Diversity and Distributions, (Diversity Distrib.) (2011) 17, 1132–1140

BIODIVERSITY RESEARCH

Keep collecting: accurate species distribution modelling requires more collections than previously thought

Kenneth J. Feeley[1,2*] and Miles R. Silman[3,4]

[1]Department of Biological Sciences, Florida International University, Miami, FL, USA, [2]Center for Tropical Plant Conservation, Fairchild Tropical Botanic Garden, Coral Gables, FL, USA, [3]Department of Biology, Wake Forest University, Winston-Salem, NC, USA, [4]Biodiversity and Ecosystem Services Group, Center for Energy, Environment, and Sustainability, Wake Forest University, Winston-Salem, NC, USA

*Correspondence: Kenneth J. Feeley, Department of Biological Sciences, Florida International University, 1200 SW 8th ST, Miami, Fl 33199, USA.
E-mail: kjfeeley@gmail.com

ABSTRACT

Aim Species distribution models (SDMs) use the locations of collection records to map the distributions of species, making them a powerful tool in conservation biology, ecology and biogeography. However, the accuracy of range predictions may be reduced by temporally autocorrelated biases in the data. We assess the accuracy of SDMs in predicting the ranges of tropical plant species on the basis of different sample sizes while incorporating real-world collection patterns and biases.

Location Tropical South American moist forests.

Methods We use dated herbarium records to model the distributions of 65 Amazonian and Andean plant species. For each species, we use the first 25, 50, 100, 125 and 150 records collected and available for each species to analyse changes in spatial aggregation and climatic representativeness through time. We compare the accuracy of SDM range estimates produced using the time-ordered data subsets to the accuracy of range estimates generated using the same number of collections but randomly subsampled from all available records.

Results We find that collections become increasingly aggregated through time but that additional collecting sites are added resulting in progressively better representations of the species' full climatic niches. The range predictions produced using time-ordered data subsets are less accurate than predictions from random subsets of equal sample sizes. Range predictions produced using time-ordered data subsets consistently underestimate the extent of ranges while no such tendency exists for range predictions produced using random data subsets.

Main conclusions These results suggest that larger sample sizes are required to accurately map species ranges. Additional attention should be given to increasing the number of records available per species through continued collecting, better distributed collecting, and/or increasing access to existing collections. The fact that SDMs generally under-predict the extent of species ranges means that extinction risks of species because of future habitat loss may be lower than previously estimated.

Keywords
Amazon, Andes, collecting biases, conservation biogeography, MAXENT, range maps.

INTRODUCTION

Species distribution models (SDMs) are a general suite of models that relate the frequency of species occurrences (presence only or presence/absence) to sets of environmental variables. These relationships can then be used to generate predictions of the geographic areas where the species are expected to occur, making SDMs powerful and widely used tools in conservation biology, biogeography and ecology (Franklin, 2009; Richardson & Whittaker, 2010). SDMs are also increasingly used to predict where species may occur in the future under different climate change scenarios. These predictions can then be used to predict extinction risks because of changes in habitat area as species "migrate" from their

DOI:10.1111/j.1472-4642.2011.00813.x
http://wileyonlinelibrary.com/journal/ddi

A Journal of Conservation Biogeography
Diversity and Distributions

current to future ranges (e.g. Thomas *et al.*, 2004; Feeley & Silman, 2010a). SDMs have rapidly gained in popularity given the potential value of their output combined with their ease of implementation using freely available user-friendly software (such as GARP, MODECO, BIOMOD, MAXENT, DIVA) and the growing number of extensive online species occurrence databases (e.g. the millions of species collection and observation records available online through GBIF, SpeciesLink, Mantis, and elsewhere)(Franklin, 2009).

SDMs have been previously evaluated based on how the accuracy of their range predictions scales with number of occurrence records, or sample size (e.g. Stockwell & Peterson, 2002; Kadmon *et al.*, 2003; Elith *et al.*, 2006; Hernandez *et al.*, 2006; Wisz *et al.*, 2008). These studies, which serve as guidelines for the minimum sample size required for species to be included in SDMs used for conservation planning, have used various measures to assess the accuracy of the range maps produced by the SDMs with reduced sample sizes. For example, accuracy can be assessed as the degree of accordance between range maps predicted using reduced versus full data sets, or alternatively as the ability of range maps produced using data subsets to correctly predict the presence (or presence versus absence) of the species as recorded in independent data sets. However, it is important to note that all of these studies produce their data subsets by selecting samples at random from the larger pool of species occurrence records and then iterating this random subsampling process many times to produce multiple range predictions per species per sample size (Stockwell & Peterson, 2002; Kadmon *et al.*, 2003; Elith *et al.*, 2006; Hernandez *et al.*, 2006; Wisz *et al.*, 2008).

While randomly sub-sampling the data may allow for the assessment and comparison of power between different SDMs, it is likely a poor indicator of the actual number of collections that are required to accurately map and characterize species' distributions. This is because species are rarely, if ever, collected at random. Instead, collecting efforts and field studies suffer from many documented biases, both intrinsic and unintentional (Kadmon *et al.*, 2004; Moerman & Estabrook, 2006; Schulman *et al.*, 2007; Tobler *et al.*, 2007; Loiselle *et al.*, 2008). For example, studies have shown that plants tend to be collected close to research stations and/or along routes of relatively easy access such as roads and rivers (Kadmon *et al.*, 2004). Furthermore, scientific studies are not conducted at random, but rather tend to focus on specifically chosen areas that are systematically explored. This may result in geographic biases being temporally autocorrelated as a result of collection campaigns in which researchers collect many specimens from a small area and then move to another area within that region. Given the nature of these biases, subsets of records ordered by collection date will likely come from just a limited, non-random portion a species' full range.

