Journal of Biogeography (J. Biogeogr.) (2017)

METHODOLOGICAL APPLICATIONS

Implications and alternatives of assigning climate data to geographical centroids

Daniel S. Park* and Charles C. Davis

Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA

ABSTRACT

Aim When precise coordinate data for training species distribution models (SDMs) are lacking, climatic variables are often assigned to centroids of geopolitically defined regions, frequently counties. This is problematic because approximations using centroids may not be representative of the regional climate or the locality from where species actually occur, thus leading to spurious conclusions. We evaluated county centroid climate versus simple alternatives for assigning climate to species observations in the absence of precise occurrence data.

Location United States of America.

Methods We assessed the disparity between the actual climate of all points within a county and metrics estimating county climate using the climate of geographical centroid, mean county climate and median county climate. To further evaluate the performance of these metrics, we generated SDMs of four common species using these estimates and compared the results with observed species distributions (red trillium, Pacific trillium, tall thistle and annual fleabane). Finally, we projected future ranges for annual fleabane to examine the difference in predicted range change between models.

Results Mean and median climate metrics were significantly better fits for approximating the climate of specimen records than climate of the geographical centroid. Moreover, county mean climate SDMs were the most similar to SDMs using actual coordinate data. In contrast, models applying climate to county centroid significantly overpredicted species range. This had implications for future projections of annual fleabane SDMs: the county centroid model predicted a decrease in suitable habitats for this species while other models predicted an increase.

Main conclusions County centroid climate, although commonly applied, is not suitable for SDMs as a means to approximate species climate when locality data are less precise. When only county level data are available, and more computationally intensive methods of accounting for spatial uncertainty cannot be readily implemented, we suggest considering mean county climate as an alternative.

Keywords

climate change, county centroid, ecological niche modelling, *Erigeron annuus*, georeferencing error, niche, species distribution modelling, positional uncertainty

*Correspondence: Daniel S. Park, Department of Organismic and Evolutionary Biology, Harvard University Herbaria, 22 Divinity Ave. Harvard University, Cambridge, MA 02138, USA.
E-mail: danielpark@fas.harvard.edu

Journal of Biogeography

INTRODUCTION

Characterizing the spatial distribution of species and their associated environmental requirements is a central component of biogeography, conservation biology and ecology. Species distribution models (SDMs), also known as ecological niche models, represent a powerful tool for accomplishing these goals (Hannah *et al.*, 2002; Pearson & Dawson, 2003; Soberón & Peterson, 2005; Elith & Leathwick, 2009; Peterson & Soberón, 2012). Despite their utility, however, a major challenge to developing SDMs has been the availability and quality of input data (Wolf *et al.*, 2011; Beck *et al.*, 2014; Fourcade *et al.*, 2014). Although species occurrence data, which are provided largely by natural history museums (Graham *et al.*, 2004; Newbold, 2010), are increasingly available via massive open-access databases such as the Global Biodiversity Information Facility (GBIF; www.gbif.org), these records often exhibit strong geographical, temporal and taxonomic biases (Dennis *et al.*, 2000; Wolf *et al.*, 2011; Isaac & Pocock, 2015; Meyer *et al.*, 2016). Despite recent efforts to categorize these biases (Wolf *et al.*, 2011; Meyer *et al.*, 2016), one key issue has received comparatively less attention: Whether geographical occurrence data commonly used to train SDMs accurately reflect the location of the specimen record, and thus the climate, of where it was collected/observed (Naimi *et al.*, 2011, 2014). Using occurrence records that are unrepresentative of the climatic conditions species occupy can lead to potentially spurious conclusions. This is further confounded by the fact that species occurrence records are commonly available at less fine scale resolution than environmental variables (Keil *et al.*, 2013), and the majority of biological collections and survey information do not include precise coordinate data (Wieczorek *et al.*, 2004; Naimi *et al.*, 2014).

Due to the lack of accurate occurrence data, records with imprecise locality information are routinely approximated to the centroids of high-order geopolitical regions, such as states, provinces, municipalities, townships and especially counties. Alternatively, imprecise localities can also be georeferenced to the centroid of a polygon or circle associated with a locality and/or degree of uncertainty (e.g. point-radius method; Wieczorek *et al.*, 2004). These centroids often serve as the location for which climate is extracted and assigned to the collection record (Fitzpatrick *et al.*, 2007). This practice commonly ignores the uncertainty inherent in georeferenced location data by treating these centroids as precise point occurrences, and by extension assumes accurate knowledge of the climate variables associated with the locality (Feeley & Silman, 2010). Such positional uncertainty can lead to spurious estimations of species–environment relationships (Dormann *et al.*, 2008; Beale & Lennon, 2012; Tulowiecki *et al.*, 2014), the magnitude of which is determined by the level of local spatial autocorrelation in the environmental variables (Naimi *et al.*, 2011, 2014). For example, Montrose County in Colorado spans the depths of the Black Canyon to the peaks of the Rocky Mountains, and daily temperatures can vary over 20 °C between these locations. In such cases, not only is it unlikely that the centroid accurately represents the environmental conditions where the specimen was collected, but it is unlikely to be a suitable representation of the average climatic conditions of the region.

Despite these pitfalls, using county centroid values to represent the climatic conditions appears to have become an increasingly common practice in SDMs when more precise coordinate data are lacking (Fitzpatrick *et al.*, 2007; Medley, 2010; Duehl *et al.*, 2011; Zhu *et al.*, 2012; Escobar *et al.*, 2013; Harrigan *et al.*, 2014; Wells & Tonkyn, 2014). Perhaps, more pervasive is the unintentional use of county centroids given their apparent prevalence in data portals like GBIF. Indeed, an examination of GBIF reveals that nearly half the coordinate data of certain major groups of organisms in the United States fall roughly within 20 km of the county centroid (Table 1). In addition, depending on taxonomic group, 40–99% of the coordinates were not assigned quantifiable metrics of uncertainty. Despite recent calls by GBIF for better inclusion of indicators of uncertainty for georeferenced data (Anderson *et al.*, 2016), these results suggest that a potentially large number of studies that use such data are estimating climate based on localities that are suspect at best. Indeed, many SDMs are projected at fine spatial resolutions without accounting for uncertainty, asserting confidence in outputs that may be misleading (Refsgaard *et al.*, 2007; Sinclair *et al.*, 2010; Wenger *et al.*, 2013; Gould *et al.*, 2014). Importantly, the potential effect of using county centroid data is seldom, if ever, considered (Costa *et al.*, 2010; Beck *et al.*, 2014).

