs are large and environmental gradients steep. If possible, central tendency values should be considered as a more robust alternative. As our results suggest that effects of using ASU centroids outweigh effects of grain size, it is important for modellers to recognize this source of error. In order to enable researchers to assess whether the use of centroid locations is appropriate for a specific project, new methods and guidelines need to be developed.

ACKNOWLEDGMENTS

Yanchao Cheng was funded by the China Scholarship Council, No. 201506040059. Stephanie Thomas was funded by the Bavarian State Ministry of the Environment and Consumer Protection and the Bavarian State Ministry of Health and Care through the BayVirMos project (TKP 01KPB-73560).

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

AUTHOR CONTRIBUTIONS

Y. Cheng, N. B. Tjaden, S. M. Thomas, A. Jaeschke, and C. Beierkuhnlein conceptualized the study. Y. Cheng and N. B. Tjaden processed and analyzed the data and prepared the figures. Y. Cheng wrote the first draft and prepared the appendices. N. B. Tjaden wrote parts of the results section. C. Beierkuhnlein, S. M. Thomas and A. Jaeschke supervised the analyses. All authors discussed the methodology and results, and reviewed and edited the manuscript.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the WorldClim dataset at https://www.worldclim.org/. Details about how to recreate the simulated data are provided in the Supporting Information Appendices; R source code is available upon request from the authors.

ORCID

Yanchao Cheng https://orcid.org/0000-0003-2649-876X
Nils Benjamin Tjaden https://orcid.org/0000-0002-7685-1659
Anja Jaeschke https://orcid.org/0000-0001-8361-0960
Stephanie Margarete Thomas https://orcid.org/0000-0003-0507-2006
Carl Beierkuhnlein https://orcid.org/0000-0002-6456-4628

REFERENCES

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, *43*, 1223–1232.

Baldwin, R. A. (2009). Use of maximum entropy modeling in wildlife research. *Entropy*, *11*, 854–866.

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*, 481–492.

Collins, S. D., Abbott, J. C., & McIntyre, N. E. (2017). Quantifying the degree of bias from using county-scale data in species distribution modeling: Can increasing sample size or using county-averaged environmental data reduce distributional overprediction? *Ecology and Evolution*, *7*, 6012–6022.

Connor, T., Hull, V., Viña, A., Shortridge, A., Tang, Y., Zhang, J. D., Wang, F., & Liu, J. G. (2018). Effects of grain size and niche breadth on species distribution modeling. *Ecography*, *41*, 1270–1282.

Cornes, R. C., van der Schrier, G., van den Besselaar, E. J. M., & Jones, P. D. (2018). An ensemble version of the E-OBS temperature and precipitation data sets. *Journal of Geophysical Research: Atmospheres*, *123*, 9391–9409.

De Marco, P. J., & Nóbrega, C. C. (2018). Evaluating collinearity effects on species distribution models: An approach based on virtual species simulation. *PLoS ONE*, *13*, e0202403.

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., Jacob McC, M., Overton, A. T., … Zimmermann, N. E. (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, *29*, 129–151.

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*, 677–697.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, *17*, 43–57.

Escobar, L. E., & Craft, M. E. (2016). Advances and limitations of disease biogeography using ecological niche modeling. *Frontiers in Microbiology*, *7*, 1174.

Evans, J. M., Fletcher, R. J., & Alavalapati, J. (2010). Using species distribution models to identify suitable areas for biofuel feedstock production. *Global Change Biology Bioenergy*, *2*, 63–78.

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*, 4302–4315.

Fois, M., Cuena-Lombraña, A., Fenu, G., Cogoni, D., & Bacchetta, G. (2018). Does a correlation exist between environmental suitability models and plant population parameters? An experimental approach to measure the influence of disturbances and environmental changes. *Ecological Indicators*, *86*, 1–8.

Fourcade, Y., Besnard, A. G., & Secondi, J. (2018). Paintings predict the distribution of species, or the challenge of selecting environmental predictors and evaluation statistics. *Global Ecology and Biogeography*, *27*, 245–256.