As such, it can be hypothesized that time-ordered data subsets will be less representative of the full climatic conditions under which species occur. Furthermore, SDMs based on time-ordered data subsets will result in less accurate distribution estimates than if the same number of randomly selected collections were used (Dormann *et al.*, 2007). If this is the case, accurately estimating species ranges will require that more records be collected and incorporated into SDMs than previously suggested. The potential effects of temporal autocorrelations on the ability of SDMs to map species ranges with small sample sizes have been noted (Wisz *et al.*, 2008) but not evaluated.

Here, we assess the climatic representativeness of collections and evaluate the accuracy of SDMs in predicting species ranges on the basis of different sample sizes while incorporates real-world assumptions about how samples are collected in the field, including potential effects of temporal autocorrelations and collection biases. Specifically, we use a database of dated herbarium collection records to model the distributions of 65 well-collected Amazonian and Andean plant species (each with ≥ 200 dated records available after data filtering). We then use the popular SDM software, MAXENT, to produce range estimates using the first 25, 50, 100, 125 and 150 records that were collected for each species. We assess the accuracy of the results generated from these time-ordered subsets to those generated from the full data set of all available records for each species by quantifying accordance between the predicted range estimates. We compare the accuracy of range estimates generated with the time-ordered subsets to the accuracy of range estimates generated using the same number of collections but randomly subsampled from all available records. By comparing the accuracy of models generated from the time-ordered and randomly subsampled data sets, we can assess the impacts of collection biases (i.e. collector behaviour and the nature of field projects) on range estimates. This information will aid in the design of field inventories designed to understand biodiversity distribution and its conservation, and also improve recommendations of the minimum number of collections that should be included when estimating species current (and future) ranges. This is a key challenge in conservation biogeography (Richardson & Whittaker, 2010).

METHODS

We downloaded all available plant herbarium records for tropical South America through the Global Biodiversity Information Facility data portal (GBIF, http://www.gbif.org/; specific databases accessed are listed in the Appendix S1 in the Supporting Information) and SpeciesLink (http://splink.cria.org.br; Appendix S1). All records were then screened using several standard data filters (Feeley & Silman, 2010b, 2011). First, we only included records with geographic collection coordinates and 'without known coordinate issues' (GBIF) or coordinates that were 'not suspect' (SpeciesLink). We also excluded any records with obvious geographic or elevational errors (e.g. those occurring over bodies of water or at elevations > 6000 m). Furthermore, we only included specimens collected from the 'Tropical and Subtropical Moist Broadleaf Forests' and 'Montane Grasslands and Shrublands' biomes of South America as defined by the World Wildlife Fund (Olson *et al.*, 2001), thereby effectively limiting our focus

to just Amazonian and Andean plant species by excluding any records from other geographic areas or biomes. Finally, we excluded all records that did not include a date of collection.

For each plant species with ≥ 200 dated collection records (n = 65, Table S1), we estimated their potential range using the MAXENT species distribution model. MAXENT is an SDM based on machine learning and the principle of maximum entropy (Phillips *et al.*, 2006; Phillips & Dudík, 2008). MAXENT is used to estimate 'the multivariate distribution of suitable habitat conditions (associated with species occurrences) in environmental feature-space' (Franklin, 2009). MAXENT has consistently been shown to be one of the most robust SDMs and to perform very well using presence-only data even with limited sample sizes (Elith *et al.*, 2006; Hernandez *et al.*, 2006; Wisz *et al.*, 2008; Elith & Graham, 2009). Indeed, some previous studies have indicated that it can produce accurate range maps using data from as few as 20–30 occurrences (Hernandez *et al.*, 2006; Wisz *et al.*, 2008). As a result of its strong performance, availability and ease of implementation, MAXENT is one of the most popular SDMs currently being employed to estimate species ranges in relation to environmental predictors (Phillips & Dudík, 2008; Franklin, 2009).

We used three climatic variables to estimate species distributions in MAXENT: (1) mean annual temperature, (2) total annual precipitation and (3) seasonality of precipitation (coefficient of variation of monthly rainfall). We selected these climatic variables as they have previously been shown to play important roles in the performance of tropical plants and are believed to be strongly related to individual species distributions as well as continental-scale gradients of plant species diversity and composition (Gentry, 1988; Ter Steege *et al.*, 2003, 2006; Kreft & Jetz, 2007). Climatic data were downloaded from the WorldClim database (http://www.worldclim.org/) at a resolution of 2.5 arc min (*c.* 5 km at the equator) (Hijmans *et al.*, 2005).

The output of MAXENT is a continuous cumulative probability field (Phillips *et al.*, 2006). We transformed this probability field to a binary map of the 'Suitable' versus 'Unsuitable' habitat within tropical South America by thresholding. For each MAXENT run, we set the threshold as the cumulative probability at which the sum of sensitivity and specificity is maximized (habitat labelled as suitable when probability ≥ threshold). In validation tests, this threshold criterion has been found to perform well and to have a high degree of accuracy in transforming probability fields to binary range maps (Jiménez-Valverde & Lobo, 2007).