Given the critical importance of associating climate to geographical data especially for SDMs, a broader assessment of these assumptions needs to be evaluated. Here, we investigate how well the geographical centroid represents the climatic conditions of counties and compare these results to simple alternatives, including utilizing county climate averages (Iverson & Prasad, 1998; Ulrichs & Hopper, 2008; Loeb & Winters, 2013). We then extend our analyses to include a performance assessment of SDMs generated for four common species in the United States based on centroid climate and alternative metrics: Red trillium (*Trillium erectum* L.), Pacific trillium (*T. ovatum* Pursh), tall thistle (*Cirsium altissimum* (L.) Hill) and annual fleabane (*Erigeron annuus* (L.)

Table 1 Proportion of GBIF occurrence data approximating to US county centroids. The last column depicts the percentage of records that do not have any associated measure of uncertainty.

Taxon	2.5 min	5 min	10 min	Uncertainty not recorded
Mammalia	3%	7%	22%	39%
Tracheophyta	3%	8%	25%	52%
Aves	6%	16%	45%	99%
Fungi	3%	8%	23%	82%
Insecta	6%	15%	38%	70%

Pers). Finally, we present a case study involving future SDM projections of annual fleabane to illustrate the different conclusions reached using different estimates of county climate.

MATERIALS AND METHODS

Evaluating the prevalence of county centroid occurrence data

To assess the proportion of occurrence data that may have been approximated to county centroids, we examined GBIF records from the United States. The United States is by far the largest contributor of primary biodiversity data to GBIF (Anderson *et al.*, 2016). We collected coordinate data from GBIF for the following major groups of taxa: mammals (class Mammalia), vascular plants (Tracheophyta), birds (class Aves), insects (class Insecta) and fungi (kingdom Fungi). For each of these groups, we then determined the proportion of these occurrences located within 2.5, 5 and 10 arc-minutes (roughly corresponding to 5, 10 and 20 km, respectively) of geographical county centroids, matching the degrees of map resolutions frequently used in SDMs (e.g. Knowles *et al.*, 2007; Zhang *et al.*, 2014; Park & Potter, 2015a,b).

Climatic layer analyses

We obtained global data on 19 bioclimatic variables (Nix, 1986; Busby, 1991) from the WorldClim dataset (Hijmans *et al.*, 2005), at 2.5′, 5′ and 10′ resolutions. WorldClim bioclimatic layers are among the most frequently used data in SDM studies. Lower resolutions were not considered because cell sizes become larger than most counties at coarser grains. All the geographical points (cells) within each county were separated onto axes of the 19 bioclimatic variables. We then calculated the Euclidian distance in this multidimensional climate space between all points occurring in every US county and (1) the centroid of the county, (2) the mean climate of the county, (3) the mean of climate values within the 95th percentile range in each county, (4) the median climate of the county and (5) the median of climate values within the 95th percentile range in each county. The 95th percentile mean and median climates exclude the more extreme climate values that occur in each county and were examined to assess the effects of climatic outliers. These values were compared with a Kruskal–Wallis one-way analysis of variance as the data were not all normally distributed. A post hoc multiple comparison Dunn test was used to further examine the nature of the significant differences we identified using the Kruskal–Wallis test. We then performed linear mixed models to determine which factors predict the degree of environmental heterogeneity present in counties. Here, we assigned the average distance in climate space between all points occurring in each county as the dependent variable. To account for the spatial autocorrelation present in county climate heterogeneity, we fit the spatial process to a Gaussian covariance model and used the model parameters to krige the data. The kriged projections were then subtracted from the observed data to yield a de-trended dataset of residuals on which subsequent models were fit. Predictors examined include standard deviation of county elevation, mean elevation, county area, longitude, latitude and distance to coast (Figs S1–S2 in Appendix S1 in Supporting Information). Counties comprising less than three cells at each resolution, as well as those with centroids falling in bodies of water, were excluded from consideration. We also repeated this analysis on each individual bioclimatic variable separately. All analyses were performed using R v3.3.2 (R Core Team, 2016).

Species distribution modelling

To assess performance of county climate estimates in an SDM framework, we generated SDMs for four easily identified, common native plant species of North America: red trillium (*Trillium erectum* L.), Pacific trillium (*T. ovatum* Pursh), tall thistle (*Cirsium altissimum* (L.) Hill) and annual fleabane (*Erigeron annuus* (L.) Pers). These plant species were selected to represent the heterogeneity present in county size and climate based on their distributions in the United States. *Trillium erectum* L. is an herbaceous perennial widespread in the northeastern United States; *T. ovatum* Pursh, also a perennial, is distributed across the west coast; *Cirsium altissimum* (L.) Hill, a biennial to short-lived perennial species, is mainly distributed in the central United States; and *Erigeron annuus* (L.) Pers., an annual herb, is widespread across 43 of the 48 contiguous states of the United States and is considered a weed in many areas.

Maxent (v3.3.3k; http://www.cs.princeton.edu/~schapire/maxent/) is a machine-learning method that searches for the probability distribution that maximizes entropy in a dataset of geographical occurrence points in relation to background environmental variables and can be used to project relative occurrence probabilities (Phillips *et al.*, 2006). We applied this approach to model the distribution of species because it has proven to be effective for presence-only records and small sample sizes (Hernandez *et al.*, 2006; Wisz *et al.*, 2008; Guo *et al.*, 2011), and has become the most widely used method for generating SDMs (Merow *et al.*, 2013; Fourcade *et al.*, 2014). Also, it has been found to be robust to moderate levels of georeferencing error and uncertainty (Graham *et al.*, 2008). We cleaned our GBIF occurrence records by discarding observations with coordinates that did not match the listed country/state/county, absent coordinate data, or coordinates that mapped to open water outside land boundaries. We selected the following seven bioclimatic variables strongly associated with the distribution of species in North America using a correlation coefficient below 0.75 (following Rissler *et al.*, 2006) to prevent multicollinearity and overfitting (Dormann *et al.*, 2013): isothermality (BIO2), minimum temperature of the coldest month (BIO6), mean temperature of the wettest quarter (BIO8), precipitation of wettest month (BIO13), precipitation

seasonality (BIO15), precipitation of the warmest quarter (BIO18) and precipitation of the coldest quarter (BIO19). For modelling, we used Maxent as implemented in the 'dismo' (Hijmans *et al.*, 2011) package in R with default settings and 5000 random background points. While the default settings for Maxent are not necessarily optimal for all species, they are commonly used, and our goal was not to assess the performance of Maxent *per se,* but to instead assess performance of models based on imprecise location data according to general modelling practices. We conducted up to 5000 simulations to allow the models adequate time to converge. For each species, SDMs were generated based on (1) all actual occurrence coordinates, (2) the centroids, (3) the mean, (4) the mean of climate values within the 95th percentile range in each county, (5) the median climate and (6) the median of climate values within the 95th percentile range in each county. To assess whether the models based on these metrics were significantly different from random, SDMs were also generated from 100 sets of random points from each county of occurrence to serve as a null model.