Gábor, L., Moudrý, V., Barták, V., & Lecours, V. (2020). How do species and data characteristics affect species distribution models and when to use environmental filtering? *International Journal of Geographical Information Science*, *34*, 1567–1584.

Gao, X., & Cao, Z. (2019). Meteorological conditions, elevation and land cover as predictors for the distribution analysis of visceral leishmaniasis in Sinkiang province, Mainland China. *Science of the Total Environment*, *646*, 1111–1116.

García-Callejas, D., & Araújo, M. B. (2016). The effects of model and data complexity on predictions from species distributions models. *Ecological Modelling*, *326*, 4–12.

Guisan, A., Graham, C. H., Elith, J., Huettmann, F., Dudík, M., Ferrier, S., Hijmans, R., Lehmann, A., Li, J., Lohmann, L. G., Loiselle, B., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., McC, J., Overton, A. T., Peterson, S. J., Phillips, K. R., … Zimmermann, N. E. (2007). Sensitivity of predictive species distribution models to change in grain size. *Diversity and Distributions*, *13*, 332–340.

Hefley, T. J., Brost, B. M., & Hooten, M. B. (2017). Bias correction of bounded location errors in presence-only data. *Methods in Ecology and Evolution*, *8*, 1566–1573.

Hernandez, P. A., Graham, C. H., Master, L. L., & Albert, D. L. (2006). The effect of sample size and species characteristics on performance of different species distribution modeling methods. *Ecography*, *29*, 773–785.

Hijmans, R. J. (2019). *raster: Geographic data analysis and modeling. R package version 3.0-7*. https://CRAN.R-project.org/package=raster

Johnson, T. L., Haque, U., Monaghan, A. J., Eisen, L., Hahn, M. B., Hayden, M. H., Savage, H. M., McAllister, J., Mutebi, J. P., & Eisen, R. J. (2017). Modeling the environmental suitability for Aedes (Stegomyia) aegypti and Aedes (Stegomyia) albopictus (Diptera: Culicidae) in the contiguous United States. *Journal of Medical Entomology*, *54*, 1605–1614.

Kramer-Schadt, S., Niedballa, J., Lindenborn, J., Reinfelder, V., Stillfried, M., Heckmann, I., Scharf, A. K., Hofer, H., Wilting, A., Pilgrim, J. D., Schröder, B., Augeri, D. M., Cheyne, S. M., Hearn, A. J., Ross, J., Macdonald, D. W., Mathai, J., Eaton, J., Marshall, A. J., … Belant, J. L. (2013). The importance of correcting for sampling bias in MaxEnt species distribution models. *Diversity and Distributions*, *19*, 1366–1379.

Lauzeral, C., Grenouillet, G., & Brosse, S. (2013). Spatial range shape drives the grain size effects in species distribution models. *Ecography*, *36*, 778–787.

Leroy, B., Meynard, C. N., Bellard, C., & Courchamp, F. (2016). Virtualspecies, an R package to generate virtual species distributions. *Ecography*, *39*, 599–607.

Liu, C. R., Berry, P. M., Dawson, T. P., & Pearson, R. G. (2005). Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, *28*, 385–393.

Liu, C. R., Newell, G., & White, M. (2016). On the selection of thresholds for predicting species occurrence with presence-only data. *Ecology and Evolution*, *6*, 337–348.

Liu, C. R., White, M., & Newell, G. (2018). Detecting outliers in species distribution data. *Journal of Biogeography*, *45*, 164–176.

Lobo, J. M., & Tognelli, M. F. (2011). Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data. *Journal for Nature Conservation*, *19*, 1–7.

Manzoor, S. A., Griffiths, G., & Lukac, M. (2018). Species distribution model transferability and model grain size – Finer may not always be better. *Scientific Reports*, *8*, 7168.