We generated two subsamples of the full data set, a 'random' subset and a 'time-ordered' subset. For the former, we randomly selected subsets of 25, 50, 75, 100, 125 and 150 collections from each species' full data set. The data subsets were then used to create range maps in MAXENT. We iterated this process 30 times to generate distributions of possible range predictions for each species per sample size. We assessed the accuracy, or per cent overlap, of each of the 180 range maps generated per species (six sample sizes × 30 draws per sample size × 65 species = 11,700 range maps total) against range maps produced using the corresponding species' full collection data set. Accuracy was assessed using confusion matrixes (i.e. error matrixes) and the kappa statistic (Monserud & Leemans, 1992; Fielding & Bell, 1997; Franklin, 2009) which indicates the per cent accordance between predicted range maps.

We next generated a set of additional MAXENT range maps for each species using the time-ordered lists of geographic collection locations. The time-ordered subsets represented the oldest 25, 50, 75, 100 and 125 available specimens. As above, we assessed the accuracy of the resultant range maps against the range maps produced using the corresponding full data sets using the kappa statistic. We tested the hypothesis that the range estimates generated using these time-ordered collection data subsets will perform significantly worse than the random data subsets by comparing the accuracy (kappa) of the maps generated from the time-ordered collections to the distribution of kappa values generated using the equal number of randomly selected collections. A time-ordered subset was deemed to have performed significantly worse than the random subsets if its kappa value was less than the 5% quantiles of the distribution of kappa values generated with the same number of random collections for the corresponding species.

To characterize collection patterns and how the degree of spatial aggregation, or clumping, in collections may change through time, we quantified the degree of aggregation in each of the time-ordered subsets. Degree of aggregation was estimated by calculating the mean nearest neighbour distance between collection points ($MNND_{to}$) and comparing this to

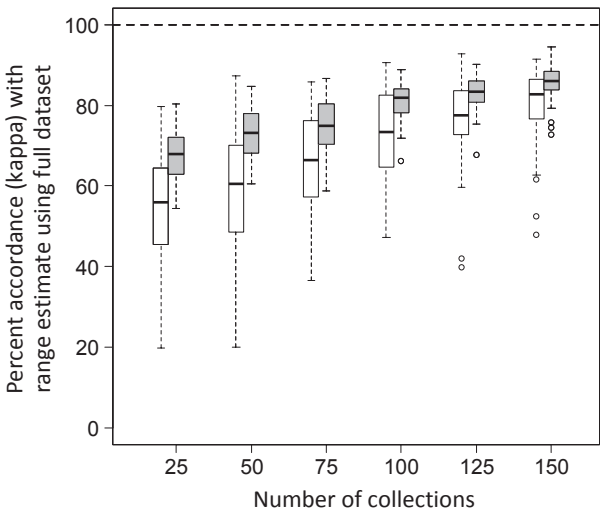

Figure 1 Box-and-whisker plots showing the distribution of the median per cent accordance values (kappa) of species ranges estimated using all available collections (minimum = 200 collections per species) versus ranges estimated using time-ordered data subsets (white) and ranges estimated using random subsets (grey). At all sample sizes, median kappa values for time-ordered subsets are significantly less than for random subsets (Welches two sample t-test; in all cases $P < 0.00005$). Boxes indicate the median and interquartile ranges; whiskers extend to the most extreme value which is ≤ 1.5 times the interquartile range from the box ends.

the expected mean nearest neighbour distance if equal numbers of collections had been sampled at random from throughout the full species range ($MNND_r$; species range based on the complete collection data set as above). Random sampling was simulated 1000 times per species per sample size to generate distributions of $MNND_r$ values. A standardized measure of spatial aggregation was then calculated as: $(MNND_{to} - mean(MMD_r))/stdev(MMD_r)$. Negative values indicate closer than expected nearest neighbour distances and hence a greater degree of clumping in collection locations.

Finally, we analysed the climatic representativeness of the time-ordered subsets by calculating the range of climatic conditions (mean annual temperature, annual precipitation and seasonality of rainfall) sampled with each data subset. Climatic ranges were calculated as the difference in maximum and minimum values extracted for each climatic variable at the sites of collection standardized as a proportion of full climatic ranges covered by each species in the full collection data set.

RESULTS

The mean accuracy of range maps increased with increasing number of samples regardless of the technique used to generate the data subsets (Fig. 1). However, at all sample sizes, the accuracy of the SDMs generated using time-ordered data subsets was significantly worse than when using subsets of collections taken randomly from the species' complete data set (Welches two sample *t*-test; in all cases $P < 0.00005$; Fig. 1). At the individual species level, the percentage of study species for which range estimates generated from time-ordered subsets performed significantly worse than range estimates generated using random data subsets was 31, 46, 38, 29, 21 and 17% using 25, 50, 75, 100, 125 and 150 samples, respectively (Fig. 2; Table S2).

The type of error differed between range maps produced using the time-ordered versus random data subsets. When using random subsets, the SDMs tended to produce different

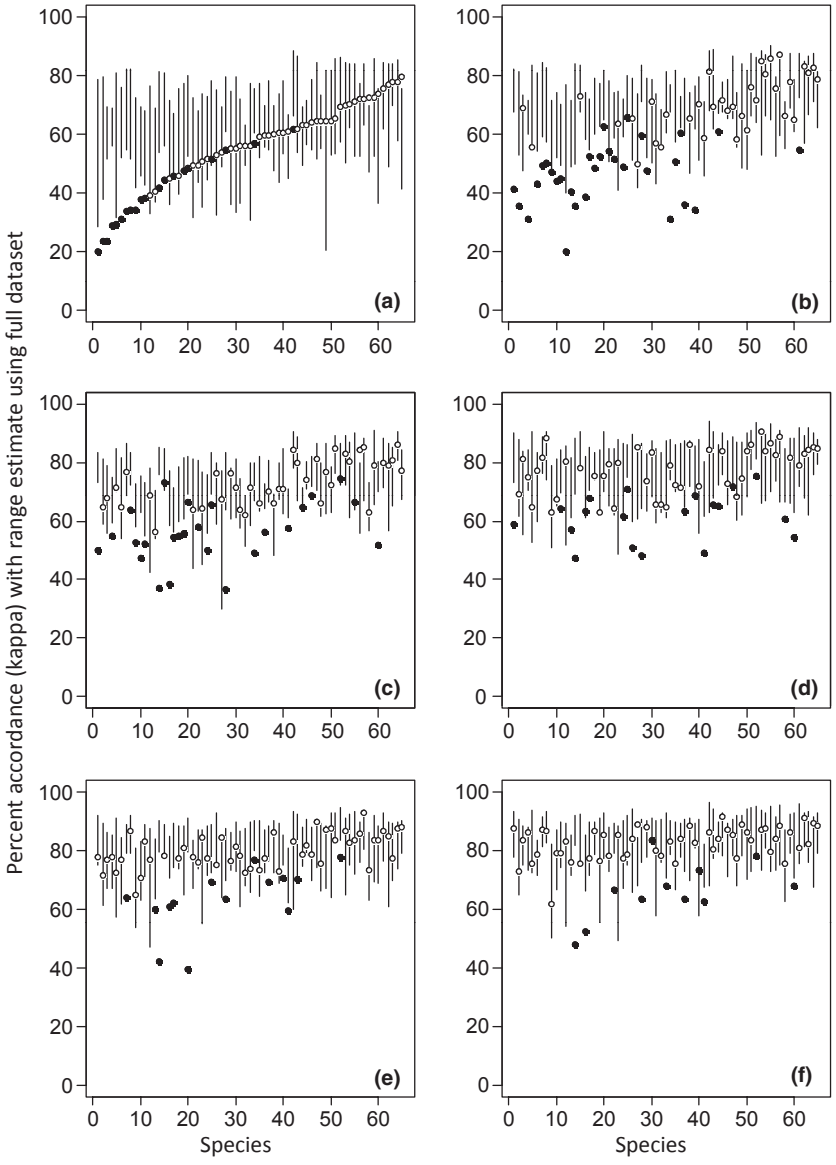

Figure 2 The per cent accordance (kappa) of species ranges estimated using all available collections (minimum = 200 collections per species) versus ranges estimated using time-ordered data subsets of the first (a) 25, (b) 50, (c) 75, (d) 100, (e) 125 and (f) 150 samples (points) and the 90% quantiles of kappa values for ranges estimated using random subsets of the same sample sizes (bars). Black points indicate species whose range estimates produced with time-ordered data subsets have Kappa values significantly less than the corresponding range estimates using random data subsets. Species are ordered according to their time-ordered data subset kappa value with 25 samples (see Table S1 in the supporting online materials for species IDs).

range predictions but did not have any consistent tendency to over- or under-predict habitat area (the size of area deemed suitable) when compared with the range estimates produced using the full data sets (Fig. 3; Table S3). In fact, predictions of habitat area differed significantly from the habitat areas of the full ranges in only 5, 8, 8, 0, 2 and 2% of species using 25, 50, 75, 100, 125 and 150 samples, respectively. In contrast, when using the time-ordered data subsets, habitat areas were consistently under-predicted. Specifically, range estimates generated using time-ordered subsets under-predicted habitat area in 66, 63, 63, 63, 63 and 54% species using 25, 50, 75, 100, 125 and 150 samples, respectively. Except at the highest sample size ($n = 150$), this is significantly more under-predictions than expected by random (binomial probability, $P \leq 0.001$).

The standardized degree of spatial aggregation increased with larger time-order data subsets (Fig. 4; Table S4). The proportion of the climatic niche breadth represented by collections also increased with larger time-ordered data subsets (Fig. 5; Tables S5–S7). This pattern was strongest in the case of mean annual temperature for which the data subsets of the first 25 collections represented on average just 53.3% (95% CI = 48.0–58.6%) of the species' full temperature ranges.

DISCUSSION

Results from this study show that the use of randomly subsampled data records disregards the temporally autocorrelated spatial biases that are ubiquitous in herbarium collection databases. Species occurrence data are not collected at random but rather collections tend to be clumped around specific areas because of the nature of collecting campaigns and ease of access to sites within a particular habitat (Kadmon *et al.*, 2004; Moerman & Estabrook, 2006; Schulman *et al.*, 2007; Tobler *et al.*, 2007; Loiselle *et al.*, 2008). Interestingly, through time,

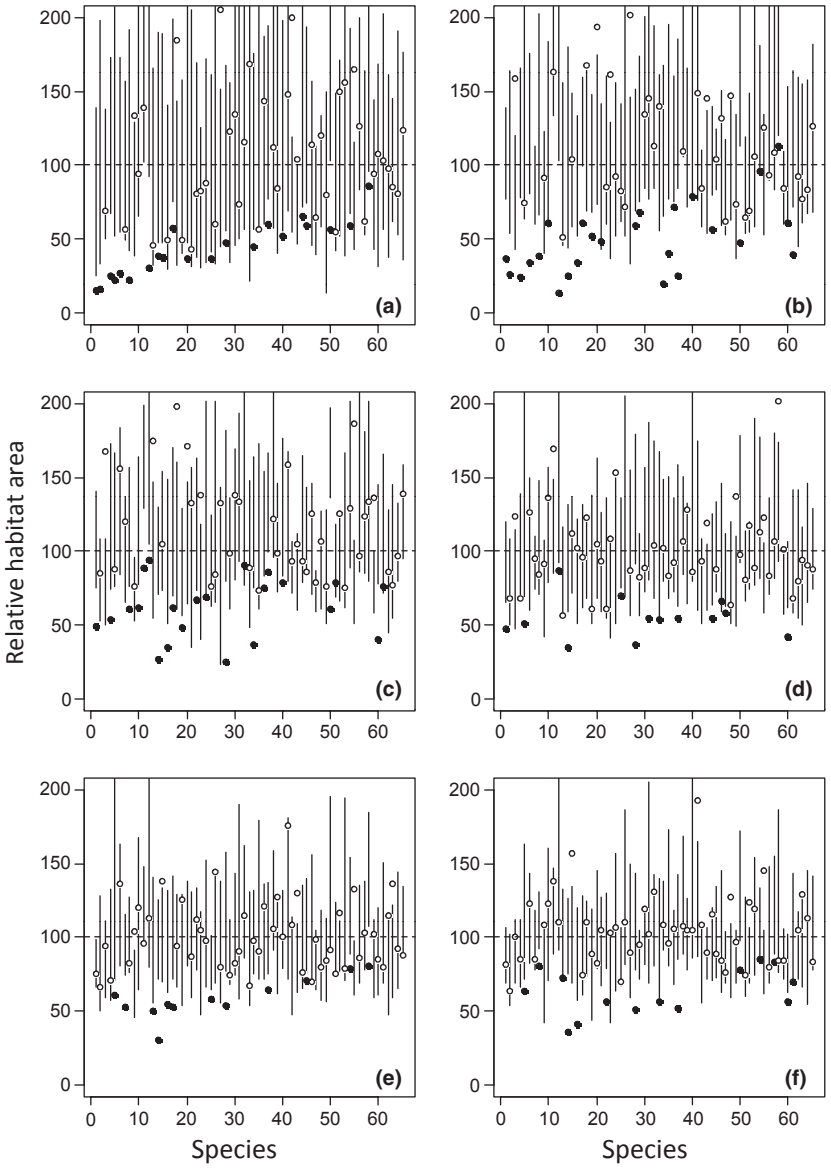

Figure 3 Relative habitat areas (in relation to habitat areas predicted using full data sets) estimated using time-ordered data subsets of the first (a) 25, (b) 50, (c) 75, (d) 100, (e) 125 and (f) 150 samples (points) and the 90% quantiles of relative areas for ranges estimated using random subsets of the same sample sizes (bars). Black points indicate species whose estimate of habitat area as produced with time-ordered data subsets is significantly less than areas estimated using random data subsets. Species order is maintained from Fig. 2 (see Table S1 in the supporting online materials for species IDs).

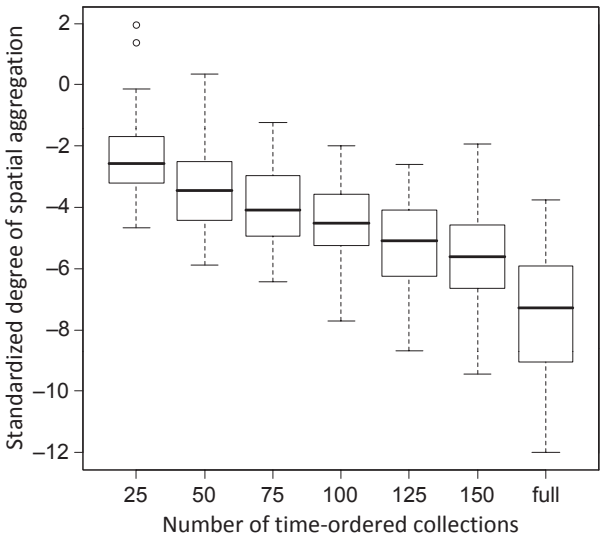

Figure 4 Box-and-whisker plots showing the distribution of the standardized degree of spatial aggregation, or clumping (see text), occurring in the time-ordered collection data sets versus number of collections. More negative values indicate a greater degree of clumping. Box and whiskers as in Fig. 1.

collections actually become increasing spatially aggregated or clumped as areas of interest are revisited (Fig. 4; Table S4). However, additional collecting sites are added through time resulting in better representations of the full climatic niches of species (Fig. 5; Tables S5–S7). Of particular interest within the context of global warming, the climatic representativeness with small sample sizes is lowest in the case of mean annual temperature (Fig. 5; Tables S5–S7).

As a result of the spatio-temporal aggregation of collections, many more records will be required to accurately characterize the 'true' range of conditions under which species may occur than if the collections had been collected at random across the species' ranges. For example, 75–100 collections are required to produce range estimates that are as accurate on average as the range estimates produced using 25 randomly subsampled collections (Fig. 1). This result has important implications for our estimates of the effort required to gather enough data to accurately characterize of species' ranges and for best-practice guidelines for future collections. While testing the accuracy of SDMs using random subsets of data gives us means of assessing the power of various models (Wisz *et al.*, 2008), the results of these studies should not be used as a guideline for how many samples should actually be included when attempting to model species ranges. Nor should the results of studies based on random subsets of collections data be used as guidelines for deciding when species are sufficiently collected to deemphasize future collections and stop funding efforts aimed at obtaining new field records.

The fact that more collections are required to accurately predict species distributions than has been suggested by previous studies also has important implications for the utility and interpretation of SDMs. This will be especially true for taxa

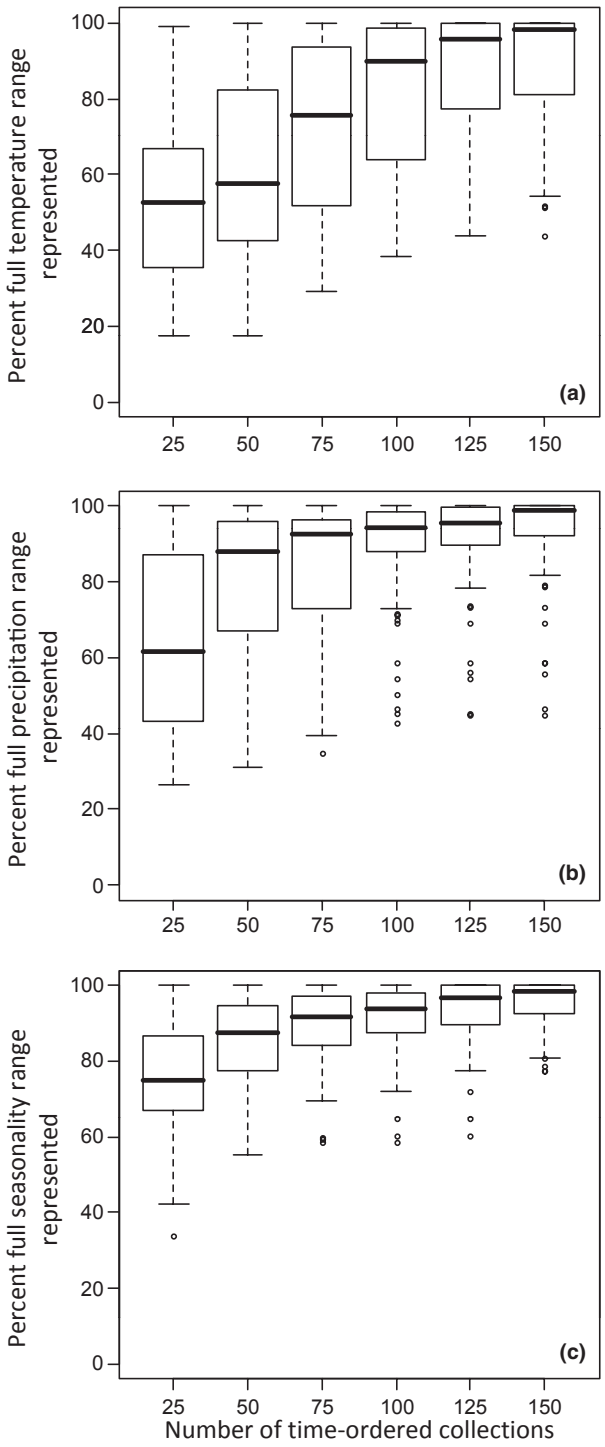

Figure 5 Box-and-whisker plots showing the percentage of full climatic niche breadth represented by time-ordered collection data sets versus number of collections: (a) mean annual temperature, (b) annual precipitation and (c) seasonality of precipitation. Box and whiskers as in Fig. 1.

or geographic areas for which the number of available occurrence records is limited. For example, *c.* 5% of tropical plant species are represented by ≥ 20 records available through the combined GBIF and SpeciesLink repositories (Feeley &

Silman, 2011). If, however, the minimum sample size is increased to ≥ 100, only 0.3% of species have enough available data (Feeley & Silman, 2011).

The amount of data required could possibly be reduced if collection biases are reduced, for example through detecting and quantifying biases in past collections (Robertson *et al.*, 2010). Biases in future collecting efforts can be minimized through systematic sampling in the field at either random, regular or strategically situated locations. Ideally, national sampling grids or inventories would be established that would minimize geographic and taxonomic collection biases. The use of regular inventories would also result in presence and absence data allowing for different sets of SDMs and more accurate predictions of species distributions. National inventory programmes are already implemented in several developed nations (e.g. The USA's Forest Inventory Analysis: http://www.fia.fs.fed.us/). Similar efforts should be implemented globally and especially in developing tropical nations which house the vast majority of terrestrial diversity and which face some of the greatest conservation challenges. Given the logistical and physical difficulty of working in Earth's highest biodiversity areas, though, the benefits of a random or systematic sampling strategy may be outweighed in many cases by simply adding additional sampling efforts in areas with marginal predictive certainty in existing models, along environmental gradients (Austin & Heyligers, 1989; Wessels *et al.*, 1998), or as selected through survey-gap analyses (Funk *et al.*, 2005).

The potentially good news arising from this study is that time-ordered subsets of records tend to under-predict species ranges (as may be expected with clumped sampling; Fig. 3; Table S3). As such, it is likely that the ranges currently being predicted for most species are underestimates. It is therefore possible that we are over-predicting the threat of extinction for many species as risk of extinction due to habitat loss (e.g. because of climate change and/or deforestation and land use change) is generally negatively associated with range extent (i.e. specialized species with small ranges will generally be a greater risk of extinction than species with larger ranges)(Jetz *et al.*, 2008; Feeley & Silman, 2010b). The overestimation of habitat loss may be magnified in spatially explicit analyses (e.g. Feeley & Silman, 2009) if disturbances and collection efforts are both concentrated around areas of easy access. Conversely, estimates of habitat loss may be artificially lowered if collectors oversample from parks and other protected areas.

We conclude by stressing that the multiple biases in SDMs as revealed in this and other papers do not invalidate or undermine the general methodology. Quite the opposite, all models have biases and uncertainties, and understanding the biases allows policy and management decisions that are both sensitive to and robust to uncertainty (Richardson & Whittaker, 2010).

ACKNOWLEDGEMENTS

We thank the Global Biodiversity Information Facility, SpeciesLink and all contributing herbaria for making their data publicly available and facilitating studies of ecology and biogeography. This research was supported by the Fairchild Tropical Botanic Garden's Herbarium and Center for Tropical Plant Conservation, the Gordon and Betty Moore Foundation Andes to Amazon Program and NSF DEB-0743666.

REFERENCES

Austin, M.P. & Heyligers, P.C. (1989) Vegetation survey design for conservation: gradsect sampling of forests in north-eastern New South Wales. *Biological Conservation*, **50**, 13–32.

Dormann, C.F., Mcpherson, J.M., Araujo, M.B., Bivand, R., Bolliger, J., Carl, G., Davies, R.G., Hirzel, A., Jetz, W., Kissling, W.D., Kuhn, I., Ohlemuller, R., Peres-Neto, P.R., Reineking, B., Schroder, B., Schurr, F.M. & Wilson, R. (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, **30**, 609–628.

Elith, J. & Graham, C. (2009) Do they? How do they? Why do they differ? – on finding reasons for differing performances of species distribution models *Ecography*, **32**, 66.

Elith, J., Graham, C.H., Anderson, R.P. *et al.* (2006) Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, **29**, 129–151.

Feeley, K.J. & Silman, M.R. (2009) Extinction risks of Amazonian plant species. *Proceedings of the National Academy of Sciences USA*, **106**, 12382–12387.

Feeley, K.J. & Silman, M.R. (2010a) Land-use and climate change effects on population size and extinction risk of Andean plants. *Global Change Biology*, **16**, 3215–3222.

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

Feeley, K.J. & Silman, M.R. (2011) The data void in modeling current and future distributions of tropical species. *Global Change Biology*, **17**, 626–630.

Fielding, A.H. & Bell, J.F. (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environmental Conservation*, **24**, 38–49.

Franklin, J. (2009) *Mapping species distributions: spatial inference and predictions*. Cambridge University Press, New York.

Funk, V.A., Richardson, K.S. & Ferrier, S. (2005) Survey-gap analysis in expeditionary research: where do we go from here? *Biological Journal of the Linnean Society*, **85**, 549–567.

Gentry, A.H. (1988) Changes in plant community diversity and floristic composition on environmental and geographic gradients. *Annals of the Missouri Botanical Garden*, **75**, 1–34.

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.

Jetz, W., Sekercioglu, C.H. & Watsono, J.E.M. (2008) Ecological correlates and conservation implications of overestimating species geographic ranges. *Conservation Biology*, **22**, 110–119.

Jiménez-Valverde, A. & Lobo, J.M. (2007) Threshold criteria for conversion of probability of species presence to either-or presence-absence. *Acta Oecologica*, **31**, 361–369.

Kadmon, R., Farber, O. & Danin, A. (2003) A systematic analysis of factors affecting the performance of climatic envelope models. *Ecological Applications*, **13**, 853–867.

Kadmon, R., Farber, O. & Danin, A. (2004) Effect of roadside bias on the accuracy of predictive maps produced by bioclimatic models. *Ecological Applications*, **14**, 401–413.

Kreft, H. & Jetz, W. (2007) Global patterns and determinants of vascular plant diversity. *Proceedings of the National Academy of Sciences*, **104**, 5925–5930.

Loiselle, B.A., Jorgensen, P.M., Consiglio, T., Jimenez, I., Blake, J.G., Lohmann, L.G. & Montiel, O.M. (2008) Predicting species distributions from herbarium collections: does climate bias in collection sampling influence model outcomes? *Journal of Biogeography*, **35**, 105–116.

Moerman, D.E. & Estabrook, G.F. (2006) The botanist effect: counties with maximal species richness tend to be home to universities and botanists. *Journal of Biogeography*, **33**, 1969–1974.

Monserud, R.A. & Leemans, R. (1992) Comparing global vegetation maps with the kappa statistic. *Ecological Modelling*, **62**, 275–293.

Olson, D.M., Dinerstein, E., Wikramanayake, E.D., Burgess, N.D., Powell, G.V.N., Underwood, E.C., Amico, J.A., Itoua, I., Strand, H.E., Morrison, J.C., Loucks, C.J., Allnutt, T.F., Ricketts, T.H., Kura, Y., Lamoreux, J.F., Wettengel, W.W., Hedao, P. & Kassem, K.R. (2001) Terrestrial ecoregions of the world: a new map of life on earth. *BioScience*, **51**, 933–938.

Phillips, S.J. & Dudík, M. (2008) Modeling of species distributions with maxent: new extensions and a comprehensive evaluation. *Ecography*, **31**, 161–175.

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

Richardson, D.M. & Whittaker, R.J. (2010) Conservation biogeography – foundations, concepts and challenges. *Diversity and Distributions*, **16**, 313–320.

Robertson, M.P., Cumming, G.S. & Erasmus, B.F.N. (2010) Getting the most out of atlas data. *Diversity and Distributions*, **16**, 363–375.

Schulman, L., Toivonen, T. & Ruokolainen, K. (2007) Analysing botanical collecting effort in amazonia and correcting for it in species range estimation. *Journal of Biogeography*, **34**, 1388–1399.

Stockwell, D.R.B. & Peterson, A.T. (2002) Effects of sample size on accuracy of species distribution models. *Ecological Modelling*, **148**, 1–13.

Ter Steege, H., Pitman, N., Sabatier, D. *et al.* (2003) A spatial model of tree alpha-diversity and tree density for the Amazon. *Biodiversity and Conservation*, **12**, 2255–2277.

Ter Steege, H., Pitman, N.C.A., Phillips, O.L., Chave, J., Sabatier, D., Duque, A., Molino, J.-F., Prevost, M.-F., Spichiger, R., Castellanos, H., Von Hildebrand, P. & Vasquez, R. (2006) Continental-scale patterns of canopy tree composition and function across amazonia. *Nature*, **443**, 444–447.

Thomas, C.D., Cameron, A., Green, R.E., Bakkenes, M., Beaumont, L.J., Collingham, Y.C., Erasmus, B.F.N., De Siqueira, M.F., Grainger, A., Hannah, L., Hughes, L., Huntley, B., Van Jaarsveld, A.S., Midgley, G.F., Miles, L., Ortega-Huerta, M.A., Townsend Peterson, A., Phillips, O.L. & Williams, S.E. (2004) Extinction risk from climate change. *Nature*, **427**, 145–148.

Tobler, M., Honorio, E., Janovec, J. & Reynel, C. (2007) Implications of collection patterns of botanical specimens on their usefulness for conservation planning: an example of two neotropical plant families (moraceae and myristicaceae) in Peru. *Biodiversity and Conservation*, **16**, 659–677.

Wessels, K.J., Jaarsveld, A.S.V., Grimbeek, J.D. & Linde, M.J.V.D. (1998) An evaluation of the gradsect biological survey method. *Biodiversity and Conservation*, **7**, 1093–1121.

Wisz, M.S., Hijmans, R.J., Li, J., Peterson, A.T., Graham, C.H. & Guisan, A. (2008) Effects of sample size on the performance of species distribution models. *Diversity and Distributions*, **14**, 763–773.

SUPPORTING INFORMATION

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

Appendix S1 List of herbaria contributing tropical plant collection records accessed through the Global Biodiversity Information Facility and SpeciesLink.

Table S1 Number of collections in full dataset for each of the 65 study species.

Table S2 Percent accordance (kappa) between species range maps produced using time-ordered data subsets versus the corresponding range maps produced using the full datasets. The 90% quantiles of kappa values for maps produced using random data subsets are listed in parentheses.

Table S3 Area of range maps produced using time-ordered data subsets relative to the corresponding range maps produced using the full datasets. The 90% quantiles of relative areas for maps produced using random data subsets are listed in parentheses.

Table S4 Degree of spatial aggregation exhibited by each species for each time-ordered data subset sample size.

Table S5 Proportion of full thermal niche breadth represented by time-ordered collection datasets.

Table S6 Proportion of full precipitation niche breadth represented by time-ordered collection datasets.

Table S7 Proportion of full seasonality of precipitation niche breadth represented by time-ordered collection datasets.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

BIOSKETCHES

Kenneth J. Feeley is an Assistant Professor in the Department of Biological Sciences at Florida International University and a Conservation Biologist at the Fairchild Tropical Botanic Garden in Miami, FL, USA. His research focuses on understanding the responses of tropical forest ecosystems to large-scale anthropogenic disturbances such as deforestation, fragmentation and climate change.

Miles R. Silman is a community ecologist interested in understanding the factors that influence species distributions and the forces that promote and maintain diversity in tropical ecosystems. He is an Associate Professor in the Department of Biology at Wake Forest University and a principal investigator in the Andes Biodiversity and Ecosystem Research Group.

Author Contributions: K.J.F. and M.R.S. conceived the study; K.J.F. collected the data; K.J.F. analyzed the data; K.J.F. and M.R.S. wrote the paper.

Editor: Jessica Hellmann