Evaluation of model performance

The commonly used SDM evaluation metric, the area under the receiver operating curve (AUC), is not directly comparable between models generated from disparate datasets (Elith *et al.*, 2011). Thus, for each species we evaluated each of the six models proposed above by projecting them onto a map of the United States and comparing the predicted occurrence probabilities for (1) actual occurrence points and (2) locales without collection records (i.e. potential non-occurrence points). These values were compared among models with a nonparametric Kruskal–Wallis one-way analysis of variance and a post hoc multiple comparison Dunn test to further examine the nature of the observed difference. Furthermore, for the species for which we were able to collect the most extensive occurrence data, *E. annuus*, we inferred its future distribution in 2050 based on the prediction of the Hadley Centre Global Environment Model version 2 (HadGEM2-ES, rcp60; Collins *et al.*, 2011). Next, we compared the difference between the predicted future and current distributions of *E. annuus* for each of our six alternative models. These results were compared to the predictions of the best-practice model generated using comprehensive coordinate data, by examining the sum of all predicted probabilities greater than 0.1, discarding points with lower occurrence probabilities as highly unlikely.

RESULTS

Climatic layer analyses

The standard deviation of elevation explained a significant proportion ($R^2 > 0.50$) of the variation in both the degree of climatic disparity between county centroids and the rest of the county, as well as the climatic heterogeneity of the county in general (Table 2). Mean elevation and longitude also had significant effects on county climate heterogeneity ($R^2 > 0.20$), but these variables were highly correlated with the standard deviation of elevation (Figs S1–S2 in Appendix S1). The climatic distance between all points in the county increased with the elevational heterogeneity of the landscape, and simultaneously the climate of the county centroid was increasingly a poor representative of county climate (Fig. 1). This effect was significant across all spatial scales and applied to other estimates of county climate as well, albeit to a lesser degree (Fig. S3 in Appendix S1). The effect of county elevation standard deviation on climatic heterogeneity varied among individual bioclimatic variables, generally being greater in temperature-related variables than precipitation-related ones (see Table S1 in Appendix S1). Other variables, such as distance to coast and county size, had minimal effects on county climate heterogeneity (Fig. S1 in Appendix S1).

Table 2 Results of six separate linear mixed models analysing the effect of the standard deviation of US county elevation on the degree of disparity between geographical centroid climates and that of all other areas as well as overall county climate heterogeneity.

Response	Marginal R^2	Conditional R^2	χ^2	*P*-value
Heterogeneity of county climate				
2.5 min	0.5755	0.9166	2651.8	< 0.001
5 min	0.5627	0.9083	2541.8	< 0.001
10 min	0.5403	0.8739	2193.0	< 0.001
Disparity of county centroid				
2.5 min	0.5690	0.9313	2610.4	< 0.001
5 min	0.5830	0.9214	2689.7	< 0.001
10 min	0.5551	0.8491	2281.1	< 0.001

The mean climate of the county, the mean of climate values within the 95th percentile range in each county, the median climate of the county and the median of climate values within the 95th percentile range in each county were closer in climate space to all points of the county than was the climatic value of the county geographical centroid (Fig. 2). The same applied to individual bioclimatic variables at all resolutions. On average, the county mean was the closest to all points in the county (Table S2 in Appendix S1).

Species distribution modelling

The SDMs based on the county centroid climate, county mean climate, 95th percentile county mean climate, county median climate and the 95th percentile county median climate projected higher relative climatic suitability onto known occurrence points on average than models created with actual coordinate data (Fig. 3). However, there was no significant difference in the projections between these models except for *Trillium ovatum*, where models based on county centroid data predicted significantly lower relative probabilities at known occurrence points. However, models based on

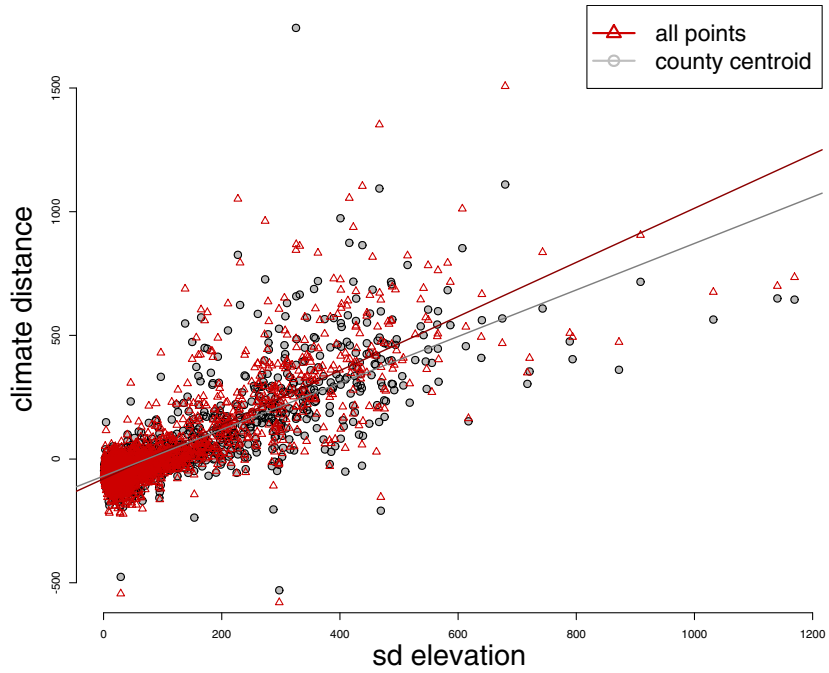

Figure 1 Distance between points in US counties by standard deviation of county elevation in 19-dimensional climate space at 2.5-minute resolution. The lines represent linear regression results for the climatic disparity between the county centroid and other points in the county (grey circles) and that between all points within the county (red triangles).

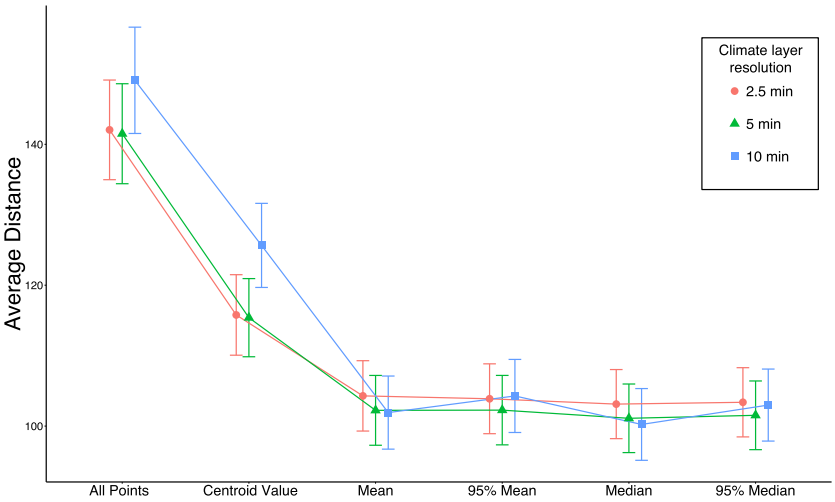

Figure 2 Average distance between metrics representing US county climates and all other points in the county in 19-dimensional climate space. The values for 'all points' represent the average pairwise distance between all points in a county in climate space. Shapes represent the average value across all counties and bars represent the 95% confidence interval.

county centroid data always assigned significantly higher relative occurrence probabilities ($P < 0.001$) to non-occurrence points (i.e. where there are no known collections of the species) than all other models. Models generated with all available coordinate data were the most conservative in this regard. In addition, county centroid models did not perform better than SDMs generated from random points within each occurrence county on average (Fig. 4, Fig. S4 in Appendix S1). The relative importance of different bioclimatic variables to each SDM varied depending on the data used to train the model (Table S3 in Appendix S1). However, the variables with the largest influence on species distributions were generally found to be in common.

When the different SDMs of *Erigeron annuus* were projected onto a future climate scenario (2050), the SDM based on county centroid data predicted that the suitable range for the species would decrease (Fig. 5). In contrast, all other models, including the model generated using actual occurrence point data, predicted an increase. The prediction using the county mean climate was the closest model to that of the model using coordinate data.

DISCUSSION

Our analyses indicate that a potentially significant percentage of the coordinate data in GBIF have been approximated to county centroids (Table 1). The proportions of coordinates approximating to county centroids were especially high for birds and insects, where 15% of the occurrences were roughly within 10 km of the geographical centre of the

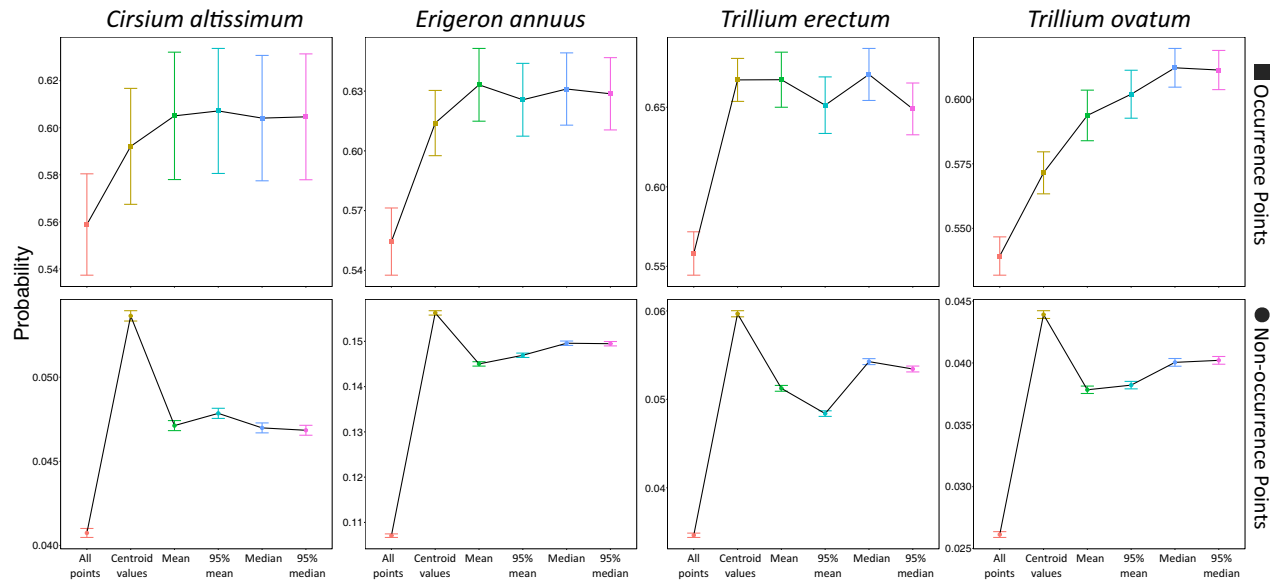

Figure 3 Predicted species distribution probabilities at species occurrence points and all other non-occurrence points (presumed absences) in the United States. Squares and circles represent the mean predicted probability, and bars indicate the 95% confidence interval. The *x*-axis depicts the different types of climate data used to create each species distribution model: that from county centroid values, county means, the mean of the 95th percentile of county values, county medians, the median of the 95th percentile of county values and all occurrence points.

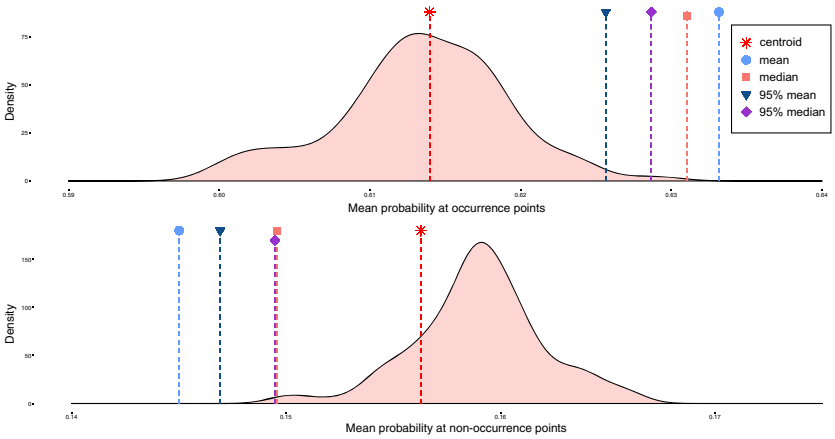

Figure 4 Distribution of occurrence probabilities in the United States predicted by 100 random point models for *Erigeron annuus*. Shapes represent the mean occurrence probabilities predicted by models based on different county climate metrics.

county. In addition, our results demonstrate that using the climate of geographical centroids is problematic and that simple alternative estimates of climate perform significantly better. Lastly, we show that the use of centroid climate to model and predict the future distribution of species can differ dramatically from assessments based on more appropriate approximations of climate. We outline these findings in more detail in the following.

Geographical centroids of county are not representative climate proxies

Our results demonstrate that the climate of the county centroid is a poor representative of county climate. In contrast, we demonstrate that the mean/median climate of a county is significantly closer to the centre of the county climate space and therefore better reflects county climate. Not surprisingly, the relative performance of these metrics is strongly influenced by the degree of environmental heterogeneity within a county. While it has been suggested that climate estimates for larger counties may be less accurate when applying climate to county centroid (Harrigan *et al.*, 2014), we find instead that the standard deviation of elevation best explains the majority ($R^2 > 0.50$) of climatic heterogeneity. Whereas many large political areas can show relatively little variation in climate, such as throughout the Midwestern United States, elevational gradients often accompany corresponding variation in key abiotic features, including temperature and moisture. One concern that arises in light of these findings is that climatically variable

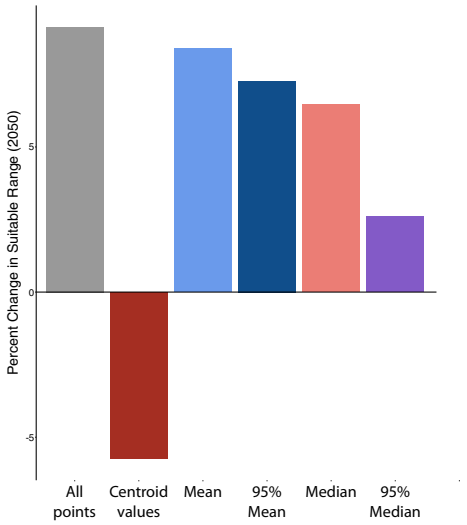

Figure 5 Percent change in suitable range for *Erigeron annuus* in 2050 in the United States. Bars represent the predicted change in the sum of all occurrence probabilities above 0.1. The *x*-axis depicts the different types of climate data used to create each species distribution model.

montane regions harbour the majority of the biodiversity in the world and are expected to experience elevated rates of climate change (Rahbek, 1995; Jenkins *et al.*, 2013; Mountain Research Initiative EDW Working Group, 2015; Osborne & Mclanahan, 2016). Thus, it is especially important to understand the climatic niches of species in these spatially heterogeneous regions. However, large-scale patterns of species turnover occur across relatively small areas in montane environments (see Hoorn *et al.*, 2013; Hughes & Atchison, 2015), and approximating locality data to regional centroids are likely to result in highly spurious outcomes. Our results suggest that extra caution should be exercised with geographical approximations from regions encompassing highly variable elevations with high regional species diversity.

SDMs based on geographical centroid climate data are unreliable

We demonstrated that SDMs based on county mean/median climate variables perform significantly better than those applying geographical centroid values and more closely resemble models generated with actual coordinate data. Indeed, SDMs based on geographical centroid values tended to greatly overestimate species range and were often no better than models created from random point data (Fig. 4). This is presumably due to the fact that the geographical climate centroid can be more or less a random draw from the environmental heterogeneity present in (especially) large politically defined geographical areas such as counties (Fig. 6). Associating randomly assembled environmental data with species can make them appear capable of tolerating a wider range of environmental conditions, potentially portraying them as more generalist than they may be. SDMs based on such data may overestimate the conditions in which species persist, yielding generally high probability scores for both suitable and unsuitable habitats. It is possible that presumed overestimation of SDMs could be an artefact of false absences (Gu & Swihart, 2004). Indeed, biological collection/survey data may not reflect the full distribution of a species. In this case, we would expect higher mean predicted probabilities at non-occurrence points as the number of false absences in the data increases. However, in our study, all four species have well-known ranges and have been collected extensively, likely minimizing the degree of false absences. In addition, the number of true absences across the United States most certainly outnumber undocumented occurrences, making the effect of false absences minimal when examining SDM predictions at non-occurrence points. Furthermore, we demonstrate that the climate at geographical centroids is less representative of county climatic conditions than mean county climate (Fig. 2); hence, centroid-based SDMs are likely to be more inaccurate in their predictions.

These differences can have marked effect on the projection of future species distributions. As we demonstrate in

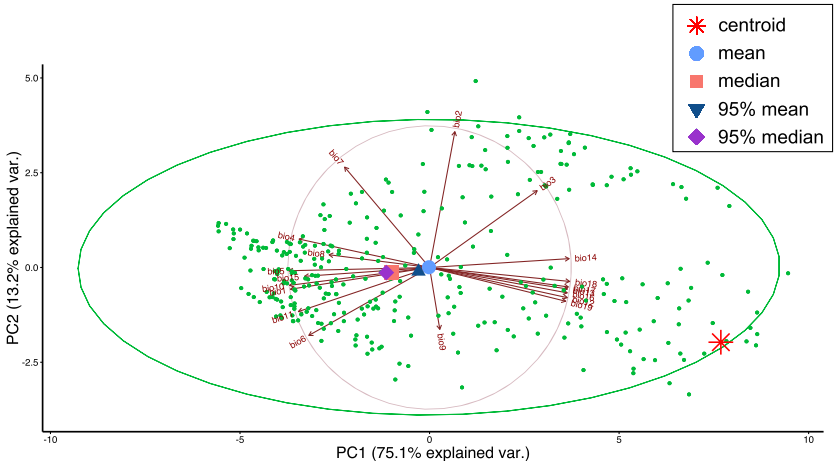

Figure 6 Ordination of points in Montrose County, Colorado, in climate space. Small green dots depict each cell of the county at 2.5-minute resolution. Shapes represent different metrics of county climate. The large green oval comprises 95% of the county and arrows depict eigenvectors (bioclimatic variables).

our SDM projections of the future range of the weedy *Erigeron annuus*, the use of geographical centroid results in dramatic departures from the preferred alternative of using actual coordinate data (Fig. 5). Along these lines, only the SDM trained with county centroid climate predicted a decrease in range. Such discrepancies can have serious downstream consequences for conservation planning and/or invasive species management. Unlike county centroid climate, the mean climate is never a truly incorrect estimation of the climate of any point within a county. Using county mean climates in SDMs is akin to modelling species distributions with precise coordinates on lower resolution climate maps, where the small variations in specific habitat climates are lost, but the results can still be relevant for large-scale predictions. All things being equal, the effects of locational uncertainty on SDM performance are likely to be more severe for specialist species, as approximate locations, such as county centroids, will frequently misrepresent their more specific environmental requirements compared with generalists with broader climatic niches (Tulowiecki *et al.*, 2014).

Discarding imprecise data may not be a viable option

One solution to handling the issues we raise regarding county-level occurrence data, and uncertain occurrence data in general, is to discard records beyond a given error threshold (Feeley & Silman, 2010). However, due to the rapid influx of digitized museum collections (Baird, 2010; Beaman & Cellinese, 2012; Balke *et al.*, 2013), the global accumulation of occurrence records is likely to far outpace our ability to efficiently curate and georeference these data manually. Moreover, it is often impossible to determine whether a specimen was actually collected at the given location, or georeferenced to the location of a geographical approximation like county centroid. For example, in the effort to map all mammal collections as part of the Mammal Networked Information System (http://manisnet.org/), 78.4% of the records that had coordinate data were found to have no associated metadata regarding the localities, nor did they include information about the methods and assumptions involved in assigning coordinates (Wieczorek *et al.*, 2004). We further confirm this issue in other groups as well (Table 1). Perhaps most importantly, the paucity of accurate occurrence records for many species makes discarding records unpalatable for most researchers, and resources are likely to be scarce for broad systematic field surveys to better document biodiversity (Phillips *et al.*, 2009). This is especially so for studies with a temporal component such as those using specimens to examine phenological change (Davis *et al.*, 2015) or species distributional changes over time (Lütolf *et al.*, 2006).

A better approach going forward is to revamp our models to improve workflows for handling coarse-level occurrence data. Various methods have been suggested towards this end, from restricting ranges using *a priori* knowledge of species habitat preferences (Jetz *et al.*, 2007; Niamir *et al.*, 2011; Rondinini *et al.*, 2011) or physiological tolerances (Kearney & Porter, 2009), to bootstrapping approaches (Fernández *et al.*, 2013). Among these, Bayesian methods that account for uncertainty in occurrence data are especially promising (e.g. McInerny & Purves, 2011; Beale & Lennon, 2012; Keil *et al.*, 2013; Velásquez-Tibatá *et al.*, 2016). These approaches represent the best current standard of practice but are often more computationally demanding compared with alternative SDM approaches, making them relatively difficult to implement on a large scale (Velásquez-Tibatá *et al.*, 2016). Bayesian methods also require additional information, such as estimations of uncertainty or absence data, which are unavailable for many species (Table 1). Furthermore, incorporating such methods with commonly used SDM approaches, such as Maxent, remain a challenge (Keil *et al.*, 2013).

Along these lines, when dealing with large-scale SDM analyses, a simpler approach could be to summarize covariate data at the unit of sampling (e.g. county) by using measures of central tendency, such as the mean or median of continuous data (Young *et al.*, 2009). While not preferable to Bayesian approaches or more intricate means of georeferencing, we have demonstrated that the mean is a better option than using geographical centroid values to represent the climate of an area and exhibits superior performance in SDMs. A bootstrapping approach where random locations within a county are repeatedly sampled would yield similar results, with the added benefit of propagating uncertainty through the modelling approach. However, in large-scale modelling efforts such as those involving thousands of different species with large distributions (e.g. Zhang *et al.*, 2016), bootstrapping could become impractical. Nonetheless, issues of computational efficiency regarding Bayesian or bootstrapping approaches may be of little consequence in the near future with improving processor speeds, availability of multithreading processing technology and more efficient sampling algorithms (Velásquez-Tibatá *et al.*, 2016). In the meantime, for the purposes of large-scale SDM studies, mean climate can provide a simple alternative that is significantly better than the climate of geographical centroids, especially when *a priori* knowledge of the species and computational resources are limited. Above all, our results further highlight the need for better data standards regarding the accuracy and treatment of georeferenced locality data.

ACKNOWLEDGEMENTS

The authors thank D. S. Barrington, C. G. Willis and A. M. Ellison for their insightful comments on the project and manuscript. We are also grateful to D. S. Chapman, L. Barwell and all anonymous referees for their invaluable comments on previous versions of the manuscript. This work was supported by the Harvard University Herbaria.

REFERENCES

Anderson, R.P., Araújo, M.B., Guisan, A., Lobo, J.M., Martínez-Meyer, E., Peterson, A.T. & Soberón, J. (2016) Are species occurrence data in global online repositories fit for modeling species distributions? The case of the Global Biodiversity Information Facility (GBIF). Final Report of the Task Group on GBIF Data Fitness for Use in Distribution Modelling. Version 1.1, 1–27.

Baird, R.C. (2010) Leveraging the fullest potential of scientific collections through digitisation. *Biodiversity Informatics*, **7**, 130–136.

Balke, M., Schmidt, S., Hausmann, A. *et al.* (2013) Biodiversity into your hands – a call for a virtual global natural history "metacollection". *Frontiers in Zoology*, **10**, 55.

Beale, C.M. & Lennon, J.J. (2012) Incorporating uncertainty in predictive species distribution modelling. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 247–258.

Beaman, R.S. & Cellinese, N. (2012) Mass digitization of scientific collections: new opportunities to transform the use of biological specimens and underwrite biodiversity science. *ZooKeys*, **209**, 7–17.

Beck, J., Böller, M., Erhardt, A. & Schwanghart, W. (2014) Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. *Ecological Informatics*, **19**, 10–15.

Busby, J.R. (1991) BIOCLIM A Bioclimate Analysis and Prediction System. *Plant Protection Quarterly*, **6**, 8–9.

Collins, W.J., Bellouin, N., Doutriaux-Boucher, M., Gedney, N., Halloran, P., Hinton, T., Hughes, J., Jones, C.D., Joshi, M., Liddicoat, S., Martin, G., O'Connor, F., Rae, J., Senior, C., Sitch, S., Totterdell, I., Wiltshire, A. & Woodward, S. (2011) Development and evaluation of an Earth-System model – HadGEM2. *Geoscientific Model Development*, **4**, 1051–1075.

Costa, G.C., Nogueira, C., Machado, R.B. & Colli, G.R. (2010) Sampling bias and the use of ecological niche modeling in conservation planning: a field evaluation in a biodiversity hotspot. *Biodiversity and Conservation*, **19**, 883–899.

Davis, C.C., Willis, C.G., Connolly, B., Kelly, C. & Ellison, A.M. (2015) Herbarium records are reliable sources of phenological change driven by climate and provide novel insights into species' phenological cueing mechanisms. *American Journal of Botany*, **102**, 1599–1609.

Dennis, R., Thomas, C., Thomas, C.D. & Sciences, M. (2000) Bias in butterfly distribution maps: the influence of hot spots and recorder's home range. *Journal of Insect Conservation*, **4**, 73–77.

Dormann, C.F., Purschke, O., Garcia Marquez, J.R., Lautenbach, S. & Schroder, B. (2008) Components of uncertainty in species distribution analysis: a case study of the Great Grey Shrike. *Ecology*, **89**, 3371–3386.

Dormann, C.F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J.R.G., Gruber, B., Lafourcade, B., Leitão, P.J., Münkemüller, T., Mcclean, C., Osborne, P.E., Reineking, B., Schröder, B., Skidmore, A.K., Zurell, D. & Lautenbach, S. (2013) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, **36**, 027–046.

Duehl, A.J., Koch, F.H. & Hain, F.P. (2011) Southern pine beetle regional outbreaks modeled on landscape, climate and infestation history. *Forest Ecology and Management*, **261**, 473–479.

Elith, J. & Leathwick, J. (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., Peterson, A.T., Favi, M., Yung, V., Pons, D.J. & Medina-Vogel, G. (2013) Ecology and geography of transmission of two bat-borne rabies lineages in Chile. *PLoS Neglected Tropical Diseases*, **7**, 1–10.

Feeley, K.J. & Silman, M.R. (2010) Modelling the responses of Andean and Amazonian plant species to climate change: the effects of georeferencing errors and the importance of data filtering. *Journal of Biogeography*, **37**, 733–740.

Fernández, M., Hamilton, H. & Kueppers, L. (2013) Characterizing uncertainty in species distribution models derived from interpolated weather station data. *Ecosphere*, **4**, 1–17.

Fitzpatrick, M.C., Weltzin, J.F., Sanders, N.J. & Dunn, R.R. (2007) The biogeography of prediction error: why does the introduced range of the fire ant over-predict its native range? *Global Ecology and Biogeography*, **16**, 24–33.

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**, 1–13.

Gould, S.F., Beeton, N.J., Harris, R.M.B., Hutchinson, M.F., Lechner, A.M., Porfirio, L.L. & Mackey, B.G. (2014) A tool for simulating and communicating uncertainty when modelling species distributions under future climates. *Ecology and Evolution*, **4**, 4798–4811.

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C. & Peterson, A.T. (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution*, **19**, 497–503.

Graham, C.H., Elith, J., Hijmans, R.J. *et al.* (2008) The influence of spatial errors in species occurrence data used in distribution models. *Journal of Applied Ecology*, **45**, 239–247.

Gu, W. & Swihart, R.K. (2004) Absent or undetected? Effects of non-detection of species occurrence on wildlife-habitat models. *Biological Conservation*, **116**, 195–203.

Guo, Q., Li, W., Liu, Y. & Tong, D. (2011) Predicting potential distributions of geographic events using one-class data: concepts and methods. *International Journal of Geographical Information Science*, **25**, 1697–1715.

Hannah, L., Midgley, G.F. & Millar, D. (2002) Climate change-integrated conservation strategies. *Global Ecology and Biogeography*, **11**, 485–495.

Harrigan, R.J., Thomassen, H.A., Buermann, W. & Smith, T.B. (2014) A continental risk assessment of West Nile virus under climate change. *Global Change Biology*, **20**, 2417–2425.

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., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hijmans, R.J., Phillips, S., Leathwick, J. & Elith, J. (2011) dismo: Species distribution modeling. R package version 0.6-3.

Hoorn, C., Mosbrugger, V., Mulch, A. & Antonelli, A. (2013) Biodiversity from mountain building. *Nature Geoscience*, **6**, 154–154.

Hughes, C.E. & Atchison, G.W. (2015) The ubiquity of alpine plant radiations: from the Andes to the Hengduan Mountains. *New Phytologist*, **207**, 275–282.

Isaac, N.J.B. & Pocock, M.J.O. (2015) Bias and information in biological records. *Biological Journal of the Linnean Society*, **115**, 522–531.

Iverson, L.R. & Prasad, A.M. (1998) Predicting abundance of 80 tree species following climate change in the eastern United States. *Ecological Monographs*, **68**, 465–485.

Jenkins, C.N., Pimm, S.L. & Joppa, L.N. (2013) Global patterns of terrestrial vertebrate diversity and conservation. *Proceedings of the National Academy of Sciences USA*, **110**, E2602–10.

Jetz, W., Wilcove, D.S. & Dobson, A.P. (2007) Projected impacts of climate and land-use change on the global diversity of birds. *PLoS Biology*, **5**, 1211–1219.

Kearney, M. & Porter, W. (2009) Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. *Ecology Letters*, **12**, 334–350.

Keil, P., Belmaker, J., Wilson, A.M., Unitt, P. & Jetz, W. (2013) Downscaling of species distribution models: a hierarchical approach. *Methods in Ecology and Evolution*, **4**, 82–94.

Knowles, L.L., Carstens, B.C. & Keat, M.L. (2007) Coupling genetic and ecological-niche models to examine how past population distributions contribute to divergence. *Current Biology*, **17**, 940–946.

Loeb, S.C. & Winters, E.A. (2013) Indiana bat summer maternity distribution: effects of current and future climates. *Ecology and Evolution*, **3**, 103–114.

Lütolf, M., Kienast, F. & Guisan, A. (2006) The ghost of past species occurrence: improving species distribution models for presence-only data. *Journal of Applied Ecology*, **43**, 802–815.

McInerny, G.J. & Purves, D.W. (2011) Fine-scale environmental variation in species distribution modelling: regression dilution, latent variables and neighbourly advice. *Methods in Ecology and Evolution*, **2**, 248–257.

Medley, K.A. (2010) Niche shifts during the global invasion of the Asian tiger mosquito, Aedes albopictus Skuse (Culicidae), revealed by reciprocal distribution models. *Global Ecology and Biogeography*, **19**, 122–133.

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.

Meyer, C., Weigelt, P., Kreft, H. & Lambers, J.H.R. (2016) Multidimensional biases, gaps and uncertainties in global plant occurrence information. *Ecology Letters*, **19**, 992–1006.

Mountain Research Initiative EDW Working Group (2015) Elevation-dependent warming in mountain regions of the world. *Nature Climate Change*, **5**, 424–430.

Naimi, B., Skidmore, A.K., Groen, T.A. & Hamm, N.A.S. (2011) Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modelling. *Journal of Biogeography*, **38**, 1497–1509.

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**, 191–203.

Newbold, T. (2010) Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. *Progress in Physical Geography*, **34**, 3–22.

Niamir, A., Skidmore, A.K., Toxopeus, A.G., Muñoz, A.R. & Real, R. (2011) Finessing atlas data for species distribution models. *Diversity and Distributions*, **17**, 1173–1185.

Nix, H.A. (1986) A biogeographic analysis of Australian Elaphid snakes. *Atlas of Elapid Snakes of Australia*. (ed. by R. Longmore), pp. 4–15. Australian Government Publishing Service, Canberra.

Osborne, C. & Mclanahan, S. (2016) Linked references are available on JSTOR for this article. *Partnership Instability and Child Well-Being*, **69**, 1065–1083.

Park, D.S. & Potter, D. (2015a) A reciprocal test of Darwin's naturalization hypothesis in two Mediterranean-climate regions. *Global Ecology and Biogeography*, **24**, 1049–1058.

Park, D.S. & Potter, D. (2015b) Why close relatives make bad neighbours: phylogenetic conservatism in niche preferences and dispersal disproves Darwin's naturalization hypothesis in the thistle tribe. *Molecular Ecology*, **24**, 3181–3193.

Pearson, R.G. & Dawson, T.P. (2003) Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? *Global Ecology and Biogeography*, **12**, 361–371.

Peterson, A.T. & Soberón, J. (2012) Integrating fundamental concepts of ecology, biogeography, and sampling into effective ecological niche modeling and species distribution modeling. *Plant Biosystems*, **146**, 789–796.

Phillips, S.J., Anderson, R.P. & Schapire, R.E. (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Phillips, S.J., Dudík, M., Elith, J., Graham, C.H., Lehmann, A., Leathwick, J. & Ferrier, S. (2009) Sample selection bias

and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, **19**, 181–197.

R Core Team (2016) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Rahbek, C. (1995) The elevational gradient of species richness: a uniform pattern? *Ecography*, **18**, 200–205.

Refsgaard, J.C., van der Sluijs, J.P., Højberg, A.L. & Vanrolleghem, P.A. (2007) Uncertainty in the environmental modelling process – A framework and guidance. *Environmental Modelling and Software*, **22**, 1543–1556.

Rissler, L.J., Hijmans, R.J., Graham, C.H., Moritz, C. & Wake, D.B. (2006) Phylogeographic lineages and species comparisons in conservation analyses: a case study of California herpetofauna. *The American Naturalist*, **167**, 655–66.

Rondinini, C., Di Marco, M., Chiozza, F., Santulli, G., Baisero, D., Visconti, P., Hoffmann, M., Schipper, J., Stuart, S.N., Tognelli, M.F., Amori, G., Falcucci, A., Maiorano, L. & Boitani, L. (2011) Global habitat suitability models of terrestrial mammals. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **366**, 2633–2641.

Sinclair, S.J., White, M.D. & Newell, G.R. (2010) How useful are species distribution models for managing biodiversity under future climates? *Ecology and Society*, **15**, 1–13.

Soberón, J. & Peterson, A.T. (2005) Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Informatics*, **2**, 1–10.

Tulowiecki, S.J., Larsen, C.P.S. & Wang, Y.C. (2014) Effects of positional error on modeling species distributions: a perspective using presettlement land survey records. *Plant Ecology*, **216**, 67–85.

Ulrichs, C. & Hopper, K.R. (2008) Predicting insect distributions from climate and habitat data. *BioControl*, **53**, 881–894.

Velásquez-Tibatá, J., Graham, C.H. & Munch, S.B. (2016) Using measurement error models to account for georeferencing error in species distribution models. *Ecography*, **39**, 305–316.

Wells, C.N. & Tonkyn, D.W. (2014) Range collapse in the Diana fritillary, Speyeria diana (Nymphalidae). *Insect Conservation and Diversity*, **7**, 365–380.

Wenger, S.J., Som, N.A., Dauwalter, D.C., Isaak, D.J., Neville, H.M., Luce, C.H., Dunham, J.B., Young, M.K., Fausch, K.D. & Rieman, B.E. (2013) Probabilistic accounting of uncertainty in forecasts of species distributions under climate change. *Global Change Biology*, **19**, 3343–3354.

Wieczorek, J., Guo, Q. & Hijmans, R. (2004) The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. *International Journal of Geographical Information Science*, **18**, 745–767.

Wisz, M.S., Hijmans, R.J., Li, J. *et al.* (2008) Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, **14**, 763–773.

Wolf, A., Anderegg, W.R.L., Ryan, S.J. & Christensen, J. (2011) Robust detection of plant species distribution shifts under biased sampling regimes. *Ecosphere*, **2**, 1–23.

Young, B.E., Franke, I., Hernandez, P.A., Herzog, S.K., Paniagua, L., Tovar, C. & Valqui, T. (2009) Using spatial models to predict areas of endemism and gaps in the protection of Andean slope birds. *The Auk*, **126**, 554–565.

Zhang, M.G., Zhou, Z.K., Chen, W.Y., Cannon, C.H., Raes, N. & Slik, J.W.F. (2014) Major declines of woody plant species ranges under climate change in Yunnan, China. *Diversity and Distributions*, **20**, 405–415.

Zhang, M.G., Slik, J.W.F. & Ma, K.-P. (2016) Using species distribution modeling to delineate the botanical richness patterns and phytogeographical regions of China. *Scientific Reports*, **6**, 22400.

Zhu, G., Petersen, M.J. & Bu, W. (2012) Selecting biological meaningful environmental dimensions of low discrepancy among ranges to predict potential distribution of bean plataspid invasion. *PLoS ONE*, **7**, 1–9.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Appendix S1. Supplementary tables and figures.

DATA ACCESSIBILITY

All environmental GIS layers are available at WorldClim (http://www.worldclim.org/) and occurrence data at the Global Biodiversity Information Facility (http://www.gbif.org/).

BIOSKETCH

Daniel Park is a postdoctoral fellow in the Department Organismic and Evolutionary Biology at Harvard University. His research focuses on the use of evolutionary frameworks, species distribution modelling and molecular genetics to understand plant invasions and the effects of climate change. Charles Davis is Professor of Organismic and Evolutionary Biology at Harvard University. His research is focused on the factors influencing plant distributions in deep and shallow evolutionary times. More recently, he has worked to integrate phylogeny with traits such as phenology to assess the impact of more recent human-influenced climate change.

Author contributions: D.S.P conceived the study idea and performed all analyses; D.S.P. and C.C.D. interpreted the results; and D.S.P and C.C.D. wrote the manuscript.

Editor: Daniel Chapman