Moudrý, V., & Simova, P. (2012). Influence of positional accuracy, sample size and scale on modelling species distributions: A review. *International Journal of Geographical Information Science*, *26*, 2083–2095.

Nenzen, H. K., & Araujo, M. B. (2011). Choice of threshold alters projections of species range shifts under climate change. *Ecological Modelling*, *222*, 3346–3354.

Nezer, O., Bar-David, S., Gueta, T., & Carmel, Y. (2017). High-resolution species-distribution model based on systematic sampling and indirect observations. *Biodiversity and Conservation*, *26*, 421–437.

Park, D. S., & Davis, C. C. (2017). Implications and alternatives of assigning climate data to geographical centroids. *Journal of Biogeography*, *44*, 2188–2198.

Pearson, R. G., Dawson, T. P., & Liu, C. (2004). Modelling species distributions in Britain: A hierarchical integration of climate and land-cover data. *Ecography*, *27*, 285–298.

Peterson, A. T. (2014). *Mapping disease transmission risk: Enriching models using biogeography and ecology*. Johns Hopkins University Press.

Qiao, H. J., Escobar, L. E., Saupe, E. E., Ji, L. Q., & Soberón, J. (2017). A cautionary note on the use of hypervolume kernel density estimators in ecological niche modelling. *Global Ecology and Biogeography*, *26*, 1066–1070.

Quiner, C. A., & Nakazawa, Y. (2017). Ecological niche modeling to determine potential niche of Vaccinia virus: A case only study. *International Journal of Health Geographics*, *16*, 28.

R Core Team. (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/

Record, S., Strecker, A., Tuanmu, M. N., Beaudrot, L., Zarnetske, P., Belmaker, J., & Gerstner, B. (2018). Does scale matter? A systematic review of incorporating biological realism when predicting changes in species distributions. *PLoS ONE*, *13*, e0194650.

Savicky, P. (2014). *pspearman: Spearman's rank correlation test. R package version 0.3-0*. https://CRAN.R-project.org/package=pspearman

Stolar, J., & Nielsen, S. E. (2015). Accounting for spatially biased sampling effort in presence-only species distribution modelling. *Diversity and Distributions*, *21*, 595–608.

Tjaden, N. B., Caminade, C., Beierkuhnlein, C., & Thomas, S. M. (2018). Mosquito-borne diseases: Advances in modelling climate-change impacts. *Trends in Parasitology*, *34*, 227–245.

Tuszynski, J. (2014). *caTools: Tools: moving window statistics, GIF, Base64, ROC AUC, etc. R package version 1.17.1.2*. https://CRAN.R-project.org/package=caTools

Warren, D. L., & Seifert, S. N. (2011). Ecological niche modeling in Maxent: The importance of model complexity and the performance of model selection criteria. *Ecological Applications*, *21*, 335–342.

Yackulic, C. B., Chandler, R., Zipkin, E. F., Royle, J. A., Nichols, J. D., Grant, E. H. C., & Veran, S. (2013). Presence-only modelling using MAXENT: When can we trust the inferences? *Methods in Ecology and Evolution*, *4*, 236–243.

Yates, K. L., Bouchet, P. J., Caley, M. J., Mengersen, K., Randin, C. F., Parnell, S., Fielding, A. H., Bamford, A. J., Ban, S., Barbosa, A. M., & Dormann, C. F. (2018). Outstanding challenges in the transferability of ecological models. *Trends in Ecology and Evolution*, *33*, 790–802.

BIOSKETCH

Yanchao Cheng works in the field of biogeography, with a focus on vector-borne diseases. She is interested in interdisciplinary modelling research. She compares and integrates ecological niche modelling and epidemiological modelling approaches, and works towards improving models' performance.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Cheng Y, Tjaden NB, Jaeschke A, Thomas SM, Beierkuhnlein C. Using centroids of spatial units in ecological niche modelling: Effects on model performance in the context of environmental data grain size. *Global Ecol Biogeogr*. 2021;00:1–11. https://doi.org/10.1111/geb.